Proceedings of the

28th International Workshop on Statistical Modelling

July 8 – 12, 2013

Palermo, Italy

Vito M.R. Muggeo, Vincenza Capursi

Giovanni Boscaino, Gianfranco Lovison

(editors)

Proceedings of the 28th International Workshop on Statistical Modelling,
Palermo, July 8 – 12, 2013,
Vito Muggeo, Vincenza Capursi, Giovanni Boscaino, Gianfranco Lovison,
(editors),
Palermo, 2013.

Editors:
Vito M.R. Muggeo, vito.muggeo@unipa.it
Vincenza Capursi, vincenza.capursi@unipa.it
Giovanni Boscaino, giovanni.boscaino@unipa.it
Gianfranco Lovison, gianfranco.lovison@unipa.it

Dipartimento di Scienze Statistiche e Matematiche, 'Vianelli'
viale delle Scienze, edificio 13
Università di Palermo
90128 Palermo, Italy

Scientific Programme Committee

- Vito M.R. Muggeo (Chair)
  University of Palermo, Italy
- Ruggero Bellio
  University of Udine, Italy
- Stefano Campostrini
  University of Venezia, Italy
- Maria Durban
  University Carlos III of Madrid, Spain
- Paul Eilers
  Erasmus Medical Centre Rotterdam, Netherlands
- Herwig Friedl
  Graz University of Technology, Austria
- Thomas Kneib
  Georg-August-University Göttingen, Germany
- Arnost Komarek
  Charles University in Prague, Czech Republic
- Kenan Matwie
  University of Western Sydney, Australia
- Mikis Stasinopoulos
  London Metropolitan University, UK

  Gerhard Tutz
  Ludwig-Maximilians-Universität, München, Germany
- Florin Vaida
  University of California, San Diego, USA

Local Organizing Committee

- Vito M.R. Muggeo (Chair)
  University of Palermo, Italy
- Giada Adelfio
  University of Palermo, Italy
- Giovanni Boscaino
  University of Palermo, Italy
- Vincenza Capursi
  University of Palermo, Italy
- Marcello Chiodi
  University of Palermo, Italy
- Andrea Consiglio
  University of Palermo, Italy
- Gianfranco Lovison
  University of Palermo, Italy
- Antonella Plaia
  University of Palermo, Italy
- Mariangela Sciandra
  University of Palermo, Italy

Preface

This year, to celebrate its 28th edition, the International Workshop on Statistical Modelling (IWSM) goes South, holding its 2013 conference in Palermo, Sicily, the southernmost region of Italy. We are particularly honoured to host this conference, renewing and strengthening a tradition which makes Palermo one of the oldest and well-established schools of Statistics in Italy.

This is going to be a particularly well-attended, and hopefully successful, edition of IWSM, with a total of 172 initial submissions for either oral presentation or poster, and a final outcome of 68 contributions accepted for oral presentation and 80 for poster presentation, probably the largest number in the history of IWSM. Of course, the success of a conference cannot be measured only in quantitative terms; the remarkable number of contributions is complemented by their quality and the variety of areas of statistical modelling covered by the submitted papers, a result which the whole IWSM community should be proud of and for which we must thank the authors and the members of the Scientific Committee, who made a great job selecting the best papers. The diversity and liveliness of our discipline is also witnessed by the invited plenary talks, for which we thank Ciprian Crainiceanu, Torsten Hothorn, Stefano M. Iacus, Geoff McLachlan and Hein Putter, and by the short course on Structural Equation Modelling taught by John Fox.

Notwithstanding the large number of contributions and the richness and variety of topics and areas involved, the conference has been able to maintain its almost unique feature of scheduling one plenary session for the whole week, an organisational choice which has always helped the IWSM to have a stimulating atmosphere, encouraging exchange of ideas and cross-fertilisation among different areas of statistics.

According to the workshop tradition, in this edition student participation has been strongly encouraged: we award three students for the best paper, the best oral presentation, and the best poster; furthermore, two student travel grants have been kindly provided by the Statistical Modelling Society.

So, it is time to begin! We welcome you in Palermo, and really hope you will be enriched by the scientific experience of taking part in the 28th IWSM, without forgetting to also enjoy some of the other delights Palermo can offer you: a sunny day at the beach, a visit to its ancient streets, buildings, markets and gardens and ... its unforgettable food!!

Vito Muggeo
Vincenza Capursi
Giovanni Boscaino
Gianfranco Lovison

Palermo, May 2013

Contents

Part I - Invited Papers

Ciprian M. Crainiceanu, Vadim Zipunnikov, Jiawei Bai: Science to data to statistical models 3

Torsten Hothorn, Thomas Kneib, Peter Bühlmann: Conditional Transformation Models by Example 15

Stefano M. Iacus, Gary King: How Coarsening Simplifies Matching-Based Causal Inference Theory 27

Geoffrey J. McLachlan, Sharon X. Leemaqz: On Finite Mixtures of Skew Distributions 33

Hein Putter, Mia Klinten Grand: Regression Models for Expected Length of Stay 45

Part II - Contributed Papers (volume 1)

Antonino Abbruzzo, Ivan Vujacic, Ernst C. Wit, Angelo M. Mineo: Model selection for penalized Gaussian Graphical Models 59

Giada Adelfio, Marcello Chiodi: Mixed estimation technique in semi-parametric space-time point processes for earthquake description 65

Robert J. Adler, Kevin Bartz, Samuel C. Kou, Anthea Monod: A Topological Approach to the Statistical Estimation of Random Field Thresholds 71

Smitha Ankinakatte, David Edwards: The analysis of discrete longitudinal data using acyclic probabilistic finite automata 77

Carmen Armero, Peter J. Diggle, Anabel Forte, Hèctor Perpiñán: Markov mixture models for analyzing the evolution of chronic kidney disease in children 83

Vincent Bremhorst, Philippe Lambert: Estimation of the Latent Distribution in Cure Survival Models using a Flexible Cox Model 87

Kevin Burke, Gilbert MacKenzie: Multi-Parameter Regression Survival Models 93

Carlo G. Camarda, Paul H.C. Eilers, Jutta Gampe: Exploratory Exponential Tilting 97

Carlo G. Camarda, Niel Hens, Paul H.C. Eilers: Modelling Social Contact Data: a smoothing constrained approach 103

Roberto Colombi, Sabrina Giordano: Marginal Parameterizations for Hidden Markov Models 109

Enrico A. Colosimo, André G.F.C. Costa, Leila Amorim: Marginal Models for the Association Structure of Hierarchical Binary Responses 115

Francisco Cueva, Emilio Porcu, Ronny Vallejos: A nonparametric study of the spatial association between forest variables .. 121

Michaela Dvorzak,Helga Wagner: Sparse Bayesian modeling of underreported count data 127

Paul H.C. Eilers: Harmonic Histograms: Smoothing of Grouped Circular Data Distributions 133

Marco Enea, Antonella Plaia, Vincenza Capursi: Modeling confidential data via modified hurdle mixed models 139

Francesco Finazzi, Claire Miller, Marian Scott: A model-based clustering approach for the analysis of environmental time series 145

Gianluca Frasso, Paul H.C. Eilers: L-surface and V-valley for multi-dimensional smoothing parameter selection 151

Jonathan Gellar, Ciprian M. Crainiceanu: Cox Regression Models with Functional Covariates 157

Edward I. George, Veronika Ročková, Emmanuel Lesaffre: Faster Spike-and-Slab Variable Selection with Dual Coordinate Ascent EM 165

Anna Gottard: A joint Bradley-Terry model for tennis tournaments via Data Cloning 171

Irene L. Hudson, Shalem Y. Leemaqz, David Darwent, Greg Roach, Drew Dawson: SOM clustering and modelling of Australian railway drivers' sleep, wake, duty profiles 177

Maria Iannario, Domenico Piccolo: A proposal for modelling overdispersion in ordinal data 183

Stijn Jaspers, Marc Aerts, Geert Verbeke: Estimation of an MIC distribution using a two-stage semi-parametric mixture model 191

Vandna Jowaheer, Brajendra Sutradhar, Rajendra Neupane: Likelihood Analysis For An Incomplete Longitudinal Hemoglobin Data 197

André Klima, Helmut Küchenhoff, Paul W. Thurner: Analysis of voter transition using ecological data: Comparison of different approaches for Munich election data 203

Thomas Kneib, Elisabeth Waldmann, Fabian Sobotka: Bayesian Expectile Regression 209

Ioannis Kosmidis, David Firth: Reduced-bias inference for multidimensional Rasch models with applications 215

Philippe Lambert: A new flexible family of conditional Archimedean copulas 221

Gwenaël G.R. Leday, Aad W. Van der Vaart, Mark A. Van de Wiel: An empirical Bayesian ridge approach to modeling the transcriptional effects of DNA copy number aberrations 227

Dae-Jin Lee, María Durbán: Spatio-temporal seasonal data modelling and forecasting with penalized smooth-ANOVA models ... 233

Nicholas T. Longford: Decision theory for some elementary statistical problems 239

Nelmarie Louw: Kernel based dimension reduction and classification of spectroscopy data for authentication of South African wines 245

Robson J.M. Machado, Cibele M. Russo: Semiparametric partially nonlinear mixed-effects models with P-splines 251

Gilbert MacKenzie, Il Do Ha: The role of Frailty in survival studies 257

Juliane Manitz, Thomas Kneib: Model-based source estimation during foodborne disease outbreaks 263

Andreas Mayr, Florian Faschingbauer, Matthias Schmid: Boosting sonographic birth weight estimation 269

James P. McKeone, Anthony N. Pettitt: Bayesian P-splines with a multiplicative term in EMG trace data 275

Nicholas Mesue, Tapio Nummi: Testing of Growth Curves using Cubic Smoothing Splines: A Multivariate Approach 281

Daniel A. Molinari, Ludger Evers, Adrian W. Bowman: Posterior approximations for Gaussian models with "non-detect" data 289

Kathakali Ghosh Mukherjee, Claire Miller, Adrian W. Bowman, Gregor Thut: Characterisation and Mixed Effects Models for EEG Signals 295

Magdalena Murawska, Dimitris Rizopoulos, Emmanuel Lesaffre: A comparison between landmarking and joint modeling for producing predictions using longitudinal outcomes 301

Ruth Nysen, Marc Aerts, Christel Faes: Model averaging quantiles for censored data 307

María Oliveira, Rosa María Crujeiras, Alberto Rodríguez–Casal: CircSiZer for exploring circular data..... 313

Wolfgang Pössnecker, Gerhard Tutz: Variable Selection and Shrinkage of Varying to Fixed Effects in Finite Mixtures of Generalized Linear Models..... 319

Xanthi Pedeli, Konstantinos Fokianos, Mohsen Pourahmadi: Cholesky decomposition for multivariate volatilities 325

Jean Peyhardi, Catherine Trottier, Yann Guédon A unifying framework for specifying generalized linear models for categorical data 331

Veronika Ročková, Emmanuel Lesaffre: Bayesian Sparse Factor Regression Approach to Genomic Data Integration 337

María Xosé Rodríguez - Álvarez, Dae-Jin Lee, Thomas Kneib, María Durbán, Paul H.C. Eilers: Fast algorithm for smoothing parameter selection in multidimensional P-splines.... 343

Erlis Ruli, Laura Ventura, Walter Racugno: Approximate Bayesian inference based on modified log-likelihood ratios..... 351

Mbéry Séne, Carine A. Bellera, Cécile Proust-Lima: Joint modeling of longitudinal and time-to-event data with application to the prediction of prostate cancer recurrence..... 357

Benjamin Saefken, Sonja Greven, Thomas Kneib: Estimating prediction error in mixed models 363

Bruno Santos, Heleno Bolfarine: A two-part model using quantile regression under a Bayesian perspective 369

Gunther Schauberger, Gerhard Tutz: DIF-LASSO: Differential Item Functioning in Rasch Models 375

Sabine K. Schnabel,Paul H.C. Eilers, Fred A. van Eeuwijk : Modelling plant height data with scaled and shifted prototype curves 381

Mariangela Sciandra, Gianfranco Lovison: Model interpretation from the additive elements of the PWRSS in GLMMs 387

Fabian Sobotka, Andreas Mayr, Thomas Kneib: Fractile Boosting: a novel approach to mode regression 393

Elizabeth M. Sweeney, Russel T. Shinohara, Joshua T. Vogelstein, Daniel S. Reich, Ciprian M. Crainiceanu: Do not use a cannon to kill a mosquito: a comparison of supervised classification algorithms in the context of MS lesion segmentation in structural MRI 399

Kukatharmini Tharmaratnam, Linn Cecilie Bergersen, Ingrid K. Glad: Nonlinear Monotone Regression for High-dimensional Data 405

Anestis Touloumis: A GEE Approach for Correlated Ordinal and Nominal Multinomial Responses 411

Matilde Trevisani, Arjuna Tuzzi: Through the JASA's Looking-Glass, and What We Found There 417

Jan Van de Kassteele, Agnetha Hofhuis, Peter Teunis, Wilfrid Van Pelt: Modeling the risk of *Borrelia* infection after a tick bite - a Bayesian approach 423

Ardo van den Hout, Graciela Muniz: Joint models for discrete longitudinal outcome and survival 429

Massimo Ventrucci, Daniela Cocchi, Marian Scott: Bayesian P-spline models for land use raster datasets 435

Helga Wagner, Regina Tüchler: Bayesian Factorization Model for Analysing Mixed Data 441

Elisabeth Waldmann, Anke Stein, Thomas Kneib: Bivariate Bayesian Quantile Regression 447

Madawa P. Weerasinghe Jayawardana Rathambalage, Georgy Sofronov: GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data 453

Sten Willemsen, Paul H.C. Eilers, Regine Steegers, Emmanuel Lesaffre: A multivariate Bayesian model for human growth 459

Ernst C. Wit, Ivan Vujacic, Javier Gonzalez: Inference of non-linear ODE dynamics 465

Part III - Contributed Papers (volume 2)

Steven Abrams, Niel Hens: Extending frailty models applied to infectious disease epidemiology 475

Giada Adelfio, Giovanni Boscaino: The student *talent* in a random effects Quantile Regression Model for university performance 479

Serena Arima: Item selection via Bayesian graded response model 485

Jiawei Bai, Bing He, Thomas A. Glass, Ciprian M. Crainiceanu: Two valid and interpretable metrics to summarize raw accelerometry data 489

Ahmed S. Bani-Mustafa, Kenan M. Matawie: Recursive Residuals Application in Linear Mixed Models 493

Paul D. Baxter, Marc A. Bailey, D. Julian A. Scott: Modelling the growth of Abdominal Aortic Aneurysms using mixed effects regression with autocorrelated residuals 497

Marco Bazzi, Paola Tellaroli: Finding profiles in time-course gene expressions 501

Ruggero Bellio, Paolo Vidoni: A note on improved random effects prediction in GLMMs 507

Betsabé G. Blas Achic, Marcos Antonio A. Pereira: Calibration model with scale mixtures of skew-normal distributions . 511

Raquel Caballero-Águila, Irene García-Garrido, Josefa Linares-Pérez: Distributed fusion filtering using correlated missing observations from multiple sensors 515

Raquel Caballero-Águila, Irene García-Garrido, Josefa Linares-Pérez: Optimal least-squares linear centralized filter for systems with autocorrelated and cross-correlated noises 521

Marcos H. Cascone, Larissa A. Matos, José A. G. Campos : Nonparametric method for treatment comparison in the wine-making process 525

Manuela Cattelan, Cristiano Varin: Marginal modeling of dependent paired comparison data 529

Jona Cederbaum, Sonja Greven: Functional Linear Mixed Models for Sparsely and Irregularly Sampled Data 533

Vasiliki Christou, Konstantinos Fokianos: Testing Linearity for Nonlinear Count Time Series Models 539

Juliana Cobre, Mário De Castro: Bayesian inference for a family based on the Weibull and the power series distributions 543

Audrey H.M.A. Cysneiros, Mariana C. Araújo: Corrected Profile Likelihood in Heteroscedastic Symmetric Nonlinear Model 547

Ivan Luciano Danesi, Steven Haberman, Pietro Millossovich: Mortality forecasting for related populations using Lee-Carter type models 551

Maria del Pilar Díaz, José M. Vargas, Margarita Díaz: Multilevel factor models: Identification of Three-level Model Parameters for the Study of Regional Development in Argentina ... 555

Vinny Davies, Dirk Husmeier: Assessing the impact of non-additive noise on modelling transcriptional regulation with Gaussian processes. 559

Margaret R. Donald, Ashwin Unnikrishnan, John E. Pimanda, Susan R. Wilson: Model Comparisons for RNA-Seq data 563

Lizandra C. Fabio, Francisco José A. Cysneiros, Gilberto A. Paula: Marginal models from exponential family mixed models with nonormal random effect distribution 567

Zhou Fang: Sparse Penalised Methods in Phenology 573

Susana Faria, Arminda Manuela Gonçalves, Rui Gomes: Modelling Occupational Stress and Burnout in Portuguese University Teachers by using Structural Equation Models 577

Susana Faria, Gilda Soromenho: Measuring the component overlapping in mixtures of linear regressions 581

Alessandro Fassò: Traffic policies and air quality in Italian cities 585

Giancarlo Ferrara, Francesco Vidoli: Beyond the threshold: the efficiency of Italian manufacturing firms..... 591

Álvaro J. Flórez, Ana Nora A. Donaldson, Mercedes Andrade, Javier Torres, Nairn Wilson: Handling missing data in longitudinal studies: an application to healthcare data..... 595

Agnes Fussl, Sylvia Frühwirth-Schnatter, Christine Zulehner: Bayesian estimation of a discrete choice model for household labor supply in Austria 599

Kelly Gallacher, Claire Miller, Marian Scott: Comparison of GAM, FDA and DFA for water quality data 603

Jan Gertheiss, Henk A.L.Kiers: Penalized Non-Linear Principal Components Analysis for Ordinal Variables 607

Viviana Giampaoli, Olga Usuga, Patricia Bertolotto: Selection of mixed beta regression model for longitudinal data..... 611

Gustavo L. Gilardoni, Maria Luiza G. De Toledo, Marta A. Freitas, Enrico A. Colosimo : Dynamics of the Optimal Maintenance Policy under Imperfect Repair Models..... 615

Dulce Gomes, Patrícia A. Filipe, Carla Nunes, Bruno de Sousa: Penalized spline smoothing for delay in Pulmonary Tuberculosis diagnosis..... 621

Arminda Manuela Gonçalves, Marco Costa, Lara Teixeira: Change-point analysis for in environmental time series.... 625

Radek Hendrych: Another View on Conditional Correlations 631

Lasse Holmström, Ilkka Launonen: Posterior Singular Spectrum Analysis (PSSA)..... 635

Samuel Iddi, Geert Molenberghs: A Zero-Inflated and Overdispersed Marginalized Model for Correlated Counts..... 641

Nadja Klein, Thomas Kneib, Stefan Lang: Bayesian generalized additive models for location, scale and shape for insurance data..... 645

Bernhard Klingenberg: On a null variance estimator for the Mantel-Haenszel risk difference and corresponding confidence interval 651

Arnošt Komárek, Tomáš Kincl, Lenka Komárková: Model based segmentation of TV advertising scheduling patterns 655

Antonio J. López-Montoya, Concepción Azorit, Irene García-Garrido, Ramón Gutiérrez, Javier Moro: Statistical Models to Density-dependence Detection in Mediterranean Deer Populations 659

Mónica López-Ratón, Carmen Cadarso-Suárez, Elisa M. Molanes-López, Emilio Letón: GsymPoint: An R Package for estimating the Generalized Symmetry Point as the optimal cutpoint in continuous diagnostic tests 663

Francisco Louzada, Vicente G. Cancho, Glagys D.C. Barriga, Dipak K. Dey: The Birnbaum–Saunders survival model with cure fraction under different of activation mechanisms 669

Benn Macdonald, Frank Dondelinger, Dirk Husmeier: Inference in complex biological systems with Gaussian processes and parallel tempering. 673

Ivana Malá: Modelling of the distribution of incomes with the use of finite mixtures of distributions 677

Valentina Mameli, Monica Musio, Luca Deiana: Study of the longevity in Sardinia: an application of the Beta skew-normal regression 681

Elizabeth Martínez-Gómez, Victor Guerrero, Francisco Estrada: An Application of The Vector AutoRegression (VAR) Model to The Analysis of The Sun–Earth's Climate Connection. 685

Larissa A. Matos, Victor H. Lachos: Influence diagnostics in mixed-effects models with censored data using the multivariate-t distribution 689

Silvia Metelli, Leonardo Grilli, Carla Rampichini: Bayesian estimation with INLA for logistic multilevel models 693

Radoslava Mirkov, Thomas Maul, Ronald Hochreiter: Modeling Credit Spreads Using Nonlinear Regression 697

Ana Moreira, Luís Machado: Estimation of the conditional survival function for successive survival times 701

Darcy Steeg Morris: A Semiparametric Approach for Multivariate Longitudinal Count Data 707

Jesús Navarro-Moreno, Rosa M. Fernández-Alcalá, Juan Carlos Ruiz-Molina, Antonia Oya: A Quaternion Widely Linear Series Expansion 713

Federica Nicolussi: Smooth Graphical models of type II: link with marginal models 717

Luigi Pace, Alessandra Salvan, Nicola Sartori: Adjusted pseudo composite likelihood ratios 723

Juan Carlos Pardo-Fernández, M. Dolores Jiménez-Gamero, Anouar El Ghouch: On the use of the characteristic function of the residuals to test for the equality of regression curves 727

Daniela Pauger, Christine Duller, Helga Wagner: Analysing Formalisation of Management Accounting by Bayesian Variable Selection in a Cumulative Logit Model 731

Gleici Castro Perdona, Francisco Louzada, Cleyton Zanardo, Hayala Cavenague: A New Weibull Family of Hazard Models for Breast Cancer Survivals 735

Anthony N. Pettitt, Xing Ju Lee: Approximate Bayesian Computation for Model Choice 739

Christian Pfeifer, Achim Zeileis, Peter Höller: Trend and regional analysis of fatal off-piste and backcountry avalanche accidents in Austria within the years 1968 and 2011 743

Aluísio Pinheiro, Pranab Kumar Sen: Functional Data Analysis via Quasi U-Statistics Based Tests 749

Holger Reulen, Thomas Kneib: Boosting Multi State Models .. 753

Silvia Rizzi, Jutta Gampe, Paul H.C. Eilers: Efficient ungrouping of coarse histograms with the penalized composite link model 757

Antonio José Sáez-Castillo, Antonio Conde-Sánchez, Ana María Martínez-Rodríguez, María José Olmo-Jiménez, José Rodríguez-Avi: A hyper-Poisson regression model for zero-truncated count data 761

Stuart J. Sharples, Deborah A. Costain, Chris Sherlock: Predicting future offending in adolescents from a longitudinal survey with missing responses 765

Danilo A. Silva, Cibele M. Russo: Comparison of estimation methods for variance components in elliptical mixed effects models 769

Koen Simons, Kaatje Bollaerts, Michel Sonck, Sébastien Fierens, André Poffijn, Lodewijk Van Bladel, David Geraerts, Pol Gosselin, Herman Van Oyen, Julie Francart, An Van Nieuwenhuyse: Childhood Leukaemia incidence around the Belgian nuclear sites: Surrogate exposure modelling . 773

Slawomir Smiech, Monika Papiez: Exploratory data analysis of energy security in the EU member countries in the period 2000-2010 779

Katia Stefanova: Statistical model for multi-environment trials: Accounting for variety by environment interaction 783

Isabella Sulis, Vincenza Capursi: Analyzing SET over time using multilevel multidimensional explanatory IRT models 789

Isabella Sulis, Mariano Porcu: Handling missing data in Item Response Theory. Assessing the accuracy in estimation of two multiple imputation procedures 795

Karin A. Tamura, Viviana Giampaoli, Alexandre Noma: Nearest Neighbors Prediction Method for mixed logistic regression 799

Célia Touraine, Pierre Joly: Illness-death model for interval-censored and left-truncated data with random effects: Application to dementia 803

Gerhard Tutz, Margret-Ruth Oelker: Modeling heterogeneity by fixed effects models 807

Yannick Vandendijck, Christel Faes, Hens Niel: Weight smoothing models to estimate survey estimates from binary data 811

Luis Hernando Vanegas, Gilberto A. Paula: Symmetric and log-symmetric regression models: a semiparametric approach 815

Alicja Wolny-Dominiak: A Score Test for Zero-adjusted Effect in Claim Severity Modeling 819

Bruce J. Worton, Chris R. McLellan: Hidden Markov modelling of diffusion with an application in entomology 823

Emily Yeend, Deborah A. Costain, Karen Broadhurst: Modelling Children's Journeys in Care: A Multistate Modelling Approach 827

Author Index 831

Part I - Invited Papers

Science to data to statistical models

Ciprian M. Crainiceanu[1], Vadim Zipunnikov[1], Jiawei Bai[1]

[1] Johns Hopkins University, USA

E-mail for correspondence: ccrainic@jhsph.edu

Abstract: Due to emerging technologies data have become abundant and ubiquitous with rapidly changing size, structure, and complexity. This paper introduces two new major data structures: 1) longitudinal ultra high dimensional brain imaging; and 2) multivariate high density wearable computing. We will argue that these types of data open entirely new areas of research. We will then follow the argument to its tautological conclusion: statistical research should start from science, ask the right questions about the data, and then do modeling.

Keywords: High resolution structural imaging; Wearable computing.

1 Diffusion tensor imaging along the corpus callosum

Multiple sclerosis, an autoimmune disease characterized by neuronal demyelination and white matter lesions, leads to significant disability in patients. A hallmark of MS is damage to and degeneration of the myelin sheaths that surround and insulate nerve fibers in the brain. Such damage results in sclerotic plaques that distort the flow of electrical impulses along the nerves (Raine et al., 2008). Other manifestations of the disease includes accelerated brain atrophy and lesion formation. Magnetic Resonance Imaging (MRI) is a first line scientific approach that could help quantify these changes. Diffusion tensor imaging (DTI) is a magnetic resonance imaging (MRI) based modality that traces the diffusion of water in the brain. Because water has specific anisotropic diffusion characteristics in brain white matter, DTI is used to generate detailed images of the white matter (Basser et al., 1994, 2000; LeBihan et al., 2001, Mori and Barker, 1999). Several measurements of water diffusion are provided by DTI, including fractional anisotropy (FA), mean and parallel diffusivity. Here we will focus on FA of the corpus callosum, the large white matter fiber tract connecting the two brain hemispheres. FA is thought to be a measure of tissue integrity and be sensitive to both axon fiber density and myelination in white matter (Mori, 2007). To build up intuition Figure 1 displays a cuboid (transparent blue hue) surrounding the corpus callosum, which bears a slight resemblance to a carpet with the ends folded. The figure contains FA measurements for more than $30,000$ voxels (three dimensional version of a pixel), where measurements correspond to various colors. FA is a measure between 0, which

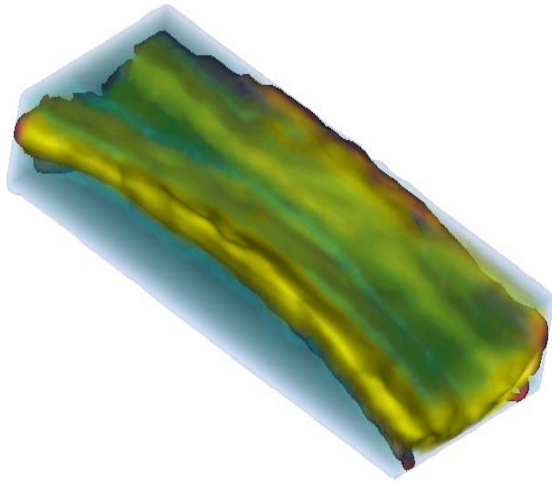


FIGURE 1. Fractional anisotropy (FA) along the Corpus Callosum. Darker shades of red: higher values of FA (FA=1 is the highest value and corresponds to perfect anisotropic movement of water molecules.) Darker shades of green: smaller values of FA (FA=0 is the smallest value and corresponds to perfect anisotropic movement of water molecules.)

corresponds to perfectly isotropic movement of water molecules (Brownian motion), and 1, which corresponds to perfectly anisotropic movement of water molecules. Darker shades of red represent higher values of FA, while darker shades of green correspond to smaller values of FA. This is one FA image at one visit for one subject. We are interested in studying the structure of the FA data across subjects and visits and its association with health outcomes.

Consider the following on-going study of multiple sclerosis (MS) patients (Reich et al., 2010). Data are derived from a natural history study of 176 MS cases drawn from a wide spectrum of disease severity. Subjects were scanned over a 6-year period up to 10 times per subject, for a total of 466 scans. The scans have been aligned using a 12 degrees of freedom transformation, meaning that we accounted for rotation, translation, scaling, and shearing, but not for nonlinear deformation. In addition to DTI data the study collected demographic and health data, including cognitive outcomes. In particular, at every visit the Paced Auditory Serial Addition Test (PASAT) was collected. PASAT is a commonly used examination of cognitive function affected by MS with scores ranging between 0 and 60.

To better understand the problem, Figure 2 displays the structure of the study where subjects are observed longitudinally and both brain DTI (see displayed 3D FA maps) and outcomes are obtained at the same visits. The study contains all the problems associated with longitudinal stud-

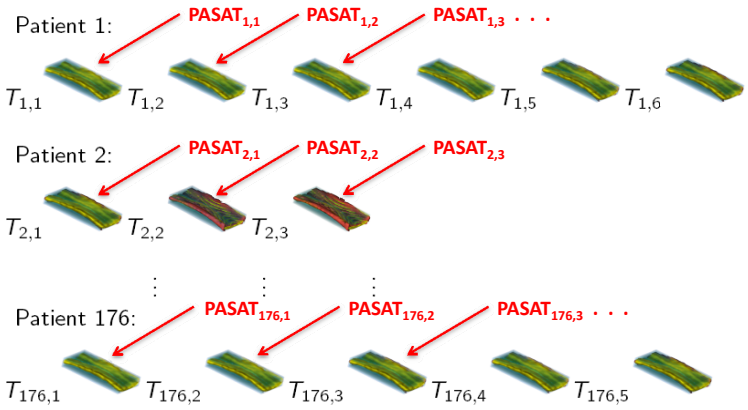


FIGURE 2. Fractional anisotropy (FA) along the Corpus Callosum observed longitudinally for 176 subjects. At each visit covariates and cognitive outcomes,including PASAT, are measured.)

ies: visits are at different subject-specific time intervals, a-priori unknown cross-sectional and dynamic behavior of variables, and heterogeneous noise structure. The major difference from standard longitudinal problems is that the exposure variable, in our case DTI imaging, is ultra-high dimensional. This type of structure has become increasingly common, while Statistical methods have lagged behind.

The data structure is of the type $[Y_{ij}, T_{ij}, \mathbf{Z}_{ij}, \{W_{ij}(v) : v \in 1, \ldots, V\}]$, where Y_{ij} is the outcome collected at time T_{ij} for subject $i = 1, \ldots, I$ at visit $j = 1, \ldots, J$. At every visit a $p \times 1$ vector of covariates is collected in addition to ultra high dimensional data, $\{W_{ij}(v) : v \in 1, \ldots, V\}$. In our case we only look at one high-dimensional exposure, though, in general, there may be several. Here we work with the vectorized image, that is the image obtained by unfolding the cuboid in Figure 1 according to the same, pre-specified, unfolding rule for all subjects and visits. In our study V is very large ($> 30,000$) and $\{W_{ij}(v) : v \in 1, \ldots, V\}$ is highly structured.

1.1 Models for longitudinal high dimensional data

We first focus on modeling the longitudinal structure of the ultra-high dimensional (UHD) process $\{W_{ij}(v) : v \in 1, \ldots, V\}$ and start with the general longitudinal UHD model (Greven et al., 2010; Zipunnikov et al., 2013)

$$W_{ij}(v) = \eta(v, T_{ij}, \mathbf{Z}_{ij}) + X_{i,0}(v) + T_{ij} X_{i,1}(v) + U_{ij}(v), \qquad (1)$$

where $\eta\{v, T_{ij}, \mathbf{Z}_{ij}\}$ is the fixed effects function, $X_{i,0}(v)$ is the image random-intercept, $X_{i,1}(v)$ is the image random slope, and $U_{ij}(v)$ is the visit-to-

visit exchangeable variability. More complex structures could be considered, though here we are primarily concerned with interpretability and ease of presentation.

The fixed effects part of model (1) contains many practically relevant models. For example, $\eta(v, T_{ij}, \mathbf{Z}_{ij}) = \alpha + T_{ij}\beta + \mathbf{Z}_{ij}^{\mathrm{T}}\gamma$ assumes that the function does not depend on the voxel location, v, and is a linear function in the covariates $(T_{ij}, \mathbf{Z}_{ij})$. This is the exact linear form of the fixed effects part of a standard linear mixed model. A slightly more complex model is $\eta(v, T_{ij}, \mathbf{Z}_{ij}) = \mu(v) + T_{ij}\beta + \mathbf{Z}_{ij}^{\mathrm{T}}\gamma$, where $\mu(v)$ is the overall mean response. In the case when there is treatment or group indicator variable, say D_i, the model can be expanded as $\eta(v, T_{ij}, \mathbf{Z}_{ij}) = \mu(v) + d(v)D_i + T_{ij}\beta + \mathbf{Z}_{ij}^{\mathrm{T}}\gamma$ to include the treatment effect, $d(v)$, which is allowed to vary with v. In spite of the importance of the fixed effects function there is very little work on general and robust methods for estimation and inference. Here we follow the simple recipe introduced by Crainiceanu et al. (2012) for estimating the difference in the means of two correlated processes. More specifically, we suggest to estimate $\eta(v, T_{ij}, \mathbf{Z}_{ij})$ under the independence assumption and obtain the distribution of the estimator using a bootstrap of subjects. We found this approach to often outperform joint inferential or pre-whitening methods that use covariance operators estimators.

Once an estimator of $\eta(v, T_{ij}, \mathbf{Z}_{ij})$ is available, we suggest focusing on the residuals $\widehat{W}_{ij}(v) = W_{ij}(v) - \widehat{\eta}(v, T_{ij}, \mathbf{Z}_{ij})$. The model for the residuals is, approximately, $\widehat{W}_{ij}(v) = X_{i,0}(v) + T_{ij}X_{i,1}(v) + U_{ij}(v)$. This model has the same form as the longitudinal functional principal component analysis (LFPCA) model introduced by Greven et al. (2010), which was a generalization of the multilevel functional principal component analysis (MFPCA) introduced by Di et al. (2009) for replicated functional data. This is the direct generalization of random intercept random slope models to the case when the observed data are UHD. It can also be viewed as an extreme case of multivariate mixed effects models. Given the difficulties associated with fitting mixed effects models, there is a need to better understand what such models can reveal about the data and exactly how to estimate them.

One can interpret $X_{i,0}(v)$ as the true baseline image. The proxy measurement, $\widehat{W}_{i1}(v)$, is corrupted by the visit-to-visit measurement error, $U_{i1}(v)$, which can be very large. Indeed, $U_{i1}(v)$ contains technical and biological variability including normal changes between replications, registration errors, changes in protocol, unexpected interactions between new data and pre-processing software, etc. Note that the relative temporal change in images between visits

$$\{\widehat{W}_{ij+1}(v) - \widehat{W}_{ij}(v)\}/(T_{ij+1} - T_{ij}) = X_{i,1}(v) + \{U_{ij+1}(v) - U_{ij}(v)\}/(T_{ij+1} - T_{ij}),$$

is a proxy of the unknown subject-specific slope $X_{i,1}(v)$. However, these differences are considerably noisier proxies of the slope, as for a one unit in time change, $T_{ij+1} - T_{ij} = 1$, the covariance of $U_{ij+1}(v) - U_{ij}(v)$ is equal

to $2K_U$, where K_U is the covariance of $U_{ij}(v)$.

Greven et al. (2010) provided a principal component approach to fitting this type of data. More precisely, they considered the mutually independent processes $\mathbf{X}_i(v) = \{X_{i,0}(v), X_{i,0}(v)\}$ and $U_{ij}(v)$ and estimated their covariance operators K_X and K_U. The Karhunen-Loéve expansion can then be used to expand the processes $\mathbf{X}_i(v) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k^X(v)$ with $\phi_k^X(v) = \{\phi_k^{X,0}(v), \phi_k^{X,1}(v)\}$ and $U_{ij}(v) = \sum_{l=1}^{\infty} \zeta_{ijl}\phi_l^U(v)$, where $\phi_k^X(v)$ and $\phi_l^U(v)$ are the eigenvectors of K_X and K_U, respectively. The LFPCA model can thus be written as

$$\widehat{W}_{ij}(v) = \sum_{k=1}^{\infty} \xi_{ik}(1, T_{i,j})\phi_k^X(v) + \sum_{l=1}^{\infty} \zeta_{ijl}\phi_l^U(v), \tag{2}$$

where ξ_{ik}, ζ_{ijl} are zero-mean mutually independent random variables with variance λ_k^X and λ_l^U, respectively. This model is reasonable in the case when there are a small number of principal components that explain the observed variability at the two levels. In these cases the two sums in the equation (2) are truncated to K and L, respectively. The truncation values are chosen using different approaches, though here we focus on thresholds on cumulative variance explained. We contend that model (2) has at least three important applications. First, it provides a structured decomposition of variability that respects the known structure of the data. Second, it provides a starting point for clustering algorithms for longitudinal high dimensional data using principles similar to those used in standard mixed effects models. Third, it provides a powerful dimensionality reduction that allows simple plug-in approaches to outcome regression.

A brute force approach to LFPCA would require calculation and diagonalization of $V \times V$ dimensional matrices. This "small detail" can derail even the best written models for UHD. Zipunnikov et al. (2012) and Zipunnikov et al. (2013) proposed a simple approach to diagonalize these operators without storing or calculating them. Here we describe the fundamental idea, henceforth referred to as the *lossless projection approach* (LPA). To better understand LPA, consider the vectorized form of the model

$$\widehat{\mathbf{W}}_{ij} = \mathbf{X}_{i0} + T_{ij}\mathbf{X}_{i1} + \mathbf{U}_{ij},$$

where the vectors (in bold) are simply obtained by binding the corresponding observations at all voxels $v = 1, \ldots, V$; all these vectors are of dimension $V \times 1$. Consider now any $M \times V$ matrix $\mathbf{A}$. By pre-multiplication to the left by $\mathbf{A}$ the model becomes

$$\mathbf{A}\widehat{\mathbf{W}}_{ij} = (\mathbf{A}\mathbf{X}_{i0}) + T_{ij}(\mathbf{A}\mathbf{X}_{i1}) + (\mathbf{A}\mathbf{U}_{ij}),$$

which has exactly the same the form because the model is linear. The main difference is that the vectors are $M \times 1$ dimensional and if $M << V$ then we have reduced a linear model for UHD to a low dimensional

linear mixed effects model. Note that the parameters are unchanged in the low dimensional model, which can be used for estimation and inference *irrespective to the choice of* $\mathbf{A}$.

An extreme case is $M = 1$, though this could oversimplify the data structure when the dimension of the subspace spanned by the data columns $\widehat{\mathbf{W}}_{ij}$ is much larger. Thus, it makes sense to consider the $V \times N$ dimensional matrix $\widehat{\mathbf{W}}$ obtained by column binding all the vectors $\widehat{\mathbf{W}}_{ij}$. Here N is the total number of visits for all subjects and $N << V$; for example, in our DTI application $N = 466$ and $V > 30,000$. Consider now the singular value decomposition $\widehat{\mathbf{W}} = \mathbf{SVD}^{\mathrm{T}}$, where $\mathbf{S}$ is the $V \times N$ matrix containing the left-singular eigenvectors corresponding to non-zero eigenvalues. With the choice $\mathbf{A} = \mathbf{S}^{\mathrm{T}}$ the projection is lossless in the sense that data vectors $\mathbf{SS}^{\mathrm{T}}\widehat{\mathbf{W}}_{ij} = \widehat{\mathbf{W}}_{ij}$ in spite of the fact that $\mathbf{SS}^{\mathrm{T}} \neq \mathbf{I}_V$. Since the singular value decomposition can be done in linear time in V the entire fitting procedure is very fast (Zipunnikov et al., 2012; Zipunnikov et al., 2013).

1.2 Statistical models for scalar-on-image regression

Another very important problem in this context is to study the association between outcomes, Y_{ij}, scalar covariates, $\mathbf{Z}_{ij}$, and UHD covariates, $W_{ij}(v)$. Probably the easiest approach is to use the high-dimensional LF-PCA decomposition discussed in Section 1.1 and use the scores ξ_{ik} and ζ_{ijl} as regressors. In the case when imaging data is not observed longitudinally one can use principal component regression. Another class of regression approaches starts from voxel-wise regression, that is perform V independent mixed effects regressions, one for every voxels. An improved version of these regressions is the locally optimal voxel estimation (LOVE) regression approach (Sweeney et al., 2013a,b), which considers small neighborhoods around the voxel as a multivariate exposure model. LOVE accounts for the local correlations, though it does not account for long range correlations, brain homotopy (symmetry), or biological correlations.

A different approach was introduced for cross-sectional scalar-on-image regression by Goldsmith et al. (2013). Consider the baseline data $[Y_i, \mathbf{Z}_i, \{W_i(v) : v \in 1, \ldots, V\}]$ (note the absence of the time variable and index j.) One could conceptualize the scalar-on-image model as a massively multivariate regression model

$$Y_i = \alpha + \mathbf{Z}_i^{\mathrm{T}}\gamma + \sum_{v=1}^{V} W_i(v)\beta_v + \epsilon_i,$$

where $\beta = (\beta_1, \ldots, \beta_V)^{\mathrm{T}}$ is a vector of coefficients for the image predictor $\mathbf{W}_i$. The main advantage of this conceptual framework is that it estimates the effect of data at every voxel while *correcting for the effect of all the other voxels*; the voxel-wise regression conducts V regressions with the effect of each voxel not being corrected for the effect of the other voxels.

While appealing, fitting the multiple linear regression model is an ill-posed problem. Indeed, the dimension of the matrix $\mathbf{W}$ is $I \times V$ with $I << V$. A standard way to make the solution identifiable is penalized regression, that is minimize a penalized least squares criterion of the type

$$(\widehat{\alpha}, \widehat{\gamma}, \widehat{\beta}) = \text{argmin}_{\alpha,\gamma,\beta}\{\sum_{i=1}^{I}(Y_i - \alpha - \mathbf{Z}_i^{\text{T}}\gamma - \mathbf{W}_i\beta)^2 + P(\beta)\}, \qquad (3)$$

where the penalty $P(\beta)$ is chosen to yield a solution to equation (3) with desirable properties. As shown in Huang et al. (2013), model (3) is statistically equivalent to the following model where the β coefficients are treated as random

$$\begin{cases} Y_i & \sim & N(\alpha + \mathbf{Z}_i^{\text{T}}\gamma + \mathbf{W}_i\beta, \sigma_\epsilon^2); \\ \beta & \sim & \exp\{-P(\beta)/2\}. \end{cases} \qquad (4)$$

The second line of the model means that β has a density function proportional to $\exp\{-P(\beta)/2\}$, where the normalizing constant is omitted. For specific forms of the penalty the distribution $\exp\{-P(\beta)/2\}$ may not be integrable (i.e. the integral of the prior values may not be finite); this need not be a problem if the posterior distribution $[\beta|\text{data}]$ is proper. Goldsmith et al. (2013) used a Casing (CAR + Ising) prior which combines an "activation map" controlled by the spatial Ising prior with a Conditionally Autoregressive (CAR) prior for the β coefficients that induces spatial smoothing among the β coefficients declared to be non-zero.

2 Wearable computing

We now change focus to a completely different type of technology, wearable computing, which is re-shaping our daily life and has the potential to change public health research. Indeed, understanding public health is fundamentally linked to environmental exposures "that go in" (e.g. food, water, air, social interactions) and "that come out" (e.g. activity, perspiration, urine) the human body. Ironically, we do not have good measurements of any of these processes. Wearable computing holds the promise to change this. For example, in the search for objective measurements of physical activity, researchers have increasingly relied on accelerometers in observational studies and clinical trials (Bai et al., 2012; Culhane et al., 2005; Grant et al., 2008; Troiano et al., 2008). A triaxial accelerometer is a wearable electromechanical sensor that records ultra-high density real-time dynamic accelerations in three mutually orthogonal directions. Accelerometers are relatively small in size and can be attached to different parts of the human body. A fundamental question is how to decipher and interpret the acceleration signals into meaningful information such as duration, intensity, and type of physical activity.

To get an idea about the complexity and richness of the signal, Figure 3 displays the raw data from a three-axial accelerometer worn at the hip.

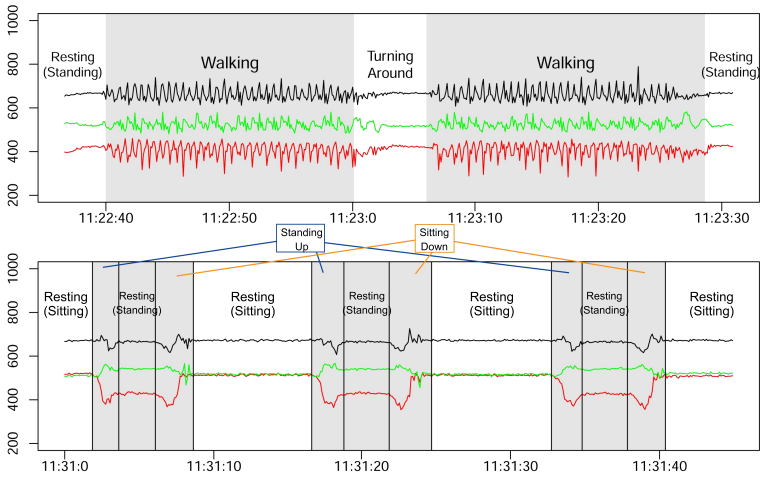


FIGURE 3. Raw data obtained from a three-axial accelerometer worn at the hip. Each axis produces a different time series (shown in black, green and red). First panel: data for normal walking followed by turning around and walking back during a roughly one minute interval. Second panel: three repetitions of resting sitting followed by standing up, resting standing, and sitting down periods.

The sampling frequency is 10Hz and the acceleration along each of the three orthogonal axes is shown in a different color (black, red, and green). Data were collected in the lab and annotated by a human observer. The first panel displays an activity period of roughly 1.5 minutes that starts with the person resting standing, walking, turning around, walking back, and resting standing. The second panel displays data for three repetitions of the following sequence of activities: resting sitting, standing up from sitting, resting standing, sitting down on a chair.

Consider the problem of recognizing such types of movements when they are not labeled by a human observer. Inspection of the data suggests that this should be possible. Indeed, it is reassuring that visual inspection of the data does not reveal any obvious problems: that is, what we know and what the accelerometer shows do not disagree in any fundamental way. Moreover, the repetitive, higher frequency characteristic of walking seems to be well represented in the data; note that one can visually identify each step taken. The slower frequency movement associated with standing up from a chair also does not exhibit any big surprises and shows a great level of consistency across the three repetitions. Most interestingly, a human observer inspecting these plots should be able to spot these patterns in the data. A fundamental problem that we face is to instruct the computer to do the "spotting" for us. This prediction problem is conceptually easy, though far from trivial. It may be seductive to think about the parallels with

speech recognition, though this is a much harder problem because: 1) speech has evolved to transmit information whereas body movement has evolved to get humans to their destination and out of trouble's way; 2) sound data patterns are virtually the same in all directions, whereas acceleration depends both on accelerometer placement, orientation as well as human body geometry; and 3) speech happens at much higher frequencies (in the kilohertz range) versus movement (in the hertz range) and has much higher signal-to-noise ratios.

Movelets, is a simple idea for prediction within the same subject proposed by Bai et al. (2012). The idea has two parts: 1) take the subject specific data, partition it into all possible overlapping 1-second intervals; and 2) for every 1-second interval predict the type of activity as the activity type of the 1-second interval whose time series is closest. Once the library of annotated movements is available at the subject level, prediction is very fast. There are many open problems that are related to this new types of data. Here we provide a non-exhaustive list.

1. Extend movelets to multiple accelerometers and devices; a first paper on this topic is available online (He et al., 2013).

2. Develop subject-level prediction without same-subject annotated data.

3. Normalization of measurements with specific emphasis on interpretability and signal extraction; moving away from black-box definitions such as "activity counts" to open source, reproducible definitions, algorithms, and software.

4. Parameterize accelerometer data corresponding to activities like "standing up from a chair" and "walking" with the ultimate goal of developing biomarkers that are sensitive to subtle changes in the ability to move.

5. Develop population level analytic methods for high density "activity intensity" and "activity type" data.

3 Discussion

The most important message of our paper is not how to analyze the types of data that we have introduced here. In fact, a critical view of our approaches is healthy and, ultimately, rewarding. In fact, here we would like to turn the table and ask the reader: how would you analyze the data? We believe that these are hard problems and principled, skeptical, and simple approaches to inference will ultimately be the most productive.

References

Bai, J., Goldsmith A.J., Caffo, B.S., Glass, T.A., Crainiceanu, C.M. (1994). Movelets: A dictionary of movement. *Electronic Journal of Statistics*, **2012**, 559 – 578.

Basser, P., Mattiello, J., and LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, **66**, 259 – 267.

Basser, P., Pajevic, S., Pierpaoli, C., and Duda, J. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine*, **44**, 625 – 632.

Crainiceanu, C.M., Staicu A.M., Ray, S., Punjabi, N.M. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in Medicine*, **31**, 3223 – 3240.

Culhane, K.M. , OConnor, M., Lyons, D., and Lyons, G. M. (2005). Accelerometers in rehabilitation medicine for older adults. *Age and Ageing*, **34**, 556 – 560.

Di, C.D., Crainiceanu, C.M., Caffo, B.S., Punjabi, N.M. (2009). Multilevel Functional Principal Component Analysis. *The Annals of Applied Statistics*, **3**, 458 – 488.

He, B. , Bai, J., Koster, A., Casserotti, P., Glynn, N., Harris, T.B., Crainiceanu, C.M. (2013). Predicting human movement type based on multiple accelerometers using movelets. *under review*

LeBihan, D., Mangin, J., Poupon, C., and Clark, C. (2001). Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, **13**, 534 – 546.

Grant, P.M., Dall, P.M., Mitchell, S.L. and Granat, M.H. (2008). Activity-monitor accuracy in measuring step number and cadence in community-dwelling older adults. *Journal of Aging and Physical Activity*, **16**, 204 – 214.

Greven, S., Crainiceanu, C.M., Caffo, B.S., Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, **4**, 1022 – 1054.

Goldsmith, J.A., Huang, L., Crainiceanu, C.M. (2013). Smooth scalar-on-image regression via spatial Bayesian selection. *Journal of Computational and Graphical Statistics, to appear*

Huang, L. , Goldsmith, J.A., Reiss, P.T., Reich, D., Crainiceanu, C.M. (2013). Bayesian Scalar-on-Image Regression with Application to Association Between Intracranial DTI and Cognitive Outcomes. *under review*

Mori, S. and Barker, P. (1999). Diffusion magnetic resonance imaging: its principle and applications. *The Anatomical Record*, **257**, 102 – 109.

Mori, S. (2007). *Introduction to Diffusion Tensor Imaging*. Elsevier.

Sweeney, E., Shinohara, R.T., Shea, C., Reich, D., Crainiceanu, C.M. (2013a). Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRIs. *American Journal of Neuroradiology*, **34**, 68 – 73.

Sweeney, E., Shinohara, R.T., Shie, N., Mateen, F., Chudgar, A., Cuzzocreo J., Calabresi, P., Pham, D., Reich, D., Crainiceanu, C.M. (2013b). OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI. *NeuroImage Clinical, to appear*.

Raine, C.S., McFarland, H., and Hohlfeld, R. (2008). *Multiple Sclerosis: A Comprehensive Text*. Saunders Ltd.

Reich, D.S., Ozturk, A., Calabresi, P.A., and Mori, S. (2010). Automated vs conventional tractography in multiple sclerosis: Variablity and correlation with disability. *NeuroImage*, **49**, 3047 – 3056.

Troiano, R.P. , Berrigan, D., Dodd, K.V., Mâsse, L.C., Tilert, T., and McDowell, M. (2008). Physical activity in the united states measured by accelerometer. *Medicine and Science in Sports and Exercise*, **40**, 181 – 188.

Zipunnikov, V., Caffo, B.S., Davatzikos, C., Schwartz, B., Crainiceanu, C.M. (2011). Multilevel functional principal component analysis for high dimensional data. *Journal of Computational and Graphical Statistics*, **20**, 852 – 873.

Zipunnikov, V., Greven, S., Caffo, B.S., Reich, D., Crainiceanu, C.M. (2013). Longitudinal high-dimensional data analysis. *Under review*.

Conditional Transformation Models by Example

Torsten Hothorn[1], Thomas Kneib[2], Peter Bühlmann[3]

[1] Universität Zürich, Switzerland
[2] Universtität Göttingen, Germany
[3] Eidgenössische Technische Hochschule Zürich, Switzerland

E-mail for correspondence: Torsten.Hothorn@uzh.ch

Abstract: The ultimate goal of regression analysis is to obtain information about the conditional distribution of a response given a set of explanatory variables. This goal is, however, seldom achieved because most established regression models only estimate the conditional mean as a function of the explanatory variables and assume that higher moments are not affected by the regressors. The underlying reason for such a restriction is the assumption of additivity of signal and noise. This common assumption can be relaxed in the framework of transformation models. A novel class of semiparametric regression models proposed by Hothorn et al. (2013, JRSS-B) allows transformation functions to depend on explanatory variables. These transformation functions are estimated by regularised optimisation of scoring rules for probabilistic forecasts. Conditional transformation models are potentially useful for describing possible heteroscedasticity, comparing spatially varying distributions, identifying extreme events, deriving prediction intervals and selecting variables beyond mean regression effects. Here, we'll illustrate conditional transformation models by applications from different domains.

Keywords: Boosting, proper scoring rules, conditional distributions, transformation models

1 A Primer on Conditional Transformation Models

Hothorn et al. (2013b) proposed a new class of conditional transformation models that allow the conditional distribution function $\mathbb{P}(Y \leq \upsilon | \boldsymbol{X} = \boldsymbol{x})$ to be estimated directly and semiparametrically under rather weak assumptions. Here, we will illustrate this class of regression models by reanalysing regression problems from different research domains.
To fix ideas, let $Y_{\boldsymbol{x}} = (Y | \boldsymbol{X} = \boldsymbol{x}) \sim \mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ denote the conditional distribution of response Y given explanatory variables $\boldsymbol{X} = \boldsymbol{x}$. We assume that $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$ is dominated by some measure μ and has the conditional distribution function $\mathbb{P}(Y \leq \upsilon | \boldsymbol{X} = \boldsymbol{x})$. A regression model describes the

distribution $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}$, or certain characteristics of it, as a function of the explanatory variables $\boldsymbol{x}$. We estimate such models based on samples of pairs of random variables $(Y, \boldsymbol{X})$ from the joint distribution $\mathbb{P}_{Y,\boldsymbol{X}}$. It is convenient to assume that a regression model consists of signal and noise, *i.e.* a deterministic part and an error term. In the following, we denote the error term by $Q(U)$, where $U \sim \mathcal{U}[0,1]$ is a uniform random variable independent of $\boldsymbol{X}$ and $Q : \mathbb{R} \to \mathbb{R}$ is the quantile function of an absolutely continuous distribution.

An attractive feature of transformation models is their close connection to the conditional distribution function. With the transformation function $h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = Q(U)$, one can evaluate the conditional distribution function of response Y given the explanatory variables $\boldsymbol{x}$ via

$$\mathbb{P}(Y \leq \upsilon|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(h(Y|\boldsymbol{x}) \leq h(\upsilon|\boldsymbol{x})) = F(h(\upsilon|\boldsymbol{x}))$$

with absolute continuous distribution function $F = Q^{-1}$. For additive transformation functions $h = h_Y + h_{\boldsymbol{x}}$, the conditional distribution function reads $F(h(\upsilon|\boldsymbol{x})) = F(h_Y(\upsilon) + h_{\boldsymbol{x}}(\boldsymbol{x}))$, *i.e.* the distribution is evaluated for a transformed and shifted version of Y. Higher moments only depend on the transformation h_Y and thus cannot be influenced by the explanatory variables. Consequently, one has to avoid the additivity in the model $h = h_Y + h_{\boldsymbol{x}}$ to allow the explanatory variables to impact also higher moments. Hothorn et al. (2013b) therefore suggest a novel transformation model based on an alternative additive decomposition of the transformation function h into J partial transformation functions for all $\boldsymbol{x}$:

$$h(\upsilon|\boldsymbol{x}) = \sum_{j=1}^{J} h_j(\upsilon|\boldsymbol{x}), \tag{1}$$

where $h(\upsilon|\boldsymbol{x})$ is the monotone transformation function of υ. In this model, the transformation function $h(Y_{\boldsymbol{x}}|\boldsymbol{x})$ and the partial transformation functions $h_j(\cdot|\boldsymbol{x}) : \mathbb{R} \to \mathbb{R}$ are conditional on $\boldsymbol{x}$ in the sense that not only the mean of $Y_{\boldsymbol{x}}$ depends on the explanatory variables. For this reason, Hothorn et al. (2013b) coined models of the form (1) *Conditional Transformation Models* (CTMs).

Conditional transformation models are fitted by direct minimization of a proper scoring rule. The loss function ℓ for estimating conditional transformation models is defined as integrated loss ρ with respect to a measure μ dominating the conditional distribution $\mathbb{P}(Y \leq \upsilon|\boldsymbol{X} = \boldsymbol{x})$:

$$\ell((Y, \boldsymbol{X}), h) := \int \rho((Y \leq \upsilon, \boldsymbol{X}), h(\upsilon|\boldsymbol{X}))\, d\mu(\upsilon) \geq 0.$$

In the context of scoring rules, the loss ℓ based on ρ_{sqe} is known as the continuous ranked probability score (CPRS) or integrated Brier score and

is a proper scoring rule for assessing the quality of probabilistic or distributional forecasts (see Gneiting and Raftery, 2007, for an overview). The risk function now reads

$$\mathbb{E}_{Y,\boldsymbol{X}} \ell((Y, \boldsymbol{X}), h) = \int \int \rho((y \leq \upsilon, \boldsymbol{x}), h(\upsilon|\boldsymbol{x})) \, d\mu(\upsilon) \, d\mathbb{P}_{Y,\boldsymbol{X}}(y, \boldsymbol{x}) \geq 0. \quad (2)$$

The corresponding empirical risk function defined by the data is

$$\hat{\mathbb{E}}_{Y,\boldsymbol{X}} \ell((Y, \boldsymbol{X}), f) = \int \int \rho((y \leq \upsilon, \boldsymbol{x}), h(\upsilon|\boldsymbol{x})) \, d\mu(\upsilon) \, d\hat{\mathbb{P}}_{Y,\boldsymbol{X}}(y, \boldsymbol{x}) \geq 0.$$

Based on an i.i.d. random sample $(Y_i, \boldsymbol{X}_i) \sim \mathbb{P}_{Y,\boldsymbol{X}}, i = 1, \ldots, N$ of N observations from the joint distribution of response and explanatory variables, we define $\hat{\mathbb{P}}_{Y,\boldsymbol{X}}$ as the distribution putting mass $w_i > 0$ on observation i ($w_i \equiv N^{-1}$ for the empirical distribution). For computational convenience, one approximates the measure μ by the discrete uniform measure $\hat{\mu}$, which puts mass n^{-1} on each element of the equi-distant grid $\upsilon_1 < \cdots < \upsilon_n \in \mathbb{R}$ over the response space. The empirical risk is then

$$\begin{aligned} \hat{\mathbb{E}}_{Y,\boldsymbol{X}} \ell((Y, \boldsymbol{X}), h) &= \sum_{i=1}^{N} w_i n^{-1} \sum_{\imath=1}^{n} \rho((Y_i \leq \upsilon_\imath, \boldsymbol{X}_i), h(\upsilon_\imath|\boldsymbol{X}_i)) \\ &= n^{-1} \sum_{i=1}^{N} \sum_{\imath=1}^{n} w_i \rho((Y_i \leq \upsilon_\imath, \boldsymbol{X}_i), h(\upsilon_\imath|\boldsymbol{X}_i)). \end{aligned}$$

This risk is the weighted empirical risk for loss function ρ evaluated at the observations $(Y_i \leq \upsilon_\imath, \boldsymbol{X}_i)$ for $i = 1, \ldots, N$ and $\imath = 1, \ldots, n$. Consequently, one can apply algorithms for fitting generalised additive models to the binary responses $Y_i \leq \upsilon_\imath$ under loss ρ for estimating model (1).
For conditional transformation models, Hothorn et al. (2013b) parameterise the partial transformation functions for all $j = 1, \ldots, J$ as

$$h_j(\upsilon|\boldsymbol{x}) = \left(\boldsymbol{b}_j(\boldsymbol{x})^\top \otimes \boldsymbol{b}_0(\upsilon)^\top\right) \boldsymbol{\gamma}_j \in \mathbb{R}, \qquad \boldsymbol{\gamma}_j \in \mathbb{R}^{K_j K_0},$$

where $\boldsymbol{b}_j(\boldsymbol{x})^\top \otimes \boldsymbol{b}_0(\upsilon)^\top$ denotes the tensor product of two sets of basis functions $\boldsymbol{b}_j : \chi \to \mathbb{R}^{K_j}$ and $\boldsymbol{b}_0 : \mathbb{R} \to \mathbb{R}^{K_0}$. Here, $\boldsymbol{b}_0$ is a basis along the grid of υ values that determines the functional form of the response transformation. The basis $\boldsymbol{b}_j$ defines how this transformation may vary with certain aspects of the explanatory variables. The tensor product may be interpreted as a generalised interaction effect. For each partial transformation function h_j, one typically wants to obtain an estimate that is smooth in its first argument υ and smooth in the conditioning variable $\boldsymbol{x}$. Therefore, the bases are supplemented with appropriate, pre-specified penalty matrices $\boldsymbol{P}_j \in \mathbb{R}^{K_j \times K_j}$ and $\boldsymbol{P}_0 \in \mathbb{R}^{K_0 \times K_0}$, inducing the penalty matrix $\boldsymbol{P}_{0j} = (\lambda_0 \boldsymbol{P}_j \otimes \mathbf{1}_{K_0} + \lambda_j \mathbf{1}_{K_j} \otimes \boldsymbol{P}_0)$ with smoothing parameters $\lambda_0 \geq 0$ and $\lambda_j \geq 0$ for the tensor product basis.

Conditional transformation models are then fitted by component-wise boosting where the base-learners are Ridge-type linear models corresponding to these partial transformation functions (the algorithm is given in Hothorn et al., 2013b). The basis functions $\boldsymbol{b}_0$ and $\boldsymbol{b}_j$ determine the form of the fitted model, and their choice is problem specific. In the simplest situation, in which the conditional distribution of Y given only one numeric explanatory variable x_1 shall be estimated, one could use the basis functions $\boldsymbol{b}_0(\upsilon) = (1, \upsilon)^\top$ and $\boldsymbol{b}_1(\boldsymbol{x}) = (1, x_1)^\top$. The corresponding base-learner is then defined by the linear function

$$((1, x_1) \otimes (1, \upsilon))\, \boldsymbol{\gamma}_1 = (1, \upsilon, x_1, x_1\upsilon)\boldsymbol{\gamma}_1.$$

For each x_1, the transformation is linear in υ with intercept $\gamma_1 + \gamma_3 x_1$ and slope $\gamma_2 + \gamma_4 x_1$, *i.e.* not only the mean may depend on x_1 but also the variance. Restricting, for example, $\boldsymbol{b}_0(\upsilon)$ to be constant, *i.e.* $\boldsymbol{b}_0(\upsilon) \equiv 1$, allows the effects of explanatory variables to be restricted to the mean alone. Assuming $\boldsymbol{b}_1(\boldsymbol{x}) \equiv 1$, on the other hand, yields a transformation function that is not affected by any explanatory variable. More flexible basis functions, *e.g.* B-spline basis functions, allow also for higher moments to depend on the explanatory variables. We illustrate appropriate choices of basis functions in the following, where we present analyses with special emphasis on higher moments of the conditional distribution, which have received less attention in previous analyses of these problems. We show that semiparametric regression using conditional transformation models is a valuable tool for detecting interesting patterns beyond the conditional mean.

2 Italian Gross-domestic Product

Here we follow Hayfield and Racine (2008) and consider Giovanni Baiocchi's Baiocchi (2006) Italian gross domestic product growth panel for 21 regions covering the period $1951 - 1998$ (millions of Lire, 1990 being the baseline). The data consist of 1008 observations for two variables, `gdp` (gross domestic product) and `year`. We first fit the conditional distribution by a conditional transformation model of the form

$$\mathbb{P}(\text{gdp} \leq \upsilon | \text{year} = x) = \Phi(h(\upsilon | \text{year} = x)).$$

The base-learner is the tensor product of B-spline basis functions $\boldsymbol{b}_0(\upsilon)$ for the gross domestic product and B-spline basis functions for time. The penalty matrices $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ penalise second-order differences, and thus $\hat{h}$ will be a smooth bivariate tensor product spline of gross domestic product and time. Since the number of observations is relatively small we used the bootstrap to determine an appropriate number of boosting iterations. Figure 1 shows the estimated conditional distribution functions on the quantile

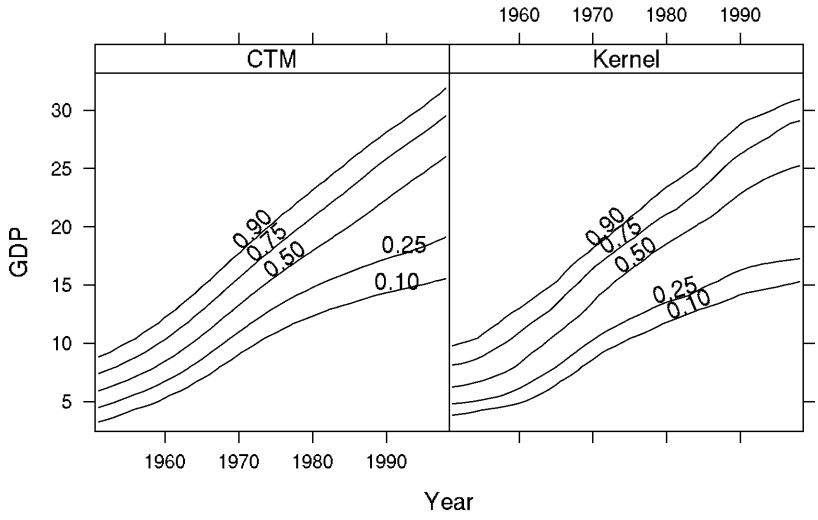


FIGURE 1. Italian Gross Domestic Product (GDP). Quantile functions obtained from a conditional transformation model (left) and a kernel estimator (right) for several quantiles.

scale. The function fitted by the kernel procedure is a little rougher than the estimate obtained from the conditional distribution model but, overall, the depicted quantiles are rather close.

3 Head Circumference Growth

The Fourth Dutch Growth Study (Fredriks et al., 2000) is a cross-sectional study that measures growth and development of the Dutch population between the ages of 0 and 22 years. The study measured, among other variables, head circumference (HC) and age of 7482 males and 7018 females. Stasinopoulos and Rigby (2007) analysed the head circumference of 7040 males with explanatory variable age using a GAMLSS model with a Box-Cox t distribution describing the first four moments of head circumference conditionally on age. The models show evidence of kurtosis, especially for older boys. We estimate the whole conditional distribution function via the conditional transformation model

$$\mathbb{P}(\text{HC} \leq v|\text{age} = x) = \Phi(h(v|\text{age} = x)).$$

The base-learner is the tensor product of B-spline basis functions $\boldsymbol{b}_0(v)$ for head circumference and B-spline basis functions for age$^{1/2}$. The root transformation just helps to cover the data better with equidistant knots. The

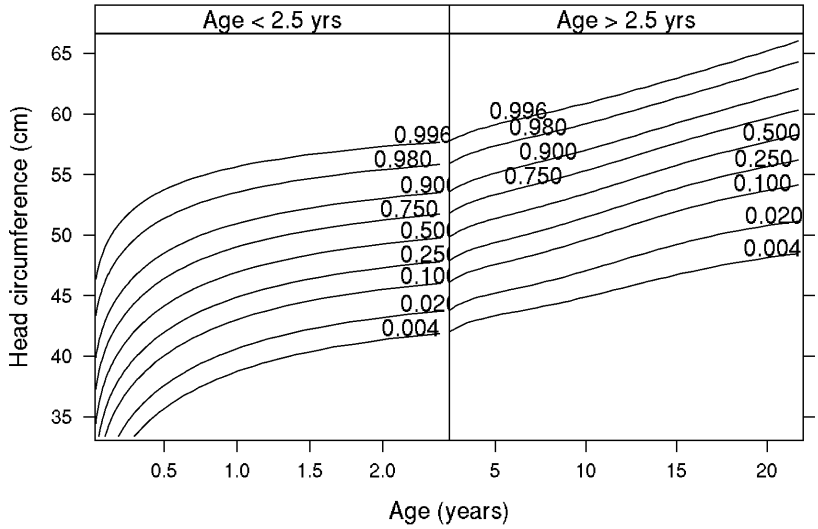


FIGURE 2. Head Circumference Growth. Observed head circumference and age for 7040 boys with estimated quantile curves for $\tau = 0.04, 0.02, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98, 0.996$.

penalty matrices $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ penalise second-order differences, and thus $\hat{h}$ will be a smooth bivariate tensor product spline of head circumference and age. It is important to note that smoothing takes place in both dimensions. Consequently, the conditional distribution functions will change only slowly with age, which is a reasonable assumption. Since the number of observations is also large, we stopped the algorithm based on the in-sample empirical risk.

Figure 2 shows the data overlaid with quantile curves obtained via inversion of the estimated conditional distributions. The figure can be directly compared with Figure 16 of Stasinopoulos and Rigby (2007) and also indicates a certain asymmetry towards older boys.

4 Deer-vehicle Collisions

Collisions of vehicles with roe deer are a serious threat to human health and animal welfare. In Bavaria, Germany, more than 40,000 deer-vehicle collisions (DVCs) take place every year. Hothorn et al. (2012) investigated the spatial distribution of the risk of deer-vehicle collisions; here we focus on the temporal aspect of the risk for two years, 2006 and 2009. For all 74,650 collisions reported to the police in these two years, we attributed each accident to the specific day of the year.

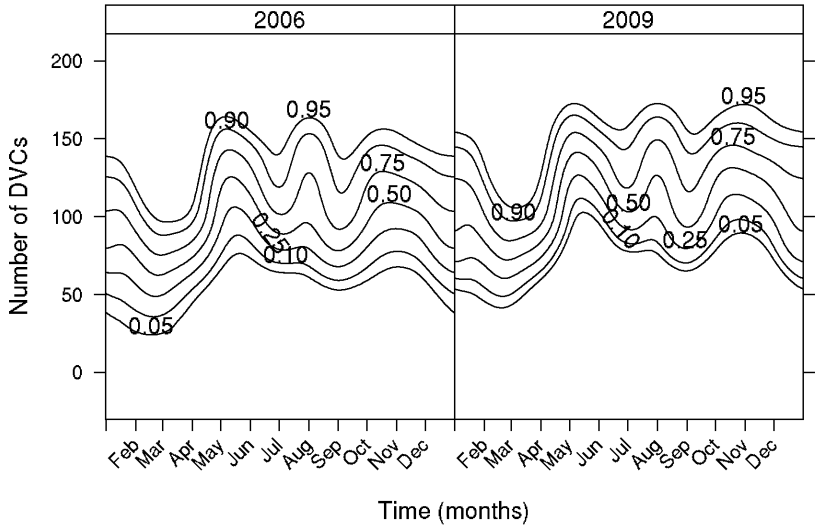


FIGURE 3. Deer-vehicle Collisions. Number of deer-vehicle collisions (DVCs) per day in 2006 and 2009 in Bavaria, Germany, with estimated quantile curves for $\tau = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$.

Although the number of DVCs is a discrete random variable, the distribution of the number of DVCs conditional on the day of the year can be estimated by means of an appropriate base-learner using the model

$$\begin{aligned}\mathbb{P}(\text{DVCs} \leq \upsilon | \text{day} = x_1, \text{year} = x_2) \quad &= \quad \Phi(h_1(\upsilon|\text{day} = x_1) \\ &\quad + h_2(\upsilon|\text{day} = x_1, \text{year} = x_2)).\end{aligned}$$

Here, $\hat{\mu}$ is the counting measure with support $\upsilon_1, \ldots, \upsilon_N$ equal to the support of the empirical distribution of the response. Conceptually, the basis function $\boldsymbol{b}_0$ should allow for $n = N$ parameters (one for each υ_i), whose first-order differences should not become too large. To restrict the number of parameters in the base-learners, we use B-splines to approximate such a discrete function on the υ-grid. It should further be noted that the day of year is a discrete cyclic random variable. Therefore, we chose $\boldsymbol{b}_1(x_1)$ as cyclic B-splines of the day, which are obtained by a simple modification of the B-spline design matrix and the difference penalty that results from fusing the two ends of the co-domain. In analogy, a cyclic B-spline is applied to the varying coefficient term $\boldsymbol{b}_2(x_1, x_2) = \boldsymbol{b}_1(x_1) \times I(x_2 = 2009)$, which captures temporal differences between the two years and yields a cyclic B-spline of the days in 2009. Since the data are discrete, we only penalise first-order differences in both base-learners.

Figure 3 shows three risk peaks. The first one occurs early in May – the

beginning of the growing and buck hunting season – and ends mid-June. A second and sharper peak is observed in the first week of August and corresponds to the mating season of roe deer. After a low-risk period of approximately six weeks, the risk starts to increase again at the beginning of October and slowly decreases until April for reasons yet unknown. Note that the distribution in 2009 has a larger median than that in 2006 but also shows less extreme peaks.

5 Beyond Mean Boston Housing Values

The Boston Housing data, first published by Harrison and Rubinfeld (1978) and later corrected and spatially aligned by Gilley and Pace (1996), have become a standard test-bed for variable selection and model choice. Almost exclusively, the 13 explanatory variables have been selected with respect to their influence on the mean or median of the conditional median house value in a certain tract. Assuming a conditional transformation model, we attempt to detect dependencies of higher moments of the conditional median house value from the explanatory variables. We focus on the 12 numeric explanatory variables and ignore the binary variable coding for Charles River boundary in the conditional transformation model

$$\mathbb{P}(\text{MEDV} \leq \upsilon | \boldsymbol{X} = \boldsymbol{x}) = \\ \Phi\left(\alpha_{\text{tract}} + h_0(\upsilon|1) + \sum_{j=1}^{12} h_j(1|x_j) + \sum_{j=1}^{12} h_j(\upsilon|x_j)\right).$$

In this model, α_{tract} is a tract-specific, spatial random effect, whose correlation structure is determined by a Markov random field defined by the neighbouring structure of the tracts capturing spatial autocorrelation and heterogeneity. The term $h_0(\upsilon|1)$ is an unconditional transformation of the median house value, *i.e.* this transformation is independent of the explanatory variables. The explanatory variables may influence the mean of the transformed median house value $h_0(\text{MEDV}|1)$ via $h_{\boldsymbol{x}}(\boldsymbol{x}) = \sum_{j=1}^{12} h_j(1|x_j)$ only or may also affect higher moments via the interaction terms $\sum_{j=1}^{12} h_j(\upsilon|x_j)$. The latter term extends the transformation model $h_0(\text{MEDV}|1) + \sum_{j=1}^{12} h_j(1|x_j)$ to a conditional transformation model. The base-learners for the transformation function $h_0(\upsilon|1)$, the effects $h_j(1|x_j)$ and the interaction terms $h_j(\upsilon|x_j)$ are constructed based on cubic B-spline basis functions supplemented with second-order difference penalty. More specifically, $\boldsymbol{b}_j(\boldsymbol{x})$ and $\boldsymbol{b}_0(\upsilon)$ are both represented in terms of a reparameterisation of the B-spline basis functions that allows separation of the non-linear terms into a constant, a linear effect and the non-linear (orthogonal) deviation from the linear effect, *i.e.*

$$\boldsymbol{b}_j(\boldsymbol{x}) = 1 + x_j + \tilde{\boldsymbol{b}}_j(x_j) \quad \text{and} \quad \boldsymbol{b}_0(\upsilon) = 1 + \upsilon + \tilde{\boldsymbol{b}}_0(\upsilon),$$

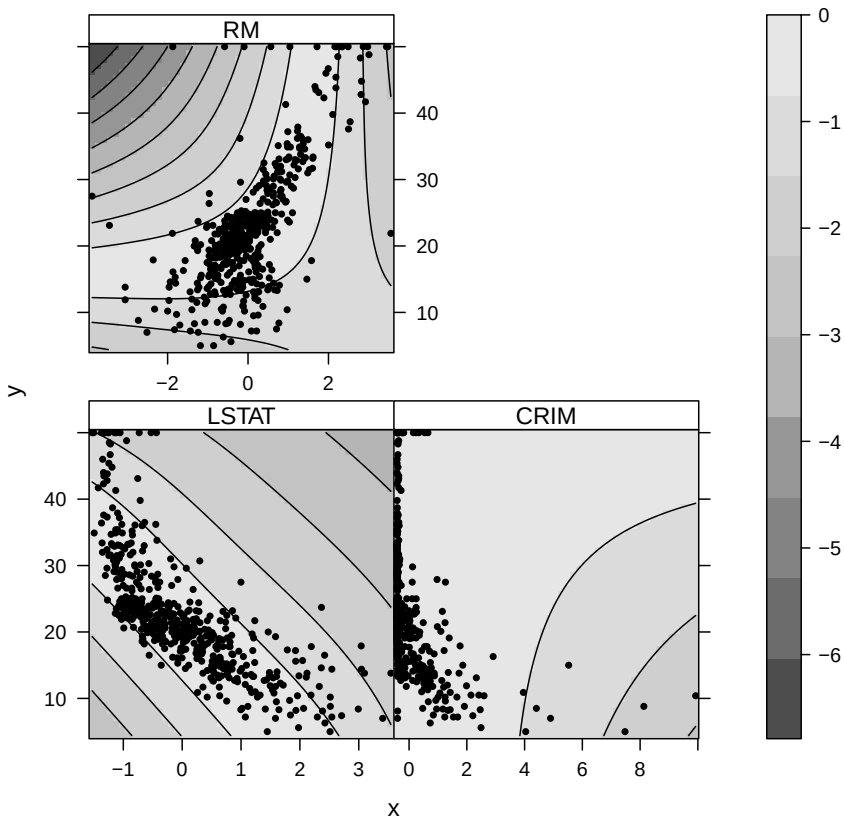


FIGURE 4. Beyond Mean Boston Housing Values. Conditional transformation model for the three selected variables per capita crime (CRIM), average numbers of rooms per dwelling (RM) and percentage values of lower status population (LSTAT). Each panel depicts the data as scatter plots along with the corresponding negative absolute values of the estimated transformation function at the probit scale. The explanatory variables were standardised prior to analysis.

where $\tilde{\boldsymbol{b}}_j(x_j)$ and $\tilde{\boldsymbol{b}}_0(\upsilon)$ are the non-linear deviation effects. Taking the tensor product after applying the decomposition yields a decomposition into linear and non-linear main effects of x_j and υ as well as linear and non-linear interaction terms (see Fahrmeir et al., 2004; Kneib et al., 2009, for technical details of this decomposition). The advantage of this expanded parameterisation is that the automatic model choice capabilities of the boosting algorithm allow us to flexibly determine whether linear or non-linear effects are required and whether there actually is an interaction between the transformation function and specific effects of explanatory variables.
Censored observations were dealt with by choosing inverse probability of

censoring weights w_i for the empirical risk function (2) derived from the Kaplan-Meier estimate of the censoring distribution. The stability selection procedure (Meinshausen and Bühlmann, 2010) selected three variables that have an influence on the conditional distribution of the median housing value (MEDV), namely per capita crime (CRIM), average numbers of rooms per dwelling (RM), and percentage values of lower status population (LSTAT). After variable selection, we refitted a conditional transformation model of the simpler form

$$\begin{aligned}&\mathbb{P}(\text{MEDV} \leq \upsilon|\text{CRIM}, \text{RM}, \text{LSTAT}) \\ &= \Phi\left(h_{\text{CRIM}}(\upsilon|\text{CRIM}) + h_{\text{RM}}(\upsilon|\text{RM}) + h_{\text{LSTAT}}(\upsilon|\text{LSTAT})\right),\end{aligned}$$

where the base-learners are tensor products of B-spline bases. The fitted functions can be conveniently depicted in the observation space. For example, a scatter plot of MEDV and CRIM and a grey-level image of the bivariate function $\hat{h}_{\text{CRIM}}(\text{MEDV}|\text{CRIM})$ can be viewed in the same coordinate system. We show negative absolute values of the fitted functions $\hat{h}$ for easier interpretation.
Figure 4 indicates that the percentage values of lower status population (LSTAT) lead to smaller values of the median housing value at almost constant variance. However, the conditional distribution will be skewed towards higher MEDV values. For tracts with small average numbers of rooms per dwelling (RM), the median housing value is small and increases with increasing numbers of rooms. The same applies to the variability, since the estimated function $\hat{h}_{\text{RM}}(\text{MEDV}|\text{RM})$ shows more spread for larger values of RM. Per capita crime seems to have an effect on variability and skewness, since for larger crime values, the distribution will be heavily skewed and less variable than small per capita crime values. However, compared to the other two variables, the influence is only of marginal value due to small absolute contributions of this model term to the full model.

6 Summary

Conditional transformation models extend the class of the classical (unconditional) transformation models by allowing interactions between the response and explanatory variables. The resulting regression models can describe the whole conditional distribution of the response as a function of the explanatory variables. Thus, if the data scientist is not only interested in studying effects on the conditional mean, conditional transformation models are an potentially interesting and helpful alternative to kernel-based methods or to GAMLSS models.

Computational Details

Conditional transformation models were fitted using an implementation of component-wise boosting in package **mboost** (version 2.2-2, Hothorn et al., 2013a). Kernel distribution estimation was performed using package **np** (version 0.50-1, Hayfield and Racine, 2013). All computations were performed using R version 3.0.0 (R Development Core Team, 2013). For further computational details we refer the reader to the R code that implements the analyses presented here, which is available in an experimental R package **ctm** at `http://R-forge.R-project.org/projects/ctm`. The results presented in this paper can be reproduced using this package.

Bibliography

Baiocchi, G. (2006). *Economic Applications of Nonparametric Methods*. Ph.D. thesis, University of York.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, **14**, 731–761.

Fredriks, A.M., van Buuren, S., Burgmeijer, R.J.F., Meulmeester, J.F., Beuker, R.J., Brugman, E., Roede, M.J., Verloove-Vanhorick, S.P., and Wit, J. (2000). Continuing positive secular growth change in The Netherlands 1955–1997. *Pediatric Research*, **47**, 316–323.

Gilley, O.W. and Pace, R.K. (1996). On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, **31**, 403–405.

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, **5**, 81–102.

Hayfield, T. and Racine, J.S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, **27**, 1–32. URL `http://www.jstatsoft.org/v27/i05`.

Hayfield, T. and Racine, J.S. (2013). *np: Nonparametric kernel smoothing methods for mixed data types*. URL `http://CRAN.R-project.org/package=np`. R package version 0.50-1.

Hothorn, T., Brandl, R., and Müller, J. (2012). Large-scale model-based assessment of deer-vehicle collision risk. *PLoS ONE*, **7**, e29510.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2013a). *mboost: Model-Based Boosting*. URL http://CRAN.R-project.org/package=mboost. R package version 2.2-2.

Hothorn, T., Kneib, T., and Bühlmann, P. (2013b). Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. doi:10.1111/rssb.12017.

Kneib, T., Hothorn, T., and Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics*, **65**, 626–634.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 417–473.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1–46. URL http://www.jstatsoft.org/v23/i07.

How Coarsening Simplifies Matching-Based Causal Inference Theory

Stefano M. Iacus[1], Gary King[2]

[1] Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7, I-20124 Milan, Italy
[2] Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge, USA

E-mail for correspondence: stefano.iacus@unimi.it

Abstract: The simplicity and power of matching methods have made them an increasingly popular approach to causal inference in observational data. Existing theories that justify these techniques are well developed but either require exact matching, which is usually infeasible in practice, or sacrifice some simplicity via asymptotic theory, specialized bias corrections, and novel variance estimators; and extensions to approximate matching with multicategory treatments have not yet appeared. As an additional option for researchers, we show how conceptualizing continuous variables as having logical breakpoints (such as phase transitions when measuring temperature or high school or college degrees in years of education) is both natural substantively and can be used in some applications to construct a relatively simple theory of causal inference. The result is a finite sample theory that is simple to understand and easy to implement by using matching to preprocess the data, after which one can use whatever method would have been applied without matching. The theoretical simplicity also allows for binary, multicategory, and continuous treatment variables from the start and for extensions to valid inference under imperfect treatment assignment. In applications where the existing theory of matching is difficult to apply, the new approach added to the existing toolkit may help some researchers in these situations make valid causal inferences, or at least better understand why they cannot.

Keywords: Causal Inference; Matching; Observational Study.

1 Introduction

Matching is a powerful nonparametric approach to improve causal inferences in observational data, that is where assignment of units to treatment and control groups is not under the control of the investigator and not necessarily random. The basic idea involves pruning observations to improve balance between the treated and control groups (solely as a function of measured pre-treatment covariates so as to avoid inducing selection bias) and then estimating the causal effect from the remaining sample.

The most prevalent theory justifying matching requires data with a large number of exact matches between the treated and control groups which may not be available. An alternative approximate matching approach works under asymptotic theory, requiring specialized bias corrections and novel variance estimators (Abadie and Imbens, 2006). We add an alternative matching theory to the existing approaches. The basic idea behind our alternative theory is to recognize that observational data often comes or can be though as coming from stratification of the original population. The alternative offered here is simpler in theory and practice, enabling researchers to use whatever estimation method they would have without matching. The approach is valid for approximate matching in finite samples; applies to binary, multicategory, or continuous treatments; and can be easily extended to allow the true and observed treatment status to diverge. Although simpler to understand, the real question is whether it is applicable to the data set at hand.

2 Statistical Framework

Consider a sample of n observations where subject i ($i = 1, \ldots, n$) has been exposed to treatment $T_i = t$, for $t \in \mathcal{T}$, where $\mathcal{T}$ is either a subset of $\mathbb{R}$ or a set of (ordered or unordered) categories, T is a random variable, and t one possible value of it. Then $\mathcal{Y} = \{Y_i(t), t \in \mathcal{T}, i = 1, \ldots, n\}$ is the set of *potential outcomes*, the possible values the outcome variable has when T takes on different values. For each observation, we observe one and only one of the set of potential outcomes, for which the treatment was actually assigned: $Y_i \equiv Y_i(T_i)$. We also observe a $p \times 1$ vector of pre-treatment covariates X_i for subject i, and for some purposes consider this to be a random variable drawn from a superpopulation, where $X \in \mathcal{X}$.
Let t_1 and t_2 be distinct values of T that happen to be of interest, regardless of whether T is binary, multicategory, or continuous (and which, for convenience we refer to as the treated and control conditions, respectively). Define the *treatment effect* for each observation as the difference between the corresponding two potential outcomes, $\mathrm{TE}_i = Y_i(t_1) - Y_i(t_2)$, of which at most only one is observed. The object of statistical inference is usually an average of treatment effects over a given subset of observations. One example is the *sample average treatment effect on the treated*, where inference is for all treated units in the sample at hand: $\mathrm{SATT} = \frac{1}{\#\{T_i = t_1\}} \sum_{i \in \{T_i = t_1\}} \mathrm{TE}_i$. The control units are used to help estimate this quantity.

2.1 Classes of Matching Methods

There exist two classes of matching methods: EPBR and MIB. In the Equal Percent Bias Reducing (EPBR) class of matching methods, the percent

reduction in expected imbalance over repeated samples from matching on one variable is intended to be the same for all variables (Rubin, 1976). The number of observations matched in EPBR methods is determined ex ante, whereas the level of imbalance achieved is not guaranteed and so must be checked ex post. In the Monotonic Imbalance Bounding (MIB) class of matching methods, reducing in-sample imbalance on one variable has no effect on the ex ante choice for the maximum in-sample imbalance on any of the other variables (Iacus *et al*, 2011). In MIB, the maximal level of imbalance is chosen ex ante, whereas the number of observations matched is a result of the method. The most relevant example in the EPBR is Propensity Score Matching (PSM) while for the MIB class is Coarsened Exact Matching (CEM).

2.2 Example of matching methods: PSM and CEM

The promise of PSM is to make matching easier by allowing a match on a lower dimensional quantity — the propensity score — rather than the original p covariates in X and still, under certain conditions described below, to produce balance on X in expectation. The generalized propensity score (function) is the probability of receiving a particular level of the treatment t given the pre-treatment variables $e(t, x) = Pr(T = t|X = x) = E\{D(t)|X = x\}$, where $D(t)$ is an indicator function for the event $T = t$ (Imbens, 2000). Under CEM, each variable is temporarily coarsened as much as the analyst is willing (this is nothing but "stratification on covariates", e.g., years of education might be coarsened into grade school, high school, college, and graduate school). Then, treated and control groups are matched exactly on the coarsened variables, and finally the uncoarsened values of the matched units are passed to the analysis stage. Since CEM is a member of the class of MIB matching methods, it inherits all the statistical properties of this class, including in-sample properties (Iacus *et al*, 2011, 2012). In contrast, EPBR is a class that applies only in expectation and only if certain assumptions about the data generation process are met, and so, to be precise, PSM is best described as only *potentially EPBR*.

2.3 Basic Assumptions for Identification

Scholars typically make three assumptions to justify causal inferences under exact matching.

Assumption 1 [SUTVA: Stable Unit Treatment Value Assumption (Rubin, 1991)] *The values of the potential outcomes* $\{Y(t), t \in \mathcal{T}\}$ *are independent of the treatment status* T.

Assumption 2 [Weak unconfoundedness (Imbens, 2000)] *The treatment assignment* T *is unconfounded, given covariates* X, *if* $D(t) \perp Y(t)|X = x$ *for each* $t \in \mathcal{T}$ *and almost every* $X = x$.

Assumption 3a [Common Support] *There exists* $\eta \in (0,1)$, *such that for all measurable sets* $B \in \mathcal{T}$ *and almost every* $X = x$: $\eta < p(T \in B|X = x) < 1 - \eta$.

Assumption 2 (or "no omitted variable") ensures that the pre-treatment covariates are sufficient to adjust for any biases that may be induced by the lack of random assignment. Therefore one can use a matched control unit to fill in for an unobserved potential outcome and this yields to the relation: $E\{Y(t)|X = x\} = E\{Y|T = t, X = x\}$, which makes the mean of the potential outcomes identified. As a result, one can average over the observations with different $X = x$, yielding $E(Y(t)) = E(E\{Y(t)|X\})$. Assumption 3 means that for any unit with observed treatment condition $T_i = t_1$ and covariates X_i, it is also *possible* to observe a unit with the counterfactual treatment, $T_i = t_2$, and the same covariate values. Taken together, Assumptions 1, 2 and 3 ensure that the treatment effect is correctly identified under exact matching.

2.4 Additional Assumptions for Propensity Score Matching

PSM allows (exact) matching on the scalar propensity score instead of the multidimensional X, but requires three additional assumptions to achieve identification (in addition to Assumptions 1, 2, and 3).

Since the propensity score function is not known in practice and sensitive to specification choices when estimated, users of PSM must make two assumptions about the existence of, knowledge about, and uniqueness regarding the true propensity score function $e_\psi(\cdot|X)$ and one that helps make it feasible.

Assumption 4 [Uniquely Parametrized Propensity Score (Imai and van Dik, 2004)]*For each* $X \in \mathcal{X}$, *there exists a unique finite-dimensional parameter* $\theta \in \Theta$, *such that* $e_\psi(\cdot|X) = e(\cdot|\theta_\psi(X))$ *and* $\int_B e_\psi(t|\theta)dt = \int_B e_\psi(t|\theta')dt$ *for all measurable sets* $B \subset \mathcal{T}$ *imply* $\theta = \theta'$. *That is,* θ *uniquely represents* $e(\cdot|\theta_\psi(X))$, *which we may therefore write as* $e(\cdot|\theta)$.

This assumption implies that $e_\psi(\cdot|X)$ depends on X only through $\theta_\psi(X)$, i.e. θ is sufficient for T. In this case, the propensity function is effectively summarized by the parameter θ, which is typically of much lower dimension than X (and scalar when T is binary), therefore matching can be done by subclassifying on θ (Imai and van Dik, 2004).

The second assumption goes beyond the existence of the propensity score in Assumption 4 to require that the specification and the specific parameter values of the specification are known as well:

Assumption 5 [Known Propensity Score]:*The propensity score* $e_\psi(\cdot, X)$ *has (i) a known functional form and (ii) known value of the parameter* ψ.

The veracity of Assumption 5(i) is more difficult to test than the usual problem of model selection in statistical analysis, since the objective function of the chosen propensity score model is optimized based on fit rather than balance.

Finally, propensity score theory requires exact matching on the (continuous) propensity score. To make this possible in finite samples, an additional assumption is routinely made (Rosenbaum and Rubin,1983), although rarely stated formally — that the propensity score function is constant over given known subsets. Let $\Pi(\mathcal{X})$ be a finite partition of the covariate space $\mathcal{X}$, and let $A_k \in \Pi(\mathcal{X})$ $(k = 1, \ldots, K < \infty)$ be one generic set of the partition, i.e. $\cup_k A_k = \mathcal{X}$ and $A_l \cap A_m = \emptyset$ for $l \neq m$.

Assumption 6 [Constant Propensity Score Intervals] *The propensity score, $e(t, x) = c_k$, is constant for all $x \in A_k$, with $A_k \in \Pi(\mathcal{X})$ and $t \in \mathcal{T}$.*

If this assumption holds, matching within blocks generated by A_k removes all bias and approximate matching on the propensity score is justified in terms of sample variability of the propensity score estimate. This is the unstated assumption behind the *blocking-on-the-propensity-score* approach.

3 An alternative approach

Assumptions 1, 2, and 3 enable the power of matching with simple point and variance estimators — in particular, letting researchers use after matching almost exactly what they would have without matching — but they only work under rare applications where treated units have exact matches. Under approximate matching, corrections for estimators and variances are available under the asymptotic theory. We present here a set of alternative assumptions that, when substantively appropriate, make simple estimators possible in finite samples under approximate matching.

It is well known to researchers that in observational studies repeated sampling from a given (super)population is just a convenient fiction to justify the results. Our idea is that observations belong to some strata A of the partition $\Pi(\mathcal{X})$ and what we observe are different replicates from these strata. The strata are defined ex-ante and remain fixed.

We now construct alternative versions of Assumptions 2 and 3 that work under approximate matching. The basic intuition is that instead of conditioning on values of the vector X, we assume that we only need to condition on an observation being in one of a set of given strata $A \subset \mathcal{X}$.

Assumption 2b [Set-wide Weak Unconfoundedness] *Assignment of the treatment variable T possesses the property of set-wide weak unconfoundedness, given pre-treatment covariate values in A, if $D(t) \perp Y(t) | A$, for all $t \in \mathcal{T}$ and each $A \in \Pi(\mathcal{X})$.*

Apart from the sampling framework, Assumption 2b happens to be a degenerate version of the Conditioning At Random (CAR) assumption introduced in (Heitjan and Rubin, 1991) in that conditioning is now fixed. Here A represents only a stratification of the reference population and is not random and so does not change from sample to sample. Assumption 2b is uncounfoundedness with respect to the set A instead of $X = x$ and hence is more stringent than Assumption 2.

Assumption 3b [Set-wide Common Support] *For all measurable sets* $B \in \mathcal{T}$ *and all sets* $A \in \Pi(\mathcal{X})$*, set-wide common support requires that* $0 < p(T \in B|X \in A) < 1$.

Assumption 3b makes the search for counterfactuals easier since those in the vicinity of, rather than exactly equal to, a given covariate vector $X \in A$ are now acceptable. Identification for treatment effect is now possible as we have $E\{Y(t)|A\} = E\{Y|T = t, A\}$. Therefore, in a single stratum A, $\tau_A^{1,2} = E\{Y(t_1) - Y(t_2)|A\} = E\{Y|T = t_1, A\} - E\{Y|T = t_2, A\}$, for any $t_1 \neq t_2 \in \mathcal{T}$. Then, $\tau_A^{1,2}$ is estimated taking the difference in means among treated and control units within each stratum A and the global treatment effect estimator is obtained by averaging over $\Pi(\mathcal{X})$. The estimator is now unbiased also in finite samples without any further assumption 4–6.

References

Abadie, A., and Imbens, G. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, **74**(1), 235 – 267.

Heitjan, D.F., and Rubin, D. (1991). Ignorability and Coarse Data. *The Annals of Statistics*, **19**(4), 2244 – 2253.

Iacus, S.M., King, G., Porro, G. (2011). Multivariate Matching Methods that are Monotonic Imbalance Bounding. *Jour. Amer. Stat. Assoc.*, **106**, 345 – 361.

Iacus, S.M., King, G., Porro, G. (2012). Causal Inference Without Balance Checking: Coarsened Exact Matching. *Political Analysis*, **20**(1), 1 – 24.

Imbens, G. (2000). The role of the propensity score in estimating the dose-response functions. *Biometrika*, **87**(3), 706 – 710.

Imai, K., van Dyk, D.A. (2004). Causal Inference with General Treatment Treatment Regimes: Generalizing the Propensity Score. *Jour. Amer. Stat. Assoc.*, **99**, 854 – 866.

Rosenbaum, P.R, and Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41 – 55.

Rubin, D. (1976). Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples. *Biometrics*, **32**(1), 109 – 120.

Rubin, D. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, **47**, 1213 – 1234.

On Finite Mixtures of Skew Distributions

Geoffrey J. McLachlan[1], Sharon X. Leemaqz[1]

[1] University of Queensland, Australia

E-mail for correspondence: g.mclachlan@uq.edu.au

Abstract: The finite mixture model has become an important tool in statistical modelling and analysis. In recent years, mixtures of skew distributions have emerged as a powerful extension to the traditional normal and t-mixture models. They have been effectively applied to model heterogeneous data with asymmetric features. We shall discuss some of the more commonly used skew symmetric distributions, in particular, the skew normal and skew t-models. Examples involving the analysis of real data sets will be given.

Keywords: multivariate skew distributions; mixture models; EM algorithm.

1 Introduction

Finite mixture distributions have become increasingly popular in the modelling and analysis of data due to their flexibility. This use of finite mixture distributions to model heterogeneous data has undergone intensive development in the past decades, as witnessed by the numerous applications in various scientific fields such as bioinformatics, biostatistics, environmetrics, financial sciences, genetics, image analysis, and medical sciences. Comprehensive surveys on mixture models and their applications can be found, for example, in the monographs by Everitt and Hand (1981), Titterington, Smith, and Markov (1985), McLachlan and Basford (1988), Lindsay (1995), Böhning (2000), McLachlan and Peel (2000), and Frühwirth-Schnatter (2006), and the edited volume of Mengersen, Robert and Titterington (2011); see also the papers by Banfield and Raftery (1993) and Fraley and Raftery (1999).

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$ be a p-dimensional random vector. For continuous features $\boldsymbol{Y}_j$, the density of $\boldsymbol{Y}$ can be modelled by a mixture of a sufficiently large enough number g of multivariate normal component distributions,

$$f(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \, \phi_p(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where $\phi_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a p-variate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The mixing proportions π_i are nonnegative

and sum to one. Here the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the mixing proportions π_i, the elements of the component means $\boldsymbol{\mu}_i$, and the elements of the component-covariance matrices $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$ known *a priori* to be distinct. Maximum likelihood estimation (MLE) of the model parameters can be obtained via the expectation-maximization (EM) algorithm (Dempster et al., 1977); see also McLachlan and Krishnan (2008).

Mixtures of multivariate t-distributions, as proposed by McLachlan and Peel (1998, 2000), provide extra flexibility over normal mixtures; see also Peel and McLachlan (2000). The thickness of tails can be regulated by an additional parameter – the degrees of freedom, thus enabling it to accommodate outliers better than normal distributions. However, in many practical problems, the data often involve observations whose distributions are highly asymmetric as well as having longer tails than the normal; for example, datasets from flow cytometry (Pyne et al., 2009).

Mixture distributions are often used in practice to provide a probabilistic clustering of a data set into a number of clusters. An outright clustering is obtained by assigning a data point to the component to which it has the greatest posterior probability of belonging. There is the question of how many components to include in the mixture model. In the typical application of normal mixture models to clustering, clusters are taken to correspond to the normal components in the mixture model. But in cases where the clusters are not elliptically symmetric, this correspondence will not hold if additional normal components are needed to allow for the asymmetry in the data and the presence of outliers. One way to enable the number of components to correspond to the number of clusters in such situations is to fit mixture models with skew normal components or skew t-components. As an illustration, we consider the Lymphoblastic cell line (LCL) data set studied by Pyne et al. (2009). Figure 1(a) shows a heatmap of two markers on some cells that were cultured in a laboratory and were known to belong to the one population. It can be seen in Figure 1(b) in modelling the density of these bivariate data by a normal mixture, we need $g = 2$ components, whereas a single skew t-distribution suffices (Figure 1(c)).

The skew normal distribution was introduced in Azzalini (1985) as an extension of the normal distribution with an additional parameter to regulate the skewness, allowing the density to take asymmetric shapes. More recently, there has been renewed interest in the development of non-normal distributions. We shall discuss some of the several proposals that have been put forward for multivariate skew distributions (Lee and McLachlan, 2011, 2013a), including software for the fitting of mixtures of them (Lee and McLachlan, 2013b). Examples involving the analysis of real data sets will be given.

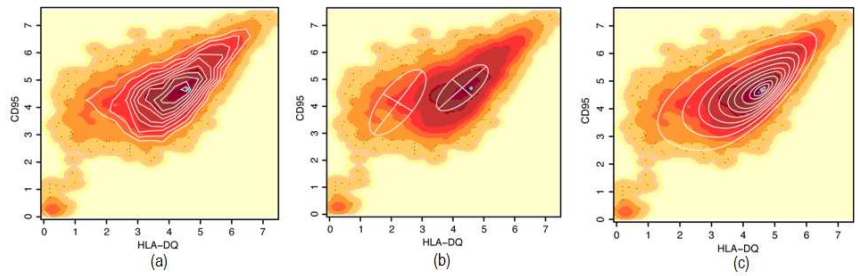


FIGURE 1. Modelling a lymphoblastic cell line (LCL) dataset derived from a single population of B cells. (a) Data contours plotted over the hue intensity of the LCL data, revealing single unimodal asymmetric population in the presence of some outliers. (b) Fitting normal mixture model results in two distinct clusters. (c) The single-component multivariate skew t-distribution can capture the asymmetry in the data and correctly identify the modes location (cyan dot).

2 Multivariate Skew distributions

The rich literature of skew distributions was initiated by the pioneering work of Azzalini (1985), who introduced the univariate skew normal distribution. Following its generalization to the multivariate case in Azzalini and Dalla Valle (1996), the number of contributions have grown rapidly. The majority of these distributions belong to the class of skew symmetric distribution, which can be further classified into four subclasses, namely, the restricted, unrestricted, extended and generalized forms (Lee and McLachlan, 2013c).

To begin, note that an asymmetric density can be generated by perturbation of symmetry (Azzalini and Capitanio, 2003). Consider a density $f(\boldsymbol{y}-\boldsymbol{\mu})$ symmetric around $\mathbf{0}$, that is, $f(-\boldsymbol{y}) = f(\boldsymbol{y})$, where $\boldsymbol{\mu}$ is a location vector in $\mathbb{R}^p$. Then a skew symmetric density can be formulated by manipulating $f(\cdot)$ through a perturbation function $h(\cdot)$, such that the product of the symmetric function and the perturbation function is a valid density; that is,

$$f(\boldsymbol{y}) = 2f(\boldsymbol{y}-\boldsymbol{\mu})h(\boldsymbol{y}-\boldsymbol{\mu}), \tag{2}$$

where $h(\cdot)$ is a function that maps $\mathbb{R}^p$ into the unit interval $[0,1]$ (Wang et al., 2004). The density f is known as the *symmetric component* of (2) and h is the *skewing component*. A typical example of $h(\cdot)$ is the distribution function corresponding to $f(\cdot)$, denoted by $F(\cdot)$.
We shall discuss several proposals of multivariate skew-symmetric distributions used in model-based clustering, in particular the multivariate skew normal (MSN) distribution and multivariate skew t (MST) distribution.

Using the terminology of Lee and mcLachlan (2013c), a random variable $\boldsymbol{Y}$ has an unrestricted multivariate skew normal (uMSN) distribution, if its density takes the form

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2^p \phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\Phi_p\left(\boldsymbol{\Delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu});0,\Lambda\right), \quad (3)$$

where $\boldsymbol{\Delta}$ is a diagonal matrix with diagonal elements given by $\boldsymbol{\delta}$, $\boldsymbol{\Lambda} = \boldsymbol{I}_p - \boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}$, $d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$ is the squared Mahalanobis distance between $\boldsymbol{y}$ and $\boldsymbol{\mu}$ with respect to $\boldsymbol{\Sigma}$, and $\Phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the distribution function corresponding to the p-variate normal density $\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$. The vector $\boldsymbol{\delta} \in \mathbb{R}^p$ is a skewness parameter. In a similar way, an unrestricted multivariate skew t (uMST) distribution can be defined as

$$\begin{aligned} f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta},\nu) &= 2^p t_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) \\ &\quad T_p\left(\boldsymbol{\Delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\sqrt{\frac{\nu+p}{\nu+d(\boldsymbol{y})}};0,\Lambda,\nu+p\right), \quad (4) \end{aligned}$$

where $t_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ denotes the p-dimensional t-density, and $T_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ is the corresponding distribution function.

The terms 'restricted' and 'unrestricted' used here were introduced by Lee and McLachlan (2012), as follows. The density (4) can be characterized by two parallel forms of stochastic representations, one via a conditioning mechanism and the other by a convolution approach. Let $\boldsymbol{Y}_0$ and $\boldsymbol{Y}_1$ be jointly distributed as

$$\begin{bmatrix} \boldsymbol{Y}_0 \\ \boldsymbol{Y}_1 \end{bmatrix} \sim t_{2p}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_p & \boldsymbol{\Delta} \\ \boldsymbol{\Delta} & \boldsymbol{\Sigma} \end{bmatrix}, \nu\right). \quad (5)$$

Then the distribution of $\boldsymbol{Y} = (\boldsymbol{Y}_1 \mid \boldsymbol{Y}_0 > \boldsymbol{0})$ takes the form of (4), where we let $(\boldsymbol{Y}_1 \mid \boldsymbol{Y}_0 > \boldsymbol{0})$ be the vector $\boldsymbol{Y}_1$ if all elements of $\boldsymbol{Y}_0$ are positive, and $-\boldsymbol{Y}_1$ otherwise. An equivalent stochastic representation of (4) is given by the convolution of a multivariate t-variable and the absolute value of another multivariate t-variable; that is,

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Delta}|\tilde{\boldsymbol{Y}}_0| + \tilde{\boldsymbol{Y}}_1, \quad (6)$$

where $|\tilde{\boldsymbol{Y}}_0|$ is the vector whose kth element is equal to the absolute value of the kth element of $\tilde{\boldsymbol{Y}}_0$ $(k = 1, \ldots, p)$, and where

$$\begin{bmatrix} \tilde{\boldsymbol{Y}}_0 \\ \tilde{\boldsymbol{Y}}_1 \end{bmatrix} \sim t_{2p}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \tilde{\boldsymbol{\Sigma}} \end{bmatrix}, \nu\right), \quad (7)$$

and $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Delta}^2$. Note that the uMSN distribution can also be generated by (5) and (7) by replacing the joint distribution of $\boldsymbol{Y}_0$ and $\boldsymbol{Y}_1$, and of $\tilde{\boldsymbol{Y}}_0$ and $\tilde{\boldsymbol{Y}}_1$ by a multivariate normal distribution.

In order to simplify the fitting of mixtures of these distributions, Pyne et al. (2009) imposed the restriction that the elements of $\tilde{\boldsymbol{Y}}_0$ are the same. That is, the latent variable $\tilde{\boldsymbol{Y}}_0$ in (7) can be replaced by a scalar Y_0. Hence, the convolution-type representation of a restricted MST (rMST) distribution can be simplified to

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\delta}|\tilde{Y}_0| + \tilde{\boldsymbol{Y}}_1, \tag{8}$$

where

$$\begin{bmatrix} \tilde{Y}_0 \\ \tilde{\boldsymbol{Y}}_1 \end{bmatrix} \sim t_{1+p}\left(\begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} 1 & \boldsymbol{\delta}^T \\ \boldsymbol{\delta} & \tilde{\boldsymbol{\Sigma}} \end{bmatrix}, \nu\right). \tag{9}$$

Note that this restriction corresponds to replacing $\boldsymbol{Y}_0$ with a scalar Y_0 in the conditioning-type representation (5).

With this characterization, the density of the rMST distribution reduces to

$$\begin{aligned} f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta},\nu) &= 2t_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) \\ &\quad T_1\left(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\sqrt{\frac{\nu+p}{\nu+d(\boldsymbol{y})}};0,\lambda,\nu+p\right), \end{aligned} \tag{10}$$

where $\boldsymbol{\lambda} = 1 - \boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ and $d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$.

Analogously, the density of the restricted MSN (rMSN) distribution is given by

$$f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\delta}) = 2\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\Phi_1\left(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu});0,\lambda\right). \tag{11}$$

Note that with the restriction that all the elements of $\tilde{\boldsymbol{Y}}_0$ in (6) be the same (that is, it is effectively replaced by a univariate random variable), the skewing component in (2) is no longer given by the p-dimensional distribution function corresponding to the symmetric component, but by a univariate distribution function.

The restricted skew normal and t-distributions given by (11) and (10), respectively, are the same, after reparametrization, as the MSN and MST distributions of Azzalini and Dalla Valle (1996), Branco and Dey (2001), and Lachos et al. (2010). Further discussion of these distributions and their characterizations can be found in Lee and McLachlan (2013a, 2013c).

3 Fitting finite mixtures of skew distributions

Adopting (11), (10), (3) and (4) as the component density of a finite mixture model (1) leads to the FM-rMSN, FM-rMST, FM-uMSN, and FM-uMST

models. The restricted and unrestricted skew normal and skew t mixture distributions admit convenient hierarchical characterizations which facilitate the computation of maximum likelihood (ML) estimates of the unknown model parameters via the EM algorithm. These E- and M-steps can be carried out in closed-form, as described by Lin (2009), Pyne et al. (2009), Lee and McLachlan (2011), and Lee and McLachlan (2013a), and their software implementations are available from the R package EMMIX-skew (Wang et al., 2009) and EMMIX-uskew (Lee and McLachlan, 2013b).

4 Other finite mixtures of skew distributions

A number of other asymmetric models have been put forward in recent years, including the multivariate normal-inverse Gaussian (MNIG) distribution (Karlis and Santourian, 2009), the multivariate shifted asymmetric Laplace (MSAL) distribution (Franczak et al., 2012), and the (restricted) multivariate skew t-normal (rMSTN) distribution (Lin et al., 2013).
The MNIG distribution is a flexible parametric family with four parameters. Like the skew t-distribution, the MNIG distribution can accommodate skewness and heavy tails in the data. Computation of the ML estimates of the parameters of the model is carried out by the EM algorithm, with closed-form E- and M-steps involving modified Bessel functions. The MSAL distribution is another alternative to the skew normal and skew t-distribution. As a three-parameter distribution, the MSAL distribution has parameters that controls its location, scale, and skewness. The EM algorithm for fitting mixtures of MSAL distributions is computationally straightforward compared to that for the FM-MNIG model and skew mixture distributions. The rMSTN distribution was introduced as a computational more feasible alternative to the MST distribution, where the skewing component of MST distribution is replaced by a (univariate) normal distribution function. The rMSTN distribution shares the same set of parameters as the rMST and uMST distributions, but considerable time is saved when implementing the E-step of the EM algorithm.

5 Model-based clustering with skew mixture models

5.1 Clustering the AIS data

Our first example on real data concerns the clustering of male and female athletes in Australian Institute of Sport (AIS) data. For illustration, we consider a bivariate subset of the data (Cook and Weisberg, 1994), consisting of the variables body mass index (BMI) and lean body mass (LBM). We fitted two component mixtures of MN, MT, rMSN, rMSTN, uMST, MSAL and MNIG distributions to the data. Table 1 list the number of misclassified

observations against the true clustering (male and female) for each model. Figure 2 shows the contours of the fitted density of each model. It can be observed from Table 1 that the mixtures with skew component distributions performed much better than the symmetric mixture models. Both the FM-uMST and FM-MSAL models have misallocated 17 observations, the smallest number among the seven models in this case. The FM-rMSN and FM-rMSTN models achieved comparable results in this example.

TABLE 1. Clustering performance of various multivariate mixture models on the AIS dataset.

Model	number of misclassified observations
FM-MN	93
FM-MT	51
FM-rMSN	18
FM-rMSTN	18
FM-uMST	17
FM-MSAL	27
FM-MNIG	17

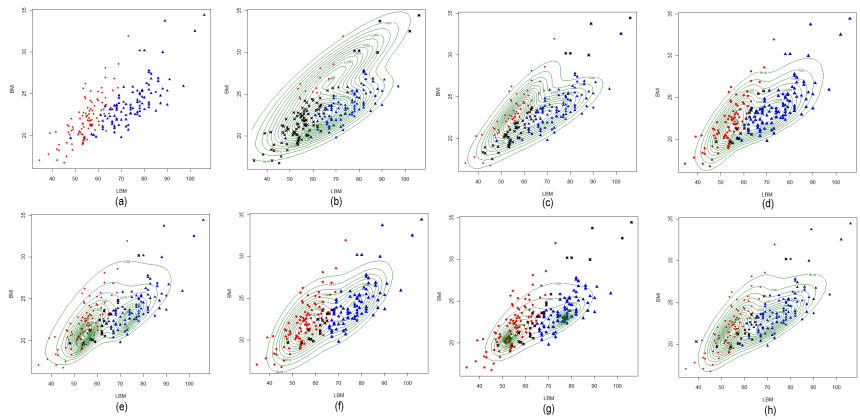


FIGURE 2. AIS dataset: Modelling the distribution of Australian male and female athletes. (a) Scatter plot of body mass index (BMI) and lean body mass (LBM) in two colours, red dots for make and blue triangles for female; (b) the fitted contours of the FM-MN model; (c) contour plot of the fitted FM-MT model; (d) the density contours of the fitted FM-rMSN model; (e) the contours of the densities of the fitted FM-rMSTN model; (f) the density contours of the fitted FM-uMST model; (g) the contours of the densities of the fitted FM-MSAL model; (h) contour plot of the fitted FM-MNIG model.

5.2 Automated gating of a DLBCL sample

We consider now the clustering of a trivariate Diffuse Large B-cell Lymphoma (DLBCL) dataset provided by the British Columbia Cancer Agency. The data contain fluorescent intensities of multiple conjugated antibodies (known as markers) stained on a sample of over 8000 cells derived from the lymph nodes of patients diagnosed with DLBCL. Each sample was stained with three markers CD3, CD5, and CD19. The task is to automatically gate the cells by clustering the data into four groups. Hence we fit four-component mixture models to the data.

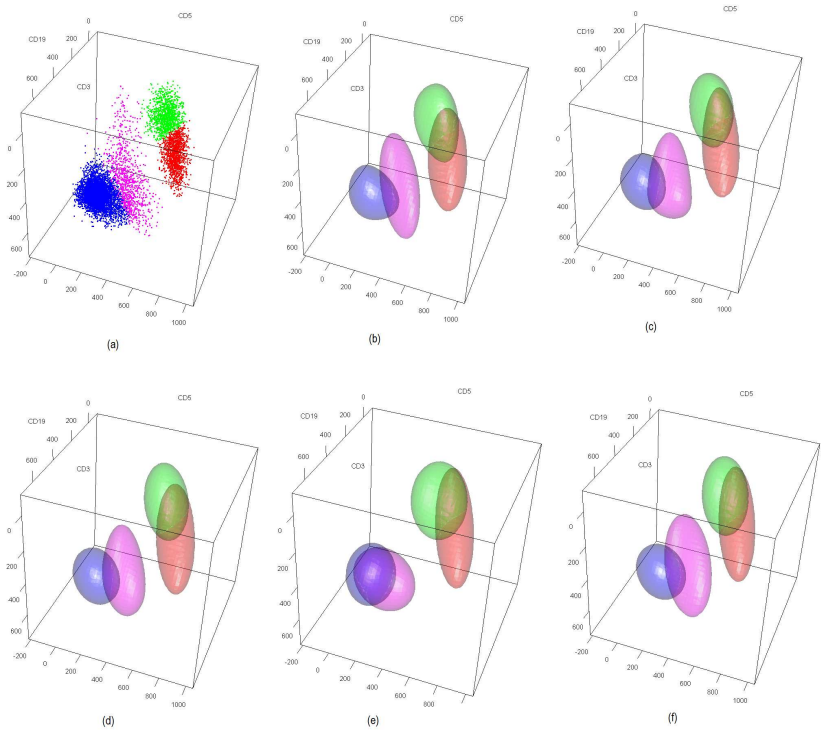


FIGURE 3. DLBCL dataset: Automated gating results of DLBCL sample using five different finite mixture models. The population of 8149 cells were stained with three fluorescence reagents - CD3, CD5, CD19. (a) manual expert clustering of the DLBCL into four groups; (b) the fitted component contours of the four-component FM-uMST model; (c) the contours of the component densities of the fitted restricted (FM-rMST) model; (d) the component contours of the fitted FM-rMSTN model; (e) the fitted component contours of the FM-MSAL model; (f) the contour plot of the fitted FM-MNIG model.

A scatterplot of the data is shown in Figure 3(a), where the dots are

coloured according to the clustering provided by human experts, which are taken as the 'true' class labels. Figure 3(b)-(e) shows the density contours of the components of the fitted FM-uMST, FM-rMST, FM-rMSTN, FM-MSAL, and FM-MNIG models respectively, which are displayed with matching colours to Figure 3(a). To assess the performance of these algorithms, we calculated the rate of misclassification against the 'true' results, given by choosing among the possible permutations of the cluster labels the one that gives the lowest value. A lower misclassification or error rate indicates a closer match between the true labels and the cluster labels given by the candidate algorithm. Note that dead cells were removed before evaluating the misclassification rate.

TABLE 2. Misclassification rates for various multivariate mixture models on the DLBCL dataset. Cells identified as dead cells were not included in the calculation of error rate.

FM-uMST	FM-rMST	FM-rMSTN	FM-MSAL	FM-MNIG
0.0464	0.0689	0.0688	0.3303	0.05948

From Table 2, it can be seen that the multivariate skew t-mixture models outperform the other methods in this dataset. This is also evident in Figure 3, where the component contours of the FM-uMST model resemble quite well the shape of the clusters identified by manual gating; see, for example, the density of the blue and pink cluster. The results from Table 2 reveal that the unrestricted model is more accurate than the restricted variant and the other three models. The FM-MNIG model also gave quite reasonable clustering results, but does not appear to be able to capture the shape of the blue cluster as well as the FM-uMST model. The FM-rMSTN model give an error rate comparable to that of the FM-rMST model. However, the FM-MSAL model does not perform as well, having difficulty in separating the lower two (blue and pink) clusters.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171 – 178.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society, Series B*, **65**, 367 – 389.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, **83**, 715 – 726.

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803 — 821.

Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Discase Mapping and Others*. New York: Chapman & Hall.

Branco, M. and Dey, D. (2001). A general class of multivariate skew-elliptical distributions. *Computational Statistics and Data Analysis*, **54**, 2926 – 2941.

Cook, R. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.

Dempster, A.P. and Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, **39**, 1 – 38.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*. London: Chapman & Hall.

Fraley, C. and Raftery, A.E. (1999). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *Computer Journal*, **41**, 578 – 588.

Franczak, B. and Browne, R., and McNicholas, P. (2012). Mixtures of shifted asymmetric Laplace distributions. *arXiv:1207.1727*.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.

Karlis, D. and Santourian, A. (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, **19**, 73 – 83.

Lachos, V., Ghosh, P., and Aerllano-Valle, R. (2010). Likelihood based inference for skew normal independent linear mixed models. *Statistica Sinica*, **20**, 303 – 322.

Lee, S.X. and McLachlan, G.J. (2011). On the fitting of mixtures of multivariate skew t-distributions via the EM algorithm. *arXiv:1109.4706*.

Lee, S.X. and McLachlan, G.J. (2013a). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*. To appear (DOI: 10.1007/s11222-012-9362-4).

Lee, S.X. and McLachlan, G.J. (2013b). EMMIX-uskew: an R package for fitting mixtures of multivariate skew t-distributions via the EM algorithm. *Journal of Statistical Software*. To appear.

Lee, S.X. and McLachlan, G.J. (2013c). On mixtures of skew-normal and skew t-distributions. *Advances in Data Analysis and Classification*. To appear.

Lin, T.-I., Ho, H.J., and Lee, C.R. (2013). Flexible mixture modelling using the multivariate skew-t-normal distribution. *Statistics and Computing*. To appear (DOI: 10.1007/s11222-013-9386-4).

Lin, T.-I. (2010). Robust mixture modeling using multivariate skew t distribution. *Statistics and Computing*, **20**, 343 – 356.

Lin, T.-I. (2009). Maximum likelihood estimation for multivariate skew-normal mixture models. *Journal of Multivariate Analysis*, **100**, 257 – 265.

Linsay, B.G. (1995). *Mixture Models: Theory, Geometry, and Applications*. Alexandria, Virginia: IMS.

McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications*. New York: Marcel Dekker.

McLachlan, G.J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Second Edition. Hoboken, New Jersey: John Wiley & Sons.

McLachlan, G.J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In *Lecture Notes in Computer Science*, Vol. 1451, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.). Berlin: Springer-Verlag, pp. 658 – 666.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*, New York: Wiley Series in Probability and Statistics.

Mengersen, K.L., Robert, C.P., and Titterington, D.M. (2011). *Mixtures: Estimation and Applications*. Chichester, United Kingdom: John Wiley & Sons.

Peel, D. and McLachlan, G.J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**, 339 – 348.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L.M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D.A., De Jager, P.L. and Mesirow, J.P.(2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences USA*, **106**, 8519 – 8524.

Titterington, D.M., Smith, A.F.M., and Markov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Wang, J., Boyer, J., and Genton, M. (2004). A skew-symmetric representation of multivariate distributions. *Statistica Sinica*, **14**, 1259 – 1270.

Wang, K., McLachlan, G.J., Ng, S.K., and Peel, D. (2009). EMMIX-skew: EM Algorithm for Mixture of Multivariate Skew Normal/t Distributions. URL: `http://www.maths.uq.edu.au/ gjm/mix_soft/EMMIX-skew`, R package version 1.0-12.

Regression Models for Expected Length of Stay

Hein Putter[1], Mia Klinten Grand[1]

[1] Leiden University Medical Center, The Netherlands

E-mail for correspondence: h.putter@lumc.nl

Abstract: In an ageing society, it is crucial to be able to estimate healthy life expectancy and expected life years spent in disability, also conditional on the current age and health/disability status of an individual. Moreover, it is important to be able to assess the effects of individual characteristics, like gender and socio-economic status, and of behavioral characteristics, like dietary habits and smoking, on these quantities. This paper proposes a simple and effective regression method, based on pseudo-observations, to address this question. An illustration based on the AHEAD study is provided.

Keywords: Survival analysis; Multi-state models; Healthy life expectancy; Pseudo-observations

1 Introduction

Over the 20th century, from the 1920's onward, the life expectancy of humans has increased an incredible 2.5 years every decade (Oeppen & Vaupel, 2002). The increase has been remarkably steady with no signs as yet that this trend is disappearing in the 21st century. Clearly this increased life expectancy will have a profound effect on modern society.
Among demographers there is a heavy debate, whether these additional life years are being spent in health (compression) or in disability (expansion). A distinction between life years spent in health or in disability is important, both for the well-being of individuals and for health resources. An important question is then how background characteristics of individuals, like gender and socio-economic status, and behavioral characteristics, such as dietary habits and smoking, influence expected (remaining) life spent in health or in disability. In a paper studying the effects of these factors on healthy life expectancy and expected life in disability, Reuser et al. (2009) summarized the most striking behavioral effects as "Smoking kills, obesity disables". To contribute to this debate there is a need for methods to assess and model expected remaining life in health and in disability for older people.

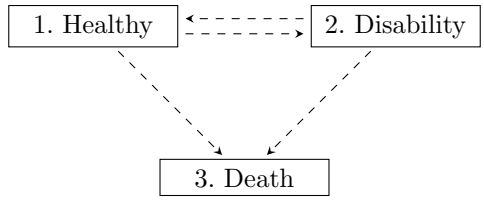


FIGURE 1. The healthy-disability-death multi-state model.

The typical approach used to address these questions is to view this in the context of a multi-state model (Putter et al., 2007). A reasonable multi-state model for the above healthy-disability debate is shown in Figure 1. It is an example of an illness-death model, with disability as the "illness" state. The illness-death model of Figure 1 is reversible, since recovery from disability is possible. Such multi-state models enable estimating the effect of explanatory factors on the transition intensities, but they do not give a direct quantification on the effect of these factors on healthy life expectancy. The objective of this paper is to propose regression models that directly quantify the effect of explanatory factors on healthy life expectancy and expected life in disability, while taking into account censoring.

2 Data

To illustrate our methods, the Asset and Health Dynamics Among the Oldest Old (AHEAD), now part of the wider US Health and Retirement Study (HRS), will be used (Juster & Suzman, 1995). The AHEAD survey includes a nationally representative sample of initially non-institutionalized persons born before 1923, aged 70 and older in 1993. The outcome of interest is survival (irrespective of the cause); the time scale is age. Table 1 shows the frequency in the AHEAD data of the time-fixed covariates considered in the illustration (body-mass index (BMI) and smoking status are assessed at entry into the study). Disability status is defined according to the Basic Activities of Daily Living (ADL) scale by Katz et al. (1963), which includes items for walking, bathing, dressing, toileting and feeding. A subject is defined to be ADL disabled here if he/she responds "with difficulty" for at least one of the ADL items.

In Section 3 we will study the dynamics of disability and recovery in the healthy-disability-death multi-state model of Figure 1. In this data, for a total of 4032 subjects, 1929 transitions from healthy to ADL disabled occurred and 679 recoveries (transitions from ADL disability to healthy). A total of 1994 deaths were observed, 922 from the healthy state and 1072 from ADL disability.

TABLE 1. Baseline covariates of the AHEAD study.

Covariate	N	(%)
Gender		
Male	1564	(39%)
Female	2468	(61%)
Education		
Less than high school	1736	(43%)
High school	1212	(30%)
Some college	1084	(27%)
BMI		
≤ 25	2244	(56%)
25 – 30	1388	(34%)
> 30	390	(10%)
Missing	10	
Smoking		
Never	1997	(50%)
Past	1683	(42%)
Current	324	(8%)
Missing	28	

3 Expected length of stay in multi-state models

Multi-state models, transition hazards and transition probabilities

A multi-state model is a random process $X(t)$ taking values in a finite state space, generally taken to be $\mathcal{K} = 1, \ldots, K$, with $X(t) = g$ meaning that the subject is in state g at time t. The multi-state model of Figure 1 has three states: 1=Healthy, 2=ADL disability, 3=Death. The two central quantities in multi-state models are the *transition intensity* or transition hazard from state g to state h, given as

$$\lambda_{gh}(t) = \lim_{dt \to 0} P(X(t + dt) = h | X(t) = g)/dt,$$

and the *transition probability* from state g to state h, given as

$$P_{gh}(s, t) = P(X(t) = h | X(s) = g).$$

Note that these quantities are only really meaningful in Markov multi-state models; otherwise the conditional probabilities will also depend on the further past.

It is straightforward to estimate transition intensities non-parametrically, by

$$d\hat{\Lambda}_{gh}(t) = \frac{dN_{gh}(t)}{Y_g(t)}.$$

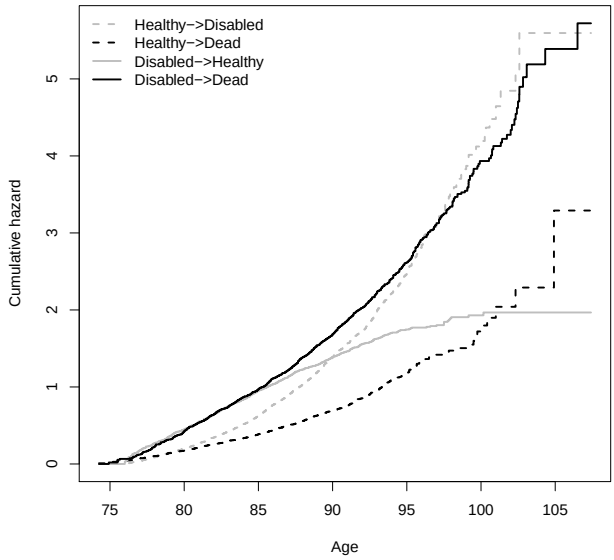


FIGURE 2. Nelson-Aalen estimates of the transition intensities in the healthy-disability-death multi-state model.

At each observed transition time t from g to h, the estimated cumulative transition hazard $\hat{\Lambda}_{gh}$ makes a jump, the size of which is the number of observed g to h transitions, $dN_{gh}(t)$, divided by the number of individuals in state g, $Y_g(t)$. Figure 2 shows the Nelson-Aalen estimates of the transition intensities in the healthy-disability-death multi-state model of Figure 1, based on the AHEAD data.

From these estimated transition hazards, transition probabilities may be estimated using the Aalen-Johansen (1978) estimator. The estimator is implemented in the R package `mstate` (de Wreede et al., 2010). Figure 3 shows the transition probabilities $P_{1h}(75, t)$ on the left (a) and $P_{2h}(75, t)$ on the right (b), for the different states h, for t between 75 and 110. The probabilities are shown in three shades of grey; the lightest corresponds to the healthy state ($h = 1$), the middle to the ADL disability state ($h = 2$), and the darkest to the death state ($h = 3$).

Expected length of stay

In this paper we are interested in the expected length of stay (ELoS) in a certain state h, when it is known that a subject is in state g at a certain time s. This expected length of stay is denoted by $E_{gh}(s, \tau)$, and can be

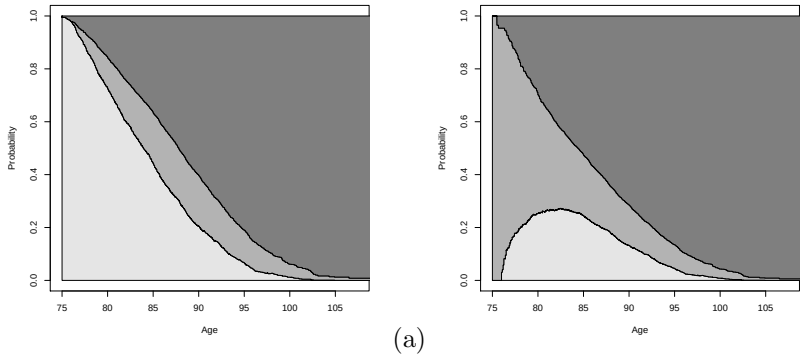


FIGURE 3. Transition probabilities in the healthy-disability-death multi-state model, given healthy at age 75 (a) and given ADL disabled at age 75 (b).

expressed as a functional of the transition probabilities $P_{gh}(s,t)$, as follows:

$$\begin{aligned} E_{gh}(s,\tau) &= E\int_s^\tau 1\{X(t)=h|X(s)=g\}dt \\ &= \int_s^\tau P\{X(t)=h|X(s)=g\}dt = \int_s^\tau P_{gh}(s,t)dt. \end{aligned} \quad (1)$$

A technical point should be mentioned. Since (estimates of) the transition probabilities $P_{gh}(s,t)$ may be positive for all t larger than some t^*, the integral to ∞ may be infinite. It is therefore common to consider the *restricted* ELoS instead; τ is some large value, for instance the last observed time point. The expected length of stay is an extension of expected (residual) life; the latter appears as a special case for the multi-state model with two states, alive and death.

Estimates of the (conditional) expected length of stay immediately present themselves by Equation (1), namely

$$\hat{E}_{gh}(s,\tau) = \int_s^\tau \hat{P}_{gh}(s,t)dt,$$

with $\hat{P}_{gh}(s,t)$ the Aalen-Johansen estimate of the transition probabilities (or indeed, some other estimate).

For illustration, we estimated the $E_{gh}(s,\tau)$, based on the AHEAD data and using the Aalen-Johansen estimates of the transition probabilities of Figure 3; we chose $\tau = 110$. Conditional on being healthy at age $s = 75$, the expected remaining healthy life until the age of $\tau = 110$ years, equals $\hat{E}_{11}(s,\tau) = 9.69$. Likewise, the expected remaining number of years in disability is much lower, as $\hat{E}_{12}(s,\tau) = 3.47$. Finally, the expected remaining number of years spent in the death state equals $\hat{E}_{13}(s,\tau) = 21.84$.

These corresponds to the areas of lightest grey (healthy), middle grey (ADL disabled) and the darkest grey (dead) parts in Figure 3(a). The total remaining life expectancy at age 75 for a healthy individual equals $\hat{E}_{11}(s,\tau) + \hat{E}_{12}(s,\tau) = 35 - \hat{E}_{13}(s,\tau) = 13.16$. Conditional on being ADL disabled at age 75, we obtain for the expected remaining years in health, disability, and death, $\hat{E}_{21}(s,\tau) = 3.52$, $\hat{E}_{22}(s,\tau) = 7.16$, and $\hat{E}_{23}(s,\tau) = 24.32$, respectively. These numbers now correspond to the areas of lightest grey (healthy), middle grey (ADL disabled) and darkest grey (dead) parts in Figure 3(b). The total expected remaining life years at age 75 for an individual who is ADL disabled equals $\hat{E}_{21}(s,\tau) + \hat{E}_{22}(s,\tau) = 35 - \hat{E}_{23}(s,\tau) = 10.68$, a difference of 2.5 years, compared to the individual who was healthy at age 75.

4 Regression models for ELoS

Covariates and ELoS in multi-state models

So far we have considered non-parametric estimates of the expected length of stay in a state, $E_{gh}(s,\tau)$. The real interest is in how certain characteristics of the subject and their behavior influence $E_{gh}(s,\tau)$. Interesting questions in the context of the AHEAD study are: is there a difference in expected healthy life between males and females and between lower and higher education? Also of interest is: can we confirm the findings of Reuser et al. (2009), "smoking kills" (life expectancy is shorter for smokers) and "obesity disables" (expected life in disability increases with higher BMI)? One way to assess the effect of covariates is to include covariates in the models for the transition intensities of the multi-state model. This may be done along the lines of the tutorial Putter et al. (2007). The basic idea is to fit a Cox model on the transition intensities

$$\lambda_{gh}(t \mid \mathbf{Z}) = \lambda_{gh,0}(t) \exp(\boldsymbol{\beta}_{gh}^{\mathrm{T}} \mathbf{Z}),$$

with separate (non-parametric) baseline hazards $\lambda_{gh,0}(t)$ and separate effects $\boldsymbol{\beta}_{gh}$ of the covariates $\mathbf{Z}$ for each transition. This model was fitted to the AHEAD data, with the additional proportionality constraint $\lambda_{23,0}(t) = \exp(\gamma)\lambda_{13,0}(t)$ between both baseline transition hazards into the death state. The hazard ratio $\exp(\gamma)$ quantifies the increase in death rate of an ADL-disabled individual, compared to a healthy individual with the same values of the other covariates. The results of this modeling approach are described in Tables 2 and 3. With respect to the disability and recovery rates in Table 2, it can indeed be seen that obesity disables, and that females have both a higher disability rate and a lower recovery rate (albeit not significant). The lower death rate of females compared to males is apparent from Table 3, as well as (at least the trend for) lower death rates of higher educated individuals. Very clear is also (smoking kills) the higher

TABLE 2. Covariate effects for the disability/recovery transitions in the multi--state model.

	Healthy → ADL		ADL → Healthy	
	HR	95% CI	HR	95% CI
Gender				
Female	1.28	1.16 – 1.42	0.88	0.73 – 1.05
Education				
High school	1.08	0.97 – 1.20	1.18	0.98 – 1.41
College	1.01	0.91 – 1.13	1.30	1.08 – 1.57
BMI				
25–30	1.18	1.07 – 1.30	1.22	1.03 – 1.43
> 30	1.38	1.18 – 1.62	0.92	0.72 – 1.17
Smoking				
Ever	1.08	0.98 – 1.20	1.02	0.86 – 1.22
Present	1.45	1.22 – 1.72	0.94	0.69 – 1.27

TABLE 3. Covariate effects for the death transitions in the multi-state model.

	Healthy → Death		ADL → Death	
	HR	95% CI	HR	95% CI
Gender				
Female	0.67	0.58 – 0.72	0.56	0.48 – 0.65
Education				
High school	0.87	0.75 – 1.02	0.93	0.81 – 1.07
College	0.77	0.66 – 0.91	0.86	0.74 – 1.00
BMI				
25–30	0.75	0.65 – 0.87	0.65	0.57 – 0.75
> 30	0.92	0.72 – 1.18	0.67	0.55 – 0.83
Smoking				
Ever	1.24	1.07 – 1.44	1.14	0.99 – 1.31
Present	1.82	1.46 – 2.28	1.37	1.09 – 1.72
ADL			3.11	2.45 – 3.95

death rate for (ever and present) smokers. Finally, ADL disability itself increases the death rate with a factor of more than 3. Interesting is that individuals with high BMI have lower death rates, compared to low BMI. It is possible that underlying disease is a confounder; subjects with severe disease both have higher death rates and tend to have very low BMI.

Interesting though these results are, it is not straightforward to extract the effect of a covariate on expected length of stay, for instance on healthy life expectancy. In case of a single covariate, if that covariate is categorical, one can estimate all model-based transition hazards, from those the transition

probabilities for each value of the covariate, and estimate ELoS based on these transition probabilities. If that covariate is continuous, however, the effect on ELoS will not be linear and difficult to assess. The situation is even harder in multivariate models. In that case the effect of a covariate of interest depends on the values of other covariates. All in all, the effect of a covariate on expected length of stay is difficult to assess and quantify.

5 Pseudo-observations

In the absence of censoring, the situation would be very simple. In that case one would actually observe y_i, the length of stay in the state of interest, for each subject i. In that case one would probably use a linear model, specifying

$$y_i = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_i + \varepsilon_i,$$

where ε_i is an error term with a certain (for instance normal) distribution. The problem is how to deal with censoring.

In such cases, pseudo-observations are very useful. They have been used to model probabilities in competing risks and multi-state models and to model (restricted) mean survival, by Andersen et al. (2003) and further papers. The basic idea is as follows: we have an estimator $\hat{\theta} = \hat{E}_{gh}(s, \tau)$ of $\theta = E_{gh}(s, \tau)$. For each subject i in the sample, we can also define the leave-one-out estimator $\hat{\theta}^{-i}$, i.e., the same estimate, but based on the whole sample except observation i. The pseudo-observation of $E_{gh}(s, \tau)$ for subject i is then defined as

$$\hat{\theta}_i = n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}^{-i}.$$

It is not hard to show that, in the absence of censoring, where $y_i =$ LoS of subject i can be observed, we have $\hat{\theta}_i = y_i$.

In the presence of censoring the incompletely observed y_i are replaced by $\hat{\theta}_i$, and $\hat{\theta}_i$ is used as outcome variable in a generalized linear model (GLM) for $\theta_i = E_{gh}(s, \tau | \mathbf{Z}_i)$,

$$\theta_i = g^{-1}(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_i) = \mu_i(\boldsymbol{\beta}),$$

Estimates of $\boldsymbol{\beta}$ may be obtained using generalized estimating equations (GEE). The estimating equations are given by

$$U(\boldsymbol{\beta}) = \sum_i \left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} V_i^{-1} (\hat{\theta}_i - \theta_i) = 0, \tag{2}$$

where V_i is a working variance matrix. After having obtained an estimate $\hat{\boldsymbol{\beta}}$ from solving (2), the covariate matrix of $\hat{\boldsymbol{\beta}}$ may be obtained by a sandwich estimator. For more details see Andersen et al. (2003). The method can be

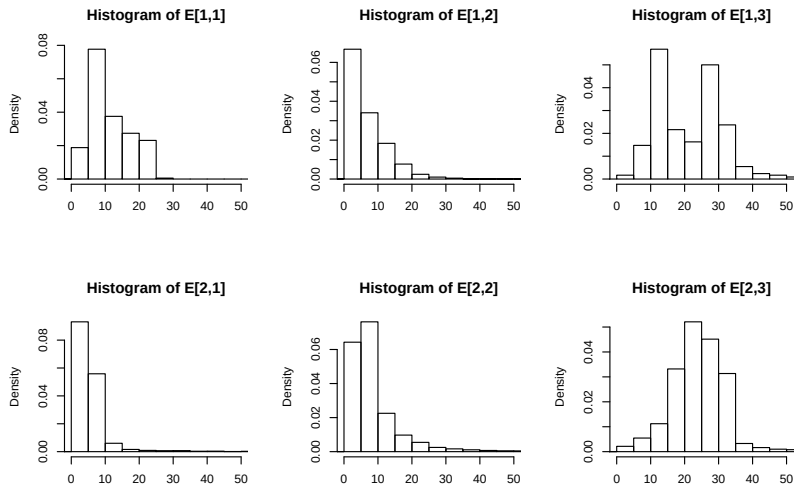


FIGURE 4. Pseudo-observations of $E_{1h}(s,\tau)$ and $E_{2h}(s,\tau)$ in the AHEAD data.

implemented in standard statistical software (for instance `PROC GENMOD` in SAS and the `geepack` package in R).

Based on the AHEAD data, pseudo-observations were calculated for $E_{1h}(s,\tau)$ and $E_{2h}(s,\tau)$, i.e., for expected remaining life years in state h, given healthy at age s, and given ADL disabled at age s, respectively, with $s = 75$, $\tau = 110$, and $h = 1$ (healthy life), $h = 2$ (life in disability) and $h = 3$ (death). Figure 4 shows histograms of these pseudo-observations.

The results of the linear modeling procedure, taking the identity link, are shown in Tables 4 and 5. An individual who is healthy at age 75 (Table 4) spends more remaining life years in health than in disability. Females spend more life years in disability, compared to males. Higher education, especially college education, is associated with more life years, mostly in health. Higher BMI is associated with more life years in disability (obesity disables), and smoking with fewer life years in health and in disability. These same general trends can be observed in Table 5, for individuals who are ADL disabled at age 75. Notice the larger standard errors in Table 5, caused by the fact that the majority of person-time is spent in health in the AHEAD data. The most striking difference with respect to the result of Table 4 is the intercept. Individuals who are ADL disabled at age 75 spend far fewer years in health and more years in disability; their total remaining life time is lower than for individuals who are healthy at age 75.

TABLE 4. Covariate effects for the expected length of stay in health $E_{11}(s,\tau)$ and disability $E_{12}(s,\tau)$, given healthy at age $s = 75$.

	ELoS in health			ELoS in disability		
	B	SE	P	B	SE	P
Intercept	9.71	0.39	< 0.0001	1.42	0.30	< 0.0001
Gender						
Female	0.51	0.32	0.11	2.14	0.23	< 0.0001
Education						
High school	0.05	0.34	0.89	0.24	0.27	0.37
College	1.25	0.32	0.0001	0.58	0.26	0.028
BMI						
25–30	0.51	0.31	0.099	1.67	0.26	< 0.0001
> 30	-0.95	0.51	0.064	1.84	0.38	< 0.0001
Smoking						
Ever	-1.04	0.31	0.0007	-0.39	0.24	0.11
Present	-3.77	0.67	< 0.0001	-0.80	0.42	0.054

TABLE 5. Covariate effects for the expected length of stay in health $E_{21}(s,\tau)$ and disability $E_{22}(s,\tau)$, given ADL disabled at age $s = 75$.

	ELoS in health			ELoS in disability		
	B	SE	P	B	SE	P
Intercept	1.47	0.74	0.046	4.63	0.71	< 0.0001
Gender						
Female	1.63	0.57	0.004	3.38	0.58	< 0.0001
Education						
High school	1.05	0.60	0.082	-0.65	0.67	0.33
College	1.20	0.62	0.055	-0.10	0.62	0.87
BMI						
25–30	1.01	0.55	0.066	2.81	0.59	< 0.0001
> 30	1.53	1.24	0.220	2.92	1.01	0.0039
Smoking						
Ever	-0.07	0.55	0.90	-1.01	0.55	0.065
Present	-0.35	1.23	0.77	-1.38	1.04	0.19

6 Discussion

In this paper we have proposed the use of pseudo-observations to obtain direct regression models for expected length of stay in health and in disability. We applied the method to the AHEAD data. The analysis presented in this paper is not comprehensive; it is only meant to illustrate the method. The use of pseudo-observations allows direct modeling of remaining life in

health/disability. It circumvents modeling of transition intensities.
We have only considered a single value of s, namely $s = 75$. Dynamic versions, evaluating how $E_{gh}(s, \tau)$ varies with s, will be investigated in future work. These models are extensions of the proportional mean residual life model (Oakes & Dasu, 1990).
Other useful measures may also be studied in much the same way. One of the possibilities is regression models for quality-adjusted (remaining) life years. A utility q_h (per time unit spent in state) is then assigned to each state h, and one is interested in $\sum_h q_h E_{gh}(s, \tau)$. In another application, q_h could be (medical) costs associated with being in state h. Another useful measure may be the proportion of remaining life spent in health; in our setting that would be $E_{g1}(s, \tau)/(E_{g1}(s, \tau) + E_{g2}(s, \tau))$.

References

Aalen, O.O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, **5**, 141 – 150.

Andersen, P.K., Klein, J.P. and Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, **90**, 15 – 27.

Juster, F.T. and Suzman, R. (1995). An overview of the Health and Retirement Study. *The Journal of Human Resources*, **30**, 7 – 56.

Katz, S., Ford, A.B., Moskowitz, R.W., Jackson, B.A. and Jaffe, M.W. (1963). The index of ADL: a standardized measure of biological and psychosocial function. *Journal of the American Medical Association* **185**, 914 – 919.

Oakes, D. and Dasu, B. (1990). A note on residual life. *Biometrika*, **77**, 409 – 410.

Oeppen, J. and Vaupel, J.W. (2002). Broken limits to life expectancy. *Science*, **296**, 1029 – 1031.

Putter, H., Fiocco, M. and Geskus, R. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* **26**, 2277 – 2432.

Reuser, M., Bonneux, L.G. and Willekens, F.J. (2009). Smoking kills, obesity disables: a multistate approach of the US Health and Retirement Survey. *Obesity*, **17**, 783 – 789.

de Wreede, L.C., Fiocco, M. and Putter, H. (2010). The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, **99**, 261 – 274.

Part II - Contributed Papers (volume 1)

Model selection for penalized Gaussian Graphical Models

Antonino Abbruzzo[1], Ivan Vujacic[2], Ernst C. Wit[2], Angelo M. Mineo[1]

[1] Department of Statistics, University of Palermo, Viale delle Scienze, Italy
[2] Johann Bernoulli Institute, University of Groningen, Groningen 9747 AG, The Netherlands

E-mail for correspondence: antonino.abbruzzo@unipa.it

Abstract: High-dimensional data refers to the case in which the number of parameters is of one or more order greater than the sample size. Penalized Gaussian graphical models can be used to estimate the conditional independence graph in high-dimensional setting. In this setting, the crucial issue is to select the tuning parameter which regulates the sparsity of the graph. In this paper, we focus on estimating the "best" tuning parameter. We propose to select this tuning parameter by minimizing an information criterion based on the generalized information criterion and to use a stability selection approach in order to obtain a more stable graph. The performance of our method is compared with the state-of-art model selection procedures, including Akaike information criterion and Bayesian information criterion. A simulation study shows that our method performs better than the AIC, BIC.

Keywords: Gaussian Graphical Model; Penalized likelihood; Information Criteria, Stability Selection.

1 Penalized Gaussian Graphical Models

A graph consists in a set of V nodes and a set of $E \subset V \times V$ links. A link is undirected if $(j,k) \in E$ and $(k,j) \in E$ whereas a link is directed from link j to vertex k if $(j,k) \in E$ and $(k,j) \notin E$. In a graphical model, the nodes of the graph are in one-to-one correspondence with a set of random variables $(X^{(1)}, \ldots, X^{(d)}) \sim P$. The links represent conditional dependence relationships between random variables. The pair (G,P) is referred to as a graphical model. Consider a multivariate normal distribution for the set of random variables $(X^{(1)}, \ldots, X^{(d)}) \sim N_d(0,\Sigma)$, that defines a *Gaussian graphical model*. The most important property of the GGMs which make their use appealing in order to analyse conditional independence between random variables is that the links in a GGM are given by the inverse of the

covariance matrix:

$$(j,k) \text{ and } (k,j) \notin E \leftrightarrow X^{(j)} \perp X^{(k)} | X^{(V\setminus\{(j,k)\})} \leftrightarrow \Sigma^{-1}_{jk} = 0.$$

Consider n i.i.d samples from $N_d(0,\Sigma)$. The negative log-likelihood for data $\mathbf{x} = (x^{(1)},\ldots,x^{(d)})^T$ is

$$l(\Sigma;\mathbf{x}) = \frac{n}{2}\left[\log\left(\det(\Theta)\right) + \operatorname{tr}\left(S\Theta\right) + D\right], \tag{1}$$

where $S = (1/n)\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$, $\Theta = \Sigma^{-1}$, and D is a constant with respect to Θ. The maximum likelihood estimator for Θ cannot be obtained when the sample size is smaller than the number of variables. However, high-dimensional data refers to this situation. In order to deal with high-dimensional data, we adopt a penalized maximum likelihood approach. Specifically, we add an ℓ_1-norm term $||\Sigma^{-1}||_1 = \sum_{j<k}|\Sigma^{-1}_{j,k}|$ to the negative log-likelihood (1). The estimator is found as follows:

$$\hat{\Sigma}(\lambda) = \underset{\Sigma^{-1}\succ 0}{\operatorname{argmin}}\left[-\log(\det\left(\Sigma^{-1}\right)) + \operatorname{tr}(S\Sigma^{-1}) + \lambda||\Sigma^{-1}||_1\right]. \tag{2}$$

The minimization is taken over the set of positive definite matrix, i.e. $\Sigma^{-1} \succ 0$. The ℓ_1-penalized GGM has the property that it shrinks some of the elements in the precision matrix to zero, so it is related to a sparsity assumption of the precision matrix. The minimization problem is convex and fast algorithms have been proposed to find a solution of (2). In this paper, we use the graphical lasso algorithm (glasso) proposed by Friedam et al (2008).

2 Model selection for penalized GGMs

Model selection for penalized Gaussian graphical model is referred to as the choice of the tuning parameter λ. The tuning parameter λ must be a positive number from zero to infinity. Specifically, we are interested in estimating the set of present links for the conditional independence graph:

$$E_0 = \{(j,k) \in V \times V : \Theta_{jk} \neq 0\}.$$

For estimating the edge set E_0, we can use the estimator:

$$\hat{E}(\lambda) = \{(j,k) \in V \times V : \hat{\Theta}_{jk}(\lambda) \neq 0\}.$$

where $\hat{E}(\lambda)$ depends on the tuning parameter λ which regulates the sparsity of the graph. The greater is this tuning parameter the sparser is the graph. The most used methods for choosing the regularization parameter are the k-fold cross-validation and measures based on information criteria such AIC and BIC (Yuan et al., 2007). Though these methods have good theoretical properties in low dimensions, they are not suitable for high-dimensional problems. Recently, Liu et al. (2010) proposed a new approach (StARS) to model selection, based on sub-sampling.

2.1 Generalized Information Criterion for penalized GGM

In this subsection, we derive a criterion to select the regularization parameter for penalized GGMs, whose penalty term can be approximated by function (3), based on the generalized information criterion (GIC) (Konishi, 1996).

In order to derive the GIC, the penalized likelihood function (2) should be twice differentiable with respect to Σ. However, the ℓ_1-norm function is not differentiable. We overcome this problem by using the smooth approximation in (3) (Schmidt et al., 2007) as penalty term instead of $||\Sigma^{-1}||_1$. The smooth approximation is:

$$P_\alpha(\Theta) = \sum_{i<j} |\Theta_{ij}|_\alpha, \tag{3}$$

where $|\Theta_{ij}^{-1}|_\alpha = (1/\alpha)[\log\{1+\exp(-\alpha\Theta_{ij})\} + \log\{1+\exp(\alpha\Theta_{ij})\}]$. Then, we write our minimization problem as:

$$\hat{\Theta}(\lambda,\alpha) = \underset{\Theta \succ 0}{\operatorname{argmin}} \left[-\log(\det(\Theta)) + \operatorname{tr}(S\Theta) + \lambda P_\alpha(\Theta)\right], \tag{4}$$

which is what will be referred to as the approximated problem. This quantity depends on α which approximates the ℓ_1 solution when it becomes large enough, i.e.

$$\lim_{\alpha\to\infty} \hat{\Theta}(\lambda,\alpha) = \hat{\Theta}(\lambda). \tag{5}$$

Estimator $\hat{\Theta}(\lambda,\alpha)$ of Θ belongs to the class of M-estimators, since it can be defined as a solution of the equation

$$\sum_{k=1}^{n} \psi(x_k, \hat{\Theta}(\lambda,\alpha)) = 0, \tag{6}$$

where $\psi(\mathbf{x}_k,\Theta) = vec\{Dl_k(\Theta;\mathbf{x}_k) - \lambda D P_\alpha(\Theta)\}$ and 0 is vector of zeros of dimension p^2. The GIC for M-estimators is given by:

$$GIC(\lambda,\alpha) = -2\sum_{k=1}^{n} l_k(\hat{\Sigma}(\lambda,\alpha);\mathbf{x}_k) + 2\operatorname{tr}\{R(\alpha)^{-1}Q(\alpha)\}, \tag{7}$$

where $R(\alpha)$ and $Q(\alpha)$ are square matrices of order p^2

$$R(\alpha) = -\frac{1}{n}\sum_{k=1}^{n}\{D\psi(\mathbf{x}_k,\Theta)\}^T\big|_{\Theta=\hat{\Theta}(\lambda,\alpha)}, Q(\alpha) = \frac{1}{n}\sum_{k=1}^{n}\psi(\mathbf{x}_k,\Theta(\lambda,\alpha))Dl_k(\Theta)\big|_{\Theta=\hat{\Theta}(\lambda,\alpha)}.$$

Applying (7) to the approximated problem, and using formula (5) we find a closed-form expression for the GIC which is given by:

$$GIC(\lambda) = -n\left[\log(\det(\hat{\Theta}(\lambda))) - \operatorname{tr}(S\hat{\Theta}(\lambda))\right] + 2\operatorname{tr}(R^{-}Q). \tag{8}$$

where Q and R^- are square matrices of order d^2:

$$R^- = 2\left[\hat{\Theta} \otimes \hat{\Theta}\right], \qquad Q = \frac{1}{4}\left[\frac{1}{n}\sum_{i=1}^{n} vecS_i vecS_i^T - vecS vecS^T\right].$$

After some algebra the trace term can be re-written as:

$$\mathrm{tr}(R^- Q) = \frac{1}{2}\left[\frac{1}{n}\sum_{k=1}^{n}\left[\oplus(S_k \odot \hat{\Theta} S_k \hat{\Theta})\right] - \left[\oplus(S \odot \hat{\Theta} S \hat{\Theta})\right]\right],$$

where $\odot$ is the componentwise product, and $\oplus$ is the componentwise sum. The calculation of GIC is feasible for high-dimensional data.

2.2 Combining stability selection with Generalized Information Criterion for penalized GGM

In this subsection, we propose to use a two steps procedure in order to increase the performance of our estimation for the precision matrix Θ. The first step is to select a tuning parameter λ by minimizing the GIC proposed in (8). The second step consists of using the StARS approach in order to obtain a more stable graph. In particular, we do a stability selection for a single regularization parameter $\hat{\lambda}_{GIC}$, i.e. the best tuning parameter according to GIC. When we do stability selection for a single parameter we obtain as many graphs as the number of sub-samplings we consider. For example, if we consider 20 sub-samples we obtain 20 estimated precision matrix. For each of the precision matrix we obtain a set $\hat{E} = \{(j,k) \in V \times V : \hat{\Theta}^{-1} \neq 0\}$. By counting the number of times a link is present in the set $\hat{E}$ and by dividing this number by the number of sub-samples we obtain an empirical probability of the presence of a link. Let A be the adjacent matrix of the empirical probability, then an element A_{jk} is 1 if the empirical probability of the presence of the link (j,k) is greater than a threshold δ and zero otherwise. In order to incorporate the threshold in our method we indicate it as GICS$_\delta$.

3 Simulation Study

In a simulation study, we evaluate the performance of AIC, BIC, StARS, GIC, and GICS$_\delta$, with respect to five measures: Precision, true positive rate (TPR), false positive rate (FPR), F1score and Accuracy. We have conducted an extensive simulation study in which we take under control the sample size n, and the number of nodes d in the graph. In this paper, we present the results only for $d = 30$ and $n = 44, 87, 130, 174$ due to lack of space. Note that, the total number of possible parameters in a Gaussian graphical model is given by $d \times (d-1)/2$ which means that for $d \geq 30$ we have already to deal with more than 435 parameters. Moreover, the total

number of present links is around $d/3$ so that the simulated graphs are sparse. In Table 1, we show the mean and standard deviation (numbers between brackets) after 50 runs.

TABLE 1. Mean and standard deviation in brackets of Precision, true positive rate (TPR), false positive rate (FPR), F1score and Accuracy after 50 runs where the number of nodes is $d = 30$ and the sample size is $n = 87$.

	Precision	TPR	FPR	F1score	Accuracy
AIC	0.15(0.01)	0.98(0.02)	0.85(0.01)	0.27(0.02)	0.27(0.06)
BIC	0.88(0.19)	0.14(0.20)	0.12(0.19)	0.19(0.15)	0.87(0.01)
GIC	0.25(0.02)	0.93(0.03)	0.75(0.02)	0.39(0.03)	0.60(0.04)
StARS	0.14(0.01)	0.99(0.01)	0.86(0.01)	0.25(0.01)	0.18(0.01)
GICS_0.85	0.42(0.04)	0.80(0.05)	0.58(0.04)	0.55(0.04)	0.82(0.03)
GICS_0.90	0.46(0.05)	0.76(0.06)	0.54(0.05)	0.57(0.05)	0.85(0.02)
GICS_0.95	0.54(0.06)	0.67(0.06)	0.46(0.06)	**0.60**(0.05)	**0.88**(0.02)

The best results in terms of F1score and Accuracy measures were obtained by GIC-StARS-0.95. GIC has higher TPR and FPR so that it performs overall poorer than GICS. In fact, by using GICS with a threshold fixed at 0.95 we are able to reduce the FPR. Obviously, the TPR is also reduced but this reduction is acceptable as shown by the higher scores both in F1score and Accuracy. When we reduced the threshold (0.90 and 0.85) both the TPR and FPR increase. The results, in terms of F1score and Accuracy, are still much better than AIC, BIC, StARS and GIC. These results confirm that the GIC is a good method to select the tuning parameter λ and that the second step in which we use StARS is necessary to reduce the FPR. Note that, StARS needs a double loop in order to find the best λ which makes it computationally intensive. Whereas GIC reduces this computational burden.

4 Conclusion

High-dimensional data arises in many fields of science. The study of the structure, in terms of conditional independence graph, can be carry out with penalized Gaussian graphical models. These models allow us to visualize with a graph the conditional independence among the set of random variables. In this paper, we presented penalized GGM in order to model high-dimensional data. These models make use of the sparsity assumption, i.e. many parameters are not statistically significant. In this setting, we proposed to approximate the ℓ_1 term with a smooth and convex function in order to find a closed-form expression for the generalized information criterion (GIC). We select the tuning parameter that minimize the GIC.

Then, we combined the GIC with a stability selection approach which allow us to obtain a more stable graph. In a simulation study, we showed the performance of GICS_δ with respect to, AIC, BIC and StARS, in terms of five measures: Precision, TPR, FPR, F1score and Accuracy. The results of GICS_δ look very promising and help us to improve the number of links correctly estimated.

References

Friedman, Jerome and Hastie, Trevor and Tibshirani, Robert (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432 – 441.

Konishi, Sadanori and Kitagawa, Genshiro (1996). Generalised information criteria in model selection. *Biometrika*, **4**, 875 – 890.

Liu, Han and Roeder, Kathryn and Wasserman, Larry (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *arXiv preprint arXiv:1006.3316*.

Mazumder, Rahul and Hastie, Trevor (2011). Exact covariance thresholding into connected components for large-scale graphical lasso. *arXiv preprint arXiv:1108.3829*.

Schmidt, Fung and Rosales (2007). Fast optimization methods for l1 regularization: Acomparative study and two new approaches suplemental material. *In Proceedings of European Conference on Machine Learning*, 286 – 297.

Yuan, Ming and Lin, Yi (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19 – 35.

Mixed estimation technique in semi-parametric space-time point processes for earthquake description

Giada Adelfio[1], Marcello Chiodi[1]

[1] Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Italia

E-mail for correspondence: giada.adelfio@unipa.it

Abstract: An estimation approach for the semi-parametric intensity function of a particular space-time point process is introduced. In particular we want to account for the estimation of parametric and nonparametric components simultaneously, applying a forward predictive likelihood to semi-parametric models. For each event, the probability of being a background event or one belonging to a seismic sequence is therefore estimated.

Keywords: nonparametric estimation; forward predictive likelihood; ETAS model; point process; earthquakes.

1 Introduction

Prediction of large earthquakes is often complicated by the presence of clusters of aftershocks, superimposed to the persistent background seismicity. Indeed, earthquake clusters, formed by the main event of each sequence, its foreshocks and its aftershocks, may complicate the statistical analysis of the background seismic activity that might be related to changes in the tectonic field. Since the seismogenic features controlling the kind of seismic release of background and clustered seismicity are not similar, sometimes a preliminary subdivision or declustering of a seismic catalog is useful to study separately the features of independent events and triggered ones (Adelfio et al., 2006).
In previous works (Adelfio, 2010; Adelfio et al., 2010) we proposed a clustering technique to separate and find out the two main components of seismicity, i.e. the background seismicity and the triggered one.
Adelfio et al. (2010) presented a seismic sequences detection technique based on MLE of parameters, that identifies the conditional intensity function of a model describing the seismic activity as a clustering-process, like ETAS model (Epidemic Type Aftershocks-Sequences model; Ogata, 1988).
In Adelfio (2010) nonparametric methods are used to estimate the inten-

sity function of a space-time point process and clustering results are interpreted by a second-order diagnostic approach (Adelfio and Schoenberg, 2009). Zhuang et al. (2002) proposed a stochastic method associating to each event a probability to be either a background event or an offspring generated by other events. A probabilistic clustering approach, providing an uncertainty about an object's class membership, can be provided by latent clustering analysis (Fraley and Raftery, 2002).
In this paper, we propose an estimation of the space-time intensity of the generating point process of the different components, that accounts simultaneously for the estimation of parametric and nonparametric components applying a forward predictive likelihood estimation approach to semi-parametric models (Chiodi and Adelfio, 2011). According to this approach we estimate, for each event, the probability of being a background event or one belonging to a seismic sequence.
In section 2 some formal definitions of point processes are recalled. A new method for nonparametric estimation is introduced in section 3; the simultaneous approach for nonparametric and parametric estimation is proposed in section 4.

2 Intensity function in point processes and ETAS model

Point process is a random collection of points, each one representing the time and space coordinates of a single event.
Let $Z^d = S^{d-1} \times T$ be a general $d-$dimensional closed region, with S^{d-1} a two or three dimensional space. Any analytic space-time point process is uniquely characterized by its associated *conditional intensity function* (Daley and Vere-Jones, 2003) defined as the frequency with which events are expected to occur around a particular location in time and space, conditional on the prior history $\mathcal{H}_t$ of the point process up to time t, i.e.:

$$\lambda(\mathbf{z}) = \lambda(\mathbf{s}, t|H_t) = \lim_{\Delta t, \Delta \mathbf{s} \to 0} \frac{\mathrm{E}\left[\#(t + \Delta t, \mathbf{s} + \Delta \mathbf{s}|H_t)\right]}{\Delta t \Delta \mathbf{s}}$$

where H_t is the space-time occurrence history of the process up to time t, $\Delta t, \Delta \mathbf{s}$ are time and space increments, $\mathrm{E}\left[\#(t + \Delta t, \mathbf{s} + \Delta \mathbf{s}|H_t)\right]$ is the history-dependent expected number of events occurring in the volume $\{[t, t+\Delta t) \times [\mathbf{s}, \mathbf{s} + \Delta \mathbf{s}]\}$. Generally, intensities $\lambda(\mathbf{z})$ depend on some unknown parameter $\boldsymbol{\psi}$, so that we have $\lambda(\mathbf{z}, \boldsymbol{\psi})$. For example, in a semi-parametric context, $\boldsymbol{\psi}$ could contains smoothing parameters.
Let denote a generic estimator of $\boldsymbol{\psi}$, based on observations until t_k, by $\hat{\boldsymbol{\psi}}(H_{t_k}) \equiv \hat{\boldsymbol{\psi}}((\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_i, ..., \mathbf{z}_k)$.
Assume that a realization of the process is observed in the space region $\Omega_{\mathbf{s}}$ and the time interval $(T_0; T_{max})$. The log-Likelihood for the point process, given the k observed values $\mathbf{z}_i$ and computed using the estimator

$\hat{\boldsymbol{\psi}}(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_i, ..., \mathbf{z}_k)$ is:

$$\log L(\hat{\boldsymbol{\psi}}(H_{t_k}); H_{t_k}) = \sum_{i=1}^{k} \log \lambda(\mathbf{z}_i; \hat{\boldsymbol{\psi}}(H_{t_k})) - \int_{T_0}^{T_{max}} \int_{\Omega_{\mathbf{s}}} \lambda(\mathbf{z}; \hat{\boldsymbol{\psi}}(H_{t_k}))\, d\mathbf{s}\, dt \tag{1}$$

In seismological context, the Epidemic Type Aftershocks-Sequences (ETAS) model is widely used (Ogata, 1988). The conditional intensity function of the ETAS model is defined as the sum of a term describing the spontaneous activity (background) and one relative to the induced seismicity:

$$\lambda_{\boldsymbol{\theta}}(x, y, t, m|\mathcal{H}_t) = \mu f(x, y) + \tau_{\boldsymbol{\phi}}(t, x, y) \tag{2}$$

with $\boldsymbol{\theta} = (\boldsymbol{\phi}, \mu)^{\mathrm{T}}$, that is the vector of parameters of the induced intensity ($\boldsymbol{\phi}$) together with the parameter of the background general intensity (μ) and

$$\tau_{\boldsymbol{\phi}}(t, x, y) = \sum_{t_j < t} g(t - t_j; \boldsymbol{\phi}) s(x - x_j, y - y_j|m, \boldsymbol{\phi}).$$

In the ETAS model, background seismicity is assumed to be stationary in time, while the occurrence rate of aftershocks at time t, following the earthquake of time t_j and magnitude m_j, is described by the following parametric model:

$$g(t - t_j|m_j) = \frac{\kappa\ e^{(\alpha-\gamma)\ (m_j - m_0)}}{(t - t_j + c)^p}, \quad \text{with} \quad t > t_j$$

where p is useful for characterizing the pattern of seismicity, indicating the decay rate of aftershocks in time.
For the spatial distribution, conditioned to magnitude of the generating event, the following distribution is often used:

$$s(x - x_j, y - y_j|m_j) = \left\{ \frac{(x - x_j)^2 + (y - y_j)^2}{e^{\gamma\ (m_j - m_0)}} + d \right\}^{-q}$$

It relates the occurrence rate of aftershocks to the mainshock magnitude m_j, through the parameters α, γ. m_0 is a given lower threshold of magnitude, d and q two parameters related to the spatial influence of the mainshock.
The simultaneous estimation of the background intensity and triggered intensity components of ETAS model is a crucial statistical issue. While the first component $f(x, y)$ is usually estimated by nonparametric techniques, $\boldsymbol{\theta}$ is estimated by ML approach. Zhuang et al. (2002) estimated the probability for each event of being a background event ($\rho_i, i = 1, ..., n$) in order to provide a random classification of events and obtain a thinned catalog, that includes events with a bigger probability of being mainshock, which spatial intensity is described by nonhomogeneous Poisson process.

In this paper, according to Console et al. (2010), we use ρ_i as weights for the kernel estimation of the background seismicity to get a simultaneous estimate of the intensity components of the ETAS model (2). For nonparametric estimation we propose the use of an estimation procedure based on the subsequent increments of likelihood obtained adding an observation one at a time, reported in the next section.

3 Forward predictive likelihood (FLP)

To estimate the nonparametric component we use an approach proposed in Chiodi and Adelfio (2011) to measure the ability of the observations and estimation until t_k to give information on the next observation.
Let $\hat{\boldsymbol{\psi}}(H_{t_k})$ be smoothing constants in a nonparametric context, based on the observed history up to t_k. Let $\log L(\hat{\boldsymbol{\psi}}(H_{t_k}); H_{t_{k+1}})$ be the likelihood computed on the first $k+1$ observations, but using the estimates based on first k. We measure the *predictive information* of the first k observations on the $k+1$-th as:

$$\delta_{k,k+1}(\hat{\boldsymbol{\psi}}(H_{t_k}); H_{t_{k+1}}) = \log L(\hat{\boldsymbol{\psi}}(H_{t_k}); H_{t_{k+1}}) - \log L(\hat{\boldsymbol{\psi}}(H_{t_k}); H_{t_k}),$$

This leads to a technique similar to cross-validation, but applied only on future observations. Therefore, we choose $\widetilde{\boldsymbol{\psi}}(H_{t_k})$ which maximizes:

$$FLP_{k_1,k_2}(\hat{\boldsymbol{\psi}}) \equiv \sum_{k=k_1}^{k_2} \delta_{k,k+1}, \tag{3}$$

where $k_1 = \left[\frac{n}{2}\right]$ and $k_2 = n-1$.
In previous applications (Adelfio and Chiodi, 2011), on the basis of the measure in (3), we observed that the bandwidths estimated by FLP approach produced better kernel estimates (in terms of MISE) of space-time intensity functions than classical methods.

4 Alternating estimation of components

In order to estimate the different components of the ETAS model (2), we here propose to alternate the standard parametric likelihood method (to estimate the parameters of the offsprings component) with the FLP approach (to estimate the background intensity).
Given a catalog of n seismic events and set $v = 1$, let $f^{(1)}(x, y)$ be a starting estimation of the background seismicity, obtained by kernel estimators. The $v - th$ iteration of the simultaneous estimation of nonparametric and parametric components proceeds as follows:

1. Get the ML estimator $\hat{\boldsymbol{\theta}}^{(v)}$ of the parameters of the ETAS model, numerically maximizing the likelihood (1).

2. Estimate $\rho_i^{(v)} = \frac{\mu f^{(v)}(x_i,y_i)}{\lambda_{\hat{\boldsymbol{\theta}}^{(v)}}(x_i,y_i,t_i,m_i|\mathcal{H}_t)}$, $i = 1, ..., n$, for each point of the catalog, on the basis of the estimated parameters. $\rho_i^{(v)}$ is used as a vector of weights for the nonparametric estimation of the background seismicity.

3. Update the estimation of the background seismicity $f^{(v+1)}(x, y)$, through weighted kernel estimator with weights $\rho_i^{(v)}$.

   - Compute the estimated triggered intensity $\tau_{\hat{\boldsymbol{\phi}}^{(v)}}(t_i, x_i, y_i)$ for each point of the catalog.
   - Estimate an optimal smoothing vector $\psi^{(v)}$ of the kernel estimator, maximizing the (3) and holding $\tau_{\hat{\boldsymbol{\phi}}^{(v)}}(t_i, x_i, y_i)$ fixed for the whole iteration.

4. Update v and start a new iteration, until some convergence rule is reached. Convergence is judged comparing the values of ETAS components in consecutive iterations, checking also the increase in the overall likelihood function.

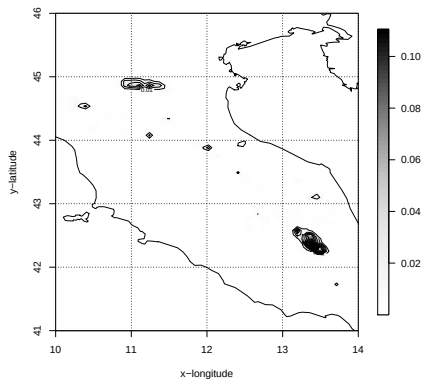


FIGURE 1. Estimated triggered intensity of Italian seismicity (2005-2012).

As an example of application we apply the proposed approach to the catalog of Italian seismic events recorded from 2005 to 2012. The estimate of the only triggered intensity function for a restricted area, reported in figure 1, shows high picks of intensity in correspondence of focal areas of the Italian seismicity, i.e. L'Aquila and Reggio Emilia, where two big sequences of events occurred in 2009 and 2012, respectively. The estimated model seems to follow adequately the seismic activity of the observed area, characterized by highly variable changes both in space and in time. Indeed, because of its flexibility, the estimation approach provides a good fitting to local

space-time changes, to analyze possible correlation between the estimated intensity function and particular distributions of some structural features (i.e. geological structures) of the studied region.

References

Adelfio, G. (2010) An analysis of earthquakes clustering based on a second-order diagnostic approach. *Data Analysis and Classification*. Springer-Verlag Berlin Heidelberg, pp. 309 – 317.

Adelfio, G. and Chiodi, M. (2011). Kernel intensity for space-time point processes with application to seismological problems. *Classification and multivariate analysis for complex data structures*. Springer-Verlag Berlin Heidelberg, pp. 401 – 408.

Adelfio, G., Chiodi, M., De Luca, L., Luzio, D. and Vitale, M. (2006) Southern-Tyrrhenian seismicity in space-time-magnitude domain. *Annals of Geophysics*, **49**, 1245 – 1257.

Adelfio, G., Chiodi, M. and Luzio, D. (2010) An algorithm for earthquake clustering based on maximum likelihood. *Data Analysis and Classification.*. Springer-Verlag Berlin Heidelberg, pp. 25 – 32.

Adelfio, G. and Schoenberg, F. P. (2009) Point process diagnostics based on weighted second-order statistics and their asymptotic properties. *Annals of the Institute of Statistical Mathematics*, **61**, 929 – 948.

Chiodi, M. and Adelfio, G. (2011) Forward Likelihood-based predictive approach for space-time processes. *Environmetrics*, **22**, 749 – 757.

Console, R., Jackson, D. D. and Kagan, Y. Y. (2010). Using the ETAS model for Catalog Declustering and Seismic Background Assessment. *Pure Applied Geophysics* **167**, 819 – 830.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes*. New York: Springer-Verlag.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation, *Journal of the American Statistical Association*, **97**, 611 – 631.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, **83**, 9 – 27.

Zhuang, J., Ogata, Y. and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, **97**, 369-379.

A Topological Approach to the Statistical Estimation of Random Field Thresholds

Robert J. Adler[1], Kevin Bartz[2], Samuel C. Kou[3], Anthea Monod[1]

[1] Technion – Israel Institute of Technology, Haifa, Israel
[2] Renaissance Technologies, New York, USA
[3] Harvard University, Cambridge, USA

E-mail for correspondence: anthea@ee.technion.ac.il

Abstract: We outline a new regression method to produce accurate $(1-\alpha)$ thresholds for signal detection in random fields that does not require knowledge of the spatial correlation structure. The idea is to fit the observed empirical Euler characteristics to the Gaussian kinematic formula via generalized least squares, which quickly and easily provides statistical estimates of Lipschitz-Killing curvatures — complex topological quantities that are otherwise extremely challenging to compute, both theoretically and numerically. With these estimates we can then make use of a powerful parametric approximation via Euler characteristics for Gaussian random fields to generate accurate $(1-\alpha)$ thresholds and p-values. We demonstrate our approach on neuroimaging fMRI data.

Keywords: Brain imaging, Gaussian kinematic formula, Lipschitz-Killing curvature, generalized least squares regression, significance level.

1 Motivation

Random field models are widely used in many scientific applications, such as the statistical analysis of brain imaging and of cosmological studies. An important problem common to most of these applications is the determination of threshold levels for the random field, which indicate that regions with values above the level are significant, while regions with values below are not. Under the setting of random fields, there is the major challenge that values are correlated in space, which makes accurate determination of threshold levels more difficult than under the simpler setting of independent observations.

The application of interest and data under study is a language priming experiment of functional magnetic resonance imaging (fMRI) carried out by Dehaene-Lambertz et al. (2006), which also appeared in the functional image analysis contest (FIAC). In this experiment, fMRI responses were measured twice for 16 subjects after each heard a sentence spoken under

two different conditions: once with the same speaker both times, and once with different speakers. After each repetition, hemodynamic activity was measured at every point (voxel) in a $64 \times 64 \times 30$ grid that encompasses the brain for each subject. The fMRI scan for the first subject is shown in Figure 1: the light and dark red domains represent mid- and high-activation regions of the brain, respectively, under same-speaker (upper left) and different-speaker conditions (lower left).

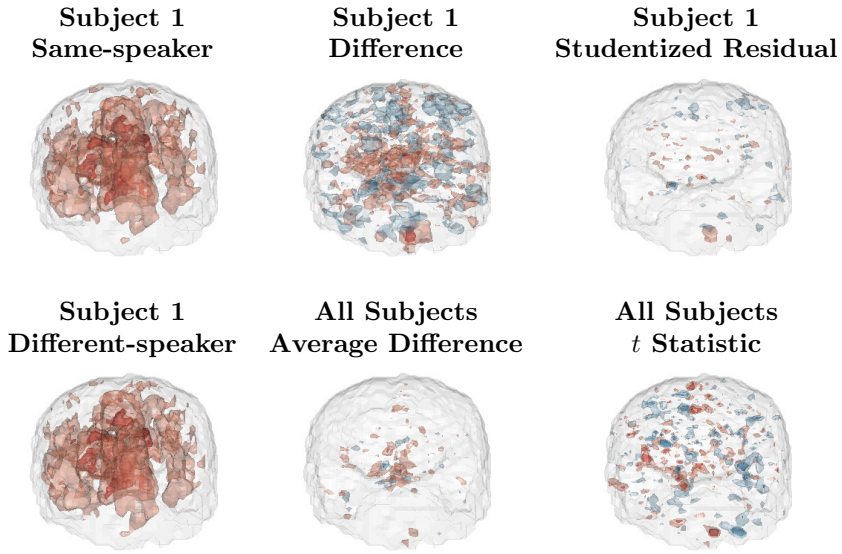


FIGURE 1. Example of fMRI brain scans from the FIAC. Responses were measured for 16 subjects after hearing the same sentence spoken twice under two conditions: both spoken by the same speaker (top left) and each by a different speaker (lower left). The grey shell gives the outline of the brain, while the light and dark red domains represent the contours at the hemodynamic activity levels $(12,000, 15,000)$. The difference brain (upper middle) shows the pointwise difference of these two scans. The pointwise average of these differences across the 16 experimental subjects is less noisy (lower middle). The residual brain (upper right) gives the difference between the first subject and the experimental average. The t brain (lower right) shows paired t statistics.

This study inspires two important research questions: Firstly, are there any significant differences at all? Secondly, if there are, where do these significant differences lie? A natural way to look for differences is to compare the two conditions at every voxel in the brain. Figure 1 (lower right) shows the result of voxelwise paired t tests: in red regions, different-speaker activity exceeds same-speaker activity; in blue regions, the opposite is true. The light- and dark-colored regions show areas where the t statistics exceed the nominal single-test 95% and 99% levels of 2.13 and 2.95, respectively. The

multiplicity of comparisons of 24,759 voxels in the brain, each with its own t test, poses a challenge in establishing significance. Using the 95% threshold, 1,273 voxels indicate significant differences, which, as a proportion of 24,759, amounts to roughly 5% and thus, comprise excessive false positives. In contrast, using the 95% Bonferroni bound of 7.23, there are no significant differences at all. This bound is known to be very conservative, and according to neurological theory, the resulting conclusion is not likely to be credible.

These preliminary exploratory results underline the importance of the motivating research questions and provide the impetus for a sound method to determine and characterize significance.

2 Method

The problem of determining thresholds can be posed as a statistical test: Let the null hypothesis H_0 assert that both brain scans under both conditions are, on average, equivalent. Under this H_0, the grid of t-statistics becomes a smooth random field $T(\cdot)$ over a region S of the brain. Since high values of T usually indicate deviation from H_0, a natural test statistic to propose is the maximum:

$$M_S := \sup_{s \in S} T(s).$$

Using M_S as a test statistic, however, requires its precise null distribution, which is practically never known. In the FIAC data, $M_S = 6.08$; testing H_0 requires the p-value, $P(M_S \geq 6.08|H_0)$. To identify the activated regions of T, we require a $(1-\alpha)$ threshold t such that $P(M_S > t|H_0) = (1-\alpha)$. Neither the null distribution nor the threshold are easy to compute because they both depend on the correlation structure of T, which itself is also unknown.

In Adler et al. (2013), we introduce Lipschitz-Killing curvature (LKC) regression, a new method to produce accurate $(1-\alpha)$ thresholds for random fields without knowledge of the correlation structure. The method borrows from the Euler characteristic heuristic (ECH) (Adler, 2000), a powerful parametric technique to determine null tail probabilities of M_S for a wide class of Gaussian and Gaussian-related random fields. It provides an accurate approximation (of the order of 1% to 2% error) of the exceedance probability $P(M_S \geq u)$ for large u (of the order of 5% or smaller).

The ECH is based on the Euler characteristic (EC) φ, an important topological invariant for many general classes of well-behaved sets. For a 3D Euclidean volume V, it counts the number of each of the three types of topological features of a manifold (solid, simply connected regions of the manifold, or connected components; visible, open holes, or handles; and invisible, closed holes, or voids) in an alternating sum. It remains invariant

under homeomorphisms of the topological space and is given by (Adler, 1981)

$$\varphi(V) = \#\text{ connected components in } V - \#\text{ handles in } V + \#\text{ voids in } V.$$

The ECH considers the ECs φ of excursion sets $A_u := \{s \in S : T(s) \geq u\}$ and provides the following result:

$$P(M_S \geq u) \approx E[\varphi(A_u)]. \tag{1}$$

Thus, we have an approximation for the tail probability for high u, which provides the required p-value to test H_0. Inverting (1) provides the threshold u_α corresponding to a $(1-\alpha)$ confidence level:

$$\hat{u}_\alpha := \max\{u : E[\varphi(A_u)] \geq \alpha\}. \tag{2}$$

The applicability of the ECH becomes apparent when considering random fields with constant mean and variance (Adler & Taylor, 2007), where the expected EC $E[\varphi(A_u)]$ takes on an explicit parametric closed form given by the Gaussian kinematic formula (GKF) (Taylor, 2006),

$$E[\varphi(A_u)] = \sum_{i=0}^{\dim(S)} \mathcal{L}_i(S)\rho_i(u). \tag{3}$$

For Gaussian random fields, the ρ_i have simple representations involving Hermite polynomials, while the $\mathcal{L}_i$ are the LKCs — complex topological quantities that are often extremely difficult to evaluate analytically as well as numerically. The idea of LKC regression is then to fit the observed empirical ECs to the GKF (3) using generalized least squares (GLS), which provides statistical estimates of the LKCs that then allow us to generate p-values by the ECH (1), and $(1-\alpha)$ thresholds by (2).

3 Application

Figure 2 provides an illustration of our method fitted to the FIAC data. The left panel shows the empirical EC $\varphi(A_u^{(i)})$ profiles for the 16 observed random fields of the FIAC data (thin grey lines) and their sample average (dashed thick line) for different values of u. We find the best-fitting LKCs through a GLS regression of (3). The fitted LKCs then produce the best-fitting profile: the solid black line. The intersection of this fitted profile and the 0.05 line yields the 95% confidence threshold as $u^*_{95\%} = 4.19$ (black dot). This threshold can then be applied to the brain of t statistics to identify significantly activated regions. The method concludes that there are only 7 voxels activated beyond 4.19, which is consistent with neurological hypotheses, and a striking contrast to the 1,273 found using the naive multiple t-test 95% threshold of 2.13, as well as the 0 found using the 95% Bonferroni bound of 7.23.

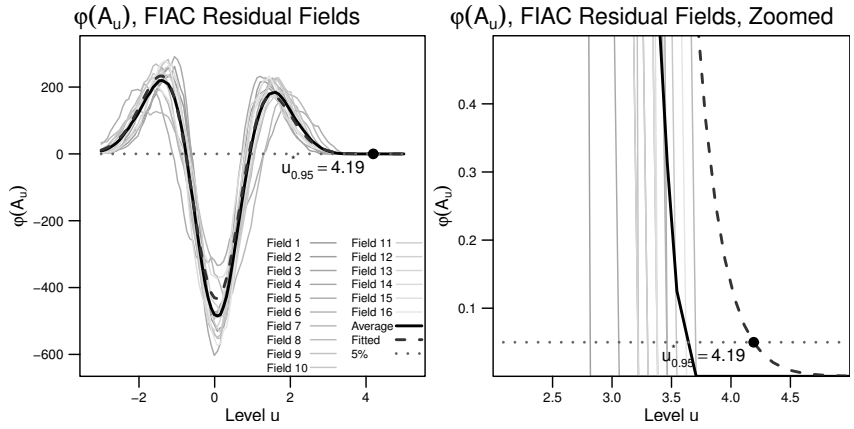


FIGURE 2. Illustration of our proposed regression method to fitting LKCs. The light grey lines show the observed EC profiles for each of the 16 observed fields from the FIAC data. The dashed line is their average. The solid black line is the expected EC estimated by our regression method. The black dot is the point where the expected EC intersects with the dotted grey line at 0.05, which denotes the 5% threshold $u^*_{95\%}$. The plots are displayed both over a broad range of levels (left) and zoomed in near the threshold (right).

4 Discussion

As a tool for hypothesis testing, our LKC regression method is powerful because it overcomes the need for the complete specification of the covariance structure of the random field. Very few assumptions about the covariance are made; even isotropy is not required. The random fields need only to be Gaussian or Gaussian-related, which is often a natural consequence of the data generating process.

The validity of the GKF (3) for all u forms the foundation of the effectiveness of our LKC regression method. For p-value and threshold calculation, one typically encounters large values of u, which correspond to small-probability events, making direct estimation unreliable. Our regression approach, instead, is grounded on the observation that when u is small or moderate, the expected EC $E[\varphi(A_u)]$ can be well estimated from the data, since these cases do not correspond to small-probability events, and so well-established statistical estimation methods are applicable. These reliable estimates then translate, via our regression, into reliable estimates of the LKCs $\mathcal{L}_i$ that do not depend on u. The $\hat{\mathcal{L}}_i$ in turn yield good approximations for p-values and threshold levels for large u. LKC regression thus leverages the precision of estimation at low levels of u to obtain accurate approximation at high levels of u.

In summary, our proposed LKC regression method features easy implemen-

tation, conceptual accessibility, and facilitated diagnostics. Furthermore, LKC regression achieves large gains over its main competitor, warping (Taylor & Worsley, 2006), without compromising accuracy. For 2D and 3D random fields, the gain in speed ranges from a factor of two to a factor of eight; for high-resolution fields, which are commonly found in practice, as well as for higher dimensional cases, this proves to be a clear advantage. A detailed numerical comparison of the two methods under simulation studies as well as under application to the FIAC data is provided in Adler et al. (2013).

Acknowledgments: This research is supported in part by the US-Israel BSF, ISF, NIH-NIGMS, and NSF. Anthea Monod's research is supported by TOPOSYS (FP7-ICT-318493-STREP). The authors would like to thank Jonathan Taylor for helpful discussions.

References

Adler, R.J. (1981). *The Geometry of Random Fields*. Chichester: John Wiley & Sons, Ltd. Reprinted in 2010 by SIAM.

Adler, R.J. (2000). On Excursion Sets, Tube Formulae, and Maxima of Random Fields. *Annals of Applied Probability*, **10(1)**, 1 – 74.

Adler, R.J., Bartz, K., Kou, S.C., and Monod, A. (2013). Estimating Thresholding Levels for Random Fields via Euler Characteristics. *In preprration*.

Adler, R.J. and Taylor, J.E. (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics.

Dehaene-Lambertz, G., Dehaene, S., Anton, J., Campagne, A., Ciuciu, P., Dehaene, G., Denghien, I., Jobert, A., LeBihan, D., Sigman, M., Pallier, C., and Poline, J. (2006). Functional Segregation of Cortical Language Areas by Sentence Repetition. *Human Brain Mapping*, **27(5)**, 360 – 371.

Taylor, J.E. (2006). A Gaussian Kinematic Formula. *Annals of Probability*, **34(1)**, 122 – 158.

Taylor, J.E. and Worsley, K. (2006). Inference for Magnitudes and Delays of Responses in the FIAC Data Using BRAINSTAT/FMRISTAT. *Human Brain Mapping*, **27(5)**, 434 – 441.

The analysis of discrete longitudinal data using acyclic probabilistic finite automata

Smitha Ankinakatte[1], David Edwards[1]

[1] Dept of Molecular Biology and Genetics, Aarhus University, Denmark

E-mail for correspondence: Smitha.AA@agrsci.dk

Abstract: We describe an approach to the analysis of discrete longitudinal data based on acyclic probabilistic finite automata models, propose a modified model selection method, and illustrate application to a set of social science data.

Keywords: state-merging, probabilistic automata, sample tree

1 Introduction

An approach to the analysis of discrete longitudinal data using *acyclic probabilistic finite automata* (APFA) was described by Ron et al (1998). APFA may be represented as directed graphs, with possibly multiple edges between vertices. The methodology has been further developed in a series of papers in the computer science and machine learning literatures. The models have been used with great success in a variety of applications, but appear to have passed unnoticed in the mainstream statistical literature. They underlie the BEAGLE program (Browning and Browning, 2007) which is widely used for phasing and imputation of DNA chip data. The class of chain event graph models (Smith and Anderson, 2008) is closely related to APFA but appears to have been developed independently.
The structure of this paper is as follows. First we give a brief sketch of APFA in statistical terms, and then propose use of a model selection algorithm based on a penalized likelihood criterion. Finally we illustrate how covariates may be included in the framework and demonstrate the practical utility of the approach by applying it to a set of social science data.

2 APFA

Automata are essentially machines that output (or input) strings of symbols. For example, Figure 1 represents a simple acyclic probabilistic finite automaton. One node (the initial state, or *root*) has only outgoing edges (*out-edges*), another node (the final state, or *sink*) has only incoming edges (*in-edges*), and the remaining nodes have at least one in-edge and at least

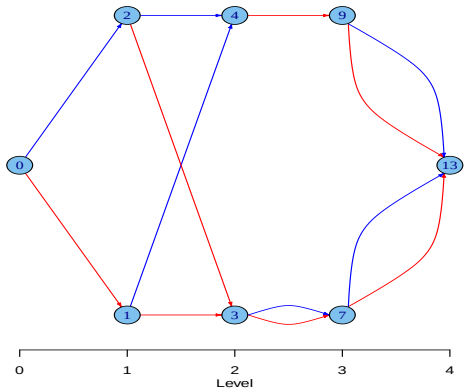


FIGURE 1. A simple APFA model.

one out-edge. All out-edges from a node have different colours: these indicate the symbols generated: for example, red for '1' and blue for '2'. Each edge has a real number associated that specifies the probability of choosing that edge (generating that symbol) having arrived at the node. All paths from the root to the sink have the same length. The nodes of the graph fall into groups $0, 1, \ldots p$ called *levels*, where the root has level 0, the children of the root have level 1, and so on. All edges go from one level to the next. In statistical terms, such a graph specifies a probability distribution over p discrete random variables, say $X = (X_1, \ldots, X_p)$, corresponding to the p levels. For example, in Figure 1 X_1 corresponds to nodes 2 and 3, and the edges from the root node 1 to 2 and 3, to the events $X_1 = 1$ and $X_1 = 2$. The joint probabilities are given by

$$\Pr(X) = \Pr(X_1)\Pr(X_2|X_1)\Pr(X_3|X_1, X_2)\Pr(X_4|X_1, X_2, X_3)$$

and the graph implies certain constraints in the probabilities: for example, Figure 1 implies that

$$\Pr(X_3 = 1|X_1 = 1, X_2 = 2) = \Pr(X_3 = 1|X_1 = 2, X_2 = 2) = 1.$$

Whenever a node at level $i < p$ has several in-edges, this implies that,

$$X_{>i} \perp\!\!\!\perp X_{\leq i} | X_{\leq i} \in \mathcal{C} \tag{1}$$

where $\mathcal{C}$ is the set of paths $\mathcal{C} = \{X^j_{\leq i}\}$ from the root to the node. Thus the graph expresses a set of conditional independence constraints on $\Pr(X)$, and is a type of graphical model. But unlike conventional graphical models (Lauritzen, 1996) here the conditional independences are given *events* rather than *variables*. Maximum likelihood estimation given a data sample is trivial: the conditional probabilities are simply estimated as the relative frequencies of the corresponding counts.

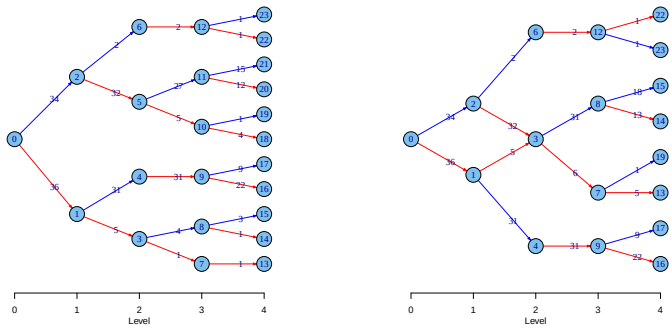


FIGURE 2. The left plot shows the sample tree, and the right plot is the result of merging nodes 4 and 6 in the second level of the sample tree

3 Model selection

Ron et al (1998) described a simple and powerful algorithm to select APFA. Suppose we have observed N observations of p discrete random variables $X = (X_1, \ldots, X_p)$. The algorithm first constructs the *sample tree* of the data, and then simplifies this in a series of state-merging operations. The idea is to merge two nodes n_1 and n_2 at level i whenever (1) is judged to hold, i.e. whenever

$$\Pr(X_{>i}|X_{\leq i} \in \mathcal{C}_1) = \Pr(X_{>i}|X_{\leq i} \in \mathcal{C}_2) \tag{2}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ are the sets of paths from the root to n_1 and n_2, respectively. Figure **??** illustrates a sample tree in which the two nodes labelled 4 and 6 are merged, together with the corresponding descendent nodes, and the edge counts are summed. Observe that the model shown in Figure 1 could be obtained by further merging nodes 4 and 6 at level 2, and nodes 8 and 7 at level three, in Figure **??**.

The algorithm proceeds from level 1 to level $p-1$. At each level nodes for which (2) is judged to hold are merged, resulting in a partition of the set of nodes at that level. All nodes at level p are merged. To judge whether (2) holds, Ron et al (1998) proposed use of a similarity measure

$$S = \max_{X_{>i}} |\Pr(X_{>i}|X_{\leq i} \in \mathcal{C}_1) - \Pr(X_{>i}|X_{\leq i} \in \mathcal{C}_2)|$$

where the maximum is taken over $X_{>i}$ of the form $(x_{i+1}, \ldots x_{i+k})$, for $k = 1 \ldots p-i$. Merging occurs when S is less than a fixed threshold μ. Browning and Browning (2007) modified the threshold μ to depend on input counts to n_1 and n_2. Other authors have proposed alternative criteria for state-merging in various classes of probabilistic automata, for example based on

multinomial tests (Kermorvant and Dupont, 2002) and on Kullback-Leibler divergence (Thollard et al,2001).
We propose instead use of a penalized likelihood approach, that trades off likelihood with model dimension:

$$\mathrm{IC} = -2\log(L) + k\dim(m) \tag{3}$$

where $\log(L)$ is the log-likelihood, $\dim(m)$ is the model dimension (number of free parameters), and k is a penalizing constant: common choices are $k = 2$ for Akaike's information criterion (AIC) and $k = \log(N)$ for the Bayesian information criterion (BIC). It is well-known that under reasonable assumptions, model choice by minimizing BIC is consistent (Ripley, 1996). In the present context we need to compute the change in the criterion when merging nodes, for example by calculating $\Delta\mathrm{BIC} = \text{deviance} + \log(N)\text{df}$. The deviance (likelihood ratio test statistic) and degrees of freedom can conveniently be calculated using standard expressions (Højsgaard et al 2012) for the deviance and adjusted degrees of freedom associated with certain $r \times c$ contingency tables: we omit the details.

4 An application

The biofam data set was constructed by Müller et al. (2007) from data obtained in a retrospective biographical survey carried out by the Swiss Household Panel (SHP) in 2002. The data describe family life courses of $N = 2000$ individuals born between 1909 and 1972, including only individuals who were at least 30 years old at the time of the survey. It contains sequences of family life states recorded once a year from age 15 to 30 and a series of covariates. Family life state is classified into 8 categories: (i) living with parents, (ii) left home, (iii) married, (iv) left home and married, (v) have children, (vi) left home and have children, (vii) left home, married and have children, and (viii) divorced. In addition, a large number of covariates were recorded. Here for the sake of simplicity we only include sex and religion, the latter coded as 'catholic', 'protestant' or 'other'.
To illustrate the method we apply it to the biofam data. To include the covariates in the model we construct a single factor with six levels encoding the possible combinations of sex and religion and include this in the model framework as the first variable. The following 16 variables are the family life states from age 15 to 30, so $p = 17$.
The sample tree of the data begins with the root node, which forms the parent node for the 6 covariate nodes at level one. Starting from this level, the algorithm looks for nodes at the same level to merge. At each level, the algorithm checks the all possible combinations of node pairs, and two nodes at a level are merged if they are sufficiently similar, using the BIC criterion (3). The selected model is shown in Figure 3. Each path from root to sink node represents an individual life course, that is, a sequence of family life

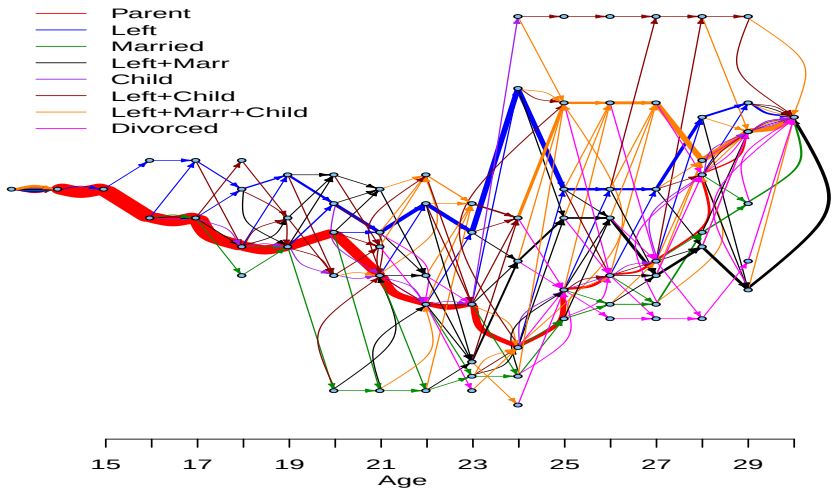


FIGURE 3. APFA model for biofam data. The width of the edges indicates the size of the counts.

states. The legend shows the colour-coding of the life state variables, and the width of the edges is proportional to the corresponding sample counts. The x-axis shows the ages corresponding the different levels in the graph. The three most common states are: staying with parents, left home, and got married.

We first note that all the 6 covariate nodes in the sample tree are merged into one node in the graph. This implies that sex and religion do not affect the future life courses. The red edges represent children staying with their parents. We observe that a large number of children remain living with their parents until the age of 20 and from then the number gradually decreases. The blue edges, indicating those who left home, increase correspondingly. The different parts of the plot show the life courses of those that left home without getting married (blue edges), got married (green edges), got divorced (purple edges) and got married, left home and had children (orange edges).

The R package TraMineR implements other useful methods for the analysis of this type of data: see Gabadinho et al. (2011).

5 Conclusion

As the example illustrates, APFA constitute a rich and expressive class of models for longitudinal data. They may be represented as graphs that are easily interpretable (although for high-dimensional data these may be very complex). We believe that they may useful for the analysis of discrete longitudinal data arising in a variety of application areas.

References

Browning, B., and Browning, S. (2007). Efficient multilocus association testing for whole genome association studies using localized haplotype clustering Genetic epidemiology. *Wiley Online Library*, **31**, 365 – 375.

Gabadinho, A., Ritschard, G., Müller, N. S., Studer, M. (2011), Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, **40**, 1 – 37.

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012), Graphical models with R. *Springer*

Kermorvant, C., Dupont, P., (2002), Stochastic Grammatical Inference with Multinomial Tests. *Grammatical Inference: Algorithms and Applications, Springer Berlin Heidelberg*, **2484**, 149 – 160

Lauritzen, S. L. (1996), Graphical Models. *Oxford Science Publications*

Müller, N. S., M. Studer, G. Ritschard (2007). Classification de parcours de vie à l'aide de l'optimal matching. *In XIVe Rencontre de la Société francophone de classification, Paris*, 157 – 160.

Ripley, B.D (1996), Pattern recognition and neural networks, *Cambridge Univ Press*

Ron, D., Singer, Y., and Tishby, N. (1998). On the Learnability and Usage of Acyclic Probabilistic Finite Automata. *Journal of Computer and System Sciences*, **56**, 133 – 152.

Smith, J. Q. and Anderson, P. E. (2008) Conditional independence and chain event graphs. *Artificial Intelligence*,**172**, 42 – 68.

Thollard, F. (2001), Improving Probabilistic Grammatical Inference Core Algorithms with Post-processing Techniques. *Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc.,* 561 – 568.

Markov mixture models for analyzing the evolution of chronic kidney disease in children

Carmen Armero[1], Peter J. Diggle[2], Anabel Forte[3], Hèctor Perpiñán[1]

[1] Universitat de València (Spain).
[2] Lancaster University (UK).
[3] Universitat Jaume I de Castelló (Spain).

E-mail for correspondence: carmen.armero@uv.es

Abstract: Markov mixture models are considered for assessing the progression of chronic kidney disease in children. The observational process consists of repeated measurements of the estimated filtration glomerular rate (eGFR). In our models, the distribution of eGFR depends on a latent continuous-time process that describes transitions amongst three different states of the disease process: crisis; recovery; stability.

Keywords: Bayesian inference; Hidden models; Longitudinal data.

1 Introduction.

Chronic kidney disease (CKD) is characterized by a progressive loss of renal function which ends in the so-called end-stage renal disease, at which point the subject needs life-saving renal replacement therapy (dialysis or transplantation). CKD in children has a strong influence on their physical, psychological, social and intellectual growth as well as reducing their life expectancy.
The evolution of this disease in children, despite its importance for Public Health, is poorly understood because most studies have focused on adult populations. The main goal of this study is to assess the progression of this disease in children through a longitudinal study of the estimated glomerular filtration rate (eGFR), the most widely used variable for quantifying renal function.

2 The longitudinal data.

Data for our analysis come from ReVaPIR, a study of CKD in València, Spain. The study includes all patients (n=168) living in the Comunitat Valenciana from 1st January 2005 until 31st December 2010 who had been

diagnosed with the disease during this period or before. The data-base includes for each patient eGFR measurements and other information at the time of diagnosis and at consecutive follow-up visits. The data are unbalanced: follow-up times are not common to all patients, and the number of follow-up visits on which data are available varies between one and 13. The time-origin for each patient's sequence of measurements is their date of diagnosis. The response variable is the logarithm of eGFR.
Figure 1 shows time-plots of log(eGFR) values against time since diagnosis, with consecutive observations on each child connected through line segments. Interesting features of Figure 1 include the wide variation in the initial values of log(eGFR) and in the subsequent trajectories for different children.

FIGURE 1. Response profiles of log(eGFR) against time since diagnosis. Colours indicate results for different children.

3 First longitudinal models.

We have considered different longitudinal models for explaining the progression of log(eGFR), using a parametric approach to accommodate the unbalanced nature of the data (Diggle, 2002):

- A mixed effect lineal model (MLM) with a random intercept for each child.
- A MLM with a random intercept and a random slope.
- A MLM with a random intercept and an Ornstein-Uhlenbeck (OU) process to capture serial correlation between repeated measurements on the same child.

- A MLM with a random intercept, a random slope and an OU process for serial correlation.

All these models include gender, etiology of the disease and age of the child at diagnosis time as baseline covariates, and hypertension medication as a time-varying covariate registered at each follow-up examination.
The models were fitted using Bayesian methods with conventional non-informative priors using the WinBUGS software (Lunn *et al.*, (2000)). Unfortunately, none of the models seems able to capture the heterogeneity in the log(eGFR) data, especially for children whose eGFR sequences show high variability.

4 Markov mixture models.

We have noticed, and also discussed with the medical team, some different patterns in the evolution of the disease, not only between different children but also within the same child in different time-periods. For example, there are some children who maintain a stable renal function for a time before experiencing an unstable period. In some cases, this situation represents a crisis from which the child subsequently recovers and returns to a period of relative stability. In others, an irreversible decline in renal function leads to a requirement for long-term renal replacement therapy, ideally transplantation.
We assume that the progression of the disease depends on a latent disease state, ant it would be desirable to examine a model that takes into account this element in the log(eGFR) measurement process. For this reason we move towards Markov mixture models (Rabiner and Juang, 1986; Cappé *et al.*, 2005; Frühwirth-Schnatter, 2006) which are better able than linear models to capture the full range of heterogeneity that we see in our data.
The basic structure of this class of models includes a bivariate process whose components are a hidden categorical-valued Markov chain and a continuous-valued measurement process. In our application we define these elements as follows.

The hidden process:

A continuous-time Markov chain describing the state of the disease of a child at time t, where t is time since CKD diagnosis. The chain has three states: stable; crisis; recovery.

The measurement process given the hidden process:

A mixed linear model for the longitudinal observations of log(eGFR) for which:

- The trend is defined in terms of the state of the disease, baseline and temporal covariates, and a piecewise-linear function with slope-changes at the times of transition between states;

- Random variation includes:
  - a random intercept and an OU process realised independently for each child, both depending on the hidden disease-state;
  - a measurement error term to account for the imprecision in measured eGFR, independent of the disease-state.

Bayesian computation is based on data augmentation and requires proper prior distributions (Frütwirth-Snatter, 2006). In particular, we are using Markov Chain Monte Carlo methods and are currently exploring some different approaches introduced by Hahn and Sass (2009) and Hahn *et al.*, (2010), including: a continuous-time state process; a discrete-time version; and a combination of the two.

Acknowledgments: This work was partially supported by the research grant MTM2010-19528 from the Spanish Ministry of Education and Science.

References

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models.* Springer.

Diggle, P.J., Heagerty, P.J., Liang, K-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data.* Second Edition. Oxford University Press.

Frühwirth-Snatter, S. (2006). *Finite Mixture and Markov Switching Models.* Springer.

Hahn, M., Frühwirth-Snatter, S., and Sass J. (2010). Markov Chain Monte Carlo Methods for Parameter Estimation in Multidimensional Continuous Time Markov Switching Models. *Journal of Financial Econometrics*, 8, 1, 88-121.

Hahn, M. and Sass, J. (2009). Parameter Estimation in Continuous Time Markov Switching Models: A Semi-Continuous Markov Chain Monte Carlo Approach. *Bayesian Analysis*, 4, 1, 63-84.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.

Rabiner, L.R. and Juang, B.M. (1986). An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 1-16.

Estimation of the Latent Distribution in Cure Survival Models using a Flexible Cox Model

Vincent Bremhorst[1], Philippe Lambert[1,2]

[1] Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Belgium
[2] Institut des sciences humaines et sociales, Méthodes quantitatives en sciences sociales, Université de Liège, Belgium

E-mail for correspondence: vincent.bremhorst@uclouvain.be

Abstract: A common hypothesis in the analysis of survival data is that any observed unit will experience the monitored event if it is observed for a sufficient long time. Alternatively, one can explicitly acknowledge that an unknown and unidentified proportion of the patient population under study is cured and will never experience the event of interest. The promotion time model, which is motivated using biological mechanisms in the development of cancer, is one of the survival models taking this feature into account. The promotion time model assumes that each subject is exposed to N carcogenic cells. Given this number of carcogenic cells, we define latent event times $(Y_1, ..., Y_N)$, which are independent with a common distribution $F(t) = 1 - S(t)$ independent of N and that can be seen as incubation time. Since we assume that 1 out of N latent factors needs to be activated, the observed failure time is defined as the minimum of the latent event times.
In this work, we propose an extension which allows the covariates to influence simultaneously the probability of being cured and the latent distribution $F(t)$. We estimate the latent distribution $F(t)$ using a flexible Cox proportional hazard model where the logarithm of the baseline hazard function is specified using Bayesian P-splines. The identification issues of the related model are also investigated. A simulation study evaluating the accuracy of our methodology is presented.

Keywords: Survival analysis; Cure fraction; Promotion time model; Cox Model; Poisson Generalized Linear Model; Bayesian P-splines.

1 Introduction

A common hypothesis in the analysis of survival data is that any observed unit will experience the monitored event if it is observed for a sufficient long time. For example, in a cancer clinical trial, one implicitly assumes that all patients will be observed to have a relapse if their follow up is long enough. Hopefully, this is not always a realistic assumption and the consequences

of such a wrong hypothesis on the results of the analysis is more and more questioned in the survival literature.
Alternatively, one can explicitly acknowledge that an unknown and unidentified proportion of the patient population under study is cured and will never experience the event of interest. Such models are refered as cure survival models. There are two well known families of cure survival models. The first one, often refered as the standard mixture cure models, was introduced by Berkson and Gage (1952). The population survival function of such models is obtained as a mixture of contributions due to suceptible and cured individuals :

$$S_p(t) = pS_u(t) + (1-p) \tag{1}$$

where p is the probability of being suceptible and $S_u(t)$ is the survival function of the suceptible individuals. The estimating procedures and the way to enter covariates in the model were discussed by many authors, see for example Wang, Du and Liang (2012).
The second family, refered as the promotion time (cure) model, is motivated using biological mechanisms in the development of cancer as explained in Chen, Ibrahim and Sinha (1999). The model argues that each subject is exposed to a number $N \sim P(\theta)$ of carcogenic cells. For each cell, Y is defined as the time necessary for it to yield a detectable cancer mass. The $Y_i's$ are often refered as the latent event times. We assume that the cancer mass in each cell is detected independently from each other and that only one cell needs to be activated for a subject to fail. The latent event times $\{Y_1, ..., Y_N\}$ are independent with a common proper distribution $F(t)$ independent of N and the observed failure time is defined as $T = \min_i\{Y_i\}$. If the subject is not exposed to carcogenic cells (if $N = 0$), (s)he is considered as cured. Using the biological derivation of the model, one can show that the population survival function is given by :

$$S_p(t) = \exp\left[-\theta F(t)\right] = \exp\left[-\theta(1 - S(t))\right]. \tag{2}$$

Note that, since $F(t)$ is a proper cumulative distribution function, the probability of being cured is given by $P(N = 0) = \lim_{t\to\infty} S_p(t) = \exp(-\theta)$.
A mathematical motivation of the model was first proposed by Tsodikov (1998). It has been studied by many authors, see for example Ibrahim, Chen and Sinha (2001) in a Bayesian framework.
Starting from model (2), we propose an extension which allows the covariates to influence simultaneously the probability of being cured and the latent distribution $F(t)$.

2 Model Specification

Let $\boldsymbol{x}$ and $\boldsymbol{z}$ be two sets of covariates (they can share some components or be identical). When the covariates influence only the probability of being

cured in the promotion time model, it is usual to use the following link on the parameter θ :

$$\theta(\boldsymbol{x}) = \exp(\beta_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$$

Since the covariates might jointly influence the probability to be cured and the time necessary for a cell to yield a detectable tumor, we suggest in addition a Cox proportional hazard model for the latent distribution $F(t|\boldsymbol{z})$:

$$1 - F(t|\boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\gamma})}$$

where $S_0(t)$ is the baseline survival function.
Introducing these two covariates structures in (2), the population survival function becomes :

$$\begin{aligned} S_p(t|\boldsymbol{x},\boldsymbol{z}) &= \exp\left[-\theta(\boldsymbol{x})F(t|\boldsymbol{z})\right] \\ &= \exp\left[-\exp\left(\beta_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}\right)\left(1 - S_0(t)^{\exp(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\gamma})}\right)\right] \end{aligned} \tag{3}$$

3 Identification issues

When working with a Cox proportional hazard model, the two following assumptions are usual :

i) The vector $\boldsymbol{z}$ of covariates does not include an intercept to ensure the identifiability of the Cox proportional hazard model.

ii) The baseline cumulative distribution function $F_0(t) = 1 - S_0(t)$ is proper : $\lim_{t\to\infty} F_0(t) = 1$.

Under i) and ii) we have shown that :

1) If the follow up of the study is sufficently long, then model (3) is identifiable.

2) If the follow up of the study is not sufficently long, then model (3) is not identifiable, except if vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ do not share the same components.

4 Flexible estimation of the baseline distribution

In order to estimate the baseline survival function $S_0(t)$ in (3), we suggest to use a linear combination of cubic B-splines on the baseline log-hazard function:

$$h_0(t) = \exp\left(\sum_{k=1}^{K} b_k(t)\phi_k\right)$$

where $(b_k(.), k = 1, \ldots, K)$ denotes the cubic B-splines basis associated to a predefined number of equidistant knots on $[0, t_{Rcens}]$, where t_{Rcens} is the upper bound of the follow up.
As suggested by Eilers and Marx (1996), we choose a large number of B-splines and counterbalance the flexibility by introducing a (roughness) penalty on finite differences of adjacent B-spline parameters : $\tau \sum_k (\Delta^r \phi_k)^2 = \tau \boldsymbol{\phi}' \boldsymbol{D}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{\phi}$, where τ is the penalty parameter.

5 Bayesian Model

5.1 Prior distributions

In a Bayesian setting, the roughness penalty is translated into a prior distribution for the spline parameters (Lang and Brezger (2004)):

$$\pi(\boldsymbol{\phi}|\tau) \sim \tau^{\frac{\rho(P)}{2}} \exp\left(-\frac{\tau}{2} \boldsymbol{\phi}' P \boldsymbol{\phi}\right)$$

where $\rho(P)$ is the rank of $\boldsymbol{P} = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{D}$ and $\boldsymbol{D}$ is the r^{th} difference penalty matrix. For the penalty parameter τ, we use a common non-informative gamma prior distribution and an improper uniform prior distribution is specified for all regressors.

5.2 Posterior distribution of the spline parameters

The posterior distribution of the spline parameters is not related to a well known family. In order to generate a sample from the posterior, we will use an adaptive univariate Metropolis step as recommended by Haario and al. (2001). Lambert (2007) shows that if we apply the adaptive Metropolis step on a reparametrized posterior distribution then the mixing of the chains will be improved. To reach that goal, an estimation of the correlation structure of the spline parameters is derived using the link between survival data and the Poisson GLM.

6 Simulation study

We evaluate the accuracy of our methodology when the sufficiently long follow up assumption is (and is not) respected (see Section 3). Different percentages of cured and (random) right censored individuals were investigated. For each setting, two covariates are taken into account : $W_1 \sim N(0, 1)$ and $W_2 \sim Bin(1, 0.5)$ and the baseline distribution is related to a Weibull distribution with mean 10.8 and standard deviation 5.64. In this paper, we report only the results when 20% of cured subjects are present in the data without and with 18% of random right censoring and when the assumption of sufficient follow up is respected. Since the follow

TABLE 1. Simulation results. For $S = 200$ replications and sample size $n = 500$. 20% of cured subjects are present in the data. The biais, the coverage of the 90% and 95% credible intervals, the posterior standard deviation and the RMSE of the posterior median of the regression parameters are presented.

Random cens.	True value	Biais	$CV90$	$CV95$	Sd_{post}	RMSE
0%	$\beta_0 = 0.65$	-0.022	91.5	94.5	0.124	0.017
	$\beta_1 = 1.2$	-0.048	89	92	0.119	0.017
	$\beta_2 = 0.5$	0.011	93	97	0.151	0.021
	$\gamma_1 = -1$	0.079	83.5	89	0.122	0.114
	$\gamma_2 = 2.5$	-0.104	86.5	92.5	0.219	0.236
18%	$\beta_0 = 0.65$	-0.017	92	97.5	0.143	0.020
	$\beta_1 = 1.2$	-0.021	91.5	94.5	0.132	0.018
	$\beta_2 = 0.5$	0.020	93	96	0.169	0.026
	$\gamma_1 = -1$	0.049	90.5	95	0.135	0.139
	$\gamma_2 = 2.5$	-0.085	89	94.5	0.237	0.251

up is sufficiently long, $\boldsymbol{x}$ and $\boldsymbol{z}$ can share the same components. We define $\boldsymbol{x} = (W_1, W_2) = \boldsymbol{z}$. The values of $\beta_0 = 0.65$ and $\boldsymbol{\beta} = (\beta_1, \beta_2) = (1.2, 0.5)$ were chosen to ensure 20% of cured subjects in the data. The choice of the values of $\boldsymbol{\gamma} = (\gamma_1, \gamma_2) = (-1, 2.5)$ suggests to set the upper bound of the follow up at $t_{Rcens} = 25$. The censoring distribution is related to an exponential distribution with mean 40. Other scenarios will be discussed during the oral presentation. Simulations were performed on 200 replications of size 500.

Table 1 summarizes the simulation results for the regression parameters. One can see that the posterior medians (as estimators) of the regression coefficients have similar (small) biaises in both settings with a sligh decrease of the biais for γ_1 and γ_2 when random right censoring is introduced. The posterior standard deviations and the RMSE increase slighly with the introduction of random right censoring. The coverage probabilities of the 90% and 95% credible intervals are close to their nominal value with even better results for γ_1 and γ_2 when random right censoring is introduced.

We expect that the biais of the coefficient estimates will increase when the percentage of cured subjects increases. This can lead to some deterioration of the coverage probabilities.

The estimations of the baseline survival function (grey curves) are plotted on Figure 1. The introduction of random right censoring does not affect the estimation of the baseline distribution except in the right tail of the distribution where the uncertainty is getting larger.

These are promising results. More research is needed to understand the favourable effect of right censoring on the estimation of the regression coefficients in the Cox model.

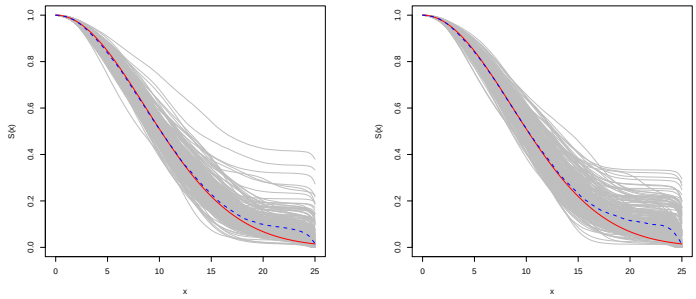


FIGURE 1. Estimation of $S_0(t)$ (One grey curve per data set). For $S = 200$ replications and sample size $n = 500$. 20% of cured subjects are present in the data. The solid line is the true survival function and the dashed line is the pointwize median of the 200 obtained curves. Left : without random right censoring. Right : With 18% of random right censoring.

Acknowledgments: The authors acknowledge financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain.

References

Chen, M.-H., Ibrahim, J.G. and Sinha, D. (1999). A New Bayesian Model for Survival Data with a Surviving Fraction. *Journal of the American Statistical Association*, **94**, 909 – 919.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89 – 121.

Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernouilli*, **7**, 223 – 242.

Ibrahim, J.G., Chen, M.-H. and Sinha, D. (2001). Bayesian Semiparametric Models for Survival Data with a Cure Fraction. *Biometrics*, **57**, 383 – 388.

Lambert, P. (2007). Archimedean copula estimation using Bayesian splines smoothing techniques. *Computational Statistics and Data Analysis*, **51**, 6307 – 6320.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183 – 212.

Tsodikov, A. (1998). A Proportional Hazard Model Taking Account of Long-Term Survivors. *Biometrics*, **54**, 1508 – 54.

Wang, L., Du, P. and Liang, H. (2012). Two-Component Mixture Cure Rate Model with Spline Estimated Nonparametric Components. *Biometrics*, **68**, 726 – 735.

Multi-Parameter Regression Survival Models

Kevin Burke[1], Gilbert MacKenzie[1,2]

[1] Centre for Biostatistics, University of Limerick, Ireland.
[2] CREST, ENSAI, Rennes, France.

E-mail for correspondence: kevin.burke@ul.ie gilbert.mackenzie@ul.ie

Abstract: The *proportional hazards* (PH) assumption in survival analysis may not always be appropriate. If data do not obey the assumption then we will reach incorrect conclusions by making it. For example we may find a covariate to be statistically insignificant when in fact it is important, but on a non-PH scale. Even if a PH model *does* pick up the statistical significance of such a covariate, the nature of the effect of the covariate on survival, as determined by this simplistic model, will clearly be incorrect. We introduce a regression-based extension of parametric PH modelling which we call *multi-parameter regression*, MPR, modelling

Keywords: Multi-parameter regression survival models, non-PH models, shape and scale regression, time-dependent hazards

1 Introduction

Generally, when modelling data parametrically we will have multiple parameters. Typically, we choose only to regress one of these parameters on covariates. For example, in GLMs (McCullagh and Nelder, 1989) the location parameter, $g(\mu) = X\beta$, is regressed whilst the dispersion parameter, σ, is often treated as a nuisance parameter. In more recent times, models have been developed in which multiple parameters are regressed simultaneously on covariates, for example, in structural dispersion (Lee & Nelder, 2001), generalized additive models for location, scale and shape (Rigby & Stasinopoulos, 2005) or joint mean-covariance modelling in longitudinal data analysis (Pan & MacKenzie, 2003). We refer to models such as these as "multi-parameter regression" (MPR) models.
In survival analysis the most widely used model is the *proportional hazards*, PH, model. The routine use of this model has inevitably led to it being imposed on data which do not obey the PH assumption. The PH model is equivalent to regressing the *scale* parameter, say λ, in a model which possesses the proportional hazards property. We propose a multi-parameter regression approach whereby the *shape* parameter, say γ, is regressed simultaneously with the scale parameter. This innovation thus generalizes

the PH model to non-PH status and affords much more flexibility. The influence of covariates on the hazard ratio, which is constant in a PH model, is now time-dependent. The ability to relate covariates to the shape of the hazard will give rise to scientific insights previously unavailable in PH analyses which may be of interest in their own right.

2 MPR Weibull

We focus on the Weibull model for illustrative purposes, although the methodology can easily be applied to other models. The particular form of the Weibull distribution we will use is that presented in Collett (2003) which has hazard function $\lambda(t) = \lambda\gamma t^{\gamma-1}$ where λ,$\gamma > 0$. The hazard is decreasing for $\gamma < 1$, constant for $\gamma = 0$ and increasing for $\gamma > 1$.
We propose the following multi-parameter regression:

$$\log(\lambda) = x^{\mathrm{T}}\beta, \qquad \log(\gamma) = z^{\mathrm{T}}\alpha, \tag{1}$$

where the log-link is used to ensure positivity of both parameters, $x = (1, x_1, \ldots, x_p)^{\mathrm{T}}$ and $z = (1, z_1, \ldots, z_q)^{\mathrm{T}}$ are covariate vectors which may or may not contain covariates in common and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_q)^{\mathrm{T}}$ are unknown regression coefficients The hazard ratio for a binary covariate, common to both regressions, say, $x^* = x_1 = z_1$, is

$$\frac{\lambda(t \,|\, x^* = 1)}{\lambda(t \,|\, x^* = 0)} = \exp(\beta_1 + \alpha_1) t^{\exp(\tilde{z}^{\mathrm{T}}\alpha)\{\exp(\alpha_1)-1\}}. \tag{2}$$

where $\tilde{z}^{\mathrm{T}}\alpha = z^{\mathrm{T}}\alpha - x^*\alpha_1$. When $\alpha_1 = 0$, the hazard ratio is $\exp(\beta_1)$ which is the familiar PH case. Thus the MPR directly generalizes the PH model.

3 Hypothesis Testing in MPR Models

We may ask if a certain covariate, x^*, has an effect on the scale, or the shape or on both the scale and shape parameters leading to three null hypotheses, namely: (i) $H_0 : \beta_1 = 0$, (ii) $H_0 : \alpha_1 = 0$ and (iii) $H_0 : \beta_1 = \alpha_1 = 0$. However, due to correlation between $\hat{\beta}_1$ and $\hat{\alpha}_1$, it is inadequate to consider testing (i) and (ii) separately using the standard Wald test approach. We must test hypothesis (iii) first. To do this, we can assume a bivariate normal distribution for the two parameters (based on standard maximum likelihood theory) from which joint confidence regions for the two parameters can be computed. Variable selection methods must also take this parameter correlation into account, i.e. we cannot consider variable selection for the each regression (scale and shape) separately.

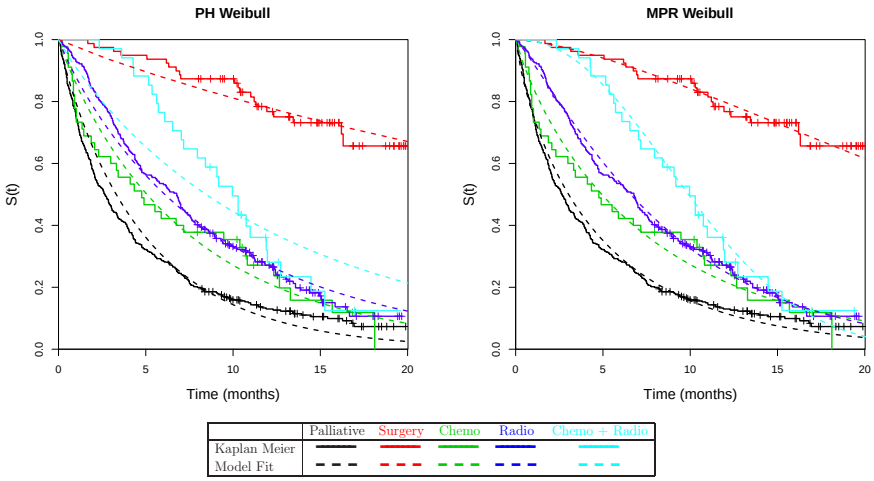


FIGURE 1. Comparison of PH and MPR Weibull Models

4 Example

We can see the flexibility of the MPR model compared with the PH model using a lung cancer data set collected in Northern Ireland between October 1991 and September 1992 (Wilkinson, 1995), in which we find some non-PH covariates, thus invalidating a PH analysis. For example, in Figure 1 we see that the treatment covariate (a factor with five levels) does not seem to obey the PH assumption based on the Kaplan Meier survivor curves. Comparing the model fits to these Kaplan Meier curves, it is visually clear that the PH model (γ constant) does not fit the data as well as the MPR model ($\gamma = e^{z^T \alpha}$). More formally, this can be confirmed by performing a likelihood ratio test (p-value < 0.001) or some selection criterion e.g. AIC or BIC. The example here is only a one factor model. In a multi factor model the improvement in fit for the MPR over the PH model, in terms of AIC for example, will be even greater because the MPR model is more general than the PH model and in the worst case just reduces to the PH model.

5 Discussion

It has been found that the multi-parameter regression Weibull model indeed affords great flexibility and leads to better fits when compared with the standard proportional hazards model. This can be verified both graphically or more formally using likelihood theory. The extra generality leads of course to additional hypothesis testing and model selection considerations.

Acknowledgments: This work was supported, in part, by the SFI's (www.sfi.ie) BIO-SI research programme, grant number, **07MI012**. The first author is an IRCSET Scholar (www.ircset.ie) and the second is the Principal Investigator of BIO-SI (www.ul.ie/bio-si).

References

Collett, D. (2003) *Modelling Survival Data in Medical Research*, 2nd ed., Chapman & Hall/CRC.

Lee, Y. & Nelder, J. (2001) Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and strucured dispersions. *Biometrika*, **88**, 987 – 1006.

McCullagh, P. & Nelder, J. (1989) *Generalized Linear Models*, Chapman & Hall/CRC.

Pan, J. & MacKenziem G. (2005) On model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, **90**, **1**, 239 – 244.

Rigby, R. A. and Stasinopoulos, D. M. (2003) Generalized additive models for location, scale and shape. *Appl. Statist.*, **54**, 507 – 554.

Wilkinson, P. (1995) *Lung Cancer in Northern Ireland 1991 – 1992*, PhD thesis, Queen's University Belfast.

Exploratory Exponential Tilting

Carlo G. Camarda[1], Paul H.C. Eilers[2], Jutta Gampe[3]

[1] Institut National d'Études Démographiques, Paris, France
[2] Dept. of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands
[3] Max Planck Institute for Demographic Research, Rostock, Germany

E-mail for correspondence: gampe@demogr.mpg.de

Abstract: We propose a new technique to summarize several distributions parsimoniously by employing ideas from Exponential Tilting (ET). We assume that the observed data are generated by densities that can be derived from a single (latent) reference distribution by ET, i.e., while preserving the sample means they have minimal Kullback-Leibler distance to the reference distribution. We show how the reference density and the resulting Lagrange multipliers can be estimated by penalized likelihood in a GLM setting. We also suggest an extension of the model, if simple ET does not lead to a satisfactory summary. We illustrate the methodology by two applications.

Keywords: Exponential Tilting; Lagrange multipliers; Latent density; Smoothing.

1 Introduction

Exponential Tilting (ET) results if, for a given reference distribution g, we look for the density f that has the shortest Kullback-Leibler distance to g and has a given expected value. ET has found applications in bootstrapping and statistical testing, but here we use the idea of ET in an exploratory setting.

We reverse the question of ET in the following way: If we have samples from a sequence of n distributions $f_j, j = 1, \ldots, n$, which we assume to evolve over j (which may, for example, be time or another ordered index), then we would like to identify a common latent reference distribution g, so that the $\hat{f}_j$ are ET estimates of g, preserving the sample means. The reference distribution g together with the resulting Lagrange multipliers summarize the f_j parsimoniously, and we call the resulting method Exploratory Exponential Tilting (EET).

As may be expected, the problem can only be solved if some restrictions are put on g. If we assume that g is smooth, we can estimate the reference density and the Lagrange multipliers by penalized likelihood in a GLM setting.

In situations where the simple EET fit is not satisfactory, we propose to extend the model by adding another sequence of constraints, which adds a bilinear component to the model and leads to extended exploratory exponential tilting (E3T). We show the approach in two applications which demonstrate that EET and E3T lead to interesting and useful results.

2 Background and Model

2.1 Exponential Tilting for Discrete Distributions

Let X be a discrete random variable taking values $x_i, i = 1, \ldots, m$, and denote by g_i a probability function on the values x_i. Exponential tilting (ET) estimates a new probability function $f = (f_1, \ldots, f_m)'$, which minimizes the Kullback-Leibler distance to $g = (g_1, \ldots, g_m)'$ and has a specified expected value $a = \sum_{i=1}^{m} x_i f_i$. This constrained optimzation problem

$$\min_f KL(f, g) = \min_f \sum_{i=1}^{m} f_i \ln \frac{f_i}{g_i} \quad \text{s.t.} \quad \sum_{i=1}^{m} f_i = 1 \quad \text{and} \quad \sum_{i=1}^{m} x_i f_i = a$$

can be solved by a Lagrange multiplier approach. If we let $\eta_i = \ln f_i$ and $u_i = \ln g_i$ we can write the Lagrangian as

$$L(\eta, \lambda_0, \lambda_1) = \sum_{i=1}^{m} e^{\eta_i}(\eta_i - u_i) + \lambda_0 \left(1 - \sum_{i=1}^{m} e^{\eta_i}\right) + \lambda_1 \left(a - \sum_{i=1}^{m} x_i e^{\eta_i}\right),$$

where λ_0 and λ_1 are the Lagrange multipliers. If we take the partial derivatives with respect to the $\eta_k, k = 1, \ldots, m$, we obtain

$$\frac{\partial L}{\partial \eta_k} = e^{\eta_k} \{\eta_k - u_k + 1 - \lambda_0 - \lambda_1 x_k\}. \tag{1}$$

Stationary points of $L(\eta, \lambda_0, \lambda_1)$ hence need to solve the system

$$\eta = u + X\lambda, \tag{2}$$

where $u = \ln g$, and $X \in \mathbb{R}^{m \times 2}$ is

$$X = \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_m \end{pmatrix}'. \tag{3}$$

The vector $\lambda = (\lambda_0, \lambda_1)'$ holds the Lagrange multipliers; for simplicity λ_0 absorbed the -1 in (1).
To estimate η from a sample of counts $y = (y_1, \ldots, y_m)'$, we consider the y_i as realizations of Poisson variates with means μ_i, and the constraints become

$$\sum_i \mu_i = \sum_i y_i \quad \text{and} \quad \sum_i \mu_i x_i = \sum_i y_i x_i, \tag{4}$$

respectively. Again we obtain

$$\ln \mu_i = \eta_i = u_i + (1\,,\, x_i)\,\lambda$$

or as vector equation

$$\ln \mu = \eta = u + X\lambda, \tag{5}$$

which can be estimated in a standard GLM setting.
If n distributions $f_j = (f_{j1}, \ldots, f_{jm})'$, $j = 1, \ldots, n$, are to be estimated, for the same reference distribution g, then the resulting n independent systems (5) can be collected in a single large system

$$\ln(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{u} + \boldsymbol{X}\,\boldsymbol{\lambda} \tag{6}$$

with appropriately concatenated offset vector $\boldsymbol{u}$, holding the same $u = \ln g$ for each sample, a design matrix $\boldsymbol{X}$ that replicates (3) n times, and a vector $\boldsymbol{\lambda}$ that combines the n pairs of Lagrange multipliers $(\lambda_{0j}, \lambda_{1j}), j = 1, \ldots, n$. Collecting the n independent systems in a single equation may look superfluous, but will become useful in the following section.

2.2 Exploratory Exponential Tilting

If we observe a sequence of samples $y_j = (y_{j1}, \ldots, y_{jm})'$, $j = 1, \ldots, n$, we can ask whether the n distributions may be generated by ET from a common (latent) reference distribution g. If the resulting fit is good, this would allow us to summarize the n distributions parsimoniously by g and the sequence of Lagrange multipliers $\boldsymbol{\lambda}_1 = (\lambda_{1j})$. We call this novel tool for exploring data the exploratory exponential tilting (EET) model.
To make the problem identifiable, we have to introduce constraints on $u = \ln g$, namely $\sum_i u_i = 0$ and $\sum_i x_i u_i = 0$. Also, we assume that $u = \ln g$ is smooth. If we denote by $\boldsymbol{Y} = (y_{ij})$ and let $\boldsymbol{y} = \mathtt{vec}(\boldsymbol{Y})$, we can rewrite (6) as

$$\ln(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{V}\,\boldsymbol{\beta} \tag{7}$$

where $\boldsymbol{\beta} = [u, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1]' \in \mathbb{R}^{m+2\,n}$, and $\boldsymbol{\lambda}_0 = (\lambda_{0j})$ and $\boldsymbol{\lambda}_1 = (\lambda_{1j})$ are the sequences of Lagrange multipliers. The design matrix $\boldsymbol{V}$ is constructed appropriately by Kronecker products. Smoothness is enforced by a difference penalty on the elements of u.
The resulting constrained iteratively re-weighted least-squares (IRWLS) algorithm is given by:

$$\begin{pmatrix} \boldsymbol{V}^{\mathrm{T}}\tilde{\boldsymbol{W}}\boldsymbol{V} + \boldsymbol{P} & \boldsymbol{H}^{\mathrm{T}} \\ \boldsymbol{H} & 0 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\omega}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{V}^{\mathrm{T}}\tilde{\boldsymbol{W}}\boldsymbol{z} \\ \boldsymbol{\kappa} \end{pmatrix},$$

where $\tilde{\boldsymbol{W}} = \mathtt{diag}(\tilde{\boldsymbol{\mu}})$ and $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{W}}^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) + \boldsymbol{\eta}$. The penalty matrix $\boldsymbol{P}$ has a block-diagonal structure and it measure the roughness of u by dth order differences, multiplied by a positive regularization parameter. The matrix $\boldsymbol{H}$ implements the two constraints.

2.3 Extended Exploratory Exponential Tilting

It can easily be anticipated that simple EET, as described in the previous section, will not provide a satisfactory fit in all situations. Therefore we consider the following extension. We imagine that, besides (4), a third constraint applies across the n distributions:

$$\sum_i \phi_i y_{ji} = \sum_i \phi_i \mu_{ji}. \tag{8}$$

The values ϕ_i vary smoothly across the range of x_i. The n Lagrange multipliers are collected in the vector $\boldsymbol{c} = (c_1, \ldots, c_n)'$. Equation (7) becomes

$$\boldsymbol{\eta} = \boldsymbol{V}\,\boldsymbol{\beta} + \boldsymbol{\phi}\,\boldsymbol{c}, \tag{9}$$

which is a bilinear extension of the simple EET. To make the model identifiable we constrain the elements of $\boldsymbol{\phi}$ to $\sum_i \phi_i = 0$ and of $\boldsymbol{c}$ to $\sum_j c_j^2 = 1$. Again a penalized likelihood can be maximized by an appropriately adapted constrained IRWLS algorithm.

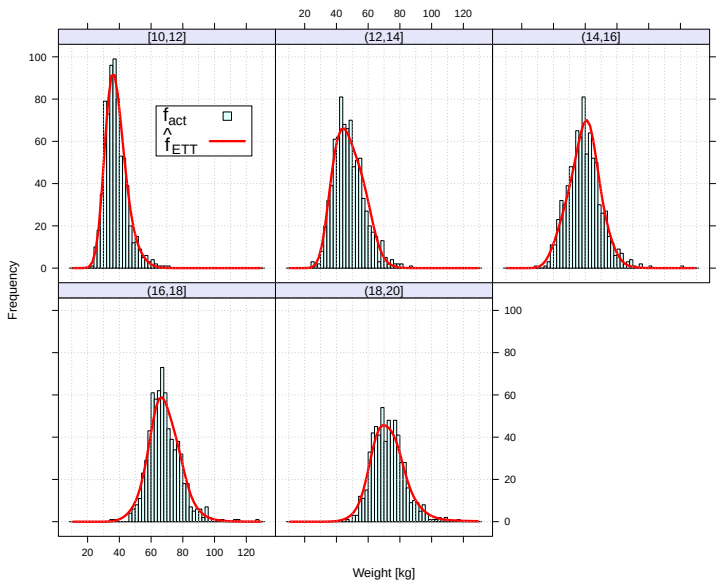


FIGURE 1. Weight distribution for Dutch children in 1997 in five different age-groups ranging from 10 to 20 years. A simple EET mode provides a good fit to these data.

3 Applications

3.1 Weight data

Figure 1 presents our first example: The weight (in kg) of Dutch children for different age-groups (in 1997; data from Fredriks et al., 2000). There is a clear shift to higher weights during puberty and an increasing variance as well as a tendency toward a normal distribution. Despite these several changing features a simple EET model provides a good fit. These data can be summarized well by a reference density g and five additional parameters (Lagrange multipliers $\boldsymbol{\lambda}_1$).

3.2 Dutch fertility

In the second example we look at the age at first birth for Dutch women born between 1938 to 1953. The data were taken from the Human Fertility Database (2013) and selected cohorts are shown in Figure 2. While the quantum of fertility has clearly decreased, the associated variability around the modal age at first birth has increased over time.

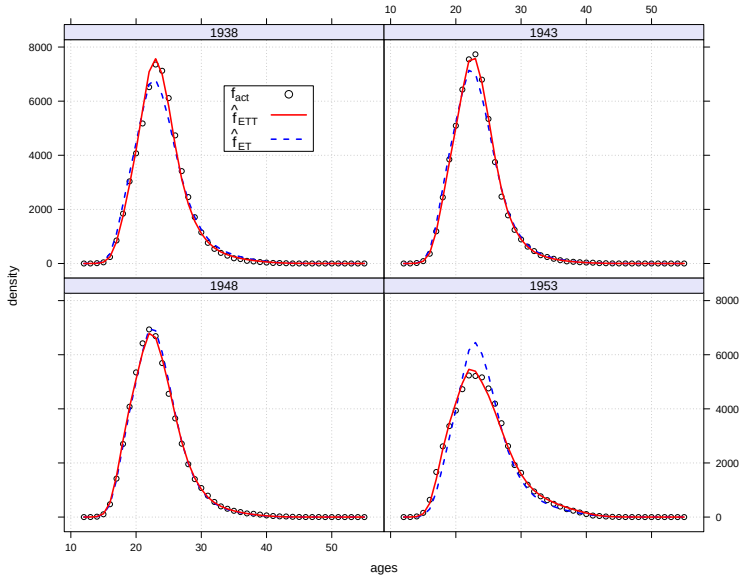


FIGURE 2. Age at first birth for selected cohorts of Dutch women (ages 12-54, cohorts born between 1938 and 1953). Fit of simple and extended EET model.

The simple EET model presented in (7) is not able to capture the development of fertility over the cohorts properly (dotted lines). Therefore

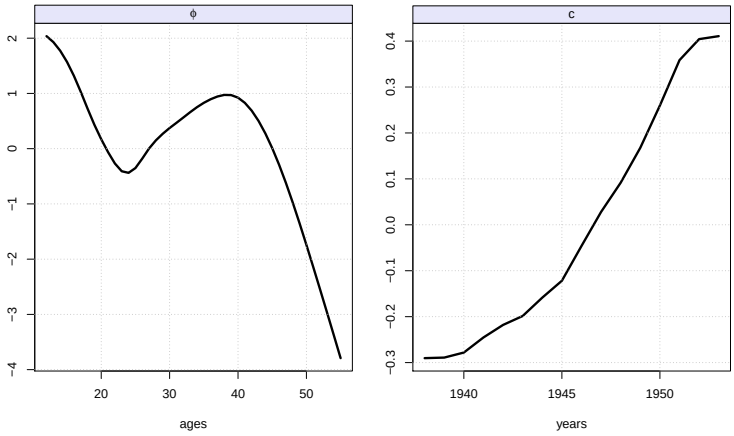


FIGURE 3. Estimated bilinear component of the E3T model, see equation (9), for the age at first birth distributions.

we fitted the extended model (9), which improves the results considerably (solid lines). Figure 3 shows the estimated values of $\boldsymbol{\phi}$ over age and of $\boldsymbol{c}$ across the birth cohorts.

References

Human Fertility Database (2013). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at `www.humanfertility.org`.

Fredriks, A.M., van Buuren, S., Wit, J.M., and Verloove-Vanhorick, S.P. (2000). Body mass index measurements in 1996-7 compared with 1980. *Archives of Disease in Childhood*, **82**, 107 – 112.

Modelling Social Contact Data: a smoothing constrained approach

Carlo G. Camarda[1], Niel Hens[2,3], Paul H.C. Eilers[4]

[1] Institut National d'Études Démographiques, Paris, France
[2] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Hasselt, Belgium
[3] Centre for Health Economics Research and Modeling Infectious Diseases, Vaccine & Infectious Disease Institute, University of Antwerp, Antwerp, Belgium
[4] Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands

E-mail for correspondence: carlo-giovanni.camarda@ined.fr

Abstract: Estimating age-specific contact rates from social contact surveys has proven valuable for modelling infectious disease spread via the respiratory or close-contact route. Here, we present a smoothing constrained approach to estimate these contact rates between people of possibly different ages. We use a two-dimensional approach where contact rates are assumed smooth from a cohort perspective as well as from the age distribution of contacts. The proposed method uses a combination of penalized likelihood for rectangular arrays for smoothing the contact rates and linear constraints to ensure reciprocity of contacts. We illustrate our approach using Belgian social contact data.

Keywords: Smoothing; Social Contact Data; Constraints; Symmetry.

1 Introduction

Mathematical modelling of infectious diseases transmitted by the respiratory or close-contact route (e.g. influenza) is increasingly being used to determine the impact of possible interventions. More recently, several authors have shown that informing mathematical models with social contact data is of great value avoiding making a priori contact assumptions with little or no empirical basis (Wallinga et al. 2006, Ogunjimi et al. 2009, Goeyvaerts et al. 2010). Goeyvaerts et al. (2010) have also shown that estimating mixing patterns from social contact data is not without difficulties and model choice is likely to effect the mathematical model outcome substantially.
In this paper we focus on modelling social contact data from a population-based contact survey that has been carried out in Belgium over the period March-May 2006 as part of POLYMOD, a European Commission project funded within the sixth framework programme (Mossong et al. 2008). Participants kept a paper diary with information on their contacts over one

day. A contact was defined as a two-way conversation of at least three words in each other's proximity. The contact information included the age of the contact, gender, location, duration, frequency, and whether or not touching was involved. Our interest goes to estimating the age-specific per capita contact rates nonparametrically while taking the reciprocal nature of contacts into account.

2 The model

Let $\boldsymbol{Y} = (y_{ij})$ denote the total number of contacts by all participants of age i, $i = 1, \ldots, m$ with contacts of age j, $j = 1, \ldots, n = m$ as measured in the sample. The vector $\check{\boldsymbol{e}} = (\check{e}_i)$ is the total number of participants at age i. We arrange the matrix of contacts by column order into a vector $\boldsymbol{y}$. Likewise we arrange the matrix of exposures $\boldsymbol{e} = \mathtt{vec}(\boldsymbol{E})$, where $\boldsymbol{E} = \check{\boldsymbol{e}}\, \mathbf{1}_{1,m}$.
The actually observed contacts are assumed to be realizations from a Poisson distribution: $\boldsymbol{y} \sim \mathcal{P}(\boldsymbol{\mu})$. The expected values $\boldsymbol{\mu}$ are the product of the exposure $\boldsymbol{e}$ and the actual contact rates, γ_{ij}, which are assumed to be smooth.
Additional to smoothness, we need to account for the reciprocal nature of contacts which acts at the population level. Let $\boldsymbol{p}$ denote the age-structure of the population in which the survey is conducted. When estimating the social contact matrix, we will enforce symmetry as follows:

$$\gamma_{ij}\, p_i = \gamma_{ji}\, p_j\,, \tag{1}$$

which means that the total number of contacts from age i to age j must equal the total number of contacts from age j to age i.
The contact rate matrix $\boldsymbol{\Gamma} = (\gamma_{ij})$ should be interpreted from a cohort perspective: people age through time and we assume contact rates for consecutive time points to be similar. Thus we aim to model our data over the diagonal component (so including all sub-diagonals). Unlike for the age of respondents, it does make sense to smooth over the dimension of the contacts' ages, e.g. children will meet their parents and grandparents who are $\pm$ 28 and $\pm$ 56 years older. The distribution of the age of (grand)parents is assumed smooth.
Figure 1 offers a graphical representation of the original data (left panel) and the equivalent scheme for the re-structure data over the age of contacts and cohort of the respondents.
Changing the coordinates allows us to reproduce a rectangular grid and take advantage of computational methods for rectangular arrays. We only need to create dummy data for the points missing from the grid and use a weight matrix to remove the dummy data from influencing the fitting procedure.
We define $\breve{\boldsymbol{y}}$, $\breve{\boldsymbol{e}}$ and $\breve{\boldsymbol{w}}$ contacts, exposure and weights (equal zero at dummy data), respectively, over the new co-ordinates in a column vector. We aim

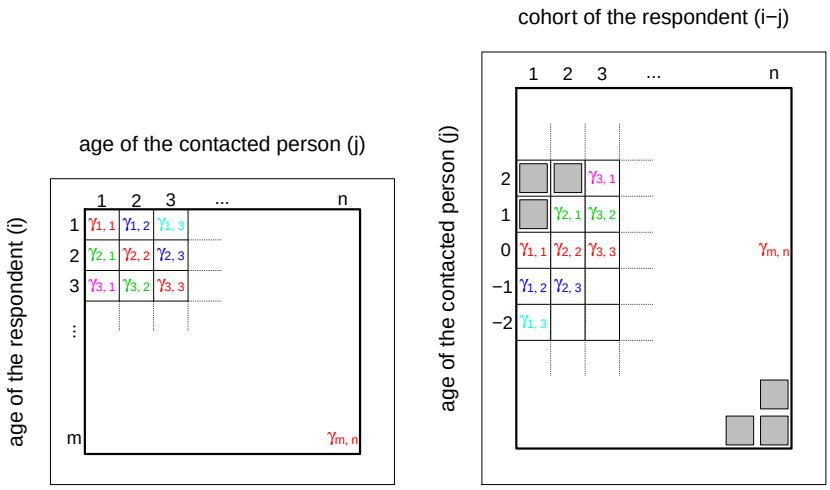


FIGURE 1. Schematic representation of the original data structure over ages of respondents and ages of contacts (left panel) and the re-arranged structure over cohort of the respondents and ages of the contacted person (right panel). Missing points are depicted with gray squares.

to smooth $\ln(\breve{\boldsymbol{\gamma}}) = \breve{\boldsymbol{\eta}}$ given the mentioned constraints. We can express these constraints as follows

$$\boldsymbol{H}\,\breve{\boldsymbol{\eta}} = \boldsymbol{\kappa}\,, \tag{2}$$

where the constraints matrix $\boldsymbol{H}$ allocates the vector $\breve{\boldsymbol{\eta}}$ to suit the constraints in (1) with

$$\begin{aligned}\boldsymbol{\kappa}^{\mathrm{T}} \;=\; & (p_2 - p_1, p_3 - p_1, \ldots, p_m - p_1,\\ & p_3 - p_2, p_4 - p_2, \ldots, p_m - p_2, \ldots\\ & p_m - p_{m-1})\,.\end{aligned}$$

The unique solution for $\breve{\boldsymbol{\eta}}$ subject to (2) is given by solving

$$\begin{pmatrix} \boldsymbol{W} + \boldsymbol{P} & \boldsymbol{H}^{\mathrm{T}} \\ \boldsymbol{H} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \breve{\boldsymbol{\eta}} \\ \boldsymbol{\omega} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}\boldsymbol{z} \\ \boldsymbol{\kappa} \end{pmatrix}, \tag{3}$$

where $\boldsymbol{W} = \mathtt{diag}(\breve{\boldsymbol{\gamma}} * \breve{\boldsymbol{e}} * \breve{\boldsymbol{w}})$ and $\boldsymbol{z} = \breve{\boldsymbol{\eta}} + (\breve{\boldsymbol{y}}/(\breve{\boldsymbol{\gamma}} * \breve{\boldsymbol{e}}) - 1)$. Here, $*$ and $/$ indicate element-by-element multiplication and division, respectively.
The penalty term is given by

$$\boldsymbol{P} = \lambda_1\, \boldsymbol{I}_n \otimes \boldsymbol{D}_{d,1}^{\mathrm{T}} \boldsymbol{D}_{d,1} \;+\; \lambda_2\, \boldsymbol{D}_{d,2}^{\mathrm{T}} \boldsymbol{D}_{d,2} \otimes \boldsymbol{I}_{m+n-1}\,,$$

where λ_1 and λ_2 are the smoothing parameters for the two new dimensions. The matrices $\boldsymbol{D}_{d,1}$ and $\boldsymbol{D}_{d,2}$ calculate d-th order differences for the

domains of $\breve{\boldsymbol{\eta}}$ (Currie et al., 2004). In the following, we use second order differences and smoothing parameters were chosen based on minimization of the Bayesian Information Criterion:

$$\mathrm{BIC}(\lambda_1, \lambda_2) = \mathrm{DEV}(\breve{\boldsymbol{y}}|\breve{\boldsymbol{\gamma}}) + \ln\left(\sum \breve{w}_i\right) \mathrm{ED}\,,$$

where $\mathrm{DEV}(\breve{\boldsymbol{y}}|\breve{\boldsymbol{\gamma}})$ is the deviance of the Poisson model. We take the trace of the hat-matrix for the estimated linearized smoothing problem in (3).

2.1 Computational note

Our model does not employ any regression basis such as B-splines because we need an exact link between constraints and linear predictor. This leads to the number of coefficients equal to the length of $\breve{\boldsymbol{y}}$. For instance, in the following application we will have $(m + n - 1) \times n = 153 \times 77 = 11781$ coefficients.

This is practically intractable on a regular personal computer. The system of equations in (3) is thus programmed within the sparse matrix `R`-package `Matrix` (Bates and Maechler, 2011).

Furthermore, the computation of the effective dimension of the model involves the inverse of an extremely huge matrix which leads to storage issues also in a sparse-matrix environment. Since the diagonal of the hat-matrix is the only object needed, we opted to save space by repeating the inverse for each column of an identity matrix with suitable dimensions and entries, and store the relevant elements of the result.

3 Belgian Contact Data

Figure 2 presents the outcomes of the proposed approach on the Belgian social contact data over the original domains. We analyze data from age 0 to 76 and both sexes. In the upper-left panel, we see the actual contact rates. Although it is rather noisy, the surface shows a general tendency of meeting coeval people, especially during school ages. Additional high rates are evident for inter-generational contacts (e.g. parents with children, teachers with pupils).

The top-right panel of Figure 2 shows the results of the smoothing constrained approach: we simultaneously smooth the per-capita expected number of contacts, retaining the symmetry at the population level. These estimates were obtained from the optimal combination of (λ_1, λ_2) as picked from the BIC-profile shown in the bottom-left panel. The image of the bottom-right panel finally presents the per-capita expected number of contacts for the Belgian population, which is symmetric around the main diagonal.

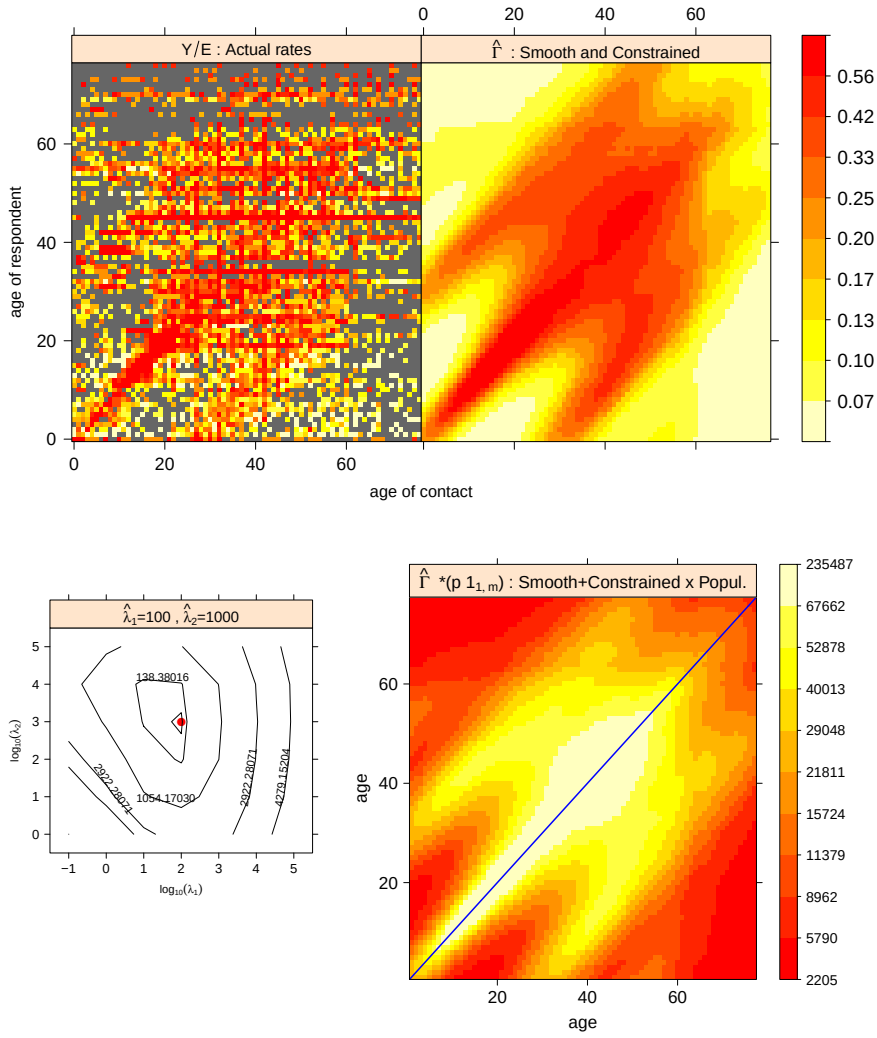


FIGURE 2. Belgian survey 2006, ages from 0 to 76. Actual and fitted contact rates (top panels). BIC-contour plot (bottom-left). Population level mixing (bottom-right).

4 Discussion

We used a two-dimensional approach to smooth age-specific contact rates. Smoothing was done from both a cohort perspective as well as from the age distribution of contacts. The proposed method uses a combination of penalized likelihood for rectangular arrays for smoothing and linear constraints

to ensure reciprocity of contacts at the population level. We illustrate our approach using Belgian social contact data.
Compared to existing methods our method is computationally efficient and facilitates taking contact reciprocity into account. The resulting contact surface shows clear like-with-like and between generation mixing. Further research focuses on using a negative binomial distribution to account for overdispersion, relaxing smoothness for school-aged children and including covariates within the algorithm, e.g. sex and duration.
Finally population-based contact surveys present two common features: contacts may be reported in age-groups and often people tend to round the age of their contacts. We intend to explore these issues in future work.

References

Bates, D. and Maechler, M. (2011). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.0-2.

Currie, I.D., Durban, M. and Eilers, P.H.C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*; **4**, 279 – 298

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., Van Damme, P. and Beutels, P. (2010). Estimating infectious disease parameters from data on social contacts and serological status. *Journal of the Royal Statistical Society Series C*, **59**, 255 – 277.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Scalia Tomba, G., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M. and Edmunds, J. (2008). Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS Medicine*, **5**, 381 – 391.

Ogunjimi, B., Hens, N., Goeyvaerts, N., Aerts, M., Damme, P. V. and Beutels, P. (2009). Using empirical social contact data to model person to person infectious disease transmission: an illustration for varicella. *Mathematical Biosciences*, **218**, 80 – 87.

Wallinga, J., Teunis, P. and Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology*, **164**, 936 – 944.

Marginal Parameterizations for Hidden Markov Models

Roberto Colombi[1], Sabrina Giordano[2]

[1] Dept of Engineering, University of Bergamo, Italy
[2] Dept of Economics, Statistics and Finance, University of Calabria, Italy

E-mail for correspondence: sabrina.giordano@unical.it

Abstract: We propose to parameterize the two components of a hidden Markov model (HMM), the observation model and the latent model, using the marginal parameterization (Bergsma and Rudas, 2002, Bartolucci *et al.*, 2009). We will show that many interesting hypotheses on the probabilities of the observable variables given the latent states and on the transition probabilities of the latent processes of the HMM can be verified by testing constraints on marginal parameters.

Keywords: Granger noncausality; Markov chains; Marginal interactions.

1 Introduction

In several applications involving time series, it is of interest to describe how the evolution of variables over time depends on latent characteristics or the focus may be on the dynamics of unobservable characteristics measured by variables observed at consecutive time occasions. These issues are addressed by hidden Markov models (e.g. MacDonald and Zucchini, 1997, Cappé et al., 2005).
Basically, a hidden Markov model (HMM) assumes that an observed time series depends on an unobservable Markov chain in such a way that the joint process is also Markovian. The main assumption of HMMs is that the observable variables, at different times, are independent given the latent states of a first order Markov chain with a finite state space.
In this work, we focus on discrete hidden Markov models with a multivariate categorical observable process and a multivariate latent chain, so we observe more variables at each time and assume that their distributions can be affected by one or more latent variables.
Our aim is to parameterize both observation and latent models as marginal models and then verify hypotheses which make these models more parsimonious by constraining marginal interactions.

2 Multivariate Hidden Markov models

Let $\mathbf{E}_{\mathcal{U}}$ be a r-variate first order Markov chain, $\mathbf{E}_{\mathcal{U}} = \{E_{\mathcal{U}}(t) : t \in \mathbb{N}\} = \{E_{it} : t \in \mathbb{N}, i \in \mathcal{U}\}$, $\mathcal{U} = \{1, ..., r\}$, $\mathbb{N} = \{0, 1, 2, ..., \}$ and let $\mathbf{F}_{\mathcal{V}}$ be a s-dimensional process of categorical variables $\mathbf{F}_{\mathcal{V}} = \{F_{\mathcal{V}}(t) : t \in \mathbb{N}\} = \{F_{jt} : t \in \mathbb{N}, j \in \mathcal{V}\}$, $\mathcal{V} = \{1, ..., s\}$. We assume that the joint process ($\mathbf{E}_{\mathcal{U}}$, $\mathbf{F}_{\mathcal{V}}$) is a hidden Markov process where $\mathbf{F}_{\mathcal{V}}$ is the observed process and $\mathbf{E}_{\mathcal{U}}$ is the latent Markov chain.
The marginal components $\{E_{it}\}$, $\{F_{jt}\}$ take values in finite sets $\mathcal{E}_i$ and $\mathcal{F}_j$, $i \in \mathcal{U}, j \in \mathcal{V}$.
One realization of the process $\mathbf{F}_{\mathcal{V}}$ at a given time is denoted by $\boldsymbol{f} = (f_1, f_2, ..., f_s) \in \mathcal{F} = \times_{j \in \mathcal{V}} \mathcal{F}_j$, and one state of the Markov chain $\mathbf{E}_{\mathcal{U}}$ is $\boldsymbol{e} = (e_1, e_2, ..., e_r)$ in $\mathcal{E} = \times_{i \in \mathcal{U}} \mathcal{E}_i$. For every subset $\mathcal{T} \subset \mathcal{U}$ and $\mathcal{R} \subset \mathcal{V}$, marginal processes of the latent Markov chain and the observed variables are represented by $\mathbf{E}_{\mathcal{T}} = \{E_{it} : i \in \mathcal{T}, t \in \mathbb{N}\}$ and $\mathbf{F}_{\mathcal{R}} = \{F_{jt} : j \in \mathcal{R}, t \in \mathbb{N}\}$ taking values on the sets $\times_{i \in \mathcal{T}} \mathcal{E}_i$ and $\times_{j \in \mathcal{R}} \mathcal{F}_j$.
The time-homogeneous joint transition probabilities are denoted by $\phi(\boldsymbol{e}|\boldsymbol{e}')$ for every pair of states $\boldsymbol{e}' \in \mathcal{E}$, $\boldsymbol{e} \in \mathcal{E}$.
Moreover, $\varphi(\boldsymbol{f}|\boldsymbol{e})$ indicates the conditional probabilities of the observable variables given the latent state $\boldsymbol{e}$ and $\boldsymbol{f}_{\mathcal{R}}$ denotes the vector with components $f_j : j \in \mathcal{R} \subset \mathcal{V}$. Furthermore, $\varphi_{\mathcal{R}}(\boldsymbol{f}_{\mathcal{R}}|\boldsymbol{e})$ represents the marginal probability of the observable variables in the set $\mathcal{R}$ given the latent state $\boldsymbol{e}$.
Finally, $\phi_{\mathcal{T}}(\boldsymbol{e}_{\mathcal{T}}|\boldsymbol{e}')$ is the marginal transition probability from state $\boldsymbol{e}' \in \mathcal{E}$ to state $\boldsymbol{e}_{\mathcal{T}}$ with components $e_i : i \in \mathcal{T} \subset \mathcal{U}$.

3 Constrained HMMs

Two components of a HMM should be distinguished, the observation model and the latent model, concerning the distribution of observable variables given the latent states and the transition probabilities of the chain, respectively.
We use the marginal parameterization (Bergsma and Rudas, 2002, Bartolucci *et al.*, 2009) to model the probabilities of the observable variables given the latent states and the transition probabilities of the latent process. We will show that many interesting hypotheses on observable and latent components of the HMM can be verified by testing constraints on marginal parameters.
We now briefly outline the basic concepts of complete hierarchical marginal models.
Consider c categorical variables denoted by the first c integers. The set of all variables is $\mathcal{C} = \{1, 2, ..., c\}$.
A marginal distribution is identified by a subset of $\mathcal{C}$. The set $\mathcal{M}$, that identifies a marginal distribution, is called marginal set.

In the complete hierarchical marginal models, the parameters are called *marginal interactions*. In particular, marginal interactions are log-linear parameters defined in different marginal distributions in Bergsma and Rudas models, while Bartolucci *et al.* (2009) use more generale marginal interactions which are meaningful when the variables have an ordinal nature.
As in log-linear models, a family of interactions is defined for every subset $\mathcal{S}$ of variables, $\mathcal{S} \subset \mathcal{C}$, which is called interaction set.
For every interaction set $\mathcal{S}$, the marginal interactions are defined within a marginal distribution identified by a marginal set $\mathcal{M}(\mathcal{S})$ belonging to a family of non-decreasing sets $\mathcal{H} = \{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$, $\mathcal{M}_k = \mathcal{C}$.
More specifically, in complete hierarchical marginal models, the interactions, involving all the variables in $\mathcal{S}$, are defined in one and only marginal distribution $\mathcal{M}(\mathcal{S})$ (completeness) and $\mathcal{M}(\mathcal{S})$ is the first marginal set of $\mathcal{H}$ which contains $\mathcal{S}$ (hierarchy), see Bergsma and Rudas, 2002, Bartolucci *et al.*, 2009. For every interaction set $\mathcal{S}$, the marginal interactions are denoted by $\eta^{\mathcal{S};\mathcal{M}(\mathcal{S})}(\boldsymbol{i}_{\mathcal{S}})$, where $\boldsymbol{i}_{\mathcal{S}}$ is the vector of indexes of variables which the interactions depend on. When $\mathcal{M}(\mathcal{S}) = \mathcal{S}$ the symbol $\eta^{\mathcal{S}}(\boldsymbol{i}_{\mathcal{S}})$ is used.

3.1 Marginal parameterizations of HMMs

For every observable or latent categorical variable, the first category is called baseline. Any observation $\boldsymbol{f} = (f_1, f_2, ..., f_s)$ which includes categories at the baseline level for variables $j \notin \mathcal{J}$, $\mathcal{J} \subset \mathcal{V}$, is denoted by $(\boldsymbol{f}_{\mathcal{J}}, \boldsymbol{f}^*_{\mathcal{V}\setminus\mathcal{J}})$. A similar notation holds for the latent state $\boldsymbol{e} = (e_1, e_2, ..., e_r)$.
For every non-empty subset $\mathcal{P}$ of the observable variables $\mathcal{V}$ and for every $\boldsymbol{f}_{\mathcal{P}} \in \times_{j\in\mathcal{P}}\mathcal{F}_j$, the baseline interactions $\eta^{\mathcal{P};\mathcal{M}(\mathcal{P})}(\boldsymbol{f}_{\mathcal{P}}|\boldsymbol{e})$, $\mathcal{M}(\mathcal{P}) \in \mathcal{H}_{obs}$, of a marginal model for the observable variables are contrasts of logarithms of the marginal probabilities of the observations given the latent states

$$\eta^{\mathcal{P};\mathcal{M}(\mathcal{P})}(\boldsymbol{f}_{\mathcal{P}}|\boldsymbol{e}) = \sum_{\mathcal{K}\subseteq\mathcal{P}} (-1)^{|\mathcal{P}\setminus\mathcal{K}|} \log \varphi_{\mathcal{M}(\mathcal{P})}(\boldsymbol{f}_{\mathcal{K}}, \boldsymbol{f}^*_{\mathcal{M}(\mathcal{P})\setminus\mathcal{K}}|\boldsymbol{e}).$$

In order to model the dependence of the distribution of the observable variables on the states $\boldsymbol{e}$, we adopt the usual factorial expansion

$$\eta^{\mathcal{P};\mathcal{M}(\mathcal{P})}(\boldsymbol{f}_{\mathcal{P}}|\boldsymbol{e}) = \sum_{\mathcal{Q}\subseteq\mathcal{E}} \theta^{\mathcal{P},\mathcal{Q}}(\boldsymbol{f}_{\mathcal{P}}|\boldsymbol{e}_{\mathcal{Q}}). \tag{1}$$

The Möbius inversion theorem ensures that $\theta^{\mathcal{P},\mathcal{Q}}(\boldsymbol{f}_{\mathcal{P}}|\boldsymbol{e}_{\mathcal{Q}}) = \sum_{\mathcal{K}\subseteq\mathcal{Q}} (-1)^{|\mathcal{Q}\setminus\mathcal{K}|}\ \eta^{\mathcal{P},\mathcal{M}(\mathcal{P})}(\boldsymbol{f}_{\mathcal{P}}|\boldsymbol{e}_{\mathcal{K}}, \boldsymbol{e}*_{\mathcal{E}\setminus\mathcal{K}})$.
When $\mathcal{M}(\mathcal{S}) = S = \{j\}$ is a singleton, the marginal parameters are marginal logits, which will be denoted by $\eta^{\{j\}}(f_j|\boldsymbol{e})$.
Analogously, in the marginal model for the latent component of HMMs, we define the marginal parameters $\lambda^{\mathcal{P};\mathcal{M}(\mathcal{P})}(\boldsymbol{e}_{\mathcal{P}}|\boldsymbol{e}')$ for every $\mathcal{P} \subseteq \mathcal{U}$,

$\mathcal{M}(\mathcal{P}) \in \mathcal{H}_{lat}$, on the marginal transition probabilities $\phi_{\mathcal{M}(\mathcal{P})}(\boldsymbol{e}_{\mathcal{P}}|\boldsymbol{e}')$ and the factorial expansion

$$\lambda^{\mathcal{P};\mathcal{M}(\mathcal{P})}(\boldsymbol{e}_{\mathcal{P}}|\boldsymbol{e}') = \sum_{\mathcal{Q}\subseteq\mathcal{U}} \delta^{\mathcal{P},\mathcal{Q}}(\boldsymbol{e}_{\mathcal{P}}|\boldsymbol{e}'_{\mathcal{Q}}). \tag{2}$$

3.2 Constrained observation and latent models

In the framework of hidden Markov models with several latent and observable variables, several interesting hypotheses on the latent and the observable models can be easily formulated.
These hypotheses reduce the number of parameters needed to parameterize the transition and observation probabilities.
We illustrate some hypotheses for the observation model using a marginal parametrization whose interactions are defined in the univariate distributions or in the joint distribution.
A useful restriction that considerably simplifies the observation model is the hypothesis of *additivity* of the effects of the latent variables on the marginal logits of the observable variables.
This marginal additive dependence allows the logits $\eta^{\{j\}}(f_j|\boldsymbol{e})$, $j \in \mathcal{V}$, to be expressed by the factorial expansion

$$\eta^{\{j\}}(f_j|\boldsymbol{e}) = \theta^{\{j\}}(f_j) + \sum_{k\in\mathcal{U}} \theta^{\{j\},\{k\}}(f_j|e_k). \tag{3}$$

Note that under this hypothesis, the parameters $\theta^{P,Q}(\boldsymbol{f}_{\mathcal{P}}|\boldsymbol{e}_{\mathcal{Q}})$ described in (1) are null if $P = \{j\}$ and $|Q| > 1$.
Another hypothesis is that of *invariant association* corresponding to the constraints $\eta^{\mathcal{P};\mathcal{V}}(\boldsymbol{f}_{\mathcal{P}}|\boldsymbol{e}) = \eta^{\mathcal{P};\mathcal{V}}(\boldsymbol{f}_{\mathcal{P}})$, if $|P| = 2$ and $\eta^{\mathcal{P};\mathcal{V}}(\boldsymbol{f}_{\mathcal{P}}) = \mathbf{0}$ if $|P| > 2$.
According to this hypothesis the odds ratios of two observable variables do not depend on the states of the latent variables and on the levels of the other observable variables.
Finally constraining to zero some of the remaining parameters $\eta^{\mathcal{P};\mathcal{V}}(\boldsymbol{f}_{\mathcal{P}})$, $|P| = 2$, it is possible to allow the probability functions of the observations given the states to satisfy a list of conditional independencies.
Other relevant hypotheses can be formulated using different marginal parameterizations as the Gloneck-McCullagh one according to which $\mathcal{M}(\mathcal{S}) = \mathcal{S}$ for each $\mathcal{S} \subseteq \mathcal{V}$.
It is possible to simplify also the latent model by constraining the parameters $\delta^{\mathcal{P},\mathcal{Q}}(\boldsymbol{e}_{\mathcal{P}}|\boldsymbol{e}'_{\mathcal{Q}})$ defined in (2).
Under the *Granger noncausality* hypothesis for first order Markov chains: $E_{\mathcal{T}}(t) \perp\!\!\!\perp E_{\mathcal{U}\setminus\mathcal{T}}(t-1)|E_{\mathcal{T}}(t-1)$, the marginal process $\mathbf{E}_{\mathcal{T}}$ is a Markov chain (see Colombi and Giordano, 2012) and if $\mathcal{M}(\mathcal{P}) \subseteq \mathcal{T}$, for all $\mathcal{P} \subseteq \mathcal{T}$, this condition is satisfied when $\delta^{\mathcal{P},\mathcal{Q}}(\mathbf{e}_{\mathcal{P}}|\mathbf{e}'_{\mathcal{Q}}) = 0$ for all $\mathcal{P} \subseteq \mathcal{T}$, $\mathcal{Q} \not\subseteq \mathcal{T}$, $\mathbf{e}_{\mathcal{P}} \in \times_{i\in\mathcal{P}}\mathcal{E}_i$, $\mathbf{e}'_{\mathcal{Q}} \in \times_{i\in\mathcal{Q}}\mathcal{E}_i$.

Another interesting hypothesis is that of conditional contemporaneous independence which corresponds to the constraints $\lambda^{\mathcal{P};\mathcal{M}(\mathcal{P})}(\boldsymbol{e}_{\mathcal{P}}|\boldsymbol{e}') = \mathbf{0}$, if $|P| \geq 2$, according to which the transition probabilities $\phi(\boldsymbol{e}|\boldsymbol{e}')$ factorize in the product of marginal transition probabilities $\phi_{\{i\}}(e_i|\boldsymbol{e})$.

4 Example

In this section, constraints on marginal parameters in latent and observation models of HMMs are tested on the data set of a soft-drink company (available in the R-package hmmm by Colombi et al., 2012).
The data consists of a one-year time series of daily sales of soft-drinks: lemon tea, orange juice and apple juice, all with categories: low, medium, high level.
Changes in sale outcomes over time can depend on time-varying unobserved factors and we consider a HMM with two dichotomous latent variables to model these data.
Table 1 reports the likelihood ratio tests (LRT), degrees of freedom (df) and p-values for models restricted under the hypotheses of additivity (3), invariant association and Granger noncausality.
The EM algorithm used in this context is described in Colombi and Giordano (2011) and implemented in the R-package hmmm.

TABLE 1. Constrained latent and observation models for soft-drink data

latent model	*obs. model*	*LRT*	*df*	*p-value*
noGranger	saturated	3.6478	4	0.4557
noGranger + inv ass	saturated	11.023	7	0.1376
saturated	addit	14.897	26	0.959
noGranger	addit	17.767	30	0.862
noGranger + inv ass	addit	18.969	33	0.976
saturated	addit + inv ass	56.241	74	0.939
noGranger	addit + inv ass	76.490	78	0.527
noGranger + inv ass	addit + inv ass	81.735	81	0.456

References

Bartolucci, F., Colombi, R. and Forcina, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, **17**, 691 – 711.

Bergsma, W. P. and Rudas, T. (2002). Marginal models for categorical data *The Annals of Statistics*, **30**, 140 – 159.

Cappé, O., Moulines, E. and Rydén T. (2005). *Inference in Hidden Markov Models*. New York: Springer.

Colombi, R. and Giordano, S. (2011). Testing lumpability for marginal discrete hidden Markov models. *Advances in Statistical Analysis*, **95**, 293 – 311.

Colombi, R. and Giordano, S. (2012). Graphical models for multivariate Markov chains. *Journal of Multivariate Analysis*, **107**, 90 – 103.

Colombi, R., Giordano, S. and Cazzaro, M. (2012). R-package hmmm: Hierarchical Multinomial Marginal Models $http://CRAN.R-project.org/package = hmmm$.

MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman & Hall.

Marginal Models for the Association Structure of Hierarchical Binary Responses

Enrico A. Colosimo[1], André G.F.C. Costa[1], Leila Amorim[2]

[1] Federal University of Minas Gerais, Brazil
[2] Federal University of Bahia, Brazil

E-mail for correspondence: enricoc@est.ufmg.br

Abstract: Marginal models for correlated binary data are presented and compared in this paper. Marginal probabilities and odds ratios are estimated in a real data application involving a four levels hierarchical clustering. ALR and ORTH models proved to be useful for modeling a complex association structure in the presence of large cluster size.

Keywords: ARL; correlated binary response; odds ratio; ORTH.

1 Introduction

Clustering structures in binary responses are often found in epidemiological and biological studies, requiring a more sophisticated approach to statistical modeling. When the structure of the association is the scientific focus, Prentice (1988), Lipsitz, et al. (1991), Liang, et al. (1992) extended the idea of GEE (Liang and Zeger, 1986) introducing a second estimation equation for the parameters of association. However, numerical methods proposed can be computationally infeasible if the amount of measures within the cluster is large. Carey, et al. (1993) proposed the ALR (Alternating Logistic Regression) and Zink (2012) proposed the ORTH (Orthogonalized Residuals) as a solution to the problems related to computational methods. ALR is structurally different from other approaches, since to avoid the computationally burdens of other methods, the second estimation equation is definded in terms conditional residuals. The ORTH model is a new approach for the second estimation equation, replacing the strategy of conditional residuals to orthogonalized ones.

2 Real Data Motivation

This paper was motivated by a study related to fungal endophytes distribution where the association structure is of primary research interest. The data set presents four levels of hierarchical clustering: fragment, leaf, individual host tree and collection site. Presence or absence in five different

collection sites, two in Brazil and three in Argentina, was used to measure fungal endophytes distribution. At each collection site a transect was obtained and twenty individual host trees were selected each one approximately five meters from the next one. For each individual host tree five leaf were selected, and the sample of fungal was carried out in six different fragments of the leaf: one from the base (C, near petiole), two from the middle vein (E and F), one from the left margin (D), one from the right margin (B) and one from the tip (A). Thus, it has 600 measures within each collection site, a total of 3000 measurements for the whole study.
Main goal of the study is to estimate measures related to fungal endophytes association, within site, within individual host tree, within leaf and within fragment. It is also of interest to test the hypothesis that as you increase the distance between individual host tree of the same site collection, the association of the fungal endophytes decreases. A secondary goal of the study is related to the mean prevalence of the fungal endophytes in comparing Brazil and Argentina.

3 Marginal Models

Suppose N independent clusters, each one with n_i observations. Consider the index i identifies the i-th cluster, $i = 1, ..., N$, and j and k identifying two observations within the same cluster, with $1 \leq j < k \leq n_i$. For the i-th cluster the response vector is given by $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^{'}$, such that Y_{ij} follows a Bernoulli distribution with mean $\mu_{ij} = pr(Y_{ij} = 1)$. GEE1 estimator proposed by Liang and Zeger (1986) for the marginal model is obtained by solving the following equation

$$S(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i^{'}}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta)) = 0, \tag{1}$$

where $\mu_i = (\mu_{i1}, ..., \mu_{in_i})$, $V_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}$, $R_i(\alpha)$ is a working correlation matrix $n_i \times n_i$ and A_i is a diagonal matrix with the elements of diagonal given by $Var(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$. Mean function $\mu_i(\beta)$ depend on a p-vector of covariates X_{ij} through a link function g as $\mu_{ij}(\beta) = g^{-1}(X_{ij}'\beta)$. Quantities α's are taken as nuisance parameters and estimated by the method of moments (Diggle et al., 2002).
On the other side, when α is the scientific focus, Prentice(1988) extended the idea of GEE1 introducing the second estimation equation given by

$$S(\alpha) = \sum_{i=1}^{N} \frac{\partial \rho_i(\alpha)^{'}}{\partial \alpha} W_i^{-1}(Z_i - \rho_i(\alpha)) = 0, \tag{2}$$

where $Z_i = \{Z_{ijk}\}$ is the correlation between Y_{ij} and Y_{ik}

$$Z_{ijk} = \frac{(Y_{ij} - \mu_{ij}(\beta))(Y_{ik} - \mu_{ik}(\beta))}{\sqrt{\mu_{ij}(\beta)(1-\mu_{ij}(\beta))(\mu_{ik}(\beta)(1-\mu_{ik}(\beta))}}$$

and $\rho_i = \{\rho_{ijk}\}$ is a vector of dimension $m_i = n_i(n_i - 1)/2$. W_i is a working matrix of Z_i, usually assumed $W_i = diag(w_{i12}, ..., w_{i1n_i}, w_{i23}, ...)$, where $w_{ijk} = Var(Z_{ijk})$ is expressed as

$$w_{ijk} = 1 + (1-2\mu_{ij})(1-2\mu_{ik})[(\mu_{ij}(1-\mu_{ij}))(\mu_{ik}(1-\mu_{ik}))]^{\frac{-1}{2}}\rho_{ijk} - \rho_{ijk}^2.$$

Taking X_{ijk} as a matrix $m_i \times q$ where q is the number of covariates used to model the correlation structure, the $E(Z_{ijk}) = Corr(Y_{ij}, Y_{ik}|X_{ijk}) = \rho_{ijk}(\alpha)$, can depend on covariates through a link function $h(\rho_{ijk}) = X'_{ijk}\alpha$. However, for binary responses the correlation coefficient as a measure of association is not widely used, mainly due to the difficulty in interpretation. Lipsitz, et al. (1991) and Liang, et al. (1992) proposed modifications in the second estimation equation proposed by Prentice (1988) using the odds ratio to account for the association.
Let's denote $\mu_{ijk} = E(Y_{ij}Y_{ik})$, $1 \leq j < k \leq n_i$. Marginal probability Y_{ij} assumes a logistic link function and the association between the pairs Y_{ij}, Y_{ik} is defined as

$$logOR(Y_{ij}, Y_{ik}) = log\left(\frac{\mu_{ijk}(1-\mu_{ij}-\mu_{ik}+\mu_{ijk})}{(\mu_{ij}-\mu_{ijk})(\mu_{ik}-\mu_{ijk})}\right) = X'_{ijk}\alpha, \quad 1 \leq j \leq k \leq n_i.$$

Defining $\xi_{ijk} = E(Y_{ij}|Y_{ik} = y_{ik})$, we have that:

$$\xi_{ijk} = \mu_{ij} + \frac{\sigma_{ijk}}{\sigma_{ikk}}(y_{ik} - \mu_{ik}),$$

where $\sigma_{ijk} = Cov(Y_{ij}, Y_{ik}) = \mu_{ijk} - \mu_{ij}\mu_{ik}$ and $\sigma_{ikk} = Var(Y_{ik}) = \mu_{ik}(1 - \mu_{ik})$.
Denoting the m_i-vector of conditional residuals by C_i with elements $Y_{ij} - \xi_{ijk}$ and S_i a diagonal matrix with elements $\xi_{ijk}(1 - \xi_{ijk})$, we have that the ALR estimator for $\theta = (\beta, \alpha)$ is the simultaneous solution of the first estimating equation given in (1) and

$$S_{\alpha,ALR} = \sum_{i=1}^{N} \frac{\partial \xi_i'}{\partial \alpha} S_i^{-1} C_i = 0. \tag{3}$$

However, the stochastic nature of S_i and $\partial\xi_i/\alpha$ does not allow a theoretical investigation of (3) through the standard theory of estimation equation. Another drawback is that $S_{\alpha,ALR}$ is invariant to permutations of the vector Y_{ij} (Kuk, 2004) whereas the robust variance estimator is not.
Zink (2003) presented the ORTH model as an alternative one to the ALR. Orthogonalized residuals approach, again keeps the same equations to estimate the parameters of the mean and for the construction of the second

estimaton equation. In general, the latter equation is based on pairwise residuals and a weighted combination of these quantities. An approximate covariance matrix is then built in a way that is very computationally feasible for larger clusters (Qaqish, et al, 2012).
Let's define $U_{ijk} = (Y_{i1}Y_{i2}, ..., Y_{in_i-1}Y_{in_i})$. Orthogonalized residuals are defined as linear regressions of U_{ijk} on Y_{ij} and Y_{ik} specifying:

$$Q_{ijk} = U_{ijk} - [\mu_{ijk} + b_{ijk:j}(Y_{ij} - \mu_{ij}) + b_{ijk:k}(Y_{ik} - \mu_{ik})], \tag{4}$$

such that $b_{ijk:j} = \mu_{ijk}(1-\mu_{ik})(\mu_{ik}-\mu_{ijk})/d_{ijk}$, $b_{ijk:k} = \mu_{ijk}(1-\mu_{ij})(\mu_{ij}-\mu_{ijk})/d_{ijk}$, $d_{ijk} = \sigma_{ijj}\sigma_{ikk} - \sigma^2_{ijk}$.
After the definition of orthogonalized residuals Q_{ijk}, the second estimation equation is given as

$$S_{\alpha,ORTH} = \sum_{i=1}^{N} \frac{-\partial Q_i^{'}}{\partial \alpha} P_i^{-1} Q_i, \tag{5}$$

where P_i is a diagonal matrix with elements $\nu_{ijk} = Var(Q_{ijk}) =$

$$\frac{\mu_{ijk}(\mu_{ij} - \mu_{ijk})(\mu_{ik} - \mu_{ijk})(1 - \mu_{ij} - \mu_{ik} + \mu_{ijk})}{\mu_{ij}\mu_{ik}(1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}) - \mu^2_{ijk}}.$$

4 Numerical Results

It was used the following linear predictors for the mean and association for the four levels fungal endophytes study

$$logitPr(Y = 1) = \beta_0 + \beta_1 I(Country = Brazil),$$

$$LogOR(Y_j, Y_k) =$$

$$= \begin{cases} \alpha_1 \text{I(within collection site)} + \alpha_5 \text{Distance}_{jk} + \alpha_4 \text{I(within fragment)}, \\ \text{If j and k are different individual host tree in the same collection site,} \\ \alpha_1 \text{I(within collection site)} + \alpha_2 \text{I(within host tree)} + \alpha_4 \text{I(within fragment)}, \\ \text{If j and k are different leaf in the same individual host tree,} \\ \alpha_1 \text{I(within collection site)} + \alpha_2 \text{I(within host tree)} + \alpha_3 \text{I(within leaf)}, \\ \text{If j and k are different fragments in the same leaf,} \end{cases} \tag{6}$$

where distance is measured in decameter (Minimum = 0, Maximum = 11.4).
Table 1 presents the estimates of ALR and ORTH models. There is a significant association of fungal endophytes at the collection site, individual host

tree and leaf levels. The association in site collection depends on the distance between individual host tree. As the distance between trees increases, association of fungal endophytes decreases significantly. There isn't a significant association within-fragment level for the ALR model. On the other hand, p-value for this term is very close to significance in ORTH model might showing an increase of efficiency. In the mean structure, it can be observed that the chance of presence of fungal endophytes in Brazil is about 4 times the chance in Argentina.

TABLE 1. Results of ALR and ORTH models.

Models	ALR			ORTH		
Mean	β	ep(β)	P-value	β	ep(β)	P-value
Intercept	-3.343	0.364	0.000	-3.344	0.363	0.000
Country=Brazil	1.461	0.367	0.000	1.464	0.366	0.000
Association	α	ep(α)	P-value	α	ep(α)	P-value
Within site	0.025	0.205	0.904	0.031	0.059	0.599
Within host tree	1.277	0.392	0.001	1.293	0.398	0.001
Within leaf	0.918	0.351	0.009	0.898	0.340	0.008
Within fragment	0.075	0.048	0.117	0.077	0.040	0.055
Distance$_{jk}$	-0.033	0.013	0.011	-0.035	0.007	0.000

References

Carey, V., Zeger, S. L and Diggle P. (1993). Modelling Multivariate Binary Data with Alternating Logistic Regressions. *Biometrika*, **80**, 517 - 526.

Diggle, P. J., Heagerty, P., Liang, K. Y., Zeger, S. L. (2002). *Analysis of Longitudinal Data.*, Oxford University Press - 2nd. Edition.

Qaqish B.F., Zink, R. C. and Preisser J.S. (2012) Orthogonalized Residuals for Estimation of Marginally Specified Association Parameters in Multivariate Binary Data *Scandinavian Journal of Statistic*, **39**, 515 - 527.

Kuk, A.Y.C. (2004) Permutation invariance of alternating logistic regression for multivariate binary data.*Biometrika* , **91**, 758 - 761.

Liang, K. Y., Zeger, S. L. (1986) Longitudinal Data Analysis using generalized linear models. *Biometrika*, **73**, 13 - 22.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992), Multivariate regression analyses for categorical data. J. R. Statist. *Soc. B*, **54**, 3 - 40.

Lipsitz, S. R., Laird, N. M, Harrington, D. P. (1991) Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure os Association . *Biometrika*, **78**, 153 - 160.

Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033 - 48.

Zink, R. C. (2003) Correlated Binary Regression Using Orthogonalized Residuals. *PhD thesis, University of North Carolina, Chapel Hill.*

A nonparametric study of the spatial association between forest variables

Francisco Cuevas[1], Emilio Porcu[1], Ronny Vallejos [1]

[1] Universidad Técnica Federico Santa María, Chile

E-mail for correspondence: ronny.vallejos@usm.cl

Abstract: We propose a nonparametric approach to assess the spatial association between two spatial variables. Our proposal is based on a Nadaraya-Watson version of the codispersion coefficient through a suitable kernel. The proposed method is useful for quantifying spatial associations between two variables measured at the same locations. We study forest data concerning the relationship among the tree height, basal area, elevation and slope of Pinus radiata plantations. A two-dimensional codispersion map is constructed to provide insight into the spatial association between these variables.

Keywords: Kernel; Spatial association; Basal area and height.

1 Introduction

In the analysis of spatial data, the quantification of spatial associations between two variables is an important issue, and considerable effort has been devoted to the construction of appropriate coefficients and tests for the association between two correlated variables.
The codispersion coefficient (Matheron, 1965) is a measure of association between two spatial variables and has been used in several applications (Chilés and Delfiner, 1999; Vallejos, 2012). Such a measure is a normalized version of the cross-variogram, being a crucial instrument for multivariate spatial prediction (Ver Hoef and Barry, 1998). Rukhin and Vallejos (2008) studied the codispersion coefficient from both theoretical and applied viewpoints, and established, for arbitrary lags, the consistency and limiting distribution of the sample coefficient. Recently, Vallejos (2012) studied some extensions of the codispersion in a time series context, while Ojeda et al., (2012) used the codispersion coefficient to assess the similarity between two digital images.
The goal of the work is to use a nonparametric version of the codispersion coefficient to assess the spatial association between several pairs of forest variables. Such extensions of this nature have previously been considered in the spatial statistics literature (García-Soidán et al., 2004).

The study of the forest variables is based on a data set of Pinus radiata plantations in the south of Chile. Through the use of codispersion maps, we explore the spatial association of these variables.

2 Methods

Throughout the paper we shall consider intrinsically stationary random fields $\{X(\boldsymbol{s}),\ \boldsymbol{s} \in D \subset \mathbb{R}^d\}$ with semi-variogram defined as

$$\gamma_X(\boldsymbol{k}) = \frac{1}{2}\text{var}\{X(\boldsymbol{s}+\boldsymbol{k}) - X(\boldsymbol{s})\}, \tag{1}$$

where $\boldsymbol{k} \in \mathbb{R}^d$ denotes the spatial lag. For n sampling sites $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$, a natural and unbiased estimator based on the method of moments is the empirical semi-variogram given by

$$\widehat{\gamma}_X(\boldsymbol{k}) = \frac{1}{2|N(\boldsymbol{k})|}\sum_{N(\boldsymbol{k})}(X(\boldsymbol{s}_i) - X(\boldsymbol{s}_j))^2, \tag{2}$$

where $N(\boldsymbol{k}) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| \in T(\boldsymbol{k}), 1 \leq i, j \leq n\}$, $T(\boldsymbol{k})$ is a tolerance region around $\boldsymbol{k}$, and where $|\cdot|$ denotes cardinality of a set. García- Soidán (2007) proposed a Nadaraya-Watson type estimator for the semi-variogram defined as

$$\breve{\gamma}_{X_h}(\boldsymbol{k}) = \frac{\sum_{i=1}^n \sum_{j=1}^n K\left(\frac{\boldsymbol{k}-(\boldsymbol{s}_i-\boldsymbol{s}_j)}{h}\right)(X(\boldsymbol{s}_i) - X(\boldsymbol{s}_j))^2}{2\sum_{i=1}^n \sum_{j=1}^n K\left(\frac{\boldsymbol{k}-(\boldsymbol{s}_i-\boldsymbol{s}_j)}{h}\right)}, \tag{3}$$

where h represents a bandwidth parameter and $K : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a symmetric and strictly positive density function. For such an estimator, García-Soidán (2007) establishes, under regularity conditions, consistency and asymptotic normality, and addresses the inadequate behavior of estimator (3) near the endpoints.
Let $\{(X(\boldsymbol{s}), Y(\boldsymbol{s})) : \boldsymbol{s} \in D \subset \mathbb{R}^d\}$ be a bivariate intrinsically stationary random field on D with cross-variogram $2\gamma_{XY}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ defined through

$$2\gamma_{XY}(\boldsymbol{k}) = \mathbb{E}[(X(\boldsymbol{s}+\boldsymbol{k}) - X(\boldsymbol{s}))(Y(\boldsymbol{s}+\boldsymbol{k}) - Y(\boldsymbol{s}))], \tag{4}$$

for all $\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{k} \in D$, and with marginal variograms $2\gamma_X$, $2\gamma_Y$ as defined through Equation (1). The codispersion coefficient (Matheron, 1965) is a normalized version of (4) and defined through

$$\rho_{XY}(\boldsymbol{k}) = \frac{\gamma_{XY}(\boldsymbol{k})}{\sqrt{\gamma_X(\boldsymbol{k})\gamma_Y(\boldsymbol{k})}}.$$

Rukhin and Vallejos (2008) and Vallejos (2008) found a closed form for the coefficient $\rho_{XY}(\cdot)$ for spatial autoregressive processes under particular

assumptions on the correlation structure of the errors and when considering a rectangular lattice.
The analogue of Matheron's estimator for the cross-variogram is obtained through

$$\widehat{\gamma}_{XY}(\boldsymbol{k}) = \frac{1}{2|N(\boldsymbol{k})|} \sum_{N(\boldsymbol{k})} (X(\boldsymbol{s}_i) - X(\boldsymbol{s}_j))(Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)), \tag{5}$$

where $N(\boldsymbol{k})$ is defined as in Equation (2). The corresponding empirical estimator of the codispersion based on (5) is given by

$$\widehat{\rho}_{XY}(\boldsymbol{k}) = \frac{\widehat{\gamma}_{XY}(\boldsymbol{k})}{\sqrt{\widehat{\gamma}_X(\boldsymbol{k})\widehat{\gamma}_Y(\boldsymbol{k})}}. \tag{6}$$

The analogue of the Nadaraya-Watson type estimator for the cross-variogram is instead given by

$$\breve{\gamma}_{XY_h}(\boldsymbol{k}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{\boldsymbol{k}-(\boldsymbol{s}_i-\boldsymbol{s}_j)}{h}\right)(X(\boldsymbol{s}_i) - X(\boldsymbol{s}_j))(Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j))}{2\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{\boldsymbol{k}-(\boldsymbol{s}_i-\boldsymbol{s}_j)}{h}\right)}, \tag{7}$$

where $K(\cdot)$ is a kernel function as in Equation (3).
A kernel type estimator of the codispersion coefficient is

$$\breve{\rho}_{XY_{\boldsymbol{h}}}(\boldsymbol{k}) = \frac{\breve{\gamma}_{XY_{h_1}}(\boldsymbol{k})}{\sqrt{\breve{\gamma}_{X_{h_2}}(\boldsymbol{k})\breve{\gamma}_{Y_{h_3}}(\boldsymbol{k})}}, \tag{8}$$

where $\boldsymbol{h} = (h_1, h_2, h_3)$, $\breve{\gamma}_{XY_{h_1}}(\boldsymbol{k})$ is as in (7) and $\breve{\gamma}_{X_{h_2}}(\boldsymbol{k})$ is as in (3).
Cuevas et al., (2013) established the consistency of estimator (8). In addition, asymptotic expressions for the bias and mean square error were derived for estimator (7). A bandwidth selection rule for the variogram and the cross-variogram was also provided.

3 An Application

Here, we present an example of an issue that motivated the present work. Pinus radiata is one of the most widely planted species in Chile; it is planted on a wide array of soil types and in a variety of regional climates. Two important measures of plantation development are the dominant tree height and the basal area. Snowdon argues convincingly that both measures are correlated with regional climate and local growing conditions (Snowdon,

2001). The variogram was used to characterize the spatial dependence of each variable. However, the assessment of the spatial association between tree height, tree basal area and other regional climate variables is of great interest for the quantification of spatial dependence and the detection of those directions in which there is either high or low degree of spatial association.

In the present article, we consider the relationship among the tree height, basal area, elevation and slope of Pinus radiata plantations. The study site is located in the sector *Escuadrón*, south of Concepción in the southern portion of Chile (36° 54' S, 73° 54' O) and has an area of 1244.43 hectare. In addition to more mature stands, we were also interested in the area containing young (i.e., four year old) stands of Pinus radiate, with an average density of 1600 trees per hectare. The basal area and dominant tree height at the year of plantation establishment (1993, 1994, 1995, and 1996) were used to represent the stand attributes. The three variables were obtained from 200 m^2 circular sample plots and point-plant sample plots. For the latter type of sample, four quadrants are established around the sample point; the four closest trees in each quadrant (16 trees in total) are then selected and measured in a clockwise direction. The samples were located systematically using a mean distance of 150 meters between samples. The total number of plots available for this study was 468. In addition to the tree height and basal area, the coordinates, elevation and slope were recorded for each site.

In this talk, we will discuss the construction of a codispersion map based on Equation (8) to provide better insight into the spatial associations between all pairs of variables in several different directions on a two-dimensional space. Figure 1 shows the codispersion maps that were created from the variables of interest using a rectangular grid. We provide a full description of the findings. In particular, we find those directions for which the codispersion coefficient is maximum or minimum for the forest data.

Acknowledgments: This work was supported in part by Fondecyt, Grant 1120048, Chile.

References

Chilés, J. P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Cuevas, F., Porcu, E., Vallejos, R. (2013). Study of spatial relationships between two sets of variables: A nonparametric approach. Submitted.

García-Soidán, P., Febrero, M. and González, W. (2004). *Journal of Statistical Planning and Inference*, **121**, 65 – 92.

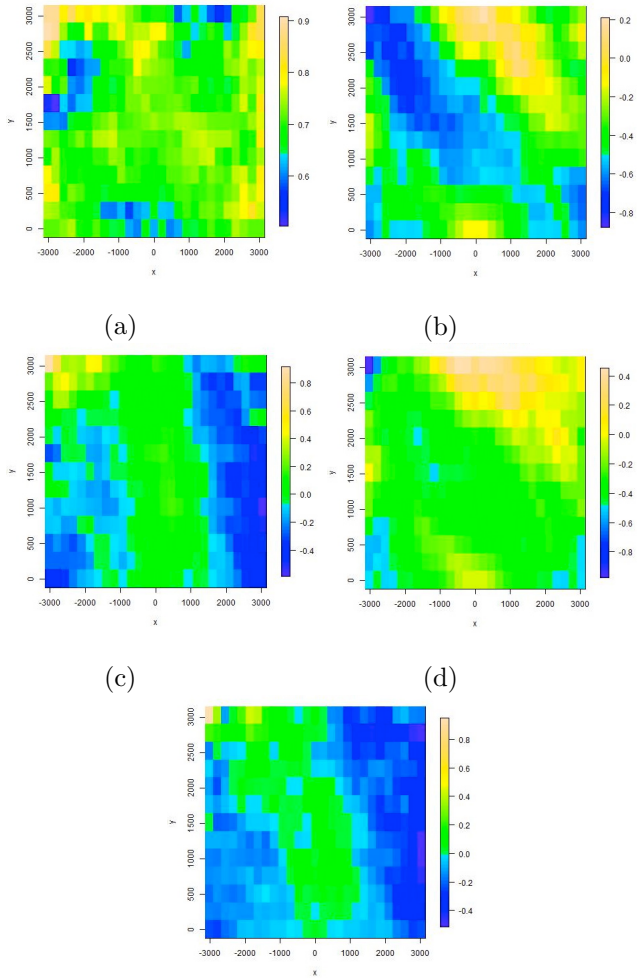


FIGURE 1. Codispersion map between all pairs of variables of interest.

García-Soidán, P. (2007). Asymptotic normality of the Nadaraya-Watson semivariogram estimators. *Test*, **16**, 479 – 503.

Matheron, P.J. (1965). *Les Variables Régionalisées et leur Estimation.* Paris: Masson.

Ojeda, S., Vallejos, R., Lamberti, P. (2012). Measure of Similarity Between Images Based on the Codispersion Coeofficient. *Journal of Electronic Imaging*, **21**, 023019.

Rukhin, A. and Vallejos, R. (2008). Codispersion coefficient for spatial and temporal series. *Statistics and Probability Letters*, **78**, 1290 – 1300.

Snowdon, P. (2001). Short-term predictions of growth of Pinus radiata with models incorporating indices of annual climatic variation. *Forest Ecology and Management*, **152** , 1-11.

Vallejos, R. (2008). Assessing the association between two spatial or temporal se quences. *Journal of Applied Statistics*, **35**, 1323 – 1343.

Vallejos, R. (2012). Testing for the absence of correlation between two spatial or temporal sequences. *Pattern Recognition Letters*, **33**, 1741 – 1748.

Ver Hoef, J. M. and Barry, R. P. (1998). Constructing and fitting models for cokriging and multivariate spatial prediction. *Journal of Statistical Planning and Inference*, **69**, 275 – 294.

Sparse Bayesian modeling of underreported count data

Michaela Dvorzak[1], Helga Wagner[2]

[1] Joanneum Research, Graz, Austria
[2] Johannes-Kepler-University, Linz, Austria

E-mail for correspondence: michaela.dvorzak@joanneum.at

Abstract: Bayesian variable selection for Poisson regression with potentially underreported counts is considered where the reporting probability itself depends on a set of covariates. Validation data on the reporting error is required to obtain parameter identification. Variable selection and parameter estimation is carried out by MCMC sampling based on auxiliary mixture sampling for Poisson and logit models. The method is applied to real data investigating the effect of reporting errors in the designation of a death cause on the estimation of cancer death rates.

Keywords: Poisson regression; underreporting; variable selection; MCMC.

1 Introduction

Any counting or register system is prone to errors in recording as not all events may be reported for various reasons. In epidemiology and public health, errors in disease classification and recording failures due to patients avoiding any diagnosis lead to incomplete register systems. Thus, it is of particular interest to account for underreporting in disease rate regression. In a Bayesian approach, we consider Poisson regression based on underreported counts where the probability of recording an event itself is related to a set of potential covariates.

2 Model specification

We assume that the total number n_i of events in category i is generated by a Poisson process with rate $\mathrm{E}_i\lambda_i$ where E_i is the amount of study time contributed by the subjects (e.g. person-years) in category i ($i = 1, \ldots, \mathrm{I}$). However, the occurrence of an event is correctly reported only with probability p_i, thus $y_i|n_i, p_i \sim \text{Binomial}(n_i, p_i)$, leading to underreported counts

$$y_i|p_i \sim \text{Poisson}(\mathrm{E}_i\lambda_i p_i),$$

with log-link $\log(\lambda_i) = \mu_\beta + \mathbf{x}_i\boldsymbol{\beta}$ for the true occurrence rate.

The reporting probability itself is related to a set of covariates $\mathbf{w}_i$ by specifying a logit model for p_i with $\log(p_i/(1-p_i)) = \mu_\alpha + \mathbf{w}_i\boldsymbol{\alpha}$, where $\mathbf{w}_i$ and $\mathbf{x}_i$ might be equal or one might be a subset of the other. Validation data containing information on the reporting error is necessary to obtain parameter identification. We therefore assume to have a small validation sample where m_i cases are diagnosed and the number of correctly classified cases c_i has a Binomial(m_i, p_i) distribution.
Bayesian variable selection is performed to identify those regressors that have a non-negligible effect and should be included in the final model. Thus, prior distributions have to be assigned to all model parameters $\mu_\beta, \boldsymbol{\beta}, \mu_\alpha$ and $\boldsymbol{\alpha}$. We use spike and slab priors as in Wagner and Duller (2012) for the parameters subject to selection with Dirac spikes for both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ where the spike is defined as a point mass at zero, $p_{\text{spike}}(\beta_i) = I_{\{0\}}(\beta_i)$ and $p_{\text{spike}}(\alpha_i) = I_{\{0\}}(\alpha_i)$, respectively. The spike and slab priors are specified hierarchically by introducing indicator variables δ_i and η_i for the elements of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ as

$$\begin{aligned} p(\beta_i|\delta_i) &= (1-\delta_i)p_{\text{spike}}(\beta_i) + \delta_i p_{\text{slab}}(\beta_i) \\ p(\alpha_i|\eta_i) &= (1-\eta_i)p_{\text{spike}}(\alpha_i) + \eta_i p_{\text{slab}}(\alpha_i) \end{aligned}$$

where $p(\delta_i|\omega_\delta) = \omega_\delta$ and $p(\eta_i|\omega_\eta) = \omega_\eta$ with hyper-priors $\omega_\delta \sim \text{Beta}(a_0, b_0)$ and $\omega_\eta \sim \text{Beta}(a_0, b_0)$. We use a normal prior for the slab, a flat but proper prior for the means μ_β and μ_α and an uninformative prior for the mixture weights ω_δ and ω_η. For the parameters in the logit model, the prior distributions are chosen appropriately in order to achieve regularization when separation is present.

3 Bayesian inference

Bayesian model selection and parameter estimation is based on MCMC sampling. Augmenting the data with the unreported occurrences

$$d_i = (n_i - y_i) \sim \text{Poisson}(\text{E}_i\lambda_i(1-p_i))$$

leads to a Gibbs sampling scheme as the complete data likelihood

$$p(\mathbf{y}, \mathbf{c}, \mathbf{d}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \prod_{i=1}^{\text{I}} (\text{E}_i\lambda_i)^{y_i+d_i} e^{-\text{E}_i\lambda_i} p_i^{y_i+c_i} (1-p_i)^{m_i-c_i+d_i}$$

is the product of a Poisson likelihood for $n_i = y_i + d_i$ with parameter $\text{E}_i\lambda_i$ and a binomial likelihood with parameters $n_i + m_i$ and p_i with $y_i + c_i$ successes and $n_i + m_i - (y_i + c_i)$ failures. Conditional on d_i, the parameters of λ_i can therefore be estimated from a Poisson model for n_i under the model $n_i \sim \text{Poisson}(\text{E}_i\lambda_i)$ and the parameters of p_i are estimated from the corresponding binomial logit model.

Posterior inference for a Poisson model as well as for a binomial logit model can be accomplished by MCMC sampling methods where the parameters $(\mu_\beta, \boldsymbol{\beta}, \boldsymbol{\delta}, \omega_\delta)$ and $(\mu_\alpha, \boldsymbol{\alpha}, \boldsymbol{\eta}, \omega_\eta)$ are sampled from their posterior distributions. Using Dirac spikes, the indicator variables $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ have to be drawn from the marginal likelihoods $p(\boldsymbol{\delta}|\mathbf{y})$ and $p(\boldsymbol{\eta}|\mathbf{y})$ integrating over the respective model parameters subject to selection. As the computation of the marginal likelihoods in each sampling iteration is computationally demanding except for normal regression models under conjugate priors, we use auxiliary mixture sampling for Poisson models as in Frühwirth-Schnatter et al. (2009) and data augmentation involving latent utilities in the binomial logit model to obtain a normal regression model in each case. Using an individual random utility model representation of the binomial logit model based on binary observations turned out to be rather time-consuming. Therefore, data augmentation in the binomial logit model is implemented as in Fussl et al. (2013) where the random utilities are aggregated for each binomial observation.
Based on the auxiliary mixture representations of the respective selection model, MCMC sampling for the model parameters from the posterior distribution involves the following steps:

(1) Sample the number of unreported cases $d_i \sim \text{Poisson}(\text{E}_i\lambda_i(1-p_i))$ and compute the total number of cases $n_i = y_i + d_i$.

(2) Variable selection and parameter estimation in the binomial logit model with $y_i + c_i$ successes and $n_i + m_i - (y_i + c_i)$ failures to estimate the reporting probability p_i.

(3) Variable selection and parameter estimation in the Poisson model for n_i to estimate the true occurrence rate λ_i.

4 Cervical cancer death rates

We consider a data set analyzed previously in Powers et al. (2010) and Whittemore and Gong (1991) containing the number of cervical cancer deaths as well as the number of woman-years at risk for different age (four age groups: 25-34, 35-44, 45-54, 55-64) and country (four countries: England, Belgium, France, Italy) categories. Additionally, validation data is available that contains information on how likely physicians from different countries are to identify and correctly report a true cervical cancer death. In a case study, a sample of physicians in each country diagnosed a specimen death certificate for a specific female patient who had died of a cervical cancer and the number of correct death certificates was recorded. Thus, validation data is available on country level but does not provide any information on the probability of reporting errors specific for age.

The MCMC method is applied to these data to investigate the effect of country-specific reporting errors on the death rates in the various countries. The model for the reporting probabilities therefore contains three covariates, whereas the estimation of the Poisson rates is based on country as well as age effects. Interaction effects of country and age are additionally included in the Poisson model to allow for deviations of the Poisson rates from the model with only country and age effects.
As all death certificates in the validation sample were correctly coded as cervical cancer in England and thus, separation occurs, the lowest age category in Belgium is used as a reference category for the cancer death rates and Belgium is the reference category for the reporting probabilities instead of England improving convergence in the logit model. MCMC was run for 12000 iterations after a burn-in of 4000 iterations and the first 2000 draws of the burn-in period were drawn from the model including all regressors without selection.

TABLE 1. Posterior means and estimated posterior inclusion probabilities $\hat{p}(\delta_i = 1|\mathbf{y})$ and $\hat{p}(\eta_i = 1|\mathbf{y})$ for the cervical cancer data (results for non-zero effects are given in bold).

		Covariate	Post. mean	Post. prob.
Poisson model	$\hat{\beta}_0$	Intercept	2.094	-
	$\hat{\beta}_1$	**England**	0.418	**1.00**
	$\hat{\beta}_2$	France	-0.010	0.10
	$\hat{\beta}_3$	**Italy**	-0.949	**1.00**
	$\hat{\beta}_4$	**35-44**	1.605	**1.00**
	$\hat{\beta}_5$	**45-54**	2.675	**1.00**
	$\hat{\beta}_6$	**55-64**	2.801	**1.00**
	$\hat{\beta}_7$	England 35-44	-0.003	0.07
	$\hat{\beta}_8$	England 45-54	0.001	0.06
	$\hat{\beta}_9$	England 55-64	-0.008	0.11
	$\hat{\beta}_{10}$	France 35-44	0.035	0.25
	$\hat{\beta}_{11}$	**France 45-54**	-0.141	**0.70**
	$\hat{\beta}_{12}$	France 55-64	-0.070	0.37
	$\hat{\beta}_{13}$	Italy 35-44	0.004	0.07
	$\hat{\beta}_{14}$	Italy 45-54	-0.002	0.07
	$\hat{\beta}_{15}$	**Italy 55-64**	0.231	**0.94**
Logit model	$\hat{\alpha}_0$	Intercept	1.862	-
	$\hat{\alpha}_1$	**England**	3.691	**1.00**
	$\hat{\alpha}_2$	**France**	-1.026	**0.93**
	$\hat{\alpha}_3$	Italy	-0.121	0.40

Table 1 reports the model selection results for the parameters of the cancer death rates as well as for the parameters of the reporting probabilities. For the Poisson rates, all main effects of age and country except for France are selected based on a posterior inclusion probability larger than 0.5. Compared to Belgium, death rates are higher in England but lower for Italy. Death risk increases for higher age groups except for France in the third age group where the interaction has a negative, non-zero effect. For Italy, the rate in age group 55-64 is even higher than estimated only by the main effects. Figure 1 gives the estimated cancer death rates when accounting for underreporting compared to the original death rates based on the raw data.

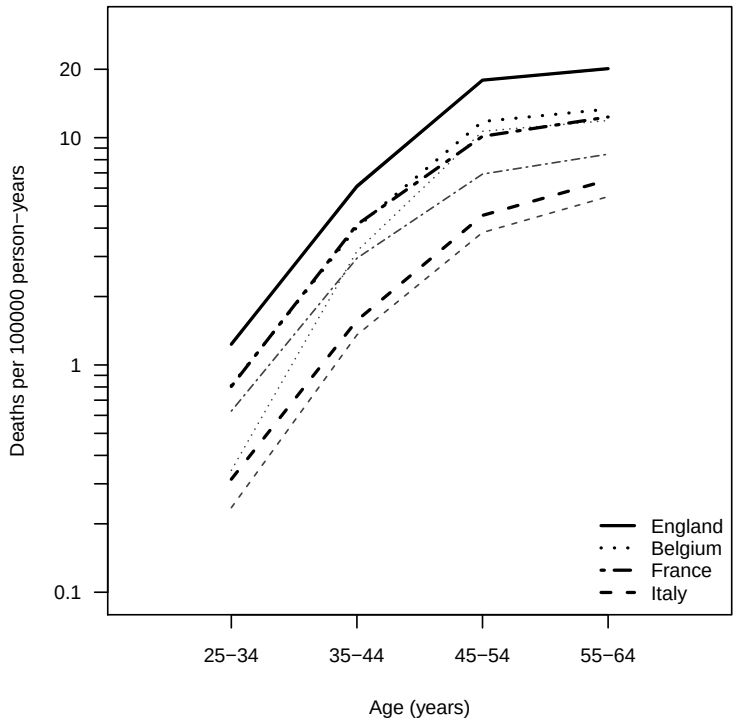


FIGURE 1. Model averaged country- and age-specific cervical cancer death rates when accounting for underreporting (bold lines) and original rates (gray lines).

The selected model for the reporting probabilities contains all country variables except for Italy. Thus, the estimated reporting probabilities differ between the countries with the highest probability obtained for England

($\hat{p}_{\text{E}} = 0.996$), followed by Belgium and Italy ($\hat{p}_{\text{B}} = 0.866$ and $\hat{p}_{\text{I}} = 0.851$) and the lowest in France ($\hat{p}_{\text{F}} = 0.698$).

5 Conclusion

Poisson regression for underreported counts is considered assuming a logit model for the reporting probability. Bayesian variable selection is performed both in the Poisson and logit model to identify those regressors that should be included in the final model. Accounting for underreporting in the Poisson model is feasible using MCMC methods if validation data on the reporting error is available. Depending on the data, the logit model can be extended by random effects to account for heterogeneity among clusters (e.g. physicians or laboratories). Moreover, the model can be modified to allow for misclassification when over- and underreporting is considered.

References

Frühwirth-Schnatter, S., Frühwirth, R., Held, R. and Rue, H. (2009). Improved Auxiliary Mixture Sampling for Hierarchical Models of Non-Gaussian Data. *Statistics and Computing*, **19**, 479 – 492.

Fussl, A., Frühwirth-Schnatter, S. and Frühwirth, R. (2013). Efficient MCMC for Binomial Logit Models. *ACM Transactions on Modeling and Computer Simulation*, **23**, 1 – 21.

Powers, S., Gerlach, R. and Stamey, J. (2010). Bayesian variable selection for Poisson regression with underreported responses. *Computational Statistics and Data Analysis*, **54**, 3289 – 3299.

Wagner, H. and Duller, C. (2012). Bayesian model selection for logistic regression models with random intercept. *Computational Statistics and Data Analysis*, **56**, 1256 – 1274.

Whittemore, A.S. and Gong, G. (1991). Poisson Regression with Misclassified Counts: Application to Cervical Cancer Mortality Rates. *Applied Statistics*, **40**, 81 – 93.

Harmonic Histograms: Smoothing of Grouped Circular Data Distributions

Paul H.C. Eilers[1]

[1] Department of Biostatistics, Erasmus University Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: p.eilers@erasmusmc.nl

Abstract: A specialized penalty for smoothing on the circle is presented. When used for smoothing a histogram, it leads to the von Mises distribution in the limit. Rounded circular data need special care. A variant of the penalized composite link model is proposed. AIC works well for selecting the amount of smoothing, unless over-dispersion is present.

Keywords: circular penalty; von Mises distribution.

1 Introduction

Penalties are useful tools for building smooth semi-parametric models. Commonly they are based on second order derivatives or differences. As a consequence a heavy penalty leads to a straight line fit. Often this is reasonable, but not for circular data. There a cosine (with the right amplitude and phase) is a desirable limit. This paper shows how to achieve that when smoothing (grouped) histograms of circular data.

Circular data occur in many places: minutes within an hour, hours within a day, days within a week and months within a year are example when we are dealing with time. In the natural sciences many examples of physical directions are known. See Fisher (1993) and the R package `circular`. A circular scale has an arbitrary origin and wherever we place it, at the left and right boundaries a curve should smoothly connect to itself.

As will be shown in the next section, it is not hard to design a penalty that gives a fitted curve the desired properties. It will be used for smoothing a high-resolution circular histogram. However, sometimes the data are not given with a high resolution and so we only have a histogram with wide bins. We can adapt the penalized composite link model (Eilers 2007, 2012) to estimate a more detailed density for such data.

The proposed smoother has as limit the von Mises distribution, the equivalent of the normal distribution on the circle.

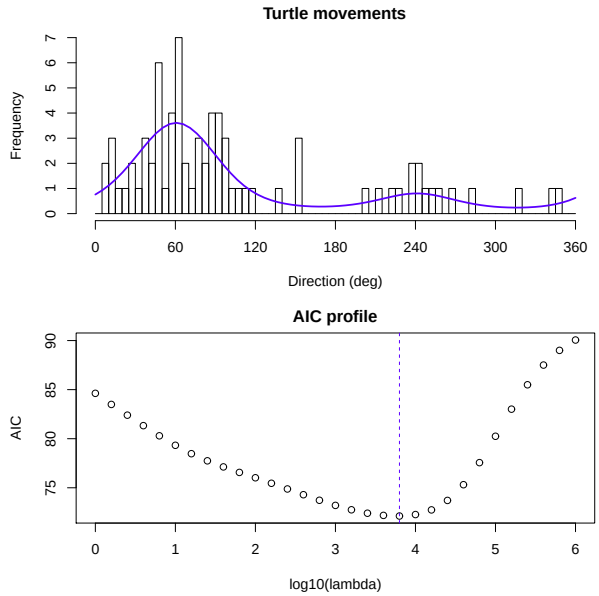


FIGURE 1. Directions of movements of turtles. Top: histogram and direct density estimate using best λ. Bottom: profile of AIC. The best value of λ (giving the lowest AIC) is marked with the vertical broken line.

2 Theory and examples

We will first consider smoothing of a non-circular histogram. An example is shown in Figure 1. It shows the directions taken by 76 turtles, see data set B.3 in Fisher (1993).

Let y_i, $i = 1, \ldots, n$ be the counts in a histogram with relatively narrow bins. We are going to fit a smooth vector μ. It is assumed that y_i is drawn from a Poisson distribution with expectation $\mu_i = \exp \eta_i$. If we did neglect the circularity, we would maximize the penalized Poisson log-likelihood

$$L = \sum_i^n (y_i \eta_i - \exp \eta_i) - \lambda \sum_i^n (\Delta^d \eta_i)^2 / 2. \tag{1}$$

Here the operator Δ^d forms d-th order differences. Common values of d are 2 or 3. The second term of L is the penalty and its influence is tuned by the parameter λ. It is convenient to introduce the matrix D such that $D\eta = \Delta^d \eta$. Then setting the derivatives of L wrt to η equal to zero, we arrive at the penalized likelihood equations

$$y - \mu = \lambda D' D \eta. \tag{2}$$

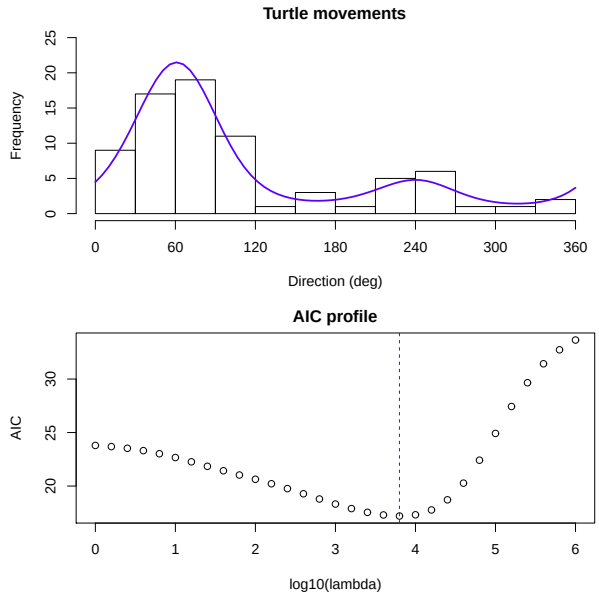


FIGURE 2. Directions of movements of turtles. Top: histogram with wide bins and PCLM density estimate using best λ. Bottom: profile of AIC. The best value of λ (giving the lowest AIC) is marked with the vertical broken line.

This is a non-linear system, because $\mu = \exp \eta$. A Taylor expansion leads to the following first order approximation, which is solved repeatedly (a tilde indicates the current approximation to the solution)

$$(\tilde{M} + \lambda D'D)\eta = y - \tilde{\mu} + \tilde{M}\tilde{\eta}, \tag{3}$$

with $M = \text{diag}(\mu)$. Good starting values are $\tilde{\eta} = \log(y + 1)$. With this choice only a handful of iterations are needed to achieve convergence.
When λ is large, it follows from (2) that $D\eta$ has to be close to zero. When $d = 2$ ($d = 3$), this is the case when η is a linear (quadratic) function of i. The polynomial limit and the fact that there is nothing that forces η to smoothly connect at left and right ends of the domain of the histograms, makes the proposed smoother unsuitable for circular data. However, that can be corrected by replacing D by the matrix Q^*, which has a structure that is illustrated by the following example

$$Q^* = \begin{pmatrix} 2\phi & -1 & 0 & 0 & -1 \\ -1 & 2\phi & -1 & 0 & 0 \\ 0 & -1 & 2\phi & -1 & 0 \\ 0 & 0 & -1 & 2\phi & -1 \\ -1 & 0 & 0 & -1 & 2\phi \end{pmatrix}. \tag{4}$$

Here $\phi = \cos(2\pi/n)$. One can show that $Q^*\eta = 0$ when $\eta = a\cos(2i\pi/n) + b\sin(2i\pi/n)$, for arbitrary values of a and b. Thus the limit of strong smoothing is periodic with period n.
Actually this penalty is not enough, because it pushes towards a (co)sine with zero mean. Generally we will need a non-zero mean. This can be solved by multiplying Q^* by a circular differencing matrix, D^* which has the pattern -1 1 in the first $n-1$ rows, with the -1 on the diagonal. It has an extra last row with 1 in the first column and -1 in the last column. With $Q = D^*Q^*$ the likelihood equations become

$$y - \mu = \lambda Q'Q\eta. \tag{5}$$

To find a suitable value of λ automatically, we use AIC, with the effective model dimension, ED, computed as

$$\text{ED} = \text{trace}[(\hat{M} + Q'Q)^{-1}\hat{M}]. \tag{6}$$

Figure 1 shows the profile of AIC for a series of values of λ and the smooth histogram obtained using the λ that minimizes AIC.
Now we turn to grouped data and the penalized composite link model. Assume that there exists a smooth density $\gamma = \exp\eta$, with length n on a fine grid (with steps of 1 degree, say). The grouping can be described by a matrix C, with m rows and n columns. Here m is the number of groups. In row i, C has a 1 in each column that contributes to group i; all other elements are zero. Then if y gives the counts in the groups, $\mu = E(y) = C\gamma$. This is an example of the composite link model of Thompson and Baker (1981). They present an algorithm to handle it as a GLM, with a working design matrix $V = M^{-1}C\Gamma$, where $M = \text{diag}(\mu)$ and $\Gamma = \text{diag}(\gamma)$. Notice that V is recomputed in each iteration of the familiar iterative weighted regression algorithm for GLMs. We want γ to be smooth and so we put a penalty, based on Q, on η. The bottom line is that we have to iteratively solve

$$(\tilde{V}'\tilde{M}\tilde{V} + Q'Q)\eta = \tilde{V}'(y - \tilde{\mu} + \tilde{M}\tilde{V}\tilde{\eta}). \tag{7}$$

To automatically determine a reasonable value for λ we use AIC as before, where now ED, the effective model dimension, is the trace of $(\hat{V}'\hat{M}\hat{V} + \lambda Q'Q)^{-1}(\hat{V}'\hat{M}\hat{V})$ after convergence.
Figure 2 shows the turtle data again, but using histogram bins with a width of 30 degrees. The proposed algorithm appears to work well.
The motivation for this works came from a data set that was kindly provided by Rosa Crujeiras. It consists of 580 azimuths of cross-beds in the Kamthi river; see Oliveira et al. (2012). The data as given are rounded to multiples of 20 degrees. If we form a histogram with bins of 5 degrees, we get a repeating pattern of three empty and one non-empty bins. AIC or any other procedure for finding the right of smoothing would get fooled. In fact this is visible for cross-validation in Figure 2 of the mentioned paper.

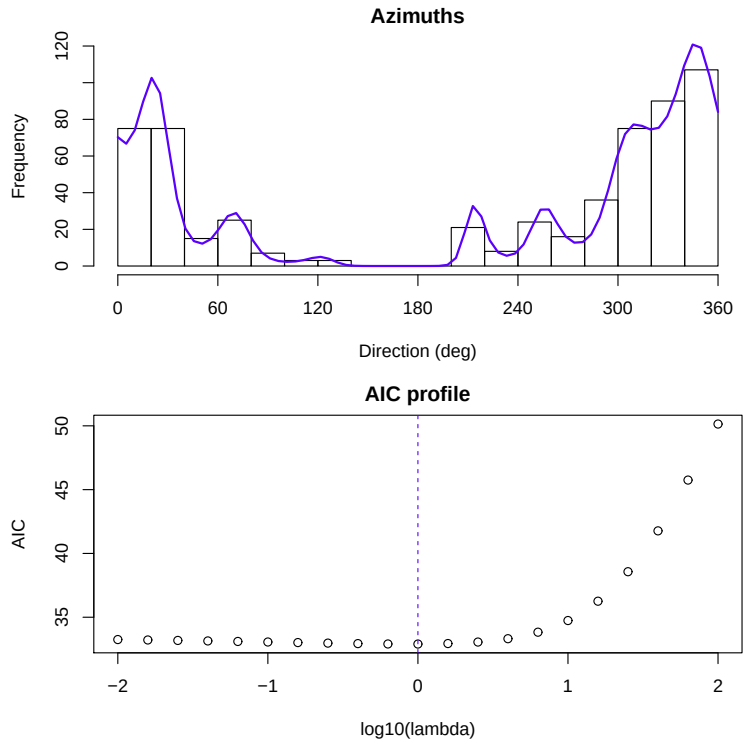


FIGURE 3. Azimuth data. Top: histogram with wide bins and PCLM density estimate using best λ. Bottom: profile of AIC. The best value (giving the lowest AIC) is marked with the vertical broken line.

Working with bins of 20 degrees and the composite link model solves this issue, but only partially. Apparently the data do not agree with a very smooth density and Poisson distributions. As Figure 3 shows, AIC indicates rather light smoothing. Actually, direct visual inspection already points in this direction. The third bin from the left contains a very low number of counts, compared to its neighbors. It is interesting to speculate about the source of the over-dispersion. A possible explanation might be digit preference in the raw data, before rounding took place.

3 Discussion

Smoothing of circular data calls for a "designer penalty", which does not have a polynomial as its limit but a cosine function with proper amplitude and phase. This can easily be achieved by replacing the usual second order differences by a slightly modified contrast. This penalty can be easily combined with a (Poisson) generalized linear model to build a circular histogram smoother. When combined with the composite link model it works well on grouped data too.
An underlying assumption is that we really are dealing with Poisson data, generated by smoothly varying expected values. When this is the case, AIC works well to tune the amount of smoothing. But as the last example made clear, it goes astray when there is over-dispersion. It will be interesting to study what can be achieved with quasi-likelihood here (Eilers at al., 2008).

References

Eilers, P.H.C. (2007) Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling* **7**, 239–254.

Eilers, P.H.C. (2012) Composite link, the neglected model. *Proceedings of the 27th International Workshop on Statistical Modelling*. Prague.

Eilers, P.H.C, Gampe J., Marx, B.D, Rau R. (2008) Modulation models for seasonal time series and incidence tables. *Statistics in Medicine* **27**, 3430-3441.

Fisher, N.I. (1993). *Statistical analysis of circular data*. Cambridge University Press.

Oliveira, M., Crujeiras, R.M., Rodriguez-Casal, A. (2012) A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics and Data Analysis* **56**, 3898–3908.

Thompson, R. and Baker, R.J. (1981). Composite link functions in generalized linear models. *Applied Statistics* **30,** 125–131.

Modeling confidential data via modified hurdle mixed models

Marco Enea, Antonella Plaia, Vincenza Capursi

[1] Dipartimento di Scienze Economiche, Aziendali e Statistiche - University of Palermo, Italy

E-mail for correspondence: marco.enea@unipa.it

Abstract: In this work we analyze event count data that show zero inflation with non-ignorable missingness due to confidentiality. The emphasis is not on imputation methods but on the choice of a suitable model. As a motivating example, we analyze a dataset on University student's mobility (SM) in Italy whose records are reported aggregately. In such data, records with less than three moving students are automatically removed. To detect the determinants of SM, similarly to a hurdle mixed model, we estimate two separate models, a binomial mixed model for the "zero" part and a two-truncated negative mixed binomial for the "non-zero" part.

Keywords: GLMM; GAMLSS; truncation; students' mobility.

1 Introduction

The analysis of event count data showing over/under-dispersion and/or too many zero counts has become very common in the literature. Popular modeling approaches are the zero-inflated (Lambert, 1992) or zero-altered (also said hurdle) Poisson regression models (Mullahy, 1986), or their negative binomial counterpart if further overdispersion is present. Additionally, the family of count data models needs adjustments in order to accommodate for non-ignorable missingness. In fact, it is common to deal with official statistics data whose feature is the limited disclosure due to privacy and confidentiality reasons. Masking is the most common technique to protect data. In Italy a common masking procedure is to fix a data truncation threshold at two, that is tables with less than three cell counts cannot be released. In this case, the fitting of the zero-inflated model or the hurdle model can be misleading. In fact, estimates could be biased if the amount and the mechanism of missingness is non-ignorable.
In this work, in order to reduce the amount of bias, we suggest to estimate two models separately, a binomial model for the zero part and a two-truncated negative binomial model for the non-zero part.

As a motivating dataset having the above outlined characteristics, we analyze the 2010/2011 cohort of the Italian freshmen in 58 Italian Universities to detect the determinants of SM within the Italian territory. The data come from the Italian university students' register (ANSU is the Italian acronym). Such data are provided in aggregate form and some information on students' characteristics is available, such as the region of origin and destination. Here the focus is on the moving freshman defined as a student who has moved from a region to another to enroll at the University. Unlikely from other European countries, mobility rates of students in Italy are intrinsically very low. From the dataset, about 173000 resident and about 19000 non-resident students are recorded, but their true number is greater because of the masking mechanism (about 240000 freshmen also including foreign students and private universities). Three regions, Sardegna, Valle d'Aosta and Trentino Alto Adige, were not considered as regions of destination. Two universities are present in Sardegna but these have no observed counts of incoming students greater than 2. The other two regions are so small that not all the study fields are represented. Thus they have been aggregated with their neighboring larger regions Piemonte and Veneto, respectively. There were no observed non-resident students in six Universities i.e. Benevento, Palermo, NapoliSeconda, NapoliParthenope, Catania and Catanzaro. Table 1 describes the covariates used in the analysis.

TABLE 1. Description of the covariates. The first category is the baseline

Variable	Categories/Range	Description
RRes	Abruzzo,...,Veneto	Origin region (20 cat.)
RUni	Abruzzo,...,Veneto	Destination region (17 cat.)
University	Bari,...,Verona	University (58 cat.)
Sex	Female, Male	Gender
HSgrade	100-; 90-99; 80-89; 70-79; 60-69;	High school final grade
HStype	liceo; magistrale; vocational; technical;	High school type
Field	Health6; Health3; Social5; Social5; Scientific5; Scientific3; Humanistic3;	Study field and course duration (in years)
INC.res	-4.5, 4.1	Per-capita average income (centered) of the origin region, in thousands of euros
NU.res	1 - 7	No. of Universities in the region of residence
Censis	-18.3, 16.9	University "quality" score (centered) according to CENSIS-Repubblica

The records are "depicted" by the covariates described above. For each record the number of non-resident students is observed. The count distri-

bution is reported in Figure 1. For graphical reasons, the pick of zeroes representing the resident students is not shown. The skewness of the distribution suggests a non-ignorable missingness for counts one and two.

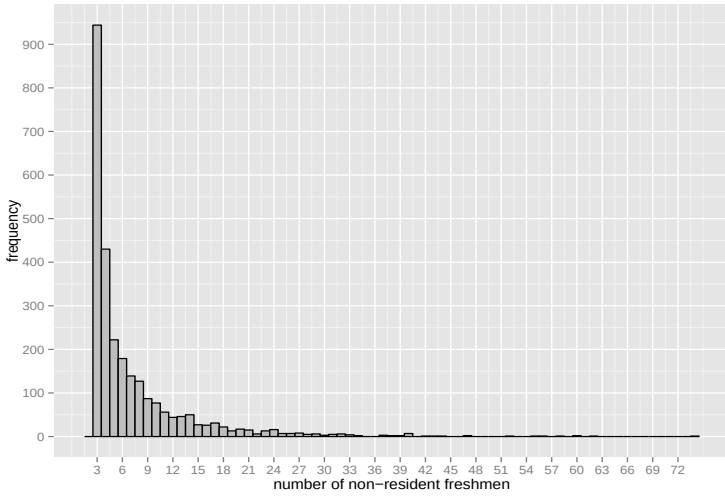


FIGURE 1. Observed non-resident freshmen records.

2 Hurdle mixed model

We first introduce the hurdle negative binomial mixed model and then we introduce the modification we suggest to accommodate for non-ignorable missingness. The choice of the negative binomial rather than the Poisson distribution is due to the likely presence of overdispersion into the non-zero part of the model. Let $Y'_{ijk} \sim NB(\mu_{ijk}, \sigma)$ with type I parameterization as in GAMLSS (Rigby and Stasinopoulos, 2005), with i ($i = 1, ..., n_j$) indexing students grouped by covariate profiles, j ($j = 1, ..., m_k$) the universities and k ($k = 1, ..., K$) the regions, respectively. Under the hurdle model, a new variable Y_{ijk} is introduced whose distribution is:

$$P(Y_{ijk} = 0) = 1 - \pi_{ijk}, \tag{1}$$

$$P(Y_{ijk} = y | y > 0) = \frac{\pi_{ijk} P(Y'_{ijk} = y)}{P(Y'_{ijk} > 0)}, \tag{2}$$

where π_{ijk} is the probability to be non-resident and $P(Y'_{ijk} = y)/P(Y'_{ijk} > 0)$ is the truncated-at-zero NB. In practice Y_{ijk} is zero-altered negative binomial (ZANB) distributed. The hurdle mixed regression model is a two-equation system:

$$\begin{aligned}\text{logit}(\pi_{ijk}) = {} & \beta_{10} + \text{Sex}_{ijk}\beta_{11} + \text{Field}_{jk}^{\text{T}}\boldsymbol{\beta}_{12} + \text{HStype}_{ijk}^{\text{T}}\boldsymbol{\beta}_{13} + \\ & \text{HSgrade}_{ijk}^{\text{T}}\boldsymbol{\beta}_{14} + \text{RRuni}_{k}^{\text{T}}\boldsymbol{\beta}_{15} + \text{INC.res}_{k}\beta_{16} + \\ & \text{NU.res}_{ijk}\beta_{17} + \text{Censis}_{jk}\beta_{18} + b_{1j}, \end{aligned} \tag{3}$$

$$\begin{aligned}\log(\mu_{ijk}) = {} & \log(n_{ijk}) + \beta_{20} + \text{Sex}_{ijk}\beta_{21} + \text{Field}_{jk}^{\text{T}}\boldsymbol{\beta}_{22} + \text{HStype}_{ijk}^{\text{T}}\boldsymbol{\beta}_{23} + \\ & \text{HSgrade}_{ijk}^{\text{T}}\boldsymbol{\beta}_{24} + \text{RRes}_{ijk}^{\text{T}}\boldsymbol{\beta}_{25} + \text{RRuni}_{k}^{\text{T}}\boldsymbol{\beta}_{26} + \text{INC.res}_{k}\beta_{27} + \\ & \text{NU.res}_{ijk}\beta_{28} + \text{Censis}_{jk}\beta_{29} + b_{2j}, \end{aligned} \tag{4}$$

where b_{1j} and b_{2j} are the random effects of the universities. It is assumed that $b_{1j} \sim N(0, \sigma^2_{b_1})$, $b_{2j} \sim N(0, \sigma^2_{b_2})$. Equation (3) uses a binomial mixed model for the zero part, whereas a zero-truncated negative binomial mixed model is employed in (4) to model the non-zero part. The offset $\log(n_{ijk})$ implies we are modeling the ratio between the number of incoming students and the number n_{ijk} of "local" students in the kth region. Estimation in (3) and (4) is usually performed via maximum likelihood. For non-longitudinal data, b_{1j} and b_{2j} are independent and the log-likelihood decomposes into the sum of the log-likelihood pertaining to the binomial part and the log-likelihood of the truncated negative binomial part. Thus both log-likelihoods can be maximized separately (Molas and Lasaffre, 2010). However, maximizing the log-likelihood under the zero-truncated negative binomial using the SM data can be misleading, since the missing data appear to be nonignorable. Since missing counts belong to the non-zero part (2), this leads to underestimate π_{ijk} and, in addition, the truncated-at-zero negative binomial does not fit well. To reduce the bias, at least for the non-zero part, we fit (4) using a two-truncated negative binomial.
Of course, that implies there is not an underlying "true" probability distribution of Y_{ijk}, and two analyses must be carried out separately, accordingly. On the other hand, the advantage is that the relationship (4) between the linear predictor and the expected value of the untruncated NB does not change. The expected value of the two-truncated distribution is $E(Y'_{ijk} = y | y > 2) = \mu_{ijk} + \mu_{ijk}(1 + 2\sigma)\frac{P(Y'_{ijk}=y|\mu_{ijk},\sigma)}{1-P(Y'_{ijk}\leq 2|\mu_{ijk},\sigma)}$. We adopt a GAMLSS framework to estimate (4).

3 Application to the SM dataset

Tables 2 and 3 show the fixed-effects estimates from the binomial mixed model and the two-truncated negative binomial mixed model, respectively. Due to lack of space we prefer to report results regarding just the fixed effects. Table 2 shows that the odds ratios of the non-resident students as opposed to the resident ones are larger in the wealthier North-Central regions. These odds ratios become larger for regions of origin with a smaller

number of universities, and " more" larger if they are considered "prestigious" (Censis). In general non-resident students have a high school diploma (liceo) while is not statistically significative the high school grade. They mainly attend five/six-year degree courses in the social and health fields. These results seem to be consistent with the literature (Bruno and Genovese, 2010).

TABLE 2. Fixed effects from the binomial mixed model.

	Estimate	Std. Error	Pr(>|z|)
Intercept	-1.71	0.79	0.03
SexMale	-0.32	0.08	0.00
HSgrade			
60_69	0.03	0.14	0.85
70_79	0.23	0.14	0.10
80_89	0.15	0.14	0.29
90_99	-0.26	0.16	0.10
HStype			
Magistrale	-1.08	0.14	0.00
Vocational	-1.29	0.19	0.00
Technical	-0.78	0.10	0.00
Field			
Health3	-0.62	0.19	0.00
Scientific5	-1.47	0.30	0.00
Scientific3	0.03	0.16	0.86
Social5	1.04	0.19	0.00
Social3	-0.12	0.16	0.44
Humanistic3	-0.18	0.17	0.28
RRuni			
BASILICATA	-3.11	1.56	0.05
CALABRIA	-6.99	1.17	0.00
CAMPANIA	-6.51	1.05	0.00
EMILIA	5.01	1.04	0.00
FRIULI	4.29	1.25	0.00
LAZIO	1.68	0.97	0.09
LIGURIA	2.47	1.57	0.12
LOMBARDIA	3.85	0.94	0.00
MARCHE	0.92	1.06	0.39
MOLISE	-1.46	1.59	0.36
PIEMONTE	1.86	1.11	0.10
PUGLIA	-5.10	1.05	0.00
SICILIA	-5.10	1.05	0.00
TOSCANA	1.78	1.10	0.11
UMBRIA	0.28	1.56	0.86
VENETO	3.21	1.01	0.00
INC.res	-1.18	0.01	0.00
NU.res	-0.29	0.01	0.00
Censis	0.11	0.04	0.00

The two-truncated negative binomial mixed model in Table 3 is fitted on data including only the non-resident students. From the estimates it results that, after Sicily, the region of destination Abruzzo (at the baseline) is the one with the higher average ratio between non-resident and resident students. However one should consider that, in this model, Sicily is represented only by the University of Messina, which takes students almost exclusively from the neighboring region Calabria, since the universities of Palermo and Catania have not non-resident students. In conclusion, the

proposed approach seems to work well in presence of non-ignorable missing data via modeling a truncated distribution. Further developments of this approach may lead to construct imputation procedures which take into account the distribution following an EM-like algorithm.

TABLE 3. Estimates from the two-truncated negative binomial mixed model.

	Est.	S.E.	P(>|t|)		Est.	S.E.	P(>|t|)
Intercept	-2.42	0.33	0.00	RRuni			
SexMale	-0.01	0.05	0.77	BASILICATA	-0.65	0.38	0.09
HSgrade				CALABRIA	-0.97	0.50	0.05
60_69	-0.24	0.09	0.01	CAMPANIA	-2.20	0.24	0.00
70_79	-0.25	0.08	0.00	EMILIA	-1.40	0.10	0.00
80_89	-0.24	0.08	0.00	FRIULI	-0.85	0.17	0.00
90_99	-0.18	0.09	0.05	LAZIO	-1.25	0.09	0.00
INC.res	-0.38	0.14	0.01	LIGURIA	-2.24	0.21	0.00
Censis	0.05	0.00	0.00	LOMBARDIA	-2.54	0.12	0.00
RRes				MARCHE	-1.18	0.16	0.00
BASILICATA	-1.08	0.18	0.00	MOLISE	0.36	0.30	0.23
CALABRIA	-0.64	0.27	0.02	PIEMONTE	-1.60	0.16	0.00
CAMPANIA	-0.67	0.36	0.07	PUGLIA	-0.70	0.19	0.00
EMILIA	2.90	0.88	0.00	SICILIA	1.00	0.16	0.00
FRIULI	2.08	0.79	0.01	TOSCANA	-1.56	0.11	0.00
LAZIO	1.81	0.58	0.00	UMBRIA	-1.76	0.16	0.00
LIGURIA	1.57	0.76	0.04	VENETO	-1.57	0.14	0.00
LOMBARDIA	2.34	0.74	0.00	HStype			
MARCHE	1.66	0.49	0.00	Magistrale	0.40	0.10	0.00
MOLISE	-0.77	0.17	0.00	Vocational	0.52	0.18	0.00
PIEMONTE	3.15	0.71	0.00	Technical	0.01	0.06	0.83
PUGLIA	-0.15	0.26	0.57	Field			
SARDEGNA	-1.43	0.26	0.00	Health3	0.23	0.12	0.05
SICILIA	-0.56	0.28	0.05	Scientific5	-0.50	0.31	0.10
TOSCANA	1.32	0.67	0.05	Scientific3	-0.43	0.08	0.00
TRENTINO	1.04	0.85	0.22	Social5	-0.27	0.12	0.02
UMBRIA	-0.08	0.48	0.87	Social3	-0.42	0.09	0.00
VDAOSTA	1.73	0.89	0.05	Humanistic3	-0.30	0.10	0.00
VENETO	2.58	0.64	0.00				

References

Bruno, G., Genovese, A. (2010). A spatial interaction model for the representation of the mobility of University students on the Italian territory. *Netw Spat Econ*, DOI 10.1007/s11067-010-9142-7

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1 – 14.

Molas, M., Lesaffre, E. (2010). Hurdle models for multilevel zero-inflated data via h-likelihood. *Statistics in medicine*, **29**, 3294 – 3310.

Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341 – 365.

Rigby, R.A., Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*,**54(3)**, 507 – 554.

A model-based clustering approach for the analysis of environmental time series

Francesco Finazzi[1], Claire Miller[2], Marian Scott[2]

[1] Dept. of Engineering, University of Bergamo, Italy
[2] School of Mathematics and Statistics, University of Glasgow, UK

E-mail for correspondence: francesco.finazzi@unibg.it

Abstract: This paper describes the application of a novel model-based clustering approach to analyse the within lake variability of physical and water quality parameters observed through remote sensing. The clustering approach enables spatially located time series to be grouped with respect to their temporal coherence and thus to identify homogeneous areas of the lake surface that behave in a similar way over time. As a case study, the analysis of the lake water surface temperature of Lake Victoria (Africa) is presented.

Keywords: Temporal coherence; Model-based clustering; Water surface temperature; Expectation Maximization;

1 Introduction

Lakes are sensitive to large-scale environmental pressures, such as regional weather patterns, and so can frequently show temporal coherence (Maberly & Elliott, 2012), defined as the degree to which lake time series' behave similarly through time. Understanding the spatial extent of coherence, both within and between lakes, for different lake characteristics is a valuable tool to extrapolate from measured to unmeasured lakes.
The five year research programme GloboLakes (Global Observatory of Lake Responses to Environmental Change, www.globolakes.ac.uk) investigates the state of lakes and their response to climatic and other environmental drivers of change at a global scale through processing of remotely sensed ecological and lake temperature data, and supported by linked auxiliary data on catchment land-use and meteorological forcing. The aim of this paper is to propose a methodology for modelling the pixel variability within lakes in order to identify homogeneous surface areas with respect to temporal coherence. Lake surface water temperature (LSWT) will be the focus of this paper. However, the methodology will be extended to water quality parameters such as chlorophyll$_a$ and phycocyanin, as part of this programme of research.

2 Case study - Data

The lake surface water temperature (LSWT) for Lake Victoria (Africa, 1S 33E) is considered in this paper. Lake Victoria has an approximately rectangular shape of 250×337 km and it has a surface of $68,800$ km^2. Data from the ARC-Lake project (`http://www.geos.ed.ac.uk/arclake/`) are considered and in particular the data product PLOBS9D_TS024SR which, for each lake, provides spatially-resolved (0.05 grid) 20-year long time series of the LSWT observed twice-monthly. The dataset related to Lake Victoria consists of 2210 time series (one for each grid pixel) and each time series is characterized by 479 timepoints (there are however missing data). The lake is large enough to exhibit variability in the spatial pattern of its LSWT and the aim of this analysis is to understand if the lake surface can be classified into homogeneous areas which differ in terms of their temporal coherence.

3 Model-based clustering

Let $\mathbf{y}(t)$ be the $N \times 1$ observation vector at time $t = 1, ..., T$, with N the number of pixels at which the LSWT is observed and T the total number of time steps. In order to cluster the N time series with respect to their temporal coherence, the following state-space model is considered

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{K}\mathbf{z}(t) + \varepsilon(t) \\ \mathbf{z}(t) &= \mathbf{G}\mathbf{z}(t-1) + \eta(t) \end{aligned} \tag{1}$$

where $\mathbf{z}(t)$ is the $p \times 1$ latent state vector, $\mathbf{K}$ is a $N \times p$ matrix of coefficients, $\varepsilon(t) \sim N\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_N\right)$ is the $N \times 1$ measurement error vector, $\mathbf{G}$ is a $p \times p$ stable transition matrix and $\eta(t) \sim N(\mathbf{0}, \Sigma_\eta)$ is the $p \times 1$ innovation vector. Assuming $\mathbf{z}(0) \sim N(\mu_0, \Sigma_0)$ with Σ_0 a known variance-covariance matrix, the model parameter set is $\Psi = \left\{\mathbf{K}, \sigma_\varepsilon^2, \mathbf{G}, \Sigma_\eta, \mu_0\right\}$. In general, if restrictions are not imposed on Ψ, model (1) is not identifiable. Since the aim is to use model (1) for clustering, the coefficients of the matrix $\mathbf{K}$ are restricted to be only 0 or 1, with the additional constraint that each row $\mathbf{k}_i$ of $\mathbf{K}$, $i = 1, ..., N$ must contain exactly a single 1. These restrictions ensure model identifiability and allow the cluster membership to be determined directly from $\mathbf{K}$. In particular, the i-th time series belongs to the j-th cluster, $j = 1, ..., p$, if the j-th element of $\mathbf{k}_i$ is equal to 1.
The model parameter set Ψ is estimated using maximum likelihood by means of a modified version of the Expectation Maximization (EM) algorithm which is able to provide an estimated matrix $\hat{\mathbf{K}}$ subject to the above mentioned constraints (Finazzi et al., 2013). Since the EM algorithm is not guaranteed to converge to a global maximum of the likelihood function, the model is estimated M times randomizing the starting values of the model parameters each time. The estimation related to the highest

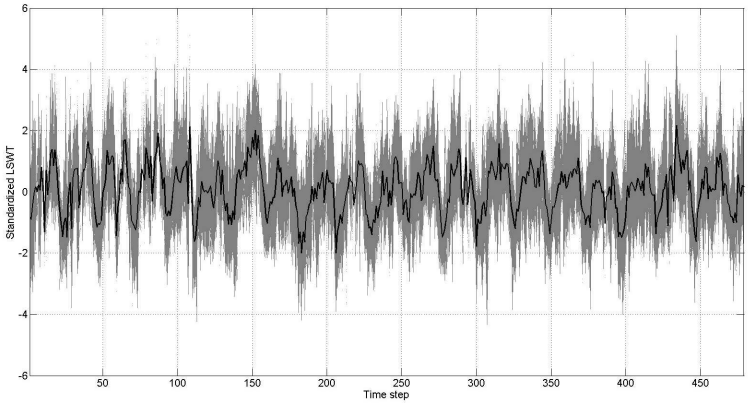


FIGURE 1. Lake Victoria standardized LSWT time series (grey line) and average time series (black line)

observed-data likelihood is retained. In particular, the elements of the matrix $\mathbf{K}$ are randomly generated from the uniform distribution $U(0,1)$ and they are normalized so that each row of $\mathbf{K}$ sums to one.
Model (1) is estimated starting from $p = 1$ (one cluster) and the number of clusters is incremented progressively until a stopping criterion is satisfied. In this work, the number of clusters is increased by one until an empty cluster is identified, that is, the matrix $\hat{\mathbf{K}}$ contains a column of zeros for cluster p. If the best solution (with respect to the smallest change in observe-data log-likelihood) includes an empty cluster, then the optimum number of clusters is given by $p - 1$.

4 Case study - Results

Due to the high percentage of missing data in some series, any spatial pattern in LSWT is not identifiable by simply looking at the spatial maps of the observed data and the temporal variability complicates the task. Moreover, the average standard deviation of the LSWT at each observation time is only 0.59 K, namely the difference in temperature between any two pixels is very low. Figure 1 shows the 2210 time series, each time series standardized with respect to its own average and standard deviation. Note that the time series exhibit a seasonal pattern but, due to the small seasonal signal in LSWT for lakes near the equator, the pattern is not very regular. The estimation procedure described in Section 3 is applied in order to estimate both the number of clusters and the cluster membership of each time series. Model (1) is a special/particular case of the more general space-time model at the basis of D-STEM (Finazzi and Fassò, 2012) and hence the D-STEM software available at `http://code.google.com/p/d-stem/`

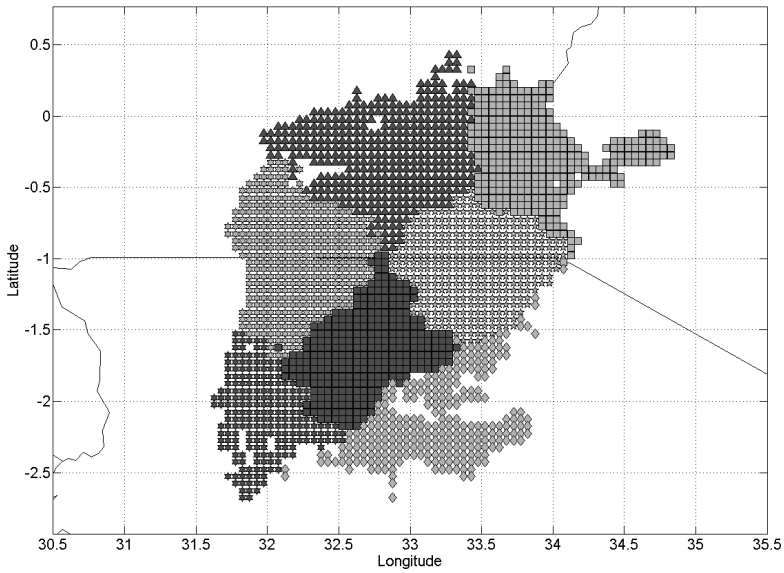


FIGURE 2. Lake Victoria LSWT clustered with respect to temporal coherence

has been used to estimate the model parameters.
Table 1 reports, for each value of p, the observed data log-likelihood and the number of empty clusters, from which it can be noted that the optimum number of clusters is 7.
The clustering result for $p = 7$ is reported in Figure 2 in terms of the spatial location of each time series. Note that the pixels (directly related to the time series) are clearly clustered in geographic space although model (1) does not include any information about the relative spatial location of the pixels. Finally, Figure 3 shows the smoothed average time series related to the seven clusters. Looking at the mean curves for each cluster, the differences are in terms of timing and amplitude of the seasonal peaks, this can vary year to year and so future work will also involve clustering year by year and following the temporal dynamic of the clustering spatial pattern.

TABLE 1. Observed data log-likelihood vs number of clusters

Clusters (p)	1	2	3	4
log-likelihood	6,398	33,886	46,065	66,969
Empty clusters	0	0	0	0
Clusters (p)	5	6	7	8
log-likelihood	80,271	88,517	96,032	96,094
Empty clusters	0	0	0	1

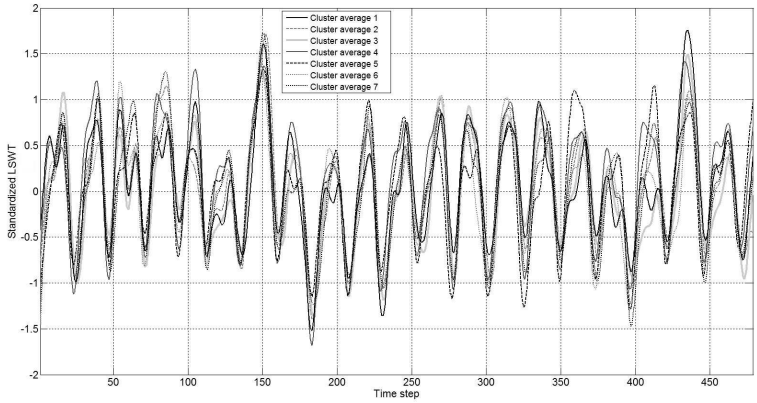


FIGURE 3. Smoothed cluster mean time series

5 Conclusions

The model-based clustering approach considered in this paper has been shown to be a valid statistical tool for the analysis of the within lake temporal coherence of time series for Lake Victoria even in the presence of missing data.
Further studies will now be conducted to explore the nature of the within lake coherence. Within the GloboLakes programme, the clustering approach will be used to study water quality parameters of more than $1,000$ lakes at a global scale.

References

Finazzi F. and Fassò A. (2012). D-STEM - A statistical software for multivariate space-time environmental data modeling. In: *Proceedings of the International Workshop on Spatio-Temporal Modelling (METMA VI)*, Guimaraes.

Finazzi F., Scott M., Miller C. and Fassò A. (2013). In the preparation paper: *A model-based clustering approach for the study of the temporal coherence of multivariate time series.*

Maberly S.C. and Elliott J.A. (2012). Insights from long-term studies in the Windermere catchment: external stressors, internal interactions and the structure and function of lake ecosystems. *Freshwater Biology*, **57**, 233 – 243

L-surface and V-valley for multi-dimensional smoothing parameter selection

Gianluca Frasso[1], Paul H.C. Eilers[2]

[1] Institut des sciences humaines et sociales, Univ. de Liége, Belgium.
[2] Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands.

E-mail for correspondence: gianluca.frasso@unina.it

Abstract: The L-surface is an attractive two-dimensional generalization of the L-curve procedure for optimal smoothing. However, one has to locate the point of maximal curvature of the surface, which is not easy. We introduce a simplified procedure that replaces maximum curvature by a distance to be minimized.

Keywords: P-splines, L-curve, tensor products, penalties

1 Introduction

The L-curve criterion was proposed by Hansen (1992) and Hansen and O'Leary (1993) for selection of the regularization parameter in ill-conditioned inverse problems. Frasso and Eilers (2012) have shown that it is an efficient and robust method to select the penalty parameter in smoothing applications. The L-curve does not require the computation of the effective model dimension and is insensitive to correlated noise.
We propose a generalization, the L-surface, for non-isotropic two-dimensional smoothing. It is a plot of the logarithm of the residual sum of squares against the logarithms of the sizes of the penalty terms, parameterized by the smoothing parameters. It appears as a surface with a nook and, in analogy with the L-curve, the optimal pair of smoothing parameters is located in the deepest point of the nook. This is the point of maximum Gaussian curvature, as defined in differential geometry. Unfortunately, locating this point is a non-trivial computational task.
On the other hand, in the nook changes of parameters lead to small displacements on the L-surface. We compute the surface for a grid of values of the two smoothing parameters. Over this grid we search for the points on the surface that are closest to each other. If we do this for the the L-curve, in the case of one-dimensional smoothing, a curve with a U-shape is obtained. Shahrak et al. (2013) already coined the name U-curve, although with a different definition. To avoid confusion, we christened our procedure

the V-valley. We describe it for two dimensions, but it works as well in one too.

2 The L-surface

Consider 2D non-isotropic smoothing with tensor product P-splines (see Eilers and Marx, 2003 and Eilers et al., 2006). We have data vectors x, y and z, where the latter is the dependent variable. Define the following quantities:

$$\psi = \log(\|z - B\beta\|^2);\ \phi_x = \log(\|P_x\beta\|^2);\ \phi_y = \log(\|P_y\beta\|^2),$$

B is the tensor product basis, based on x and y and β the vector of spline coefficient; P_x and P_y are penalty matrices, working on the coefficients in x and y direction. Although not shown explicitly here, these quantities are parameterized by λ_x and λ_y (which get chosen value on a 2D grid). Plotting these quantities in a 3D Cartesian system we obtain points on a surface with a nook (see figure 1, second panel). The profiles of the surface represent L-curves. Each L-curve shows a corner region and all curves together define a non-regular grid. The points are closest in the corner (upper right panel of figure 1). Plotting the distances between these adjacent points, a V-shaped curve is obtained (lower right panel of figure 1). All V-curves combined define the "V-valley" shown in the lower right panel of figure 1. The optimal pair (λ_x, λ_y) is located at the bottom of the valley. So we search for the minimum of

$$D(\lambda_x, \lambda_y) = \sqrt{(\nabla_\lambda \psi)^2 + (\nabla_\lambda \phi_x)^2 + (\nabla_\lambda \phi_y)^2}. \quad (1)$$

3 A real data example

We use the well known ethanol data as an example for the application of the proposed procedure. It contains 88 observations of three variables: concentration of nitrogen oxides (NOx, the dependent variable), compression ratio of the engine and equivalence ratio (C and E).
Figure 2 shows results for the V-valley and for cross-validation. The cross-validation surface does not show a clear minimum and suggests small λs leading to a wiggly surface. We suspect that this caused by serial correlation of the errors, along the E direction. The V-valley shows a clear minimum and leads to a smooth surface with intuitive appeal.

4 Discussion

We have shown how the concept of the L-curve can be extended to two-dimensional smoothing with tensor product P-splines. The L-surface has

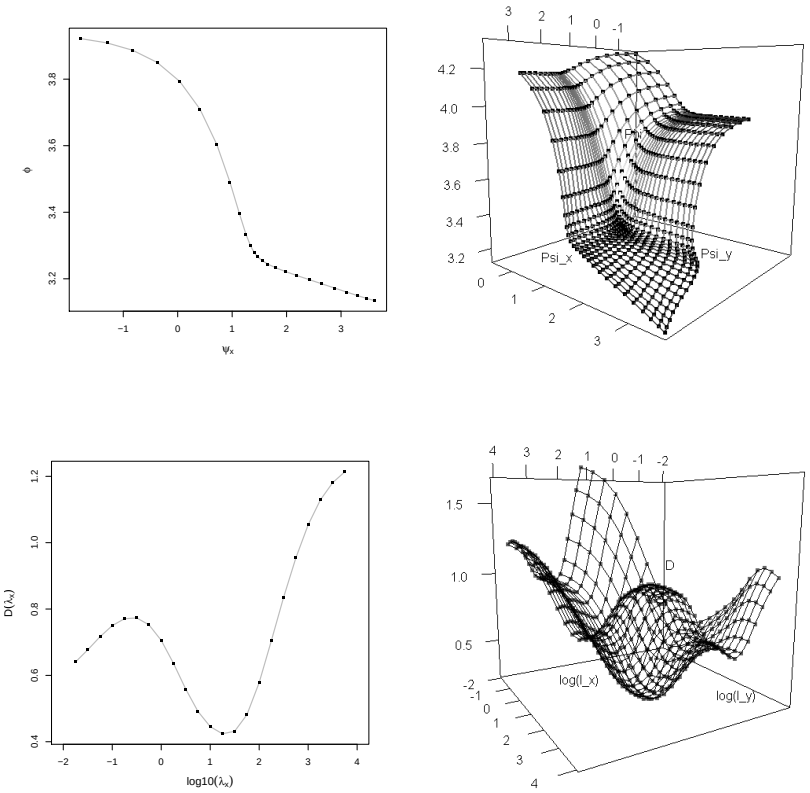


FIGURE 1. The top right panel shows the L-surface and the bottom left right panel shows the surface defining the function $D(\lambda_x, \lambda_y)$ computed on it. The left panels show the L and U curves obtained taking $\log_{10}(\lambda_2) = -2$ and varying $\log_{10}(\lambda_1)$ over the surfaces in the right part of the figure.

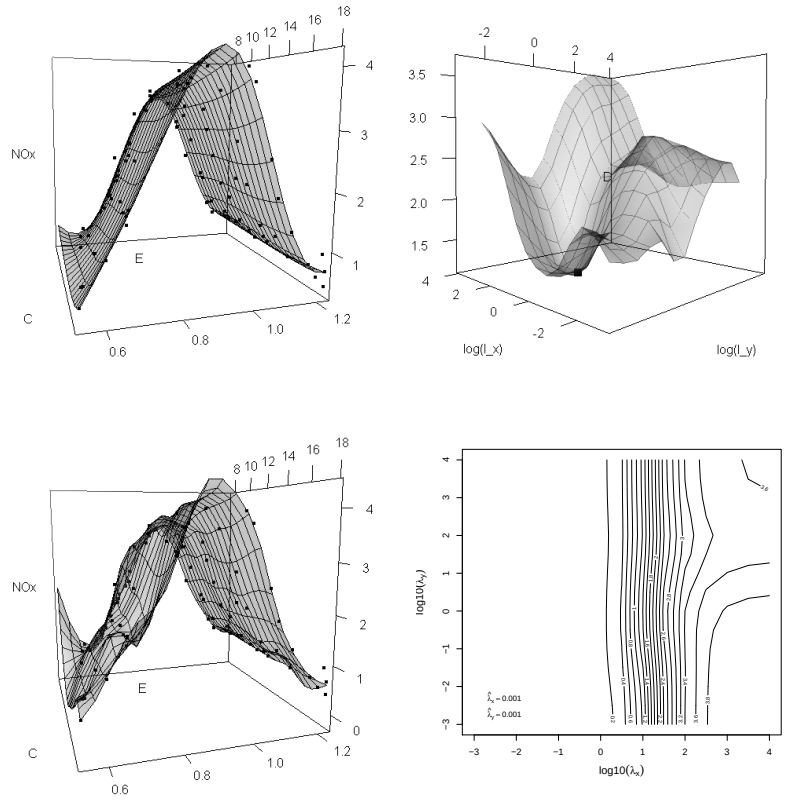


FIGURE 2. Smoothing of the ethanol data. Top left: the smooth surface obtained after selecting the λ parameters on the L-surface (upper right panel) using the V-valley criterion. The lower left panel shows the smoother estimated using cross validation while the panel on the left shows the cross validation function computed for the grid of smoothing parameters. 23 Cubic B-splines and third order penalties have been used for each dimension.

a nook in which the right balance between the residual sum of squares and two penalties is found. Instead of using local curvature to locate the right spot, we use a derived measure, based on distance, leading to simple computations.
It appears that our approach opens the road to smoothing in more dimensions, e.g. space and time. More research is needed. A more minimization method will be needed then, because brute-force exploration of a three-dimensional grid of smoothing parameters locating is time-consuming.
The example of the ethanol data suggests that the L-surface seems to be quite insensitive to serial correlation. This has to be investigated more thoroughly on real and simulated data.
There are many opportunities for further research like smoothing of non Gaussian data and penalty terms defined by vector norms different from L_2.
As a final remark we emphasize the crucial role of (the size of) penalties. Smoothing methods like kernels and local likelihood do not have an equivalent and so they cannot profit from the L-curve and its extensions.

Acknowledgments: The first author acknowledges financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy).

References

Shahrak, N. M., Shahsavand, A. Okhovat, A. Robust PSD determination of micro and meso-pore adsorbents via novel modified U curve method. *Chemical Engineering Research and Design* 91:pp. 51-62

Frasso, G., Eilers, Paul H.C (2012). Smoothing parameter selection using the L-curve.*Proceedings of the 27th International Workshop on Statistical Modelling, Prague.*

Eilers, Paul H.C. and Currie, Iain D. and Durban, Maria (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis* 50:pp. 61-76

Eilers, Paul H.C., Marx, B.D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometr. Intell. Lab. Systems* 66:pp. 159-174

M. Belge and M. Kilmer and E. L. Miller (2002). Efficient determination of multiple regularization parameters in a generalized L-curve framework. *Inverse Problems*, 18:pp. 1161.

Hansen, P.C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):pp. 561-580.

Hansen, P.C., O'Leary, D.P. (1993). The use of the L-Curve in the regularization of discrete ill-posed problems. *SIAM Journal of Scientific Computing*, 14(6):pp. 1487-1503.

Cox Regression Models with Functional Covariates

Jonathan Gellar[1], Ciprian M. Crainiceanu[1]

[1] Dept. of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

E-mail for correspondence: jgellar@jhsph.edu

Abstract: We introduce a novel method to relate a functional predictor to a survival outcome, in the presence of censoring. Our method extends the classical Cox proportional hazards model to allow for functional covariates. The functional coefficient is approximated using a spline basis, and estimated by maximizing the penalized partial likelihood. We conduct a simulation exercise to investigate the performance of the model under a variety of conditions.

Keywords: Functional Regression; Survival Analysis; Non-parametric statistics.

1 Introduction

Modern data collection techniques have increasingly given rise to data for which measurements are functions, or images, rather than scalar values. Examples of such measurements include electroencephalography (EEG) signals of the brain or images based on magnetic resonance (MR) or computed tomography (CT). Additionally, any longitudinal measurement, such as weather data or the value of a biomarker over time, may be thought of as a functional measurement. In this case, the domain of the function is the time axis, and the observations are considered to be finite sample realizations of an underlying stochastic process.

In recent years much research has been devoted to analysis techniques for such data, giving birth to the field known as functional data analysis. Much of this work has been aimed at the development of regression models for which a scalar outcome is regressed on one or more functional covariates, a technique referred to as scalar-on-function regression. These models have been extended to account for simple, gaussian-distributed outcomes to any outcome type in the exponential family.

Nonetheless, we are unaware of any methods that have been developed to account for time-to-event outcomes with functional covariates in the presence of censoring. Such a model would certainly have important clinical applications. For example, we may be interested in relating a brain scan or EEG signal conducted in stroke patients to their time to subsequent

stroke, or death. For our application, we relate a longitudinally collected biomarker, collected during a patient's hospitalization, to their risk of mortality after hospitalization. In this paper, we propose an extension of the classical Cox proportional hazards model to allow for functional covariates.
We now briefly review existing functional regression techniques that are relevant to this article. Scalar-on-function regression models the relationship between a scalar outcome and a functional covariate. Suppose that $\{Y_i\}$ are a set of scalar outcomes, $\{X_i(t)\}$ are functional covariates defined on the interval $[0, 1]$, and $\{\boldsymbol{Z}_i\}$ are non-functional covariates, where $i \in \{1, 2, \ldots, N\}$. Then the generalized functional linear model that relates Y_i to $\boldsymbol{Z}_i$ and $X_i(t)$ is

$$g(\mu_i) = \alpha + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 X_i(t)\beta(t)\,dt \tag{1}$$

where Y_i follows an exponential family distribution with mean μ_i, and $g(\cdot)$ is an appropriate link function (Cardot et al., 1999; James, 2002; Cardot and Sarda, 2005; Müller and Stadtmüller, 2005; Ramsay and Silverman, 2005; Reiss and Ogden, 2007; James et al., 2009).The key feature of this model that differentiates it from a standard (non-functional) generalized linear model is the integral term, $\int_0^1 X_i(t)\beta(t)\,dt$, which captures the contribution of the functional covariate $X_i(t)$ towards $g(\mu_i)$.
In the next section, we propose an extension of the Cox proportional hazards model that incorporates a similar integral term for functional covariates, and describe how the parameters in such a model may be estimated. Section 3 assesses the performance of our model in a simulation study. We conclude with a discussion of our findings in Section 4.

2 Cox Model with functional covariates

2.1 Proportional hazards model

Let T_i be the survival time for subject i, and C_i the corresponding censoring time. Assume that we observe only $Y_i = \min(T_i, C_i)$, and let $\delta_i = I(T_i \leq C_i)$. We also assume that for each subject, we have a collection of covariates $\boldsymbol{Z}_i = \{Z_{i1}, Z_{i1}, \ldots, Z_{ip}\}$. The Cox proportional hazards model (Cox, 1972) that relates the survival times $\{T_i\}$ to the covariates $\boldsymbol{Z}_i$ is given by

$$\log h_i(t; \boldsymbol{\gamma}) = \log h_0(t) + \boldsymbol{Z}_i\boldsymbol{\gamma}$$

where $h_i(t; \boldsymbol{\gamma})$ is the hazard at time t given covariates $\boldsymbol{Z}_i$ and $h_0(t)$ is a non-parametric baseline hazard function.
Suppose now that in addition to $\boldsymbol{Z}_i$, we have also collected a functional covariate, $X_i(s) \in \mathcal{L}^2[0, 1]$, for each subject. For convenience, we assume without loss of generality that $X_i(s)$ is centered by subtracting an estimator

of the population mean function from the observed data. We propose the following functional proportional hazards model.

$$\log h_i(t;\boldsymbol{\gamma},\beta(\cdot)) = \log h_0(t) + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 X_i(s)\beta(s)\,ds \tag{2}$$

Our method of estimating the functional parameter $\beta(s)$ follows the penalized functional regression procedure of Goldsmith, et al. (2011), which consists of two steps. The first step is a functional principal components decomposition of the predictor functions $\{X_i(s)\}$, and the second step approximates the coefficient function $\beta(s)$ using a spline basis. Define $\Sigma_X(u,v) = \text{cov}\,[X_i(u), X_i(v)]$, and let $\boldsymbol{\psi} = \{\psi_1(s), \psi_2(s), \ldots, \psi_{K_X}(s)\}$ be the collection of the first K_X eigenfunctions of the smoothed version of $\Sigma_X(u,v)$. Then based on the Karhunen-Loéve decomposition, we can approximate $X_i(s)$ by $X_i(s) = \sum_{k=1}^{K_X} c_{ik}\psi_k(s)$, where $c_{ik} = \int_0^1 X_i(s)\psi_k(t)\,ds$. Additionally, let $\boldsymbol{\phi}(s) = \{\phi_i(s), \phi_2(s), \ldots, \phi_{K_b}(s)\}$ be a spline basis over the s-domain, so that $\beta(s) = \sum_{k=1}^{K_b} b_k\phi_k(s)$. Then (2) becomes

$$\begin{aligned}
\log h_i(t;\boldsymbol{\gamma},\boldsymbol{b}) &= \log h_0(t) + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^T X_i(s)\beta(s)\,ds \\
&= \log h_0(t) + \boldsymbol{Z}_i\boldsymbol{\gamma} + \int_0^1 \boldsymbol{c}_i'\boldsymbol{\psi}(s)^T\boldsymbol{\phi}(s)\boldsymbol{b}\,ds \\
&= \log h_0(t) + \boldsymbol{Z}_i\boldsymbol{\gamma} + \boldsymbol{c}_i'\boldsymbol{J}_{\boldsymbol{\psi}\boldsymbol{\phi}}\boldsymbol{b}
\end{aligned} \tag{3}$$

where $\boldsymbol{c}_i = \{c_{i1}, c_{i2}, \ldots, c_{iK_X}\}$, $\boldsymbol{b} = \{b_1, b_2, \ldots, b_{K_b}\}$, and $\boldsymbol{J}_{\boldsymbol{\psi}\boldsymbol{\phi}}$ is a $K_X \times K_b$ dimensional matrix with (u,v)th element given by $\int_0^1 \psi_u(s)^T\phi_v(s)\,ds$ (Ramsay and Silverman, 2005). Note that $\boldsymbol{J}_{\boldsymbol{\psi}\boldsymbol{\phi}}$ is based only on the (known) basis functions, and can be solved by numerical integration.

2.2 Penalized partial likelihood approach

For notational convenience, let $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\gamma} & \boldsymbol{b} \end{bmatrix}$ and $\eta_i(\boldsymbol{\theta}) = \boldsymbol{Z}_i\boldsymbol{\gamma} + \boldsymbol{c}_i'\boldsymbol{J}_{\boldsymbol{\psi}\boldsymbol{\phi}}\boldsymbol{b}$. Then the partial likelihood for this model is

$$\begin{aligned}
L^{(p)}(\boldsymbol{\theta}) &= \prod_{i:\delta_i=1}\left\{\frac{e^{\log h_0(t_i)+\eta_i(\boldsymbol{\theta})}}{\sum_{j\in R(t_i)} e^{\log h_0(t_i)+\eta_j(\boldsymbol{\theta})}}\right\} \\
&= \prod_{i:\delta_i=1}\left\{\frac{e^{\eta_i(\boldsymbol{\theta})}}{\sum_{j\in R(t_i)} e^{\eta_j(\boldsymbol{\theta})}}\right\}
\end{aligned}$$

In order to maintain smoothness of the coefficient function $\beta(t)$, we impose a penalty on the spline coefficients, $\boldsymbol{b}$. The penalized partial log-likelihood

(PPL) is thus

$$\ell_{\lambda}^{(p)}(\boldsymbol{\theta}) \quad = \quad \sum_{i:\delta_i=1} \left\{ \eta_i(\boldsymbol{\theta}) - \log \left(\sum_{j:Y_j \geq Y_i} e^{\eta_j(\boldsymbol{\theta})} \right) \right\} - \lambda P(\boldsymbol{b})$$

where $P(\boldsymbol{b})$ is an appropriate penalty term for the spline coefficients $\boldsymbol{b}$, and λ is a smoothing parameter. The PPL has been introduced by Gray (1992) to allow for smoothing splines in the Cox model, by Verweij and Houwelingen (1994) to stabilize parameter estimates, and by Therneau et al. (2003) in frailty models. For a given λ, we may obtain parameter estimates $\hat{\boldsymbol{\theta}}$ by maximizing the PPL using a Newton-Raphson procedure. The smoothing parameter λ may be optimized by maximizing its profile likelihood, using the procedure of Ripatti and Palmgren (2000).

2.3 Inference

Gray (1992) suggested that $\boldsymbol{V} = \boldsymbol{H}^{-1}\mathcal{I}\boldsymbol{H}^{-1}$ be used as the covariance matrix of the parameter estimates, where $\boldsymbol{H}$ is the penalized Hessian of the PPL and $\mathcal{I}$ is the Fisher's information from the unpenalized Cox model. Thus, a pointwise 95% confidence interval for $\beta(s_0) = \boldsymbol{\phi}(s_0)\boldsymbol{b}$ may be constructed as $\hat{\beta}(s_0) \pm 1.96\sqrt{\boldsymbol{\phi}(s_0)'\boldsymbol{V}\boldsymbol{\phi}(s_0)}$. The utility and limitations of this confidence interval will be explored in the simulation exercise below.

3 Simulation Study

In order to assess the performance of our model under a variety of conditions, we conducted a simulation study. Of interest was our model's ability to accurately identify the coefficient function $\beta(s)$. For simplicity, we consider only the scenario where there are no non-functional covariates $\boldsymbol{Z}$, and only a single functional predictor $X_i(s)$.

Our model for simulating functional predictors is based on the procedure employed by Goldsmith et al. (2011), which was in turn adapted from Müller and Stadtmüller (2005). Let $\left\{ s_j = \frac{j}{10} : j = 0, 1, \ldots, J = 100 \right\}$ be a grid of time points over the interval $[0, 10]$. For each subject $i \in 0, \ldots, N$, we generate the survival time T_i and functional predictor $X_i(s)$ based on the model:

$$\begin{aligned} h_i(t) &= h_0(t)\exp(\eta_i) \\ \eta_i &= \frac{1}{J}\sum_{j=1}^{J} X_i(s_j)\beta(s_j) \\ X_i(s_j) &= u_{i1} + u_{i2}s_j + \sum_{k=1}^{10} \left\{ v_{ik1}\sin\left(\frac{2\pi k}{10}s_j\right) + v_{ik2}\cos\left(\frac{2\pi k}{10}s_j\right) \right\} \end{aligned}$$

where $h_i(t)$ is the hazard of T for subject i, $h_0(t)$ is the baseline hazard, $u_{i1} \sim N(0, 25)$, $u_{i2} \sim N(0, .04)$, and $v_{ik1}, v_{ik2} \sim N(0, 1/k^2)$. Random survival times are generated from this model using the procedure of Bender, et al. (2005). All subjects are censored at $C_i = 730$ days (2 years).
Additionally, we introduce possible measurement error into the simulation by assuming that we only observe $W_i(s_j) = X_i(s_j) + \delta_i(s_j)$, where $\delta_i(s_j) \sim N(0, \sigma_X^2)$. Our simulation considers two possibilities for σ_X^2 (0 or 1), corresponding to no measurement error and moderate measurement error. We also consider three different sample sizes $N \in \{200, 500, 1000\}$.
We generate data from three coefficient functions: $\beta_1(s) = 10\sin(\pi s/5)$, $\beta_2(s) = 10(s/10)^2$, and $\beta_3(s) = -10p(s|2, .3) + 30p(s|5, .4) + 10p(s|7.5, .5)$, where $p(s|\mu, \sigma)$ indicates the normal density with mean μ and standard deviation σ. These coefficient functions are similar to those used by Goldsmith, et al. (2011). Performance of our model is measured by the average mean squared error (AMSE) for each of the $R = 1000$ datasets, defined as $\text{AMSE}^{(r)}\left(\hat{\beta}_b(\cdot)\right) = \frac{1}{J+1}\sum_{j=0}^{J}\left\{\hat{\beta}_b^{(r)}(s_j) - \beta_b(s_j)\right\}^2$, where $\hat{\beta}_b^{(r)}(s_j)$ is the estimated coefficient function from dataset r at time s_j. We also measure

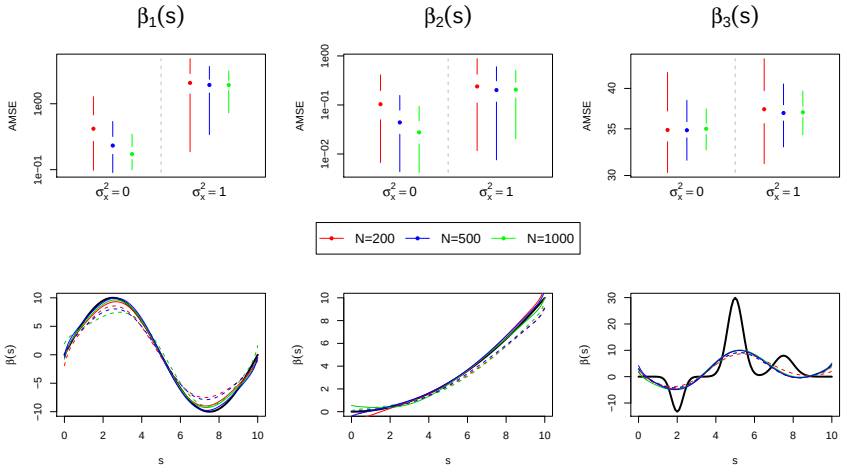


FIGURE 1. Simulation Results. The top row of figures depicts Tufte box plots of the average mean squared error (AMSE) across the 500 simulated datasets, for each combination of N, σ_X^2, and $\beta(s)$. For each box plot, the dot in the center indicates the median AMSE, while the upper line spans the 3rd quartile to the largest non-outlying point, and the lower line spans the smallest non-outlying point to the first quartile. In the lower row of figures, we plot the estimated $\hat{\beta}_1(s)$, $\hat{\beta}_2(s)$, and $\hat{\beta}_3(s)$ for the estimate with the median AMSE, for each combination of N and σ_X^2. The dark, solid line indicates the true coefficient function, and color indicates the sample size. Estimates corresponding to $\sigma_X^2 = 0$ are solid lines, while those corresponding to $\sigma_X^2 = 1$ are dashed lines.

the coverage probability of the pointwise confidence interval defined in Section 2.3.

Simulation results appear in Figure 1. We see that it is much easier to estimate the coefficient function in the absence of measurement error (i.e., $\sigma_X^2 = 0$). Additionally, estimation improves with increasing sample size, as expected. We also see that the regions of the coefficient function that are most difficult to estimate are the regions where the curvature of the true coefficient function has the highest magnitude, especially in the presence of measurement error. More specifically, measurement error appears to attenuate the estimates. $\beta_3(s)$, which contains sharp peaks that fall in regions where the variability in $\{X_i(s)\}$ is low, is much more difficult to estimate than the other two coefficient functions.

The performance of our proposed confidence interval is assessed by examining the average coverage probability of the 95% confidence intervals under each scenario (Table 1). Overall, our method underestimates the variability of our estimates, as evidenced by all mean coverage probabilities being below 95%. When $\sigma_X^2 = 0$, the confidence interval performs worst in the estimation of $\beta_3(s)$. On the other hand, when $\sigma_X^2 = 1$, the confidence interval performs very poorly in all scenarios, especially for $\beta_1(s)$.

TABLE 1. Mean coverage probability of the pointwise 95% confidence intervals, defined in Section 2.3 above.

σ_X^2	N	$\beta_1(s)$	$\beta_2(s)$	$\beta_3(s)$
0	200	93.8%	92.9%	72.6%
	500	93.2%	94.1%	81.9%
	1000	93.5%	93.8%	82.5%
1	200	44.7%	81.3%	49.5%
	500	28.3%	66.1%	56.6%
	1000	21.2%	49.6%	56.1%

4 Discussion

We develop new methodology to account for functional covariates in a Cox proportional hazards model. This model uses a spline basis to approximate the functional coefficient, and produces estimates by maximizing the penalized partial log likelihood. A simulation exercise confirms that the model does accurately identify the functional coefficient. The coefficient function is most difficult to estimate in regions with high curvature, when measurement error is present, and when the sample size is small. Our proposed confidence interval tends to underestimate the variability of our parameter estimates; thus we suggest that it should not be used. Instead, a bootstrap procedure may be used to obtain more reliable confidence intervals.

References

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, **24(11)**, 1713-1723.

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, **92(1)**, 24-41.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics & Probability Letters*, **45(1)**, 1122.

Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal Of Chronic Diseases*, **40(5)**, 373-383.

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **34(2)**, 187-220.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penalized Functional Regression. *Journal of Computational and Graphical Statistics*, **20(4)**, 830-851.

Gray, R. (1992). Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. *Journal of the American Statistical Association*, **87(420)**, 942-951.

James, G. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64(3)**, 411-432.

James, G., Wang, J., Zhu, J. (2009). Functional linear regression thats interpretable. *The Annals of Statistics*, **37(5A)**, 2083-2108.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, **33(2)**, 774-805.

Needham, D. M., Dennison, C. R., Dowdy, D. W., Mendez-Tellez, P. a., Ciesla, N., Desai, S. V., Sevransky, J., Shanholtz, C., Scharfstein, D., Herridge, M. S., and Pronovost, P. J. (2006). Study protocol: The Improving Care of Acute Lung Injury Patients (ICAP) study. *Critical care*, **10(1)**, R9.

Ramsay, J. O. and Silverman, B. (2005). *Functional data analysis*. New York: Springer.

Reiss, P. and Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71(2)**, 505-523.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **56(4)**, 1016-1022.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized Survival Models and Frailty. *Journal of Computational and Graphical Statistics*, **12(1)**, 156-175.

Verweij, P. and Houwelingen, H. V. (1994). Penalized Likelihood in Cox Regression. *Statistics in Medicine*, **13**, 2427-2436.

Zambon, M. and Vincent, J.-L. (2008). Mortality rates for patients with acute lung injury/ARDS have decreased over time. *Chest*, **133(5)**, 1120-1127.

Faster Spike-and-Slab Variable Selection with Dual Coordinate Ascent EM

Edward I. George[1], Veronika Rockova[2], Emmanuel Lesaffre[23]

[1] Department of Statistics, Wharton, University of Pennsylvania, USA
[2] Department of Biostatistics Erasmus MC Rotterdam, The Netherlands
[3] L-BioStat, KU Leuven, Belgium

E-mail for correspondence: edgeorge@wharton.upenn.edu

Abstract: Rockova and George (2012) proposed EMVS, a new approach for identifying sparse high posterior models under a Bayesian spike-and-slab formulation for variable selection uncertainty for the Gaussian linear model. An alternative to stochastic search, this approach is based on a deterministic closed form EM algorithm. In this paper we propose a version of a stochastic dual coordinate ascent algorithm which substantially speeds up a key step in the already fast EM algorithm further enhancing its potential for dynamic posterior exploration.

Keywords: Bayesian variable selection; dynamic posterior exploration; EMVS; regularization plots.

1 Introduction

Bayesian approaches to variable selection for the normal linear model under spike-and-slab priors typically use some form of Monte Carlo stochastic search to find high posterior probability models. For problems with many variables, such approaches are simply not fast enough to be practical. EMVS, an alternative approach proposed by Rockova and George (2012), avoids this problem by using a closed form EM algorithm to quickly identify sparse modal models. With this faster algorithm, it becomes feasible to conduct dynamic posterior exploration by widening the spike distribution to more clearly expose and identify sparse high probability models.
In this paper, we describe and illustrate the implementation of a stochastic dual coordinate ascent (SDCA) algorithm that substantially speeds up the key M-step of the algorithm. This M-step, which must be iterated at least several times, is the main computational burden of the algorithm. Essentially, each iteration entails computing a generalized ridge regression estimate of the regression coefficient vector, a computation that involves the inversion of a potentially large matrix. The SCDA algorithm replaces this inversion by a much faster iteration of single coordinate updates, substantially increasing the overall speed of EMVS, especially on large problems.

2 EMVS

EMVS is formulated for data which consists of $\boldsymbol{y}$, an $n \times 1$ response vector, and $\boldsymbol{X} = [x_1, \ldots, x_p]$, an $n \times p$ matrix of p potential standardized predictors that are related by a Gaussian linear model

$$f(\boldsymbol{y} \mid \alpha, \boldsymbol{\beta}, \sigma) = \mathrm{N}_n(\mathbf{1}_n \alpha + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n), \tag{1}$$

where $\mathbf{1}_n$ is a $n \times 1$ vector of 1's, α is an unknown scalar intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, and σ is an unknown positive scalar.

The cornerstone of the prior formulation for EMVS is the "spike-and-slab" Gaussian mixture prior on $\boldsymbol{\beta}$

$$\pi(\boldsymbol{\beta} \mid \sigma, \boldsymbol{\gamma}, v_0, v_1) = \mathrm{N}_p(\mathbf{0}, \boldsymbol{D}_{\sigma,\boldsymbol{\gamma}}), \tag{2}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, $\gamma_i \in \{0, 1\}$ is a vector of binary latent variables and $\boldsymbol{D}_{\sigma,\boldsymbol{\gamma}} = \sigma^2 \mathrm{diag}(a_1, \ldots, a_p)$ with $a_i = (1 - \gamma_i)v_0 + \gamma_i v_1$ for $0 \leq v_0 < v_1$. George and McCulloch (1993) introduced this prior and recommended setting the hyper-parameters v_0 and v_1 to be small and large, respectively, so that those x_i for which $\gamma_i = 1$ are to be included in the model.

Combining (2) with suitable priors on $\boldsymbol{\gamma}$, α and σ, the induced posterior distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ provides a useful summary of post-data variable selection uncertainty. For illustration, we consider the following choices here. On $\boldsymbol{\gamma}$, we consider the exchangeable beta-binomial prior obtained by coupling the iid Benoulli form $\pi(\boldsymbol{\gamma} \mid \theta) = \theta^{|\gamma|}(1 - \theta)^{p-|\gamma|}$ with a uniform prior on $\theta \in [0, 1]$. On α, we consider a uniform improper prior, proceeding from here on with the induced marginal likelihood $f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma)$. On σ^2, we consider the relatively noninfluential inverse gamma prior. $\pi(\sigma^2 \,|\gamma) = \mathrm{IG}(\nu/2, \nu\lambda/2)$ with $\nu = 1$ and $\lambda = 1$.

EMVS is based on a fast EM algorithm alternative to stochastic search of $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ which, when $v_0 > 0$, quickly finds posterior modes of $\pi(\boldsymbol{\beta}, \theta, \sigma^2|\boldsymbol{y})$ that are then thresholded to identify high posterior $\boldsymbol{\gamma}$ models under the posterior for which $v_0 = 0$. As described in detail in Rockova and George (2012), the EMVS algorithm proceeds by iteratively maximizing the objective function

$$\begin{aligned} Q\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \boldsymbol{\psi}^{(k)}\right) = & -\frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2}{2\sigma^2} - \log \sigma^{n-1+p+\nu} - \frac{\nu\lambda}{2\sigma^2} \\ & -\frac{1}{2\sigma^2} \sum_{i=1}^{p} \beta_i^2 \, \mathsf{E}_{\boldsymbol{\gamma}|\cdot} \left[\frac{1}{v_0(1 - \gamma_i) + v_1\gamma_i}\right] + \sum_{i=1}^{p} \log\left(\frac{\theta}{1 - \theta}\right) \mathsf{E}_{\boldsymbol{\gamma}|\cdot}[\gamma_i], \end{aligned} \tag{3}$$

where $\boldsymbol{\psi}^{(k)} = (\boldsymbol{\beta}^{(k)}, \theta^{(k)}, \sigma^{(k)})$ and $\mathsf{E}_{\boldsymbol{\gamma}|\cdot}[\cdot]$ denotes expectation conditionally on $[\boldsymbol{\psi}^{(k)}, \boldsymbol{y}]$. At the kth iteration, an E-step is first applied, which computes the expectations in (3), followed by an M-step that maximizes over $(\boldsymbol{\beta}, \theta, \sigma)$ to yield the values of $\boldsymbol{\psi}^{(k+1)} = (\boldsymbol{\beta}^{(k+1)}, \theta^{(k+1)}, \sigma^{(k+1)})$.

The E-step expectations are obtained quickly from the closed form expressions

$$\mathsf{E}_{\boldsymbol{\gamma}|\cdot}[\gamma_i] = \frac{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)\theta^{(k)}}{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)\theta^{(k)} + \pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 0)(1-\theta^{(k)})} \equiv p_i^\star \tag{4}$$

and

$$\mathsf{E}_{\boldsymbol{\gamma}|\cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right] = \frac{1-p_i^\star}{v_0} + \frac{p_i^\star}{v_1} \equiv d_i^\star. \tag{5}$$

For the M-step maximization, the $\boldsymbol{\beta}^{(k+1)}$ value that globally maximizes Q is obtained by the generalized ridge estimator

$$\boldsymbol{\beta}^{(k+1)} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^\star)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{6}$$

where $\boldsymbol{D}^\star = \mathrm{diag}\{d_i^\star\}_{i=1}^p$ is the $p \times p$ diagonal matrix with entries $d_i^\star > 0$, the well-known solution to the ridge regression problem

$$\boldsymbol{\beta}^{(k+1)} = \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + ||\boldsymbol{D}^{\star 1/2}\boldsymbol{\beta}||^2\}. \tag{7}$$

In problems where $p >> n$, the calculation of (6) can be enormously reduced by using the Sherman-Morrison-Woodbury formula to obtain an expression which requires a $n \times n$ matrix inversion rather than a $p \times p$ matrix inversion. The maximization of Q is then completed with the simple updates $\sigma^{(k+1)} = \sqrt{\frac{||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(k+1)}||^2 + ||\boldsymbol{D}^{\star 1/2}\boldsymbol{\beta}^{(k+1)}||^2 + \nu\lambda}{n+p+\nu}}$ and $\theta^{(k+1)} = \frac{1}{p}\sum_{i=1}^p p_i^*$.

3 Stochastic Dual Coordinate Ascent for EMVS

The key to the efficiency of the EMVS implementation is the expeditious updating of the ridge regression solutions in (6), by far the most expensive operation in the EM algorithm because of the costly matrix inversion, especially when both n and p are large. To mitigate this cost, approximate solutions to ridge and other regularized loss minimization problems can be obtained with conjugate gradient descent methods or dual coordinate ascent algorithms at only a fraction of the runtime. For this purpose, we here propose the stochastic version of the dual coordinate ascent algorithm (SDCA) of Shalev-Shwartz and Zhang (2012), which they show to possess strong theoretical guarantees. In conjunction with the already fast E-step, SDCA further enhances the potential of the EMVS procedure.

To motivate and tailor the SDCA algorithm for our context, denote the original data by $\boldsymbol{y}$, a $(n \times 1)$ response vector, and $\boldsymbol{X}^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*)^{\mathrm{T}}$, a $(n \times p)$ regression matrix, where x_i^*, the ith row of X^* has been rescaled so that $\max_i ||\boldsymbol{x}_i^*|| \leq 1$. The constrained loss function (7) to be minimized in the M-step of EMVS is the special case of

$$P^*(\boldsymbol{\beta}^*) = \left[\sum_{i=1}^n \phi_i(\boldsymbol{x}_i^{*\mathrm{T}}\boldsymbol{\beta}^*) + ||\boldsymbol{D}^{\star 1/2}\boldsymbol{\beta}^*||^2\right], \tag{8}$$

where $\phi(a) = (a - y_i)^2$. The optimizer of this generalized ridge regression problem can be obtained from a solution to a classical ridge regression after reweighing the columns of the regression matrix $\boldsymbol{X} = \boldsymbol{X}^* \boldsymbol{D}^{\star -1/2}$ so that $\boldsymbol{\beta} = \boldsymbol{D}^{\star 1/2} \boldsymbol{\beta}^*$. Thus the minimizer $\widehat{\boldsymbol{\beta}^*}$ of $P^*(\boldsymbol{\beta}^*)$ corresponds to the minimizer $\widehat{\boldsymbol{\beta}}$ of the ridge regularized loss function with unit penalty

$$P(\boldsymbol{\beta}) = \left[\sum_{i=1}^{n} \phi_i(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) + ||\boldsymbol{\beta}||^2 \right]. \tag{9}$$

Rather than minimize $P(\boldsymbol{\beta})$, SDCA operates by maximizing the dual formulation obtained by rewriting (9) in terms of $\eta_i = \phi_i(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$ and introducing Lagrange multipliers α_i for every one of the corresponding constraints $\eta_i - \phi_i(\boldsymbol{x}_i^T \boldsymbol{\beta}) = 0$. Augmenting (9) by the weighted sum of the constraint functions obtains

$$L(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\alpha}) = \left[\sum_{i=1}^{n} \eta_i^2 + ||\boldsymbol{\beta}||^2 + \sum_{i=1}^{n} \alpha_i [\phi(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) - \eta_i] \right] \tag{10}$$

and the associated dual Lagrange function $D(\boldsymbol{\alpha}) = \inf_{\boldsymbol{\beta}, \boldsymbol{\eta}} L(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\alpha})$. Differentiating the Lagrangian (10) in $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, we obtain conditions

$$\boldsymbol{\beta}(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i \quad \text{and} \quad \eta_i(\boldsymbol{\alpha}) = \frac{\alpha_i}{2},$$

which after substitution in (10) give the dual Lagrangian

$$D(\boldsymbol{\alpha}) = \left[\sum_{i=1}^{n} -\phi_i^*(-\alpha_i) + \left\| \frac{1}{2} \sum_{i=1}^{n} \alpha_i x_i \right\|^2 \right] \tag{11}$$

where $\phi_i^*(u) = \max_z(zu - \phi_i(z))$ is the convex conjugate of $\phi_i(\cdot)$. Let $\widehat{\boldsymbol{\alpha}}$ denote a maximizer of $D(\boldsymbol{\alpha})$. Then it is known that $\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}) = \widehat{\boldsymbol{\beta}}$ and $P(\widehat{\boldsymbol{\beta}}) = D(\widehat{\boldsymbol{\alpha}})$. It also holds that $P(\boldsymbol{\beta}) \geq D(\boldsymbol{\alpha})$ for all $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, which implies that the duality gap $P[\boldsymbol{\beta}(\boldsymbol{\alpha})] - D(\boldsymbol{\alpha})$ constitutes an upper bound of the sub-optimality measure $P[\boldsymbol{\beta}(\boldsymbol{\alpha})] - P(\boldsymbol{\alpha})$.
In the following, we restrict the attention to squared loss in the linear regression setting, where the dual function takes the form

$$D(\boldsymbol{\alpha}) = \left[\sum_{i=1}^{n} y_i \alpha_i - \frac{1}{4} \sum_{i=1}^{n} \alpha_i^2 + \left\| \frac{1}{2} \sum_{i=1}^{n} \alpha_i x_i \right\|^2 \right]. \tag{12}$$

A nearly optimal value $\widehat{\boldsymbol{\alpha}}$, and hence nearly optimal $\widehat{\boldsymbol{\beta}}$, can be found with a coordinate descent algorithm (CDA) on the dual Lagrangian function. SDCA is the stochastic version of this algorithm where the coordinate to be updated at each iteration is chosen at random. Initializing $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}(\mathbf{0})$, the steps of the SDCA algorithm are

(1) Iterate for $t = 1, 2, \ldots, T$

(a) Select i randomly from $\{1, \ldots, n\}$

(b) Set $\Delta\alpha_i = \frac{2(y_i - x_i^T \boldsymbol{\beta}^{(t-1)}) - \alpha_i^{t-1}}{1+||\boldsymbol{x}_i||^2}$

(c) $\alpha^{(t)} \leftarrow \alpha^{(t-1)} + \Delta\alpha_i \boldsymbol{e}_i$

(d) $\boldsymbol{\beta}^{(t)} \leftarrow \boldsymbol{\beta}^{(t-1)} + \frac{1}{2}\Delta\alpha_i \boldsymbol{x}_i$

(2) Output $\bar{\boldsymbol{\alpha}} = \frac{1}{T-T0}\sum_{t=T0}^{T} \boldsymbol{\alpha}^{(t)}$ and $\bar{\boldsymbol{\beta}} = \frac{1}{T-T0}\sum_{t=T0}^{T} \boldsymbol{\beta}^{(t)}$

For the γ-smooth loss functions (differentiable with γ-Lipschitz derivative), Schwartz and Zhang (2012) show that SDCA requires at least $T = 2(n + n\gamma/2)\log(1/\varepsilon)$ iterations in order the to have an expected duality gap $\mathsf{E}[P(\bar{\beta}) - D(\bar{\alpha})] \leq \varepsilon$ for $\bar{\beta}$ and $\bar{\alpha}$ averaged over last $T0 = T/2$ iterations. Since squared loss is 2-smooth, it suffices to perform at least $T = 4n\log(1/\varepsilon)$ iterations.

4 Timing Comparisons

We consider simulated datasets on $p = 1\,000$ explanatory variables, where only the first three are predictive with a corresponding regression vector $\boldsymbol{\beta} = (2, 3, 4, 0, \ldots, 0)^{\mathrm{T}}$. We generated three datasets with $n = 100, 500, 2\,000$ and compared computational times to obtain a single generalized ridge regression solution by (a) calculation of (6) via inversion of a $p \times p$ matrix, (b) calculation of a Woodbury-Sherman (W-S) equivalent of (6) via inversion of an $n \times n$ matrix, (c) SGDA minimization of (7) via an implementation in R, and (d) SGDA minimization of (7) via an implementation in C. The regression matrices are generated with rows drawn independently from $\mathrm{N}_p(\mathbf{0}, \Sigma)$, where $\Sigma = \left(0.6^{|i-j|}\right)_{i,j=1}^{p}$. The predictor matrices were rescaled so that $||\boldsymbol{x}_i||^2 \leq 1$, a requirement for the SGDA theoretical guarantees. The response vector for each of the three sample sizes was created according to the generating model $\mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \mathrm{I}_n)$ and further normalized so that $||\boldsymbol{y}||^2 = 1$. The vector of penalty coefficients was generated through random sampling from a Gamma distribution with shape 1 and scale 0.5. Table 1 reports runtimes in seconds obtained on a $3GHz$ server as well as distances between the exact and approximate solutions. A stopping rule $T = 4n\log(1/\varepsilon)$ was implemented to obtain an expected duality gap of at most $\varepsilon = 0.1$.

A feature of EMVS made feasible by the speed of the EM algorithm is dynamic posterior exploration as the spike variance v_0 is gradually increased to expose sparse high probability models. To illustrate how SCDA enhanced EMVS would perform on this task, we applied it to the second dataset ($n = 500$) with $v_1 = 10$ and $v_0 \in \{0.1 + k \times 0.25; 0 \leq k \leq 20\}$. We obtained the EMVS regularization plot in Figure 1a displaying the evolution of the posterior modal estimates as v_0 is increased. The log-posterior model probabilities of subsets obtained after screening out coefficients that are small

TABLE 1. Computational time of the generalized ridge regression solutions

$p = 1\,000$	Classical	W-S	SGDA(R)	SGDA(C)	$\|\|\boldsymbol{\beta} - \boldsymbol{\beta}_{ridge}\|\|^2$
$n = 100$	2.44	0.38	0.17	0.02	0.004
$n = 500$	4.41	3.69	1.27	0.16	0.005
$n = 2\,000$	9.93	10.24	5.28	0.66	0.002

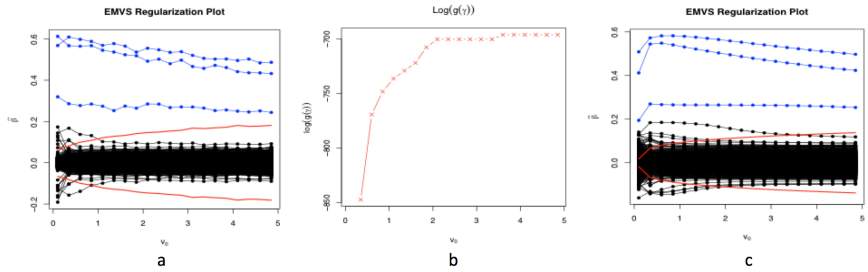


FIGURE 1. (a) Exact EMVS regularization plot, (b) model exploration based on the exact regularization plot, (c) approximated EMVS regularization plot

in magnitude (outside the threshold boundary depicted in red) are plotted in Figure 1b. The approximate regularization plot obtained using the SGDA procedure ($T = 4n\log(10)$) in the M-step is depicted in Figure 1c. The disagreement between Figures 1a and c is slight, especially for larger v_0 where sparse models are better identified.

Under the convergence criterion $\max|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}| < 0.05$, the exact evaluation of the whole regularization plot required 80 iterations taking 295 seconds using the Woodbury-Sherman updates with an R-implementation. The SGDA approximation required 94 iterations, taking 119 seconds with the R-implementation and merely 15 seconds with the C-implementation.

References

George, E.I. and McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 423 881 — 889.

Rockova, V. and George, E.I. (2012). EMVS: The EM Approach to Bayesian Variable Selection. Submitted.

Shalev-Shwartz, S. and Zhang, T. (2012). Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *Journal of Machine Learning Research* (to appear).

A joint Bradley-Terry model for tennis tournaments via Data Cloning

Anna Gottard

[1] Department of Statistics Informatics Applications, University of Florence, Italy

E-mail for correspondence: gottard@disia.unifi.it

Abstract: The paper introduces a model to study the abilities of women professional tennis players in 2012. By modelling the probability of winning a contest jointly with the number of matches played in a tournament, a possible informative cluster size effect is taken into account. As the resulting likelihood is intractable, the Data Cloning procedure is utilized to obtain the maximum likelihood estimates together with the observed Fisher's Information matrix.

Keywords: Bradley-Terry model; Data Cloning; Informative cluster size; Tennis tournament.

1 Introduction

Forecasting, ranking or analysing sports teams and players' performance is often of interest, mainly for assessing the efficiency of betting markets, sometimes as a specific research topic. The aim of this work is to study the performance of women's professional tennis (WPT), using an extension of a Bradley-Terry model (Bradley and Terry, 1952) including random effects. Several studies analysed tennis match results by different points of view. Among others, McHale and Morton (2011) present a Bradley-Terry kind model to forecast men's professional tennis, together with an extended review on the topic.

The Bradley-Terry (hence BT) model has been utilized for paired comparisons in many contexts. Applied to sports as tennis, the model assumes that the probability of a victory for a player i on a player j depends on the comparison of players' abilities, measured by specific parameters of the model. See Cattelan (2012) for an updated review and some examples of application of BT models. Covariates can be included to better explain individual abilities (Dittrich *et al.*, 1998, Firth, 2005; Turner and Firth, 2012), together with random effects taking into account for unobserved heterogeneity. The resulting model can be viewed as a multiple membership model (see Hill and Goldstein, 1998; Rasbash and Browne, 2001a; Browne et al., 2001), in which each match is a unit and each player is a gruop.

In including random effects, an important aspect has to be taken into account. Typically, ordinary models for clustered data assume that the number of observations in a cluster is independent of the response variable. However, in some particular situations, cluster size may provide important information about the distribution of the response variable. In these situations, the cluster size is *informative* and has to be adequately modelled. In fact, as shown by Neuhaus and McCulloch (2011), ignoring an informative cluster size may result in biased estimates.

Tennis tournaments are knockout tournaments, with players competing head-to-head in each round, the winners advancing to the next round and the losers being eliminated from the tournament. The knockout design often assumes that matches between top players occur at the end of the tournament. Consequently, one can figure that the number of times a tennis competitor plays at a tournament, which is related to cluster size in a random effect BT model, depends on his/her ability, influencing also the response of interest.

Several models have been proposed to incorporate informative cluster sizes for both continuous and binary variables. For example, Dunson *et al.* (2003) propose a Bayesian model for jointly modeling the cluster size and the subunit-level outcomes, assuming a latent variable taking into account the association between the outcomes and cluster size.

In this work, we propose to jointly model the probability of victory of each contest and the number of times each contender plays at each tournament. The data utilized, for the 2012 WTP season, are available from the `http://www.tennis-data.co.uk/data.php` web site. The data contain details on the date of the match, location and tournament name, surface (hardcourt, carpet, clay or grass) and series (Grand Slam, Premier, *etc.*), participants' names and match results in games and sets. Only the top women players are included in the study. As the likelihood involves a high dimensional integral, to obtain the maximum likelihood estimates, the Data Cloning (Lele et al., 2007; Lele *et al.*, 2010) algorithm has been utilized.

The paper is organized as follows. In Section 2, the proposed model for tennis tournaments taking into account the possibly informative cluster size is presented. Section 3 briefly presents the Data Cloning algorithm for estimating the proposed model. The final section includes comments and remarks.

2 A Bradley-Terry model for tennis tournaments

Considering each match as a statistical unit, BT models can typically be viewed as generalized linear models in which the response variable is a

binary variable assuming value 1 if the first contender wins the comparison:

$$Y_{ij}^k = \begin{cases} 1 & \text{if player } i \text{ beats player } j \text{ in the match } k \\ 0 & \text{if player } j \text{ beats player } i \text{ in the match } k \end{cases}$$

with $i, j = 1, \ldots, I$, and $k = 1, \ldots N$. Denoting with π_{ij}^k the probability that $Y_{ij}^k = 1$, the BT model can be then written as

$$\text{logit}(\pi_{ij}^k) = \lambda_i^k - \lambda_j^k$$

where λ_l measure the logarithm of the ability of player l, with $l = i, j$. To include player–specific explanatory variables, such as for example player age (say X), the player log-ability of player l can be modelled as

$$\lambda_l^k = \lambda_l + \beta_1 x_l + u_{lk}$$

where u_{lk} is a residual unobserved variability specific for player l in the tournament of the match k. The effect on the $\text{logit}(\pi_{ij}^k)$ of a player–specific explanatory variable is $\beta_1(x_i - x_j)$. To include contest-specific explanatory variables, such as for example the tournament surface (Z), the player log-ability can be modelled as

$$\lambda_l^k = \lambda_l + \gamma_l z_k + u_{lk}$$

including a parameter specific for each player. In this case, the effect on the $\text{logit}(\pi_{ij}^k)$ of a contest–specific explanatory variable is $(\gamma_i - \gamma_j)z_k$.
Jointly to the BT model, we can specify a model for the number of matches played in a tournament by a player. For example, a Negative Binomial model could be adequate, being more flexible than a Poisson model.
Let $t = 1, \ldots T$ be the tournament indicator. Then, we can assume that the number of matches for unit i in tournament t, say n_{it} is distributed as Negative Binomial distribution with parameters $\big(\alpha, \exp(\eta_{it})\big)$, where, for instance,

$$\eta_{it} = \lambda_i + \delta x_i + v_i$$

with v_i being a latent quantity for unobserved heterogeneity, potentially correlated with the corresponding latent variable in the BT part.

3 Data Cloning

As a multiple membership model, the proposed model has a non-hierarchical structure of the random effect, with the likelihood function obtained after integrating out the latent components. The dimension of this integral makes the likelihood function analytically intractable.
Data Cloning (DC) procedure computes maximum likelihood estimates and the inverse of the Fisher information matrix, by utilizing as an instrument the Bayesian paradigm and MCMC procedures.

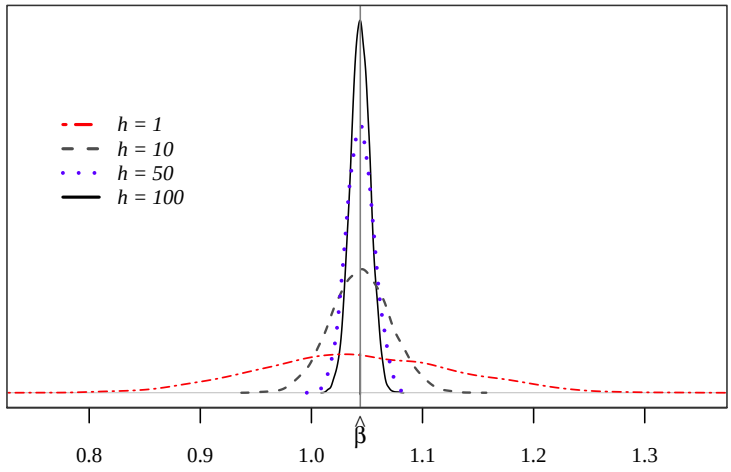


FIGURE 1. Example of pseudo–posterior distribution behaviour at increasing values of h.

Denote by $\boldsymbol{\theta}$ the entire vector of model parameters, by $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ the likelihood function of the data and by $\pi(\boldsymbol{\theta})$ an arbitrarily chosen prior distribution for the model parameters. To obtain maximum likelihood estimates, DC uses the so called *pseudo–posterior distribution*. This is the posterior distribution that is obtained by applying the Bayesian paradigm to cloned data, that is to data duplicated h times

$$\pi_{(h)}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})^h \pi(\boldsymbol{\theta}). \tag{1}$$

Under certain regularity conditions (see the Appendix A.1 in Lele *et al.*, 2010), the pseudo–posterior distribution degenerates towards maximum likelihood estimates when h tends to infinity. Moreover, for large h, the pseudo–posterior distribution tend to distribute normally, with mean equal to the maximum likelihood estimates of the model parameters and variance $1/h$ times the inverse of the Fisher information matrix. Therefore, the procedure allows the construction of asymptotic confidence intervals and the implementation of asymptotic hypothesis tests.

The DC estimates are invariant to the assumptions on the prior distribution, which can be chosen for computation convenience. Figure 1 shows an example of pseudo–posterior distribution for a logit model on simulated data of size 1 000. Note that the DC estimates converges to the maximum likelihood estimates and not to the true value of the parameter, which, in this example, was set at one.

4 Concluding remarks

The work focuses on measuring players' ability by jointly modelling matches results and number of matches played. Players' ability is conceived as not changing over time and tournaments are not modelled longitudinally. Further development of the proposal could include dynamic abilities (Cattelan *et al.*, 2012).

Acknowledgments: Special thanks to Alan Agresti and David Firth for interesting discussions on the topic.

References

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika*, **39**, 324 – 345.

Browne, W., Goldstein, H. and Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, **1**, 103 – 124.

Cattelan, M. (2012). Models for Paired Comparison Data: A Review with Emphasis on Dependent Data. *Statistical Science*, **27**, 412 – 433.

Cattelan, M., Varin, C. and Firth, D. (2012). Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**, 135 – 150.

Dittrich, R., Hatzinger, R. and Katzenbeisser, W. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**, 511 – 525.

Dunson, D.B. and Chen, Z. and Harry, J. (2003). A Bayesian Approach for Joint Modeling of Cluster Size and Subunit-Specific Outcomes. *Biometrics*, **59**, 521 – 530.

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software*, **12**, 1 – 12.

Hill, P. and Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral statistics*, **23**, 117 – 128.

Lele, S., Dennis, B. and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov Chain Monte Carlo methods. *Ecology Letters*, **10**, 551 – 563.

Lele, S., Nadeem, K. and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, **105**, 1617 – 1625.

McHale, I. and Morton, A. (2011). A Bradley–Terry type model for forecasting tennis match results. *International Journal of Forecasting*, **27**, 619 – 630.

Neuhaus, J.M. and McCulloch, C.E. (2011). Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika*, **98**, 147 – 162.

Rasbash, J. and Browne, W. (2001). Modelling non-hierarchical structures. *Multilevel modelling of health statistics*, 93 – 105.

Turner, H. and Firth, D. (2012). BradleyTerry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, **48**, 1 – 21.

SOM clustering and modelling of Australian railway drivers' sleep, wake, duty profiles

Irene L. Hudson[1], Shalem Y. Leemaqz[2], David Darwent[3], Greg Roach[3], Drew Dawson[3]

[1] School of Mathematical and Physical Sciences, University of Newcastle, NSW, Australia
[2] School of Paediatrics and Reproductive Health, University of Adelaide, Australia
[3] Appleton Institute, Central Queensland University, Adelaide, Australia

E-mail for correspondence: irenelena.hudson@gmail.com

Abstract: A Self Organizing Map (SOM) approach was adapted for time series and used for visualisation and clustering of differing patterns of sleep in a unique set of sleep/duty/work break time series profiles of 69 railway drivers (RDs) with 14 days of complete activity graph records. SOMs identified four groups of RDs with different patterns, timing and amount of attained sleep at every sleep episode. Clustering and Generalized Additive Models for Location, Scale and Shape (GAMLSS) of sleep with respect to the RD's stochastic profiles of break, sleep, duty and next duty characteristics, 1-2 sleep episodes prior and cluster membership, confirm that both the timing of sleep, break and next duty, duration of break, and hours to next duty significantly influence attained sleep. Break and sleep onset times, break duration and hours to next duty are significant effects which operate similarly across the groups. Although RDs have different sleep patterns, the amount of sleep is generally higher at night and when break duration is 2 or more days. Sleep increases for next duty onset between 10am - 4pm, and when hours since break onset exceeds 1 day - these 2 factors are found to be significant factors determining current sleep, which have differential impacts across the clusters. Some drivers catch up sleep after the night shift, while others do so before the night shift. Sleep is governed by the RD's anticipatory behaviour of next scheduled duty onset and hours since break onset.

Keywords: SOM clustering; multivariate episodic series; GAMLSS; railway drivers.

1 Introduction

Fatigue in the rail industry is an important health and safety issue in Australia and world-wide. Fatigue is affected by many factors with sleep and circadian rhythms, two of the fundamental physiological factors. For railway drivers many factors such as environmental, physical conditions,

and type of work also impact fatigue. The Australian railways shift work and workload study report (Dawson et al., 1997) suggested that RDs tend to have similar sleep patterns despite differing work schedules and personal attributes - sleeping during the night and awake during the day. Adequate sleep (5-6 hours) were reported only by RDs whose breaks began between 6 pm and 4 am, suggesting that drivers with breaks commencing outside these hours may need a longer break. The report also suggested that RDs do not tend to adapt physiologically to irregular work schedules, with alertness and performance lowest at 2.15 am if on early morning shift.

2 Data

A series of 14 field-based studies were chosen by the Rail consortium between June 1996 to June 1997. For each study drivers wore an activity monitor (actigraph) 24 hours/day for 14 days to record sleep diary details. Australia wide 253 RDs of an average age and shiftwork experience of 39.7 and 19.8 years, respectively, participated. Of these, 190 RDs with no missing sleep, wake, duty, break data, and a full 14 day sleep record were analysed in this study (n=69). This record of 69 sleep time series constitutes $>$ 1000 sleep episodes (4-23 per RD) with sleep duration from 1 to 35 hours. Six variables were calculated from the diaries: Break duration (total hours off-duty); Hours Since Break Onset; Hours To Next Duty; Next Duty, Break and Sleep onset times (on a 24-hour clock). Sleep duration (hrs) was calculated per sleep episode and socio-demographics (marital status, number of dependents, RD age and driver experience) collected.

3 Model and Methods

The Self Organizing Map (Kohonen, 2001), known as the Kohonen feature map, converts complex, nonlinear statistical relationships between high-dimensional data into simple geometric relationships on a low-dimensional display, usually a 2D map. The SOM as such is a topological map which organises itself based on the input patterns that it is trained on. A non-time series SOM approach was used recently by Nguyen et al. (2009) to map living standards in Viet Nam and for accident risk classification of Australian railway crossings (Sleep and Hudson, 2008), which developed a SOM with mixtures approach where the SOM best mapping units were clustered using model based clustering (MCLUST) of Fraley et al. (2013). The aim of this current study is to cluster and model the profiles using a SOM approach adapted recently for multivariate time series data to analyse flowering series and to derive a new metric for species synchronisation of flowering (Hudson et al., 2011b). We adopt a feature-based approach which uses SOMs to cluster the most frequently occurring patterns of profiles of episodic events. Clustering of the sleep hours is performed via VANTED (Junker

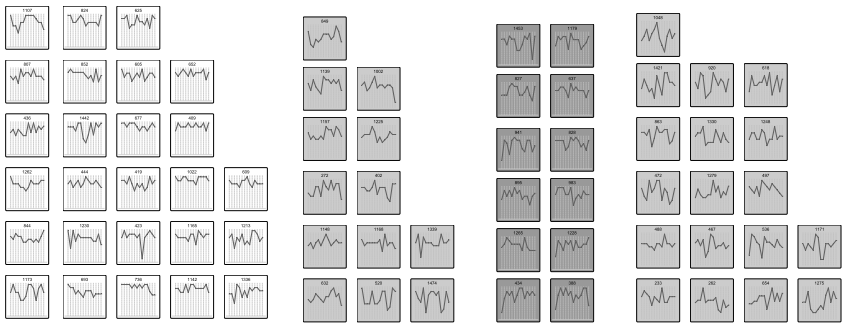


FIGURE 1. SOM clusters of sleep: clusters 4, 2, 3, and 1 from left to right.

et al., 2006). Generalized Additive Models for Location, Scale and Shape (GAMLSS) (Rigby and Stasinopoulos, 2005) of attained sleep are then used to model RDs' sleep with respect to resultant SOM cluster membership and the following predictors: hours to next duty and since break onset, sleep onset, next duty and break onset times, and break duration; and their interactions with cluster, along with sleep hours 1 or 2 episodes prior. GAMLSS models were recently used to investigate climatic effects and thresholds on the intensity of Eucalypt flowering (Hudson et al., 2011c; Hudson et al., 2010). Similarly we aim to establish which levels/thresholds of work/break and timing of sleep/break/next duty trigger fatigue. GAMLSS model optimality is based on the AIC criterion, RD is treated as a random effect and cubic splines used. Circularity of time is accommodated for. Two GAMLSS Models 1 and 2 are fitted using both stepwise and non-stepwise procedures; where Model 1 (M1) contains the predictors lag 1-2 sleep; and group factor, sleep, break and next duty onset, break duration, hours to next duty (and their interactions with group). Model 2 (M3) is Model 1 plus an additional covariate, hours since break onset.

4 Results and Discussion

SOM clustering found 4 clusters of size 18, 13, 12 and 26 across which sleep patterns are significantly different (Figure 1). RDs in cluster 1 ($n = 18$) had minimum sleep/episode (average = 6.96 hrs), cluster 2 ($n = 13$) maximum sleep (average = 7.71 hrs), cluster 3 ($n = 12$) and 4 ($n = 26$) RDs average hours sleep is 7.44 and 7.35 hours, respectively. Cluster 4 is the baseline contrast. Current sleep is highly positively related to attained sleep one episode prior ($P < 0.001$), but not at lag 2. The highly significant main effects of sleep onset time ($P < 0.000003$), break onset ($P < 0.03$), break duration ($P < 0.03$) and hours to next duty onset ($P < 0.0005$) are the same across groups. Sleep is highly significantly related to next duty onset

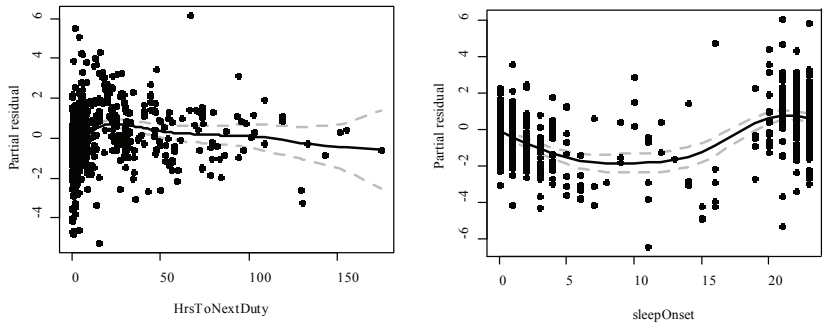


FIGURE 2. Term plots of effect of hours to next duty ($P = 0.005$) and sleep onset time ($P = 0.000003$).

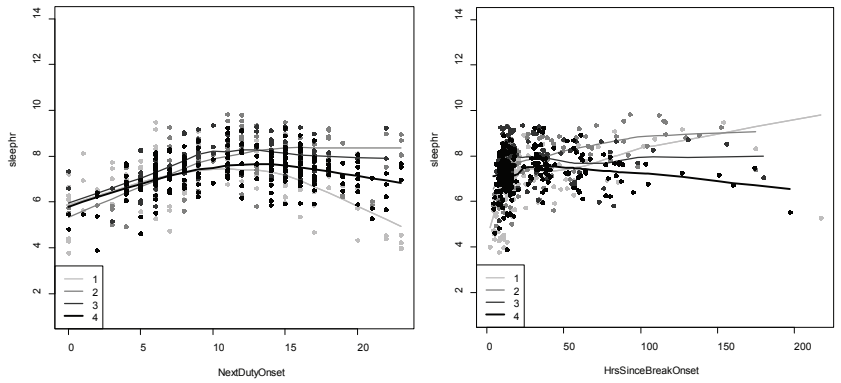


FIGURE 3. Interaction plots for next duty onset time ($P = 0.000006$) and hours since break onset ($P = 0.08$).

time ($P < 0.000006$), but hours since break onset is not a significant main effect (Figure 2). GAMLSS found significant differential effects on attained sleep of the following factors by group: next duty onset time (group 4 vs 1; $P < 0.005$) and hours since break onset (group 4 vs 1; $P < 0.08$, significant at 10%). Stepwise variants of M1 and M2 found cluster as a significant main effect, group 1 with least sleep. When hours since break onset is included (M2), break duration has a significant impact on attained sleep - not the case for M1 where break onset is a significant main effect, not break duration, as is hours to next duty (M1 stepwise) (Figure 3).

5 Conclusion

Break and sleep onset times, break duration and hours to next duty are significant effects which operate similarly across the groups. Adequate sleep (5-6 hours) is reported only by RDs with breaks between 6 pm and 4 am. Next duty onset time and hours since break onset are found to be significant factors in determining current sleep, which have differential impacts on current sleep across groups. Generally sleep increases for next duty onset between 10am - 4pm and when hours since break onset exceeds 1 day. Group 2 RDs, with maximum sleep, rapidly increase sleep as their next duty onset is later than 12 noon, group 1, with minimum sleep, have reduced sleep hours when next duty onset occurs between midnight and 6am, or from 4 pm to midnight. Group 2 and 1 RDs increase sleep as their hours since break exceeds 1 and 2 days, respectively. Group 3 and 4's sleep is decreased when hours since break onset exceeds 1 day. Generally RDs increase sleep for break duration from 1-2 days. As far as the authors are aware, this is the first study to find that sleep patterns are governed by anticipatory behaviour in relation to next duty onset time and retrospectively to hours since break onset. There was no evidence that social predictors - the presence of a partner and young kids, RD age and experience are significantly different between the identified clusters.
The results of our study reflect those from a multivariate Gaussian Hidden Markov Model analysis of the same dataset (Nur et al., 2012) and of a study using a hybrid of SOMs and model-based clustering (MCLUST) of sleep/wake/duty feature parameter vectors, obtained via a transitional state process approach - a multivariate extension of the mixture transition distribution model, which accommodates covariate interactions (Kim et al., 2009).

References

Dawson, D., Roach, G., Reid, K., and Barker, K. (1997). *Australian railways shiftwork and workload study: Final report.* Centre for Sleep Research, University of South Australia.

Farley, C., Raftery, A.E., Murphy, T.B., and Scrucca, L. (2013). *MCLUST Version 4 for R: Normal Mixture Modeling and Model-Based Clustering, Classification, and Density Estimation.* Technical Report No. 504, Department of Statistics, University of Washington.

Hudson, I.L., Kim, S.W., and Keatley, M.R. (2011a). Modelling lagged dependency of flowering on current and past climate on Eucalypt flowering: a mixture transition state approach. *International Congress of Biometeorology*, Auckland, New Zealand, pp. 239 – 244.

Hudson, I.L., Keatley, M.R., and Lee, S. (2011b). Using Self-Organising Maps (SOMs) to assess synchronies: and application to historical eucalypt flowering records. *International Journal of Biometeorology*, **55**(6), pp. 879 – 904.

Hudson, I.L., Kim, S.W., and Keatley, M.R. (2011c). Climate effects and thresholds for flowering of eight Eucalypts: a GAMLSS ZIP approach. *19th International Congress on Modelling and Simulation*, Chan, F., Marinova, D., and Anderssen, R.S. (Ed.), 12-16 December, Perth, Australia, pp. 2647 – 2653.

Hudson, I.L., Kim S.W., and Keatley, M.R. (2010). *Climatic influences on the flowering phenology of four Eucalypts: A GAMLSS approach.* In: Hudson I.L. and Keatley, M.R. (Ed.) Phenological Research: Methods for Environmental and Climate Change Analysis. Springer, Dordrecht, pp. 209 – 228.

Junker, B., Klukas, C., and Schreiber, F. (2006). VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinform*, **7**(1), pp. 109.

Kim, S.W., Hudson, I.L., and Keatley, M.R. (2009). Modelling the flowering of four Eucalypts species via MTDg with interactions. In: *18th World IMACS Congress and International Congress on Modelling and Simulation*, Anderssen, R.S., Braddock, R.D., and Newham, L.T.H. (Ed.). 13-17 July, Cairns, Australia, pp. 2625 – 2631.

Kohonen, T. (2001). *Self-Organizing Maps.* Springer Series in Information Sciences. Vol. 30, Springer, Berlin.

Nguyen, P.N., Haughton, D., and Hudson, I.L. (2009). Living standards of Vietnamese provinces: a Kohonen map. *Case Studies in Business, Industry and Government Statistics* **22**(2), 109 – 113.

Nur, D., Hudson, I.L., and Kim, S.W. (2012). Multivariate Gaussian Hidden Markov models for sleep profiles of railway drivers. *Australian Statistical Conference*, 9-12 July, Adelaide, Australia.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized Additive Models for Location, Scale and Shape (with discussion). *Applied Statististics*, **54**, pp. 507 – 554.

Sleep, J.A., and Hudson, I.L. (2008). Comparison of Self-Organising Maps, Mixture, K-means, and Hybrid approaches to risk classification of passive railway crossings. *Proceedings of the 23rd International Workshop on Statistical Modelling*, Paul H.C. Eilers (Ed.), 7-11 July, Utrecht, pp. 396 – 401.

A proposal for modelling overdispersion in ordinal data

Maria Iannario, Domenico Piccolo

[1] Department of Political Sciences, University of Naples Federico II, Italy

E-mail for correspondence: maria.iannario@unina.it

Abstract: The paper describes a mixture generated by a Beta Binomial and Uniform random variables for analysing a possible *overdispersion* in ordinal data generated by a sample survey. This distribution is very useful for fitting data generated by individual responses when subjects' covariates are available. A real case study illustrates the greater versatility of the new model compared with the standard one.

Keywords: Overdispersion, Mixture model, CUBE distribution.

1 Introduction

In categorical data analysis it is often found that data exhibit greater variability than predicted by the implicit mean-variance relationship. This phenomenon mentioned as *overdispersion* has been widely considered in literature, particulary in relation to Binomial and Poisson distributions. Failure to take account of this process can lead to serious underestimation of standard errors and misleading inference for the regression parameters. Consequently, a number of models and associated estimation methods have been proposed for handling such problem (McCullagh and Nelder, 1989). Cox (1983), for instance, showed that the heterogeneity factor usually provides an adequate correction, without further modelling, unless the data are highly unbalanced, whereas Finney (1971) took a more cautious approach, warning against corrections as a global remedy.
Models using a single parameter to account for extravariation in categorical data have been widely considered (Crowder, 1978; Altham 1976; Williams, 1982; Moore, 1987; Haseman and Kupper 1979, and Cox and Snell 1989, among others). However, only few papers concern the analysis of overdispersion in ordinal data for which a generalization of binary logistic regression is usually introduced.
In this context the possible causes of overdispersion could be the variability of experimental design (this can be thought of as individual variability of the experimental units and may give an additional component of variability which is not accounted for by the basic model), the correlation between

individual responses, cluster sampling, aggregated data (the aggregation process can lead to compound distributions). It could be also generated by scale usage heterogeneity, subjective interpretation of wording or modalities, etc.
Other mixing distributions can be used, such as the normal distribution (Hinde, 1982) and also overdispersed distributions without a mixture interpretation are also available.
In this work we assume a finite mixture distribution for ordinal variables proposed in order to emulate the process of selection of subjects' responses in rating analyses (Iannario and Piccolo, 2012; Iannario 2012a). This mixture is able to introduce the ability to take account of a possible overdispersion within the same framework.
We test the ability of the approach by means of a real case study. We apply such a model on the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy. Since the observations in this study are not a simple random sample, overdispersion due to design effects is expected to be present. Furthermore, this extra variability might also be induced on account of random variation in response probabilities due, for example, to interview effects or, more generally, to rater's selection for each response.
The paper is organized as follows. Section 2 describes the mixture for ordinal data analysis with basic notations and inferential issues. Section 3 summarizes some empirical results. Few comments end the paper.

2 Model and statistical inference

The finite mixture model introduced by Piccolo (2003) is based on the psychological process of selection among m ordered alternatives. He proposed the presence of two main components: feeling and uncertainty. To model the pattern of the responses, he considers a shifted Binomial (for feeling component) and Uniform (for uncertainty) distributions (CUB models). Then, the basic model has been extended by means of a logistic link on parameters in order to capture a possible effect of subject covariates on the responses.
When observed data exhibit overdispersion, the replacement of Beta Binomial for the feeling component leads to CUBE models (Iannario, 2012b). Specifically, to study the pattern of an ordinal trait (rating response) in which data consist of an ordinal vector response $\boldsymbol{r}_i = (r_{i1}, r_{i1}, \ldots, r_{in})'$ for a given number $m > 4$ of categories and covariates t_i, $i = 1, 2, \ldots, n$ belonging to a matrix $\boldsymbol{T}$, we define the probability mass function of a CUBE model:

$$Pr(R = r_i) = \pi_i \, \beta e(\xi_i, \phi_i) + (1 - \pi_i) \, U \,, \tag{1}$$

where $\beta e(\xi_i, \phi_i)$ is the Beta Binomial distribution implied for the feeling

component and defined for $i = 1, 2, \ldots, n$ by:

$$\binom{m-1}{r_i-1} \frac{\prod_{k=1}^{r_i} [1-\xi_i+\phi(k-1)] \prod_{k=1}^{m-r_i+1} [\xi_i+\phi_i(r_i-1)]}{[1-\xi_i+\phi_i(r_i-1)]\,[\xi_i+\phi_i(m-r_i)] \prod_{k=1}^{m-1} [1+\phi_i(k-1)]},$$

and $U = \frac{1}{m}$ is the discrete Uniform distribution assumed for the uncertainty. Parameters $\boldsymbol{\theta} = (\pi_i, \xi_i, \phi_i)'$ are related to each respondent's covariates may be by means of the relationships:

$$\pi_i = 1/[1+\exp(-y_i\boldsymbol{\beta})]\,; \xi_i = 1/[1+\exp(-w_i\boldsymbol{\gamma})]\,; \phi_i = 1/[1+\exp(-x_i\boldsymbol{\alpha})]\,;$$

where $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}$ are parameter vectors to be estimated and the rows $\boldsymbol{y}_i$, $\boldsymbol{w}_i$, $\boldsymbol{x}_i$ are subset of t_i (here $t_0 = 1$).

For this class of models it is possible to consider the same covariates for the three components (feeling, uncertainty, overdispersion) without running into multicollinearity problems.

The parameter space is the positive octant in R^3 bounded over the unit square. Moreover, CUB models are nested in CUBE models when $\phi \equiv 0$. Generally, it is worth to say that a substantial overdispersion may be obtained with small values of the corresponding parameter ($\phi < 0.3$, say).

Indeed, a difference with respect to CUB models (which have a unique modal value) is the possibility to explain two opposite modes at $R = 1$ and $R = m$, respectively. Given the shape of the Beta Binomial distribution, it is immediate to observe that unimodal (bimodal) CUBE models with an intermediate mode arise when $\phi < 0.5$ ($\phi > 0.5$).

From an inferential point of view, given a sample of ordinal data, the log-likelihood function is expressed by:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \pi_i \left[\beta e(\xi_i, \phi_i) - \frac{1}{m} \right] + \frac{1}{m} \right\}. \tag{2}$$

Asymptotic inference on the parameter vector $\boldsymbol{\theta}$ for a CUBE model has been implemented (Iannario, 2012b) and an effective procedures for ML estimates has been obtained by the EM algorithm (available in the `R` environment).

Specifically, in order to check the significance of the estimates, we refer to the asymptotic theory of ML estimators and derive the asymptotic variance-covariance matrix $V(\boldsymbol{\theta})$ of ML estimators $\hat{\boldsymbol{\theta}}$ of the parameter vector $\boldsymbol{\theta}$. This matrix is obtained by inverting the opposite of the expectation of the second derivatives matrix of the log-likelihood function $\ell(\boldsymbol{\theta})$. This is based on the *expected information matrix* $\boldsymbol{I}(\boldsymbol{\theta})$.

Predictive and likelihood-based local/global fitting measures are generally based on likelihood ratio statistics or some omnibus measures as the Bayesian Information Criterion (BIC).

3 Empirical evidence of overdispersion

We check the proposed model for assessing a possible overdispersion on the SHIW data. The survey collects data on the economic behavior of Italian households. It also gathers other information regarding job, such as the level of specialization, of work experience and of qualification required from the activity, a set of questions concerning health, perceived happiness, economic perceived conditions and so on. The number of available observations for the empirical analysis consists of 1290 individuals.
More details about these data and some results obtained by means of alternative models for the analysis of ordinal data (ordinal probit and standard CUB models) are in Gambacorta and Iannario (2013), whereas details on the survey design and on the content of the questionnaire can be found in Faiella *et al.* (2008).

TABLE 1. Estimated CUB and CUBE models with covariates.

Covariates	CUB model	CUBE model
Constant	$\beta_0 = -1.824$*(0.324)*	$\beta_0 = -1.607$*(0.442)*
Education	$\beta_1 = 1.446$*(0.130)*	$\beta_1 = 1.656$*(0.195)*
Constant	$\gamma_0 = -1.593$*(0.023)*	$\gamma_0 = -1.536$*(0.025)*
Gender	$\gamma_1 = 0.133$*(0.033)*	$\gamma_1 = 0.139$*(0.036)*
Constant		$\alpha_0 = -3.541$*(0.259)*
Health		$\alpha_1 = 0.284$*(0.096)*
$\ell(\hat{\boldsymbol{\theta}})$	−6737.1	−6660.5
BIC	13507.0	13387.0

At the end of each interview, the interviewer answers to some topics about the quality of the interview. In this context, we model the response to *Global comprehension of questions* (ranging from 1=*low comprehension* to 10=*high comprehension*) by considering the impact of education (for uncertainty), gender (for feeling) and perceived health status (for overdispersion) of the respondents. The results of estimated CUB and CUBE models with covariates are reported in Table 1.
The inclusion of the overdispersion improves the global fitting of the model: we summarize these results in Figure 1 (left panel) by comparing CUB and CUBE model on parameter space (for CUBE models, the size of the circle is proportional to $\hat{\phi}$). Moreover, CUBE model reduces the uncertainty (measured by $1-\pi$) of CUB model but it does not substantially modify the degree of feeling (measured by $1-\xi$). Actually, this analysis stresses the crucial role of overdispersion.
It is also possible to create profiles (right panel) in which we compare interviewer's profile for given levels of covariates. Specifically, we compare the perception of responses of a high educated woman who considers for

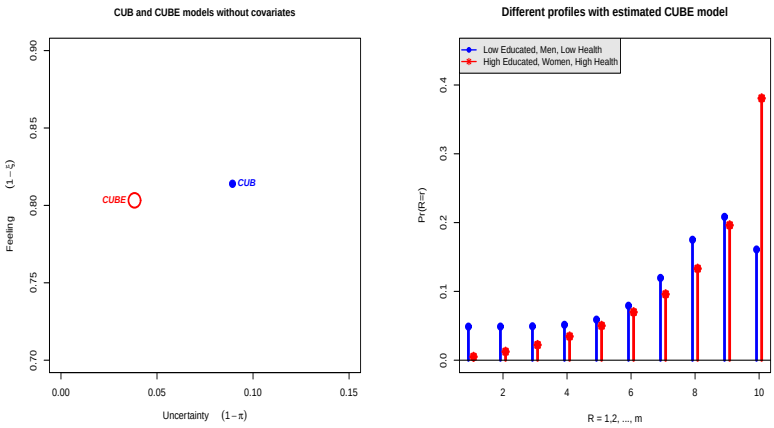


FIGURE 1. CUB and CUBE models for *Global comprehension of questions*.

herself an higher level of perceived health status (*Gender*=1, *Education*= post graduate, *Health*=8-9-10) with a low educated man who manifests a low level of perceived health (*Gender*=0, *Education*=compulsary education, *Health*=1-2-3). The right panel of Figure 1 shows how these profiles are different since the effect of covariates noticeably changes the degree of comprehension of questions.
In this way, the sample results are summarized with a direct interpretation of parameters. For instance, the perceived health reduces the overdispersion in the global perception of the quality of interview.

4 Discussion

As a final comment, we notice that the introduction of CUBE models allows for a unique framework for explaining different behaviours with a limited number of parameters.
Obviously, some critical issues arise from the circumstance that a further parameter requires a finer splitting of possible ordinal choices. Moreover, the model requires a more complex estimation procedure if we consider further extensions with the inclusion of covariates, hierarchical contexts and *shelter effect*. These improvements are to be pursued in future researches in order to fully exploit the capability of the class of CUBE models for interpreting and predicting data with a possible overdispersion.

Acknowledgments: This research has been partly supported by FIRB 2012 project "Mixture and latent variable models for causal-inference and analysis of socio-economic data" at University of Perugia.

References

Altham, P. M. E. (1976) Discrete variable analysis for individuals grouped into families, *Biometrika*, **63**, 263–269.

Cox, D. R. (1983) Some remarks on overdispersion. *Biometrika*, **70**, 269–274.

Cox, D. R. and Snell, E. J. (1989) *Analysis of Binary Data*, 2^{nd} edition. New York: Chapman and Hall.

Crowder, M. J. (1978) Beta-binomial ANOVA for proportions. *Applied Statistics*, **27**, 34–37.

Faiella, I. (2008) Accounting for sampling design in the SHIW. *Temi di Discussione*, **662**, Rome, Bank of Italy.

Finney, D. J. (1971) *Probit Analysis*, Cambridge: Cambridge University Press.

Gambacorta, R. and Iannario, M. (2013) Measuring job satisfaction with CUB models. *Labour*, **27**, DOI:10.1111/labr.12008.

Haseman, J. I. and Kupper, L. J. (1979) Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*, **35**, 281–293.

Hinde, J. (1982) Compound Poisson regression models. In GLIM82 (ed. R. Gilchrist), pp. 109–121. Berlin: Springer.

Haseman, J. I. and Kupper, L. J. (1979) Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*, **35**, 281–293.

Iannario, M. (2012a) Modelling *shelter* choices in a class of mixture models for ordinal responses. *Statistical Methods and Applications*, **21**, 1 – 22.

Iannario, M. (2012b) CUBE models for interpreting ordered categorical data with overdispersion. *Quaderni di Statistica*, **14**, 137 – 140.

Iannario, M. and Piccolo, D. (2012) CUB models: Statistical methods and empirical evidence. In: *Modern Analysis of Customer Surveys: with applications using R*. Kenett R. S. and Salini S. (ed.), Chichester: J. Wiley & Sons, pp. 231 – 258

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2^{nd} edition. London: Chapman & Hall.

Piccolo, D. (2003) On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85 – 104.

Williams, D. A. (1982) Extra-binomial variation in logistic linear models. *Applied Statistics*, **31**, 144–148.

Estimation of an MIC distribution using a two-stage semi-parametric mixture model

Stijn Jaspers[1], Marc Aerts[1], Geert Verbeke[2]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[2] Interuniversity Institute for Biostatistics and statistical Bioinformatics, KU Leuven, Leuven, Belgium

E-mail for correspondence: stijn.jaspers@uhasselt.be

Abstract: Antimicrobial resistance has become one of the main public health burdens of the last decades, and monitoring the development and spread of non-wild-type isolates has therefore gained increased interest. Monitoring is performed, based on the Minimum Inhibition Concentration (MIC) values, which are collected through the application of dilution experiments. A semi-parametric mixture model is presented, which is able to estimate the full continuous MIC distribution. The model is based on an extended and censored-adjusted version of the penalized mixture approach often used in density estimation. A data application and simulation study are presented in which the promising behaviour of the new method is illustrated.

Keywords: Antimicrobial resistance; Censoring; Penalized mixture approach; Semi-parametric

1 Introduction

Antimicrobial resistance (AMR) is the main undesirable side effect of antimicrobial use in both humans and animals. Due to the continuous positive selection of resistant bacterial clones, the population structure of microbial communities is modified. AMR has become one of the main public health burdens of the last decades and it is therefore extremely important to study and monitor the emergence of isolates with reduced susceptibility against antimicrobials. This may be performed by determining the minimum inhibition concentration (MIC), which is commonly measured via a broth dilution method. A standardized amount of the isolate is exposed to successive two-fold concentrations of the antimicrobial and the MIC is defined as the smallest concentration of the antimicrobial substance that inhibits the visible growth of the microorganism. Figure 1 shows an MIC distribution determined for 5190 isolates of *E. coli* tested for susceptibility against nalidixic acid. Note that, as a result of the dilution type laboratory experiments, MIC data are censored.

Our interest is in identifying the full continuous MIC distribution. In this regard, mixture models are ideally suited as they offer a natural framework for modelling unobserved population heterogeneity. In our context, a two-component mixture

$$f(x) = \pi f_1(x|\theta_1) + (1-\pi) f_2(x|\theta_2) \tag{1}$$

is assumed, in which f_1 and f_2 respectively represent the wild-type and non-wild-type component of the MIC distribution and the prevalence of wild-type isolates is denoted by π. The first component, representing the wild-type isolates, is assumed to be of a fixed parametric form and can hence be modelled parametrically. The second component, representing the non-wild-type isolates, is often multi-modal, and in this case, it is itself a mixture of different non-wild-type subpopulations. In order to impose as little constraints as possible, the second component will be left completely unspecified and a non-parametric estimate will be considered.

2 The semi-parametric mixture model

2.1 Estimation of the first component

Denoting by Y_i the number of times MIC value i was observed over the n trials, the observed MIC groupings can be seen as possible outcomes of $Y = (Y_1, \ldots, Y_k) \sim \text{Mult}(n,p)$, where k represents the number of different MIC categories and $p = (p_1, \ldots, p_k)$ such that $p_1 + \ldots + p_k = 1$. The maximum likelihood estimates for the multinomial probabilities p_i are just the observed relative frequencies $\frac{Y_i}{n}$. Nevertheless, the main interest remains in identifying the most suited parameters of the continuous wild-type distribution rather than those of the discrete multinomial distribution. This can be achieved by exploiting the fact that the observed groupings are actually the result of the censored readings of the dilution experiment. Hence the multinomial probabilities corresponding to a certain outcome i can be rewritten as

$$\begin{cases} \tilde{p}_i = \pi * F(u_i;\theta) \text{ if } i = 1, \\ \tilde{p}_i = \pi * [F(u_i;\theta) - F(l_i;\theta)] \text{ otherwise,} \end{cases} \tag{2}$$

where u_i and l_i are the respective upper and lower values of the ith MIC category and $F(.)$ represents the cumulative distribution function under consideration, with θ its corresponding parameters. In addition, the unknown parameter π accounts for the fact that the true MIC distribution is a mixture of the wild-type and non-wild-type component.
The idea is to replace an increasing number of the multinomial probabilities with their parametric counterparts in (2). The probabilities of the remaining outcomes are left unchanged and are thus to be estimated similar to

those of the saturated model (i.e. the observed relative frequencies). The resulting sequence of likelihoods is specified by

$$l_j(\tilde{p}_1, \ldots, \tilde{p}_{k_j}, p_{k_j+1}, \ldots, p_k) = \sum_{i=1}^{k_j} y_i \log \tilde{p}_i + \sum_{i=k_j+1}^{k} y_i \log p_i, \qquad (3)$$

with $j = 1, \ldots, k-2$ and where k_j indicates how many of the original multinomial probabilities are replaced: $k_j = j +$ length of ϕ, with $\phi = (\theta, \pi)$. This sequence can be maximized to obtain several proposal estimates for the parameters of interest. Note that as a result of the parametric assumption, less parameters are used in the construction of the likelihood when j increases. The AIC criterion can be applied to select the most appropriate parameter estimates or averaged estimates can be considered using the Akaike weights.

2.2 Estimation of the second component

The second component will be estimated using a censored-adjusted version of the penalized mixture approach by Schellhase and Kauermann (2012). Let X denote the univariate random variable of interest (i.e. the MIC value), with true density function f. The main idea is to approximate f as a mixture of densities:

$$f_K(x) = \sum_{k=-K}^{K} c_k \phi_k(x), \qquad (4)$$

where the $\phi_k(x)$ are the basis densities and the c_k are called the weights. In order to avoid constrained maximization, the weights are reparametrized:

$$c_k(\beta) = \frac{\exp(\beta_k)}{\sum_{k=-K}^{K} \exp(\beta_k)},$$

with $\beta_0 \equiv 0$ for identifiability. The basis densities are assumed to be Gaussian density functions, which are located at a fixed number of knots, corresponding to their respective means.

The number of knots plays an important role in terms of bias and variance. A compromise between smoothness and unbiasedness is obtained through the approach of Eilers and Marx(1996): a large number of basis functions is considered, but the log-likelihood is penalized for overfitting via a penalty term based on the finite differences of adjacent coefficients.

Assuming an independent sample $x_i, i = 1, \ldots, n$, the final log-likelihood to be optimized can be written as

$$l_p(\beta, \lambda) = \sum_{i=1}^{n} \log \sum_{k=-K}^{K} c_k \phi_k(x_i) - \frac{1}{2} \lambda \beta^T D_m \beta,$$

where D_m is the penalty matrix and λ is the smoothing parameter. In order to obtain estimates for the β parameters, Newton-Raphson scoring is performed, while the penalty parameter λ is updated using an estimating equation. For further details, see Schellhase and Kauermann (2012).
In order to take the censoring into account, the original basis density functions are replaced by their corresponding distribution functions:

$$l(\beta) = \sum_{i=1}^{n} \log \sum_{k=-K}^{K} [c_k \Phi_k(x_i) I(x_i \in MIC_{min}) + c_k \{\Phi_k(x_i) - \Phi_k(x_i - 1)\} I(x_i \notin MIC_{min})] .$$

Penalization and optimization of the likelihood are done similar to the original procedure.

2.3 The semi-parametric mixture model

The idea is to fix f_1 to the estimate obtained in the initial phase, using the method in Section 2.1. Information on the second component is then introduced through the censored-adjusted penalized mixture approach. More specifically, the estimator for the density of the MIC values is based on (4), to which one additional component is added:

$$f_K(x) = \pi f_1(x; \theta_1) + (1 - \pi) \sum_{k=-K}^{K} c_k \phi_k(x) = \sum_{k=-(K+1)}^{K} \tilde{c}_k \tilde{\phi}_k(x). \quad (5)$$

The additional component represents the wild-type component and will not be penalized as it is assumed to be fixed. Regarding the placement of the knots for the second component, recall that the model based on the likelihood in (3) is fitted to increasing subsets of the data. The fit with the smallest AIC value identifies the subset of the data that most likely belongs to the wild type component. In addition, due to the interval censoring, the highest MIC value in this subset identifies a possible lower bound for the MIC values of the non-wild-type isolates and will hence be used as the first knot of the basis. Finally, the estimator in (5) is used to construct the penalized likelihood. The adjustment for censored observations and the optimization occur in full similarity as above.

3 Application to real data

The two-stage procedure described in Section 2 is applied to MIC data obtained from the EUCAST website. The data concern the susceptibility of *E. coli* against nalidixic acid. Both a log-normal and a gamma distribution were assumed for the first component. The optimal mean and standard

deviation for the former were (on the $\log_2$-scale) 1.04 (se = 0.01) and 0.58 (se = 0.01), respectively. In case of a gamma first component, the shape and scale were 8.32 (se = 0.31) and 0.25 (se = 0.01), respectively. Figure 1 shows the result of the semi-parametric mixture model. The fixed gamma first component results into the lowest AIC and the estimated mixing weight ($\hat{\pi}$) is 0.87 (se = 0.01).

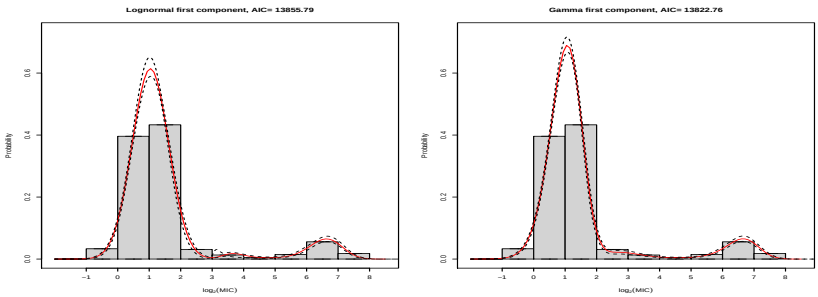


FIGURE 1. Barplot and estimates of the distribution of Minimum Inhibitory Concentrations (MIC) in *E. coli* isolates tested for susceptibility against nalidixic acid - source: EUCAST website.

4 Simulation study

Samples are taken from a mixture distribution with two main components. The wild-type component is assumed to be log-normally distributed with mean 2 and standard deviation 0.8. The non-wild-type component is a 50:50 mixture of two log-normal densities with (on the $\log_2$-scale) means equal to 4.5 and 7.5, respectively, and standard deviations equal to 0.7 and 0.6, respectively. The prevalence of wild-type isolates is taken to be 0.6. The multinomial based method is compared to the semi-parametric mixture model, assuming both a log-normal and gamma first component. The performance of the methods are compared based on the MSE values for the estimate of the prevalence of wild-type isolates. In addition, the Kullback-Leibler distance indicates the performance of the semi-parametric mixture model when estimating the entire mixture density. The considered sample sizes are 500, 1000 and 5000.

Since in real-life applications, the true underlying distribution of the first component is unknown, the averaged estimates of the two approaches should be regarded. From Table 1, it is seen that the semi-parametric mixture model outperforms the multinomial based method when estimating the prevalence of wild-type isolates. This is most pronounced in case of the larger sample size. The KL distance also indicates a promising behaviour of the semi-parametric mixture model.

TABLE 1. Summary of the simulation study. Presented are the averaged Kullback-Leibler distance (KL) and the MSE values when estimating π using the multinomial-based method (MSE_1) and the semi-parametric-mixture model (MSE_2), when assuming a log-normal and gamma first component, as well as for the most optimal fit using AIC.

Sample size	Assumed f_1	KL (s.e.)	MSE_1	MSE_2
500	Log-normal	0.022 (0.008)	0.0027	0.0020
	Gamma	0.023 (0.011)	0.0021	0.0022
	AIC	0.022 (0.008)	0.0028	0.0023
	Averaged	-	0.0017	0.0015
1000	Log-normal	0.012 (0.005)	0.0014	0.0010
	Gamma	0.015 (0.006)	0.0020	0.0024
	AIC	0.013 (0.005)	0.0016	0.0013
	Averaged	-	0.0010	0.0009
5000	Log-normal	0.004 (0.001)	0.0002	0.0002
	Gamma	0.009 (0.002)	0.0075	0.0094
	AIC	0.004 (0.001)	0.0018	0.0003
	Averaged	-	0.0014	0.0003

5 Discussion

A two-stage semi-parametric mixture model to estimate a continuous MIC distribution from censored observations was presented. In addition to an estimate for the prevalence of wild-type isolates, the model also provides an estimate for the non-wild-type distribution. Regarding the susceptibility of *E. coli* isolates against nalidixic acid, the prevalence of wild-type isolates was estimated to be 0.87 (0.01). Finally, a simulation study indicated a promising behaviour of the new method. Our ongoing research includes the simultaneous estimation of the first and second component.

Acknowledgments: Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. The research of the first author was supported by the Research Foundation Flanders (FWO), grant 11E2913N.

References

Eilers, P. and Marx, B. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, **11**, 89 – 121.

Schellhase, C. and Kauermann, G. (2012). Density estimation and comparison with a penalized mixture approach. *Computational Statistics*, **27**, 757 – 777.

Likelihood Analysis For An Incomplete Longitudinal Hemoglobin Data

Vandna Jowaheer[1], Brajendra Sutradhar[2], Rajendra Neupane[2]

[1] University of Mauritius , Reduit, Mauritius
[2] Memorial University, St. John's, Canada

E-mail for correspondence: vandnaj@uom.ac.mu

Abstract: When a univariate response such as the hemoglobin level resulting from iron intake are collected over a period of time, the repeated responses become correlated. Also it frequently happens that a few responses from an individual may be intermittently missing, making it difficult to compute the correlations of such non-patterned available responses. If one, however, ignores the correlations and uses independence assumption based likelihood approach, the resulting estimates would be inefficient. In this paper by considering a class of autocorrelations for the complete data, we demonstrate how to compute the correlations of the available responses and construct a likelihood function by pooling the unbalanced multivariate distributions of the individuals caused by the intermittent missing mechanism. This likelihood methodology is used to analyse a hemoglobin data set collected over 3 months interval for 5 time periods from 42 infants from a children hospital in St. John's, Canada.

Keywords: Correlations of repeated responses; Intermittently missing responses; Likelihood for incomplete data; Regression effects.

1 Introduction

Infants of very low birth weight (less than $1500g$) are at high risk for iron deficiency because of low stores of iron at birth (Gortem and Cross (1964)). Friel et al (1990) examined the iron status of very-low-birth-weight infants fed with iron-fortified formula during early infancy. But this study did not accommodate the longitudinal correlations among the repeated hemoglobin responses of the same infant. Further problems arise when such data are collected repeatedly but a portion of observations are missing intermittently. To address these two issues clearly, we develop a likelihood approach by blending such missing mechanism and correlation structure of the repeated data and reanalyse the longitudinal hemoglobin data. This we will do following Krisnamoorthy and Pannala (1996) except that these authors dealt with a non-regression setup whereas we deal with a linear regression model involving several covariates such as the gender, gestational age, and BHGB of the infants.

2 Incomplete Gaussian Model

Let y_{it} be the hemoglobin level recorded at the $t^{th}(t = 1, 2, \ldots, T)$ occasion for the $i^{th}(i = 1, \ldots, K)$ infant. In the present data set, $T = 5$ and $K = 42$. Also let $x_{it} = (x_{it1}, \ldots, x_{itu}, \ldots, x_{itp})'$ be the $p \times 1$ covariate vector corresponding to y_{it}. For the present hemoglobin data, there are $p = 5$ covariates, namely (i) intercept covariate (x_{it1}), (ii) gender (x_{it2}), (iii) formula (or treatment) (x_{it3}), (iv) gestation week (x_{it4}), and (v) baseline hemoglobin (BHGB) (x_{it5}), and these covariates are time independent. Let $\beta = (\beta_1, \ldots, \beta_p)'$ be the $p \times 1$ regression effects of x_{it} on y_{it}, for all $i = 1, \ldots, K;\ t = 1, \ldots, T$. It is of interest to estimate this β parameter using all available responses and by accommodating the correlations among the available repeated responses for all 42 infants. Among 42 infants, all together there are 25 infants with complete responses for 5 time points. The remaining 17 infants have at least one missing response. Let g denote the g^{th} group and n_g denote the number of infants in that g^{th} group for $g = 1, \ldots, 6$.

Let $y_{i(1)}$ be the T_1 dimensional vector containing all $T = T_1 = 5$ repeated observations for the $i^{th}(i = 1, \ldots, 25)$ infant of the first group. In general, let $y_{i(g)}$ denote the T_g dimensional vector of responses for the ith infant of the g^{th} group. Here $T_g \leq T_1 (= T)$. Let $\mu_{i(g)}$ and $\Sigma_{i(g)}$ denote the expectation and covariance matrix of $y_{i(g)}$, respectively. Suppose that $y_{i(1)}$, the full dimensional response vector, follows the Gaussian distribution as

$$Y_{i(1)} \sim N_{T_1}(\mu_{i(1)}, \Sigma_{i(1)}), \tag{1}$$

where, for $T_1 = T$,

$$\mu_{i(1)} = (\mu_{i1}, \ldots, \mu_{it}, \ldots, \mu_{iT})' \text{ and}$$
$$\Sigma_{i(1)} = (\sigma_{iut},\ u, t = 1, \ldots, T) = A_{i(1)}^{\frac{1}{2}} C_{i(1)} A_{i(1)}^{\frac{1}{2}}, \tag{2}$$

with

$$\mu_{it} = x_{it}'\beta, \text{ for } t = 1, \ldots, T;\ A_{i(1)} = \sigma_i^2 I_T,$$

σ_i^2 being the variance of the repeated responses for the i^{th} individual, I_T being the $T \times T$ identity matrix, and $C_{i(1)}$ is the full dimensional $(T \times T)$ general auto-correlation matrix defined as

$$C_{i(1)}(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{T-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \cdots & 1 \end{bmatrix}, \tag{3}$$

[Sutradhar (2011, eqn. (2.44), p. 20)] where for $\ell = 1, \ldots, T$, ρ_ℓ is known to be the ℓth lag auto-correlation. Note that there is no need of assuming any

particular correlation structure such as AR(1) (auto-regressive of order 1) and MA(1) (moving average of order 1, and so on.
For other groups $g = 2, \ldots, 6,$, the T_g dimensional mean vector $\mu_{i(g)}$ may be computed by deleting the rows of the $\mu_{i(1)} = (\mu_{i1}, \ldots, \mu_{it}, \ldots, \mu_{iT})'$ vector corresponding to the missing responses. Similarly, the $T_g \times T_g$ covariance matrix $\Sigma_{i(g)}$ and/or the correlation matrix $C_{i(g)}(\rho)$ may be computed by deleting the rows and columns of the $T \times T$ matrices $\Sigma_{i(1)}$ and/or $C_{i(1)}(\rho)$ corresponding to the missing responses. Note that the normality for $y_{i(g)}$ is still valid, i.e.,

$$Y_{i(g)} \sim N_{T_g}(\mu_{i(g)}, \Sigma_{i(g)}), \text{ for } g = 2, \ldots, 6. \tag{4}$$

For example, for the aforementioned third group of the hemoglobin data set, the $T_3 = 4$ dimensional vector $y_{i(3)} = (y_{i1}, y_{i2}, y_{i4}, y_{i5})'$ has the 4 dimensional normal distribution with mean vector $\mu_{i(3)} = (\mu_{i1}, \mu_{i2}, \mu_{i4}, \mu_{i5})'$ and correlation matrix

$$C_{i(3)}(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_3 & \rho_4 \\ \rho_1 & 1 & \rho_2 & \rho_3 \\ \rho_3 & \rho_2 & 1 & \rho_1 \\ \rho_4 & \rho_3 & \rho_1 & 1 \end{bmatrix}. \tag{5}$$

3 Incomplete Likelihood Estimation

In a simpler non-regression setup, a multivariate model similar to that of the last section was fitted by Krishnamoorthy and Pannala (1999) by using a likelihood approach. We use this approach in the present regression setup. Also note that unlike in Krishnamoorthy and Pannala (1999), we deal with a univariate problem in a longitudinal setup which lead to a specialized form of correlation structure (3) as opposed to an open multi-normal correlation structure. By denoting the total number of groups with G (where $G = 6$ for the hemoglobin data) and the number of individuals in the g^{th} group with n_g, one may write the likelihood function for all $\sum_g^G n_g = 42$ individuals as

$$L(\beta, \rho) = \Pi_{g=1}^{G} \Pi_{i=1}^{n_g} N(y_{i(g)} | \mu_{i(g)}, \Sigma_{i(g)}). \tag{6}$$

Now for known correlations, the maximization of the likelihood function (6) with respect to β is, in fact, equivalent to minimization of the quadratic function $Q = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(y_{i(g)} - \mu_{i(g)}\right)' \Sigma_{i(g)}^{-1} \left(y_{i(g)} - \mu_{i(g)}\right)$, where $\mu_{i(g)} = X_{i(g)}\beta$ with $X_{i(g)}$ as the $T_g \times p$ covariate matrix for the i^{th} individual belonging to the g^{th} group. This minimization yields the incomplete data based likelihood estimator of β as

$$\hat{\beta}_{ML,inc} = \left[\sum_{g=1}^{G} \sum_{i=1}^{n_g} X'_{i(g)} \Sigma_{i(g)}^{-1} X_{i(g)}\right]^{-1} \left[\sum_{g=1}^{G} \sum_{i=1}^{n_g} X'_{i(g)} \Sigma_{i(g)}^{-1} y_{i(g)}\right], \tag{7}$$

with its variance computed by

$$\text{var}[\hat{\beta}_{ML,inc}] = \left[\sum_{g=1}^{G}\sum_{i=1}^{n_g} X'_{i(g)}\Sigma^{-1}_{i(g)}X_{i(g)}\right]^{-1}. \tag{8}$$

Note that the lag correlations and variances involved in $\Sigma_{i(g)}$ for all $g = 1, \ldots, G$, in (7) are unknown. Also note that $C_{i(1)}$ contains all lag correlations, namely $\rho_1, \ldots, \rho_\ell, \ldots, \rho_{T-1}$, whereas $C_{i(g)}$ for $g = 2, \ldots, G$, will contain only a portion of these auto-correlations. For convenience we provide a general moment estimator for ρ_ℓ for all $\ell = 1, \ldots, T-1$, and use them as needed to construct $C_{i(g)}$ for all $g = 1, \ldots, G$. Consider the indicator variable r_{it} defined as $r_{it} = 1$, if y_{it} is present; otherwise $r_{it} = 0$, for all $i = 1, \ldots, K; t = 1, \ldots, T$. For known β and σ_i, the ℓth lag correlation estimate $\hat{\rho}_\ell$ for the larger correlation matrix $C_{i(1)}$ is given by

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^{K}\sum_{t=1}^{T-\ell} r_{it}r_{i,t+\ell}[\left(\frac{y_{it}-x'_{it}\beta}{\sigma_i}\right)\left(\frac{y_{i,t+\ell}-x'_{it,t+\ell}\beta}{\sigma_i}\right)]/\sum_{i=1}^{K}\sum_{t=1}^{T-\ell} r_{it}r_{i,t+\ell}}{\sum_{i=1}^{K}\sum_{t=1}^{T} r_{it}[\frac{y_{it}-x'_{it}\beta}{\sigma_i}]^2/\sum_{i=1}^{K}\sum_{t=1}^{T} r_{it}}, \tag{9}$$

[Sutradhar (2011, eqn. (2.40), p. 18)], where for $r_{it} = 0$, it is not necessary to compute $z_{it} = [(y_{it} - x'_{it}\beta)/\sigma_i]$ as this quantity does not contribute toward ρ_ℓ computation. Instead, for simplicity one can use $z_{it} = 0$ without any loss of generality. Furthermore, the variance σ_i^2 is computed by using the moment formula

$$\hat{\sigma}^2_{i,inc} = \frac{\sum_{t=1}^{T} r_{it}z^2_{it}}{\sum_{t=1}^{T} r_{it}}. \tag{10}$$

4 An Application to the Hemoglobin Data

We apply the likelihood method discussed in the previous sections to obtain the effects of the covariates on the incomplete hemoglobin data with $G = 6$. The results are reported in Table 1. The autocorrelations are also displayed in the same table. The lag 1 correlation 0.232 is considerably high which makes a significant contribution in regression estimation, more importantly to the standard errors of the regression estimates. Now to interpret the effects of the main covariates, it is clear that the iron intake (treatment), in general, affect the hemoglobin level over time significantly with positive coefficient 0.646. Because treatment is coded as 1, this positive value shows that the larger amount of iron intake increases the hemoglobin level as expected. The positive gender effect (2.395) indicates that the hemoglobin levels of the male infants are more than the female infants.

Note that unlike other covariates, gestation week was found to have negative regression effect (-0.265) on the hemoglobin level. This however does not mean that the infants with larger gestation week had smaller

Table 1: *Estimates for regression and autocorrelation parameters for incomplete hemoglobin data*

	Estimation Approach			
	Incomplete Lik. ($K = 42$)		**Complete Lik.** ($K = 25$)	
Param.	**Est.**	**St.Err.**	**Est.**	**St.Err.**
Interc.	123.3	0.937	117.2	0.170
Gender	2.395	0.142	4.697	0.177
Treat.	0.646	0.144	1.850	0.176
GestWeek	-0.265	0.030	-0.215	0.048
BHGB	0.004	0.003	0.008	0.004
ρ_1	0.232		0.149	
ρ_2	0.008		0.005	
ρ_3	-0.118		-0.178	
ρ_4	-0.380		–0.401	

hemoglobin level. To understand this we first notice from the raw data that the BHGB is more for infants with larger gestation week. Now it is seen by using this negative effect that the predicted hemoglobin for the infants with smaller gestation week has increased (as compared to BHGB) to a large extent, whereas the increase in hemoglobin level is moderate or small for the infants with higher gestation week. This explains the role of the negative effects of the gestation weeks. Finally, the small positive value (0.004) of baseline hemoglobin indicates that the predicted hemoglobin was higher for the infants with higher baseline hemoglobin.

In the same Table 1, we also report the regression effects computed from the first group ($G = g = 1$), that is, based on the complete data from 25 individuals. The standard errors of the estimates are generally larger as compared to the corresponding standard errors computed based on the incomplete but larger data from 42 individuals, indicating less efficiency for the complete data based estimates. Note that the aforementioned regression estimates based on more data (from all 42 individuals) are naturally more reliable than their counterparts computed from only 25 individuals.

5 Conclusion

We have developed a likelihood approach to analyze incomplete longitudinal data and implemented this approach to analyse an incomplete longitudinal hemoglobin data. For other existing approaches mainly for longitudinal monotonic missing data analysis, one may, for example, refer to Birmingham et al (2003), Sutradhar and Mallick (2010) and the references therein. As a future work, this likelihood approach can be compared with

the existing approaches.

References

Birmingham, J., Rotnitzky, A., and Fitzmaurice, G. M. (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society Series B*, **65,** 275–297.

Firel, J. K., Andrews, W. L., Matthew, J. D., Long, D. R., Cornel, A. M., Cox, M., & Skinner, C. T. (1990). Iron status of very-low-birth-weight infants during the first 15 months of infancy. *Canadian Medical Association Journal,* **143 (8)**.

Gortem, K. M. & Cross, E. R. (1964). Iron metabolism in premature infants: Prevention of iron deficiency. *Pediatrics,* **64**, 509-520.

Krishnamoorthy, K., and Pannala, M. K. (1999). Confidence estimation of a normal mean vector with incomplete data. *The Canadian Journal of Statistics,* **27,** 395-407.

Sutradhar, B. C. (2011). *Dynamic Mixed Models for Familial Longitudinal Data.* Springer, New York.

Sutradhar, B. C., and Mallick, T. S. (2010). Modified weights based generalized quasilikelihood inferences in incomplete longitudinal binary models. *Canadian Journal of Statistics, Special issue,* Eds. B. Sutradhar, **38,** 217-231.

Analysis of voter transition using ecological data: Comparison of different approaches for Munich election data

André Klima[1], Helmut Küchenhoff[1], Paul W. Thurner[2]

[1] Department of Statistics, Ludwig-Maximilians-Universität München, Germany
[2] Geschwister-Scholl-Institute of Political Science, Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: Andre.Klima@stat.uni-muenchen.de

Abstract: The standard estimation of voter transition is based on individual survey data. It is not only expensive, but also problematic due well known sources of bias. On the contrary, exact aggregate data are available for the level of electoral districts. Due to the well known ecological fallacy, direct estimates of individual voter transition is not possible without further assumptions. We present a comparison of different strategies of so-called ecological inference. We show how Bayesian methods work in this context. Further, prior knowledge and the possibilities of using extra information in this context is discussed.

Keywords: ecological inference; Bayes; voter transition, Munich

1 Introduction

In addition to the gross results of elections, a main topic of interest is voter transitions: How many voters remained loyal? How many voters chose the exit option, i.e. decided to vote for a different party, or abstained altogether? Such considerations are part of the news coverage of elections in Germany. Despite the public interest, only one big German institute, Infratest dimap, provides this information based on exit polls. Other smaller institutes also provide voter transitions using aggregated data.
Collecting individual data for estimating voter transitions in appropriate size and quality is expensive and the data may not be error free and reliable. Potential sources of bias are well known, e.g. recall bias, non response bias, false reporting and abstention. Such biases are also indicated by differences between recall data and panel data. But also with a sufficient and reliable data set for the entire election area, analysis of voter transitions for subregions can be problematic. In contrast, aggregated election data are cheap. In democratic systems such data is often provided by the statistical offices. The data include the voting behavior of all individuals. While this

data is exact and complete for each election, it lacks a direct link between both elections. Only the margins of a voter transition table are known.
In situation, when only the table margins are known, but the inner cells are the properties of interest, strategies of so-called ecological inference must be used. Estimating voter transitions using aggregated data is such a case: only the margins, the elections results for each election independently, are known, but the inner cells of the table, the combined elections behavior of the voters, are the quantities of interest. Additional assumption are necessary to perform this task and the crucial liability of them has already been highlighted in the literature. (Tam Cho 1998, Tam Cho, Manski 2008) Ecological inference is not only used for the analysis of electoral volatility and the voting behavior of social groups, but also in the context of epidemiology and consumer studies.
The presented results are from a project sponsored by the city of Munich. It had the aim to examine existing methods for ecological inference to estimate voter transitions only with aggregated data. To compare the different methods, we used the real aggregated election data from the federal elections 2005 and 2009 for Munich. Because the proposed model must be usable in the election night, the feasibility in this context was also an important factor.

2 Considered Methods for Ecological Inference

The following approaches were selected for the comparison because of their past usage, methodological developments and readiness for implementation:
Goodman's Ecological Regression is a linear regression using the shares of the parties in election one as independent variable and the share in election two as dependent variable. To ensure identifiability, constancy of the parameters over the districts is assumed. The coefficients can be interpreted as the share of voters, which voted for a respective party in election one and the party represented with the covariable in election two. It is common that some estimates are not within the logical bounds, therefore usually adjustments of the raw parameters are necessary.(Tam Cho, Manski 2008 and King 1997)
Thomsen's logit approach proposes the usage of the ecological correlation as estimate for the individual tetrachoric correlation. Thomsen derives this result from an individual logit model and an ecological logit model. His model is suitable for the 2x2 case, but Thomsen also provides an iterative algorithm, which allows inference in the RxC case. As part of this iterative algorithm, a reference party must be chosen.(Thomsen 1987)
More recently hierarchical models for ecological inference have been proposed. One of them is the Multinomial-Dirichlet model proposed by Rosen, Jiang, King and Tanner (2001). Instead of assuming constancy of the parameters over all districts, the models assumes a joint dirichlet distribution.

A multinomial distribution is assumed for the individual level. An R implementation of the model is available in the package eiPack.(Lau, Moore, Kellermann 2007)
Andreadis and Chadjipadelis (2009) proposed an iterative algorithm using 2x2 methods for the estimation of RxC voter transitions. In each iteration a fraction of the voters is explained, the considered 2x2 table is chosen by the correlation weighted by the number of voters. For their implementation, they use a 2x2 method proposed by Grofman and Merrill (2004). Thomas Kellermann proposes a slight adjustment of the algorithm: He recommends to force the algorithm to estimate all loyalty rates in the first steps one time before estimating other combinations.(Kellermann 2011)

3 Analysis using the Munich Elections results

A first objective is an assessment of the impact of the model parameters on the estimates. A second aim was the evaluation of the consequences of different ways of data pre-processing. In this first step we rely on aggregate data only. Thus, the assessment is based on a cross-model comparison and a plausibility check.
We observe a strong effect of the chosen adjustment method for the out of bounds estimates of the Goodman regression. We compared a cut off method - only the values out of bounds are cut, and a relative scaling method - all parameters are scaled relative to their extrema. Thomsen method exhibits a strong reference party dependency. The Multinomial-Dirichlet model requires high figures for burnin and thinning in the MCMC. Even with 1000 as thinning value there was still a moderate autocorrelation identifiable in some chains. The chains itself are quite stable, but sometimes a temporarily departure from the general trend is visible. Using the Multinomial-Dirichlet model in Andreadis and Chadjipadelis iterative algorithm dramaticly changes the estimates, indicating a sensitivity with regards to the chosen 2x2 method.
We also compared two easy methods treating differences in the number of eligible voters and handling of postal voters. We are able to show that the data handling impacts the estimation. Additional analyses, including the usage of higher order aggregations of the data, were also performed. While a moderate higher aggregation only influenced the estimation by the Multinomial-Dirichlet model clearly negative, a strong aggregation impacted all considered models.
Comparing the different methods, we found that the results from Goodman's regression and Thomsen's logit approach are different from the other methods, with Thomsen a bit closer to the rest. Some of the estimated loyalty rates with the iterative models were considered too high to be plausible.

4 Simulation Study

The second step of our analysis was the simulation of individual data using the original election results from 2005 and the federal voter transition published by Infratest dimap. We considered four different scenarios: Our first scenario follows the Goodman assumption and uses fixed parameters for all districts, the second follows a Multinomial-Dirichlet structure and draws the district parameter from a joint dirichlet distribution, the third assumes three subpopulation in each district differing slightly in their election behavior 2005 and their party loyalty and the last assumes a connection between the loyalty of the voters of the two big parties (SPD and CSU) and the election results 2005.
In the first scenario the Goodman regression outperforms all other models with the lowest overall error, but with exception of Thomsen's method all models have acceptable estimates. In the second scenario the best model is the Multinomial-Dirichlet model, followed by both iterative methods. The Goodman regression and Thomsen are inferior, with the latter being the worst model. In the third and fourth simulation, the Multinomial-Dirichlet model again is superior. The Goodman regression and both iterative approaches show similar performance, while Thomsen only shows in the fourth scenario comparable results. The second and the fourth scenario are the ones where all methods showed lowest performance.
With additional simulations we tested the dependency of the estimates on the aggregation level, the handling of the postal voters and changes in the electorate. The base data for this consideration were simulation two and four. The major results are that the used aggregation level has an impact, postal voters should not be ignored and changes in the electorate do not exclude sensible estimates.

5 Prior Knowledge and Survey Data

In situations with an identification problem, as in our case, the usage of prior knowledge seems suitable and justified, therefore an additional focus was the inclusion of prior knowledge to the preferred Multinomial-Dirichlet model. One option is the specification of the starting values for the dirichlet level of the model using prior knowledge. If no district knowledge is available, the starting values at the multinomial level have been specified using the same values. The performance of the models without specified starting values and the models with "true" starting values for the parameters at the dirichlet level are quite similar, no real improvement was visible.
In a second step, we also added prior knowledge at a lower level and provided "true" values for ten percent of the districts. As starting values for the remaining districts, we used the average of the known ten percent. This time there are differences, the model with used district level prior knowledge

has a much better performance, clearly outperforming the model with no prior knowledge. Additional tests with deliberately selected wrong starting values at the dirichlet level showed that such wrong prior knowledge can influence the results negatively. Adding the "truth" for ten percent of all districts while still maintaining the wrong prior knowledge at the dirichlet level lead again to results closer to the "truth".
Of course, such district prior knowledge will be more difficult to collect. But usage of external knowledge, e.g. polling data, is possible. Making use of polling data as prior knowledge for the specification of starting values is one option to incorporate individual data to ecological inference. Other alternatives are to utilize the prior knowledge for choosing informative priori distributions, which is subject to further research, or to include the individual information in so called hybrid models (e.g. Greiner, Quinn, 2010).

6 Discussion

Choosing the right method for the election night estimation of voter transitions is a difficult task. The "truth" is in general unknown and a lot of potential tables are plausible. We therefore analyzed the methods with the original data and with simulated data.
Our analysis using the original data showed strong difference between the different approaches, but the real performance of the methods could not be estimated. Some unrealistic high loyalty rates already indicated that some models do not hit the truth, but without individual data a further evaluation was not possible. The second step therefore used simulated data, in this case we knew the "truth" and could therefore evaluate the performance of the methods. The Multinomial-Dirichlet model is superior compared to the other candidate models in our simulations.
Using additional prior knowledge can boost the performance and lead to drastically improved results. If this prior knowledge is gathered through individual data, combined models can also be evaluated. But of course individual data will not always be available, be it because of financial limitations or because of the impossibility to gather such data. The former will be most likely the case for subregions while the latter is typical in historical elections analyses.

Acknowledgments: Special thanks to the town of Munich for the project and the funding of the analysis. Additional thanks go to Thomas Schlesinger and Christoph Molnar for their help conducting the simulation study.

References

Andreadis, I., Chadjipadelis, T. (2009). A Method for the Estimation of Voter Transition Rates. *Journal of Elections, Public Opinion and Parties*, **19**, 203 – 218.

Tam Cho, W. K., Manski, C. (2008). Cross Level/Ecological Inference. In: *The Oxford Handbook of Political Methodology* eds. Box-Steffensmeier, J.M., Brady, H.E., Collier, D.. New York: Oxford University Press.

Tam Cho, W. K. (1998). Iff the Assumption Fits...: A Comment on the King Ecological Inference Solution. *Political Analysis*, **7**, 143 – 163.

Greiner, D.J., Quinn, K.M. (2010). Exit Polling and Racial Bloc Voting: Combining Individual-Level and R x C Ecological Data. *The Annals of Applied Statistics*, **4**, 1774 – 1796.

Grofman, B., Merrill, S. (2004). Ecological Regression and Ecological Inference. In: *Ecological Inference: New Methodological Strategies*. eds. King, G., Tanner, M.A., Rosen, O.. Cambridge: Cambridge University Press.

Kellermann, T. (2011). Vom Wahlergebnis zur Wählerwanderung. *Stadtforschung und Statistik*, **1**, 34 – 40.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Date*. Princeton, NJ: Princeton University Press

Lau, O., Moore, R.T., Kellermann, M. (2007). eiPack: RxC Ecological Inference and Higher-Dimension Data Management. *R News*, **7**, 43 – 47.

Rosen, O., Jiang, W., King, G., Tanner, M.A. (2001). Bayesian and frequentist inference for ecological inference: the R x C case. *Statistica Neerlandica*, **55**, 134 – 156.

Thomsen, S.R. (1987). *Danish Elections 1920-79: A Logit Approach to Ecological Analysis and Inference*. Aarhus: Politica

Bayesian Expectile Regression

Thomas Kneib[1], Elisabeth Waldmann[1], Fabian Sobotka[1]

[1] Chair of Statistics, Georg-August-University Göttingen

E-mail for correspondence: tkneib@uni-goettingen.de

Abstract: Recent interest in the development of flexible regression specifications has had a specific focus on describing more complex features of the response distribution than only the mean. The standard instrument in this situation is quantile regression where conditional quantiles are related to a regression predictor. Computationally this is achieved by minimizing an asymmetrically weighted absolute residuals criterion which induces additional complexity compared to standard least squares optimization. As a consequence, expectile regression that relies on asymmetrically weighted squared residuals has gained considerable interest since expectile regression estimates can be obtained by iteratively weighted least squares fits. In this abstract, we introduce a Bayesian formulation of expectile regression that relies on the asymmetric normal distribution as auxiliary response distribution. Suitable proposal densities for the resulting Markov chain Monte Carlo simulation algorithm are proposed and the potential of the approach for extending the flexibility of expectile regression towards complex semiparametric regression specifications is discussed.

Keywords: asymmetric normal distribution, iteratively weighted least squares proposals, Markov chain Monte Carlo simulation, quantile regression, semiparametric regression

1 Expectile Regression

Suppose that regression data $(y_i, \boldsymbol{z}_i)$, $i = 1, \ldots, n$, on a continuous response variable y and a covariate vector $\boldsymbol{z}$ are given and shall be analyzed in a regression model of the form

$$y_i = \eta_{i\tau} + \varepsilon_{i\tau}$$

where η_τ is a predictor formed by the covariates and ε_τ is an appropriate error term. Unlike in mean regression where regression effects on the mean are of interest, we focus on situations where specific outer parts of the response distribution shall be studied. We will denote the extremeness of these outer parts by the asymmetry parameter $\tau \in (0, 1)$ where $\tau = 0.5$ corresponds to the central part of the distribution while $\tau \to 0$ and $\tau \to 1$ yield the lower and upper part of the distribution, respectively. The standard approach

for implementing such regression models is quantile regression where we assume that the τ-quantile of the error distribution equals zero, i.e.

$$P(\varepsilon_{i\tau} \leq 0) = \tau.$$

This implies that the predictor $\eta_{i\tau}$ corresponds to the τ-quantile of the response y_i and the regression model can be estimated by minimizing

$$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau})|y_i - \eta_{i\tau}|$$

with asymmetric weights

$$w_\tau(y_i, \eta_{i\tau}) = \begin{cases} 1-\tau & y_i \leq \eta_{i\tau} \\ \tau & y_i > \eta_{i\tau}. \end{cases}$$

To avoid numerical difficulties associated with the absolute deviations in the quantile regression specification, we will instead focus on the criterion

$$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau})(y_i - \eta_{i\tau})^2 \tag{1}$$

that yields expectile regression estimates. This criterion has the advantage to be differentiable with respect to the regression predictor so that estimates can be obtained by iteratively weighted least squares estimation. Basically, expectiles are an alternative possibility to characterize the distribution of a continuous random variable where τ indexes the "extremeness" of the part of the distribution that shall be studied, see Newey and Powell (1987).
A usual objection against expectiles as compared to quantiles is their lack of an immediate interpretation. While for quantiles, the property that $\tau 100$ percent of the data lie below the regression line and $(1-\tau)100$ percent of the data lie above the regression line is easy to understand, the extremeness of expectiles is hard to transfer to such an easy statement. However, interpretation of expectiles is still possible in the following ways:

- For i.i.d. data $y_1, \ldots, y_n$, the resulting expectile estimate will be a weighted average
$$\hat{e}_\tau = \sum_{i=1}^{n} w_i y_i$$
where the weights w_i depend on the estimated expectile. As a consequence, regression expectiles can also be considered such weighted average conditioned on a specific covariate vector.

- Expectiles are tail expectations, i.e. the τ expectile fulfills
$$\tau = \frac{\int_{-\infty}^{e_\tau} |y - e_\tau| f(y) dy}{\int_{-\infty}^{\infty} |y - e_\tau| f(y) dy}$$
showing that e_τ is characterised by a partial moment condition.

- Usually, one would not only estimate one single expectile but a whole set of expectiles for various values of τ. The collection of all estimates then gives an intuitive impression about the shape of the conditional distribution of the response and in particular allows to detect features such as heteroscedasticity, asymmetry or skewness. Moreover, conditional quantiles can still be calculated from a set of expectiles if quantile estimates are of ultimate interest.

In summary, albeit having a different (and may be less intuitive) interpretation than quantiles, expectiles are probably not more difficult to interpret than a variance.

2 Asymmetric Normal Distribution

To make expectile regression accessible in a Bayesian formulation, we require the specification of an auxiliary response distribution that yields a likelihood that is equivalent to the optimization criterion (1). For Bayesian quantile regression, this can be formalized based on the asymmetric Laplace distribution, see for example Waldmann et al. (2013). For expectile regression, the analogous distribution is an asymmetric normal distribution

$$y_i \sim \text{AN}(\eta_i, \sigma^2, \tau)$$

with density

$$p(y_i) = \frac{2}{\sqrt{\sigma^2 \pi}} \left(\sqrt{\frac{1}{1-\tau}} + \sqrt{\frac{1}{\tau}} \right) \exp\left(-\frac{1}{2\sigma^2} w_\tau(y_i, \eta_{i\tau})(y_i - \eta_{i\tau})^2 \right).$$

Maximising the likelihood arising from this distributional specification is then equivalent to minimizing (1).

3 Semiparametric Regression

Instead of only considering linear regression specifications, we are interested in applying expectile regression in the context of general semiparametric regression models with predictor

$$\eta_i = \beta_0 + \sum_{j=1}^{p} f_j(\boldsymbol{z}_i)$$

where we suppressed the index τ for notational simplicity, β_0 is an intercept representing the overall level of the predictor, and the functions $f_j(\boldsymbol{z}_i)$ reflect different types of regression effects depending on subsets of the covariate vector $\boldsymbol{z}_i$. For the regression functions f_j, we make the following assumptions:

- The functions f_j are approximated in terms of basis function representations
$$f_j(\boldsymbol{z}) = \sum_{k=1}^{K} \beta_{jk} B_k(\boldsymbol{z})$$
where $B_k(\boldsymbol{z})$ are the basis functions and β_{jk} denote the corresponding basis coefficients.

- The prior for the vector of basis coefficients $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK})'$ is a multivariate normal distribution with density
$$p(\boldsymbol{\beta}_j|\delta_j^2) \propto \exp\left(-\frac{1}{2\delta_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right)$$
where the precision matrix $\boldsymbol{K}_j$ represents different types of structural assumptions about the function f_j such as smoothness. Note that the prior may be partially improper if the precision matrix $\boldsymbol{K}_j$ is not of full rank.

This framework covers, among others, penalized splines, Markov random fields, individual-specific random effects, interaction surfaces based on either radial basis function or tensor product splines, and varying coefficient terms as special cases and therefore provides a convenient generalization of additive (mixed) models, see Fahrmeir, Kneib and Lang (2004).

4 Bayesian Inference

We complete the Bayesian specification by assuming inverse gamma priors for the error variance and the smoothing variances, i.e.

$$\sigma^2 \sim \text{IG}(a_0, b_0) \qquad \delta_j^2 \sim \text{IG}(a_j, b_j).$$

Given the model specification, this implies that the full conditionals are also inverse gamma with updated parameters. In contrast, the full conditionals for the regression coefficients $\boldsymbol{\beta}_j$ are not available in closed form since unfortunately a normal prior in combination with an asymmetric normal observation models does not induce an asymmetric normal full conditional. We therefore construct proposal densities based on the penalized iteratively weighted least squares updates that would have to be performed to compute penalized expectile regression estimates in a frequentist backfitting procedure, i.e.

$$\hat{\boldsymbol{\beta}}_{j\tau}^{[t+1]} = (\boldsymbol{B}_j'\boldsymbol{W}_\tau^{[t]}\boldsymbol{B}_j + \lambda_j\boldsymbol{K}_j)^{-1}\boldsymbol{B}_j'\boldsymbol{W}_\tau^{[t]}(\boldsymbol{y} - \boldsymbol{\eta}_{-j,\tau}),$$

where $\boldsymbol{B}_j$ is the design matrix associated with the j-th model term, $\boldsymbol{y}$ is the vector of responses, $\boldsymbol{\eta}_{-j,\tau} = \boldsymbol{\eta}_\tau - \boldsymbol{B}_j\boldsymbol{\beta}_j$ is the complete predictor

without the jth component and $\boldsymbol{W}_\tau = \text{diag}(w_\tau(y_1, \eta_{1\tau}), \ldots, w_\tau(y_n, \eta_{n\tau}))$. More precisely, we propose a new state for $\boldsymbol{\beta}_j$ from the normal distribution $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with expectation and covariance matrix given by

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j \boldsymbol{B}_j' \boldsymbol{W}_\tau (\boldsymbol{y} - \boldsymbol{\eta}_{-j,\tau}) \quad \text{and} \quad \boldsymbol{\Sigma}_j = (\boldsymbol{B}_j' \boldsymbol{W}_\tau \boldsymbol{B}_j + \lambda_j \boldsymbol{K}_j)^{-1}$$

where $\lambda_j = \sigma^2/\delta_j^2$ is the smoothing parameter obtained as the ratio of error scale parameter and smoothing variance.

5 Discussion

The Bayesian formulation of expectile regression outlined in this abstract provides both the Bayesian counterpart to frequentist expectile regression and the expectile analogue to Bayesian quantile regression. While standard semiparametric regression specifications in expectile regression can already be handled in a frequentist setting based on iteratively weighted least squares estimation, the Bayesian formulation opens up the possibility to include more complex regression specifications such as Dirichlet process mixture priors for random effects or Bayesian regularisation priors using a conditional Gaussian prior structure as suggested for Bayesian quantile regression in Waldmann et al. (2013). Moreover, Bayesian expectile regression comprises the determination of the smoothing variances δ_j^2 as an integral part of the inferential procedure and provides measures of uncertainty also for complex functionals of the model parameters. However, the asymmetric normal likelihood will usually induce a model misspecification and the impact of this misspecification will have to be studied in detail in simulations.

Acknowledgments: Financial support by the German Research Foundation (DFG), grant KN 922/4-1 is gratefully acknowledged.

References

Fahrmeir, L., Kneib, T. and Lang, S. (2004): Penalized structured additive regression for space-time data: a Bayesian perspective, *Statistica Sinica*, **14**, 731 – 761.

Newey, W.K. and Powell, J.L. (1987): Asymmetric least squares estimation and testing, *Econometrica*, **55**, 819 – 847.

Sobotka, F. and Kneib, T. (2012): Geoadditive expectile regression, *Computational Statistics & Data Analysis*, **56**, 755 – 767.

Waldmann, E., Kneib, T., Yue, Y.R., Lang, S. and Flexeder, C. (2013): Bayesian semiparametric additive quantile regression, *Statistical Modelling*, to appear.

Reduced-bias inference for multi-dimensional Rasch models with applications

Ioannis Kosmidis[1], David Firth[2]

[1] University College London, UK
[2] University of Warwick, UK

E-mail for correspondence: i.kosmidis@ucl.ac.uk

1 Multi-dimensional Rasch models

The current methodological work is motivated by an attempt to place the members of the US House of Representatives on a "liberality" scale, using multi-dimensional Rasch models and data on the voting behaviour of $S = 435$ representatives on $I = 20$ roll-calls selected by *Americans for Democratic Action* (ADA, for short). There were an extra 4 representatives that did not vote to at least a quarter of the roll calls and have been left-out of the analysis as essentially uninformative. A vote against ADA's position is noted with 0 and a vote for the ADA's position is noted with 1. Rasch models are popular models in Item Response Theory (IRT) for analysing binomial outcomes in a subject-item arrangement. Specifically, consider observations $y_{11}, \ldots, y_{1S}, \ldots, y_{1S}, \ldots, y_{IS}$ on independent random variables $Y_{11}, \ldots, Y_{1S}, \ldots, Y_{1S}, \ldots, Y_{IS}$ where Y_{is} has a Binomial distribution with probability π_{is} and index n_{is} $(i = 1, \ldots, I; s = 1, \ldots, S)$. An m-dimensional Rasch model has the form

$$\log \frac{\pi_{is}}{1 - \pi_{is}} = \eta_{is} = \alpha_i + \sum_{j=1}^{m} \beta_{ji}\gamma_{js} \quad (i = 1, \ldots, I; s = 1, \ldots, S)\,. \qquad (1)$$

Here π_{is} can be thought of as the probability that member s votes for the ADA's position on roll-call i $(i = 1, \ldots, I; s = 1, \ldots, S)$. The vectors of unknown parameters β represents the roll-calls' discrimination and, if all βs have the same sign then the vector of parameters γ represents the members' "liberalities". In this representation each of the discrimination and liberality parameters are decomposed into m dimensions. For example if $m = 2$, a potential analysis is to make a plot of the estimates of $\gamma_{11}, \ldots, \gamma_{1S}$ versus the estimates of $\gamma_{21}, \ldots, \gamma_{2S}$, in order to obtain a visualization of the liberality of the members in two dimensions.
If $m = 1$ in (1) then the log-odds is equated to $\alpha_i + \beta_i\gamma_s$ which is the two-parameter logistic model (2PL), and if $\beta_i = 1$ $(i = 1, \ldots, I)$ then the simple one-parameter logistic (1PL) model results.

2 Estimating liberality

Fitting the one-dimensional model (2PL) using maximum likelihood results to infinite parameter estimates because there are representatives voting for or against the ADA position in all roll calls. A solution would be to add a small positive constant to the responses and twice that constant to the totals and then treat the adjusted data as actual. While this can result in finite estimates i) it is then hard to quantify the effect of the adjustment to the fitted model, and ii) the choice of the constant adjustment is generally arbitrary and different choices of constants would generally give different results (see Kosmidis, 2013, for a thorough discussion).
Simulation studies on models for categorical responses illustrate that the bias reduction method in Firth (1993) offers a solution to the problems relating to boundary estimates; see, for example, Heinze & Schemper (2002), Kosmidis & Firth (2011) and Kosmidis (2013) for binomial and multinomial-response generalized linear models. The method proceeds via the adjustment of the log-likelihood derivatives (score functions).

3 The reduced-bias estimator

Let $\delta = (\alpha^T, \beta^T, \gamma^T)^T$ be the p-vector of parameters of model (1), where $p = I + m(I + S)$ is the number of model parameters. It is necessary to have $m(m+1)$ particular constraints for making the model identifiable. Let $p_E = I + m(I + S - m - 1)$ denote the number of effective parameters in the model. For notational simplicity, the dependence of quantities on the model parameters is suppressed. Furthermore, for generality of the results with respect to the many possibilities in choosing the required constraints, a set of p adjusted score functions is derived from which only the p_E that correspond to unconstrained parameters are effective. The inverse of the Fisher information on the p_E estimable parameters is extended to a $p \times p$ matrix by adding zero rows and columns for the constrained parameters. We denote the extended inverse of the Fisher information by F^{-1}. Then, using the results in Kosmidis & Firth (2009) and after some algebra, the t-th bias-reducing adjusted score function ($t = 1, \ldots, p$) is

$$U_t^* = \sum_{i=1}^{I} \sum_{s=1}^{S} \left\{ y_{is} + \frac{1}{2} h_{is} - (n_{is} + h_{is})\pi_{is} + c_{is} v_{is} \right\} z_{ist} \,. \qquad (2)$$

In the above expression h_{is} is the s-th diagonal element of the $S \times S$ projection matrix $H_i = Z_i F^{-1} Z_i^T \Sigma_i$ where Z_i is the $S \times p$ matrix with elements $z_{ist} = \partial \eta_{is} / \partial \delta_t$, and Σ_i is a diagonal matrix, with i-th diagonal element $v_{is} = \mathrm{Var}(Y_{is}) = n_{is}\pi_{is}(1 - \pi_{is})$. Also, $c_{is} = \sum_{j=1}^{m} \mathrm{AsCov}(\beta_{ji}, \gamma_{js})$, where the asymptotic covariances $\mathrm{AsCov}(\beta_{ji}, \gamma_{js})$ are obtained by the appropriate components of F^{-1} ($i = 1, \ldots, I; s = 1, \ldots, S$). For the 1PL model $m = 1$

and $\beta_1 = \ldots = \beta_I = 1$ and $c_{is} = 0$. Hence, the adjusted score functions result by adding $h_{is}/2$ to y_{is} and h_{is} to n_{is} $(i = 1, \ldots, I; s = 1, \ldots, S)$, which is exactly the result obtained in Firth (1993) for the reduction of bias in logistic regressions.
Comparing the form of U_t^* in (2) with the tth derivative of the log-likelihood $U_t = \sum_{i=1}^{I} \sum_{s=1}^{S} (y_{is} - n_{is}\pi_{is}) z_{ist}$ $(t = 1, \ldots, p)$, reduction of bias results by replacing the responses y_{is} and totals n_{is} with their adjusted counterparts

$$y_{is}^* = y_{is} + \frac{1}{2} h_{is} + n_{is} c_{is} \pi_{is} I(c_{is} \geq 0)\,,$$
$$n_{is}^* = n_{is} + h_{is} + n_{is} c_{is} \{\pi_{is} - I(c_{is} < 0)\} \quad (i = 1, \ldots, I; s = 1, \ldots, S)\,,$$

with $I(A)$ taking value one if condition A is satisfied, and 0 else.
Noting that the above adjustments satisfy $0 \leq y_{is}^* \leq n_{is}^*$ and that y_{is}^* and n_{is}^* depend on the model parameters $(i = 1, \ldots, I; s = 1, \ldots, S)$, an convinient iterative scheme for solving the adjusted-score equations involves the following two steps at each iteration:

1. Evaluate the adjusted responses and adjusted totals at the current estimates.

2. Fit model (1) to the adjusted data using maximum likelihood.

The reduced-bias estimates are a stationary point of the above procedure. The second step of each iteration can be conveniently performed using the `gnm` R package.
The maximum likelihood estimates are good starting values provided that they are finite. If at least one of those turns out being infinite then the binomial data can be adjusted by small constants and use the resultant finite estimates as starting values.

4 Liberality scales

The one-dimensional Rasch model is fitted on the voting records using the fitting procedure in Section 3. The necessary for bias reduction identifiability constraints are set by randomly selecting one of the βs and one of the γs and fixing them at their maximum likelihood estimates based on data that have been adjusted by small constants in order to ensure finiteness. All the estimates of the discrimination parameters turn out being negative. Hence, the smaller each of the liberality parameters γs is the larger the probability of voting for the ADA's position is and the more liberal is the representative.
Figure 1 shows the liberality estimates in increasing order along with the corresponding comparison intervals (Firth & De Menezes, 2004). In this case, the quasi-variance approximation is reasonable with relative errors ranging from -3% to 3.6%. So the depicted comparison intervals can safely

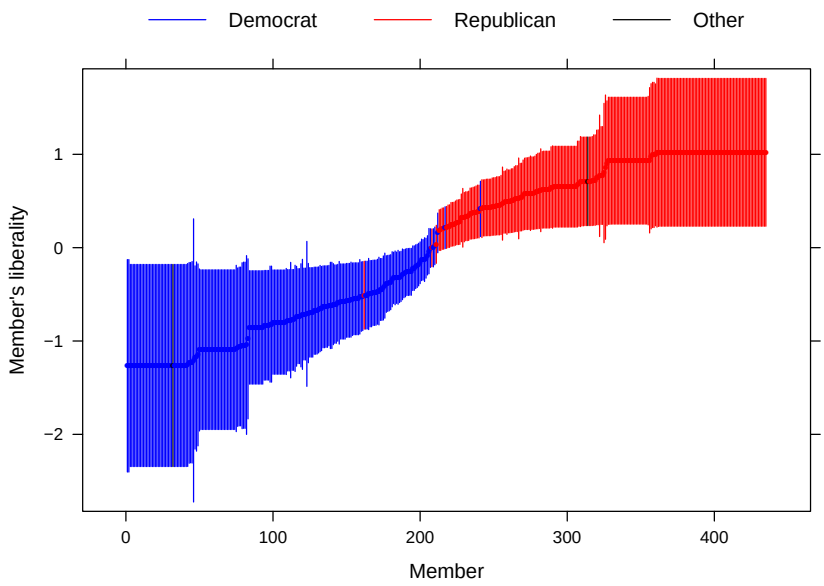


FIGURE 1. Liberality estimates with accompanying comparison intervals.

be used to check whether the difference in liberality between any two members is significant; if the comparison intervals for any pair of members are non-overlapping then there is a significant difference in their liberality. The colouring of the points and intervals is chosen according to the party information, which is also available in the data: blue is used for Democrats, red for Republicans and black for other. As is apparent, except for a few representatives, the one-dimensional Rasch model does well in placing the representatives in their respective parties.

Bias reduction is also used to fit a two-dimensional model to check whether an enhanced liberality scale can be produced. Identifiability constraints are set through the parameters β_{1i}, β_{2i}, β_{1k}, β_{2k}, γ_{1s}, and γ_{2s} for randomly chosen $i, k \in \{1, \ldots, 20\}$, $i \neq k$ and for a randomly chosen $s \in \{1, \ldots, 435\}$. These parameters are then fixed to their maximum likelihood estimates based on data that have been adjusted by small constants in order to ensure finiteness. Each point in Figure 2 represents the reduced-bias estimates of the pair $(\gamma_{1s}, \gamma_{2s})$ $(s = 1, \ldots, S)$. The two-dimensional model also performs rather well in placing the representatives into their respective parties. In fact, the two-dimensional model seems to be a better fit than the one-dimensional one; the reduction in deviance when moving from the one-dimensional model to the two-dimensional model is 850.7 on 451 degrees of freedom.

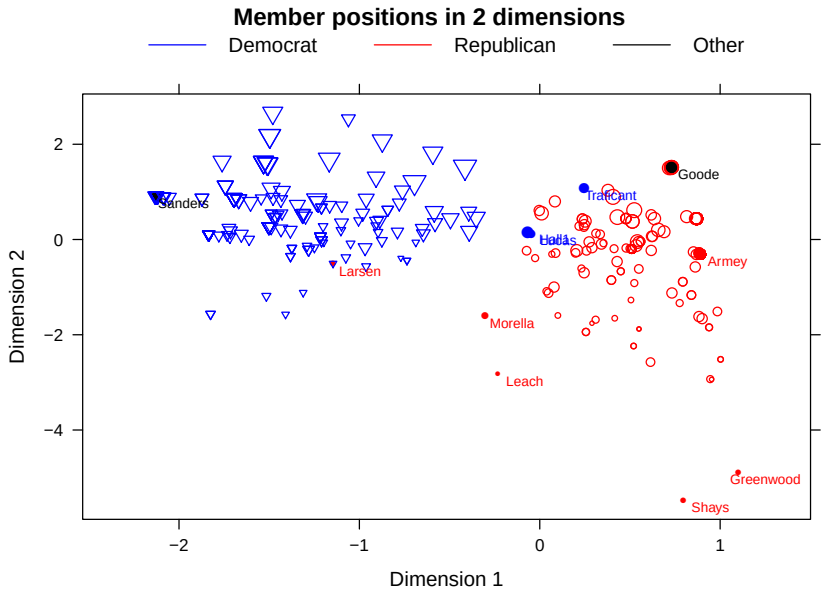


FIGURE 2. Liberality estimates from a two-dimensional Rasch model.

Some interesting points which seem to stand out from their party groups have been highlighted by printing the respective member name on the top right of each point. For example, Larsen is the only representative who was recorded as a Republican and who lies well within the group of Democrats; Larsen is actually a Democrat and ADA recorded Larsen's party incorrectly. Sanders is an independent representative who in most cases in his terms in the House of Representatives voted for Democrats. Sanders voted for ADA's poisition on all 20 roll-calls and the corresponding point overlaps perfectly with the points of all Democrats who did so. Correspondingly, Armey is placed on the other extreme of the liberality scale along with all other Republicans who voted against the ADA's position on all 20 roll calls. Goode is neither Democrat nor Republican but his is placed well within the group of Republicans in the liberality scale. Goode was a Democrat before 2000 and after 2002 he became a Republican. Between 2000 and 2002 he was independent. His position on the liberality scale according to his voting record in 2001 clearly reflects the later transitional period. Greenwood, Leach and Shays are positioned within the Republicans group but are located far from the main cloud of the points for the Republicans. Greenwood is well-known for being a Republican with moderate-to-liberal views on social matters and conservative views on economic matters.

For the specific constraints chosen and using the fact that the reduced-bias

estimator is equivariant under affine transformations, the picture in Figure 2 can be arbitrarily scaled, rotated and translated without affecting the improved bias properties of the estimators. Hence, one of the two dimensions of the plot seems to roughly capture the overall political orientation of the representative. The interpretation of the second dimension can be inferred noting the extreme position for Greenwood within the Republicans. After grouping the 20 roll calls into economic and into social matters based on their descriptions in ADA, the difference between the percentages of votes for the ADA's position on social matters and on economic matters is used to characterize the balance of the social and economic views of each representative. A large negative difference provides evidence of a representative that is quite conservative on economic matters and quite liberal on social matters, a large positive difference provides evidence for the inverse, and a zero difference refers to a representative who is as conservative/liberal on economic matters as is on social matters. The sizes of the points in Figure 2 depend on the size of that difference. Hence the plot reveals that one of the dimensions roughly captures the overall political orientation of the representative and the other the "balance" between the social and economic views of the representative.

References

Americans for Democratic Action (2013). Voting records. (`http://www.adaction.org/pages/publications/voting-records.php`)

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409 – 2419.

Firth, D. and R. X. de Menezes (2004). Quasi-variances. *Biometrika* **91**, 65 – 80.

Kosmidis, I. (2013). Improved estimation in cumulative link models. *ArXiv e-prints 1204.0105*. Accepted for publication in the *Journal of the Royal Statistical Society, Series B*.

Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family non- linear models. *Biometrika* **96** (4), 793 – 804.

Kosmidis, I and D. Firth (2011). Multinomial logit bias reduction via Poisson log-linear model. *Biometrika*, **98**, 755 – 759.

Turner, H. and D. Firth (2012). Generalized nonlinear models in R: An overview of the gnm package. (R package version 1.0-6). (`http://CRAN.R-project.org/package=gnm`).

A new flexible family of conditional Archimedean copulas

Philippe Lambert[1,2]

[1] Institut des sciences humaines et sociales, Univ. de Liège, Belgium
[2] Institut de statistique, biostatistique et sciences actuarielles (ISBA), Univ. catholique de Louvain, Belgium.

E-mail for correspondence: p.lambert@ulg.ac.be

Abstract: A new family of conditional Archimedean copula is presented. It can be used to describe how the association between variables changes with another covariate. Flexible specifications based on splines are proposed with properties investigated using simulations. Inference is made within the Bayesian paradigm with posterior distributions explored using adaptive MCMC algorithms. The modelling strategy is illustrated with the study of the association between height and weight in young boys.

Keywords: Conditional copula ; B-splines ; Growth curve.

1 Introduction

Sklar (1959) has proved that any distribution $H(y_1, \ldots, y_p)$ with marginal distributions $F_j(y_j)$ $(j = 1, \ldots, p)$ can be written as

$$H(y_1, \ldots, y_p) = C(F_1(y_1), \ldots, F_p(y_p)), \tag{1}$$

where C denotes a distribution function (named *copula*) on $(0, 1)^p$ with uniform margins. If the margins are continuous, then C is unique. Conversely, if C is a copula and $F_j(\cdot)$ are distribution functions, then Equation (1) defines a multivariate distribution with marginal distributions F_j $(j = 1, \ldots, p)$.
In most practical applications where copula are used, the marginal distributions and their potential link with covariates $\mathbf{x}$ are investigated in a first step, yielding marginal fitted quantiles, $\hat{u}_{j|x} = \hat{F}_j(y_j|\mathbf{x})$ $(j = 1, \ldots, p)$. A parametric copula is then selected to describe the dependence structure of the fitted quantiles. That copula is usually assumed to be independent of the covariates as if the strength of association between the margins did not change with subject characteristics. Then, Equation (1) becomes

$$H(y_1, \ldots, y_p|\mathbf{x}) = C(F_1(y_1|\mathbf{x}), \ldots, F_p(y_p|\mathbf{x})), \tag{2}$$

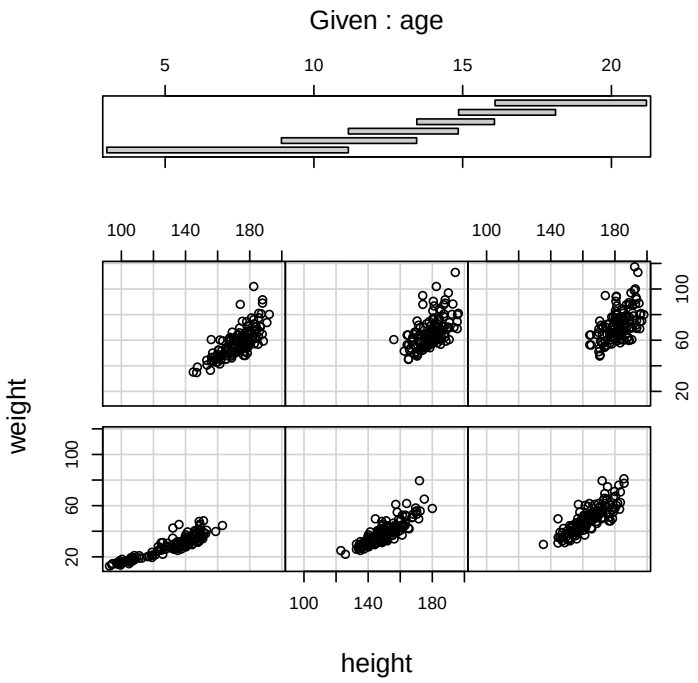


FIGURE 1. Conditional plot showing the decreasing association between weight and height as age increases.

2 Conditional copula

Unfortunately, the previous modelling assumption is not always realistic, as shown on Fig. 1 where the scatterplot of weight and height of young boys is given for different age classes. Indeed, one easily notice that the link between the marginal quantiles is getting looser as age increases. When the selected copula C_θ is parametric, one can let the copula parameter (and, hence, the underlying Kendall's tau) change with covariates, yielding

$$H(y_1, \ldots, y_p|\mathbf{x}) = C_{\theta(\mathbf{x})}(F_1(y_1|\mathbf{x}), \ldots, F_p(y_p|\mathbf{x})), \tag{3}$$

An early example of that can be found in Lambert & Vandenhende (2002) where the effect of an antidepressant on blood pressures and heart rate were studied in a longitudinal setting. Besides the effects of covariates on the marginal distributions of these 3 responses, their strengths of association were also allowed to change with sex and the presence of drug in the plasma. The same idea was used in a financial context by Patton (2006) where the name *conditional copula* for $C_{\theta(\mathbf{x})}$ was coined.
Nonparametric versions are desirable to suggest or to validate parametric specifications, or even as a substitute for these models.

3 The flex-power Archimedean family

We made a first nonparametric proposal in Lambert (2012). There, the copula was assumed Archimedean (Genest & MacKay, 1986). Such copulas are specified by the choice of a decreasing and convex function $\varphi(\cdot)$ (named the *generator*) taking values on $(0,1)$ and such that $\varphi(0^+) = +\infty$ and $\varphi(1) = 0$. Kendall's tau can be computed from the generator using

$$\tau = 1 + 4\int_0^1 \lambda(s)ds,$$

where $\lambda(\cdot) = \varphi(\cdot)/\varphi'(\cdot)$. It directly connects Kendall's tau to θ when the generator is parametric ($\varphi = \varphi_\theta$).
A (spline based) nonparametric estimate of $\varphi(\cdot)$ was proposed in Lambert (2007). An alternative version with superior properties and embedding the proposal made by Vandenhende & Lambert (2005) can be written as

$$\varphi(u) = \exp\{-g(S(u)|\theta)\} \tag{4}$$

where $S(u) = -\log(-\log(u))$ and $g'(s) = \sum_k b_k(s)\theta_k$ is a linear combination of cubic B-splines associated to a large number of equidistant knots on $(S(\epsilon), S(1-\epsilon))$.
That flexible form for the generator can be generalized by letting the spline coefficients change smoothly with covariates. A first extension was considered in Lambert (2012) by taking

$$\theta_k(x) = \mathrm{e}^{\eta(x)}\theta_k \tag{5}$$

with $\eta(x)$ expressed as a linear combination of B-splines $\{b_k^*(x)\}_{k=1}^K$ on the domain $\mathcal{X}$ of the covariate values. It directly affects Kendall's tau and relates it to the covariate by

$$\frac{1-\tau(x)}{1-\tau(x_0)} = \mathrm{e}^{\eta(x_0)-\eta(x)}.$$

Simulations suggests that it is not flexible enough with a limited ability to deal with settings where Kendall's tau oscillates between large and small values when the covariate changes.
It motivated the introduction of a new set of conditional copulas that we name the *power Archimedean family*. It relies on the following result (see e.g. Nelsen, 1999): if $\varphi(\cdot)$ is an Archimedean copula generator, then

1. $\varphi_{\alpha,1}(t) = \varphi(t^\alpha)$ is also a generator if $\alpha \in (0,1]$;
2. $\varphi_{1,\beta}(t) = (\varphi(t))^\beta$ is also a generator if $\beta \geq 1$.

Let us name these two operations *internal* and *external* transforms, respectively. The conditional model in Eq. (5) turns to be an external transform of the reference generator $\varphi(\cdot)$ where $\beta = \beta(x) = \mathrm{e}^{\eta(x)}$.
We suggest to consider both transforms to model changes of the copula generator with covariates, i.e. to take

$$\varphi_{\alpha,\beta}(t|x) = \left[\varphi(t^{\alpha(x)})\right]^{\beta(x)} \tag{6}$$

with flexible forms for

- the reference generator $\varphi(\cdot)$, see Eq. (4) ;
- the internal power

$$\alpha = \alpha(x) = \left[1 + \left(\sum_k b_k^*(x)\alpha_k\right)^2\right]^{-1} ;$$

- the external power

$$\beta = \beta(x) = 1 + \left(\sum_k b_k^*(x)\beta_k\right)^2 .$$

4 Applications

Simulations studies were performed to assess the ability of the *flex-power Archimedean family* to capture changes in the dependence structure of bivariate responses. The data pairs were generated using a parametric Archimedean copula with parameter θ forced to change with a covariate x in $(0, 1)$ to ensure that the conditional Kendall's tau is

$$\tau(x) = .5 + .3\sin(3\pi x) . \tag{7}$$

A Metropolis-within-Gibbs algorithm with adaptive proposals was used to explore the joint posterior of the spline parameters appearing in the reference copula and in the internal and external powers. One can see from Fig. 2 that the fitted Kendall's tau is a good estimate of the functional used in the simulation, even with modest sample sizes. Not surprisingly, the precision and the uncertainty respectively increases and decreases with sample size. Similar results were obtained for data generated using other Archimedean copulas.
The flex-power Archimedean copula model was also applied on the growth data mentioned in Section 1 (see also Fig. 1). The margins were first modelled using the nonparametric additive location-scale model described in Lambert (2013). The proposed conditional copula was then applied on the fitted quantiles. The resulting estimated conditional Kendall's tau is plotted with a 95% credible region in Fig. 3. It confirms and quantifies the decreasing association between weight and height suspected from Fig. 1.

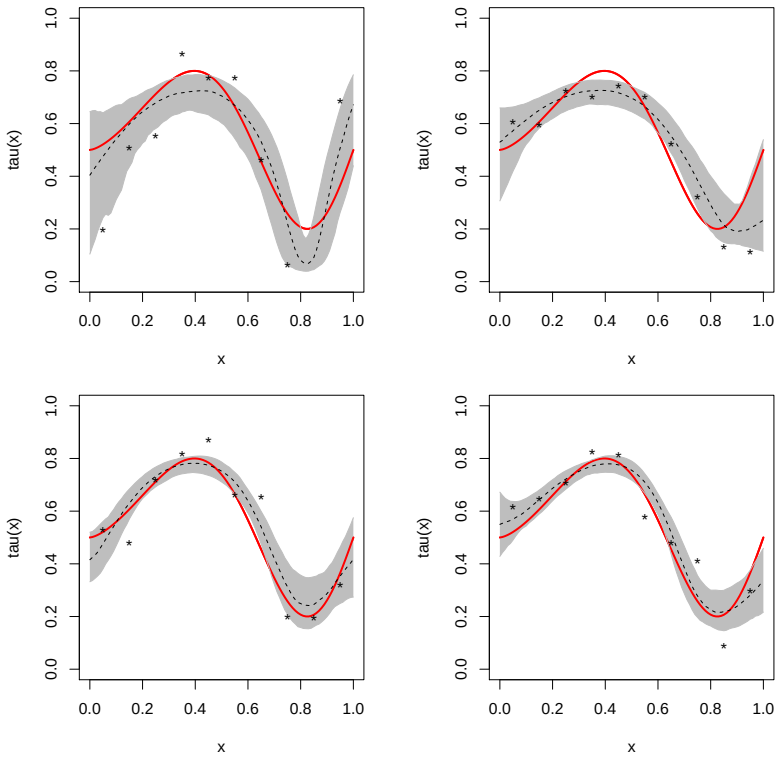


FIGURE 2. Fitted (dashed line) conditional Kendall's tau (with 95% credible region, in grey) for n pairs of data ($n = 100, 200, 400, 600$, from upper-left to bottom-right) simulated using Joe's copula with $\tau(x)$ given by Eq. (7) (solid red line). The stars show the observed values of tau for values of x between consecutive deciles.

Acknowledgments: The author acknowledges financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

References

Genest, C. and MacKay, J. (1986) Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canad. J. Statist.*, **14**: 145-159.

Lambert, P. (2013) Nonparametric additive location-scale models for interval censored data. *Statistics and Computing*, **23**: 75–90.

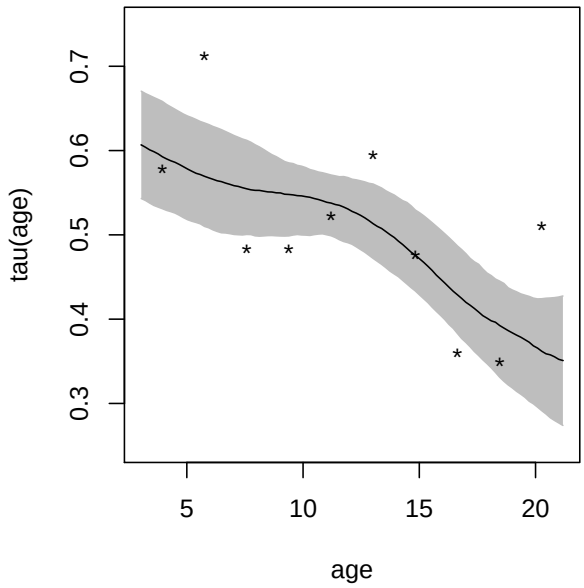


FIGURE 3. Growth dataset: fitted conditional Kendall's resulting from the flex-power Archimedean copula model.

Lambert, P. (2012) Nonparametric estimation of conditional Archimedean copula. In: *27th International Workshop on Statistical Modelling*, Prague, Czech Republic.

Lambert, P. (2007) Archimedean copula estimation using Bayesian splines smoothing techniques. *Computational Statistics and Data Analysis*, **51**: 6307-6320.

Lambert, P. and Vandenhende, F. (2002) A copula based model for multivariate non normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine*, **21**: 3197-3217.

Nelsen, R.B. (1999) *An Introduction to Copulas*. Springer, Berlin.

Patton, A.J. (2006) Modelling asymmetric exchange rate dependence. *International Economic Review*, **47**: 527-556.

Sklar, A. (1959) Fonctions de répartition à n dimensions et leurs marges *Publ. Inst. Statist. Univ. Paris 8* : 229-231

Vandenhende, F. and Lambert, P. (2005). Local dependence estimation using semiparametric Archimedean copulas *Canad. J. Statist.*, **33**: 377-388.

An empirical Bayesian ridge approach to modeling the transcriptional effects of DNA copy number aberrations

Gwenaël G.R. Leday[1], Aad W. van der Vaart[2], Mark A. van de Wiel[1]

[1] VU University, Amsterdam, Netherlands
[2] Mathematical Institute, Leiden, Netherlands
[3] VU Medical Center, Amsterdam, Netherlands

E-mail for correspondence: g.g.r.leday@vu.nl

Abstract: DNA copy number aberrations are a hallmark of cancer cells. These aberrations, focal or broad, consist in gains and losses of chromosomal DNA. These may alter directly expression levels of mRNA transcripts that map to the aberration or indirectly those that are located outside. We here present a Bayesian multivariate model for the joint estimation of direct (in cis) and indirect (in trans) transcriptional effects of DNA copy number aberrations.

Keywords: DNA copy number; mRNA expression; ridge; empirical Bayes.

1 Model

Consider gene expression and aCGH profiling of n independent tumor samples and the availability of the following data:

y_j	the vector of normalized mRNA expression values for gene j
x_j	the vector of segmented DNA copy number values for gene j
s_j	the vector of copy number states ("double loss", "loss", "normal", "gain" and "amplification", coded by -2, -1, 0, 1 and 2) for gene j .

Then, for gene $j \in \{1, \cdots, p\}$, the model is

$$y_j = f_{\alpha_j}(x_j; \theta_{j,s}) \ + \sum_{\substack{k=1 \\ k \neq j}}^{p} \beta_{j,k} y_k + \epsilon_j \tag{1}$$

$$\theta_{j,s} \sim \mathcal{N}(\mu_{j,s}, \gamma_{j,s}^2 I_2) \tag{2}$$

$$\beta_{j,k} \sim \mathcal{N}(0, \tau_j^2) \tag{3}$$

$$\epsilon_j \sim \mathcal{N}(0, \sigma_j^2 I_n) \tag{4}$$

$$\tau_j^{-2} \sim \Gamma(a_1, b_1) \tag{5}$$

$$\sigma_j^{-2} \sim \Gamma(a_2, b_2) \tag{6}$$

Gene-wise, the model contains two parts: a *low-dimensional* vector of covariates for the cis-effects and a *high-dimensional* one for the trans-effects. Cis-acting effects are modeled by piecewise linear regression splines (Leday et al., 2013) as follows (gene index j is removed for clarity reasons):

$$f_\alpha(x; \theta_s) = \theta_{0,0} + \theta_{0,1} x + \sum_{s \in \mathcal{S} \setminus \{0\}} \sum_{d=0}^{1} \theta_{s,d} \ \mathrm{sign}(s) \, (x - \alpha_s)_+^d. \tag{7}$$

Here $\mathcal{S} = \{-2, -1, 0, 1, 2\}$, θ_s is a vector of $2 \times |\mathcal{S}|$ unknown parameters, $\{\alpha_s\}$ are $|\mathcal{S}|-1$ *known* knots and $(a)_+^d$ represents the positive part $\max(a, 0)$ of a raised to the power d. This class of models combines copy number data from various steps of the preprocessing (namely the continuous segmented and discrete called data) and hence allows the effect of DNA on mRNA to differ across types of aberrations (e.g. loss, normal, gain and amplification). The model thus provides good interpretability as to how the gene copy number affects its expression. Below, Figure 1 illustrates four associations modeled by piecewise linear regression splines using Glioblastoma data from The Cancer Genome Atlas (http://cancergenome.nih.gov/; Verhaak et al., 2010).,

For the modeling of indirect trans-effects, we impose ridge priors ($\beta_{j,k} \sim \mathcal{N}(0, \tau_j^2 I_{p-1})$) to the high-dimensional vectors of parameters. Note that the amount of regularization is gene-dependent. In the next section, we describe the empirical Bayesian approach of van de Wiel et al. (2013), which we use to estimate parameters of priors.

2 Estimation of prior parameters

To estimate $\omega = \{\mu_{j,s}, \gamma_{j,s}, a_1, b_1, a_2, b_2\}$ we adopt the empirical Bayesian approach of van de Wiel et al. (2013). This consists in estimating ω by the

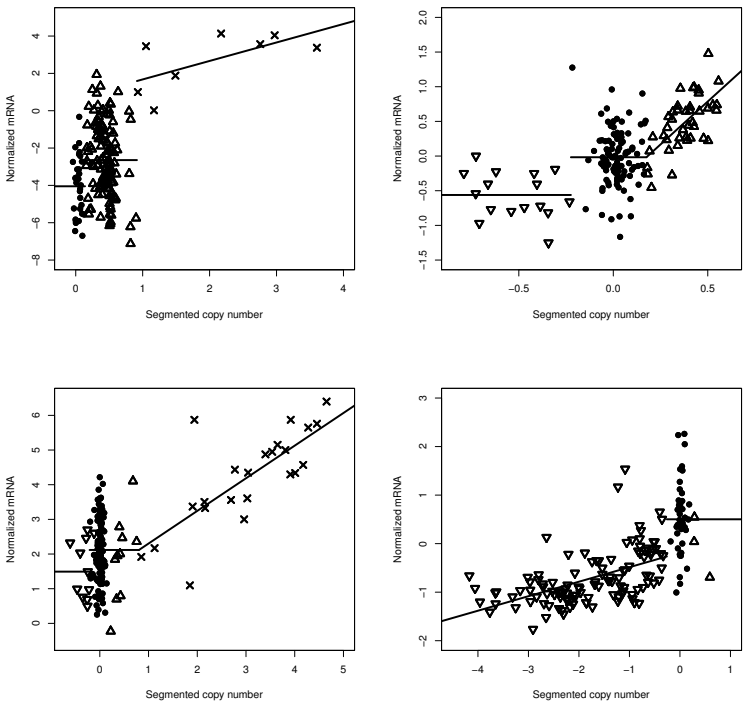


FIGURE 1. Examples of DNA-mRNA associations for four genes (top left: MET, top right: ERCC2, bottom left: AGAP2 and bottom right: CDKN2B) in the TCGA Glioblastoma data set (160 samples). X-axis: Gene dosage (segmented values), y-axis: mRNA gene expression. Copy number states are indicated by symbols: loss ($\bigtriangledown$), normal ($\bigcirc$), gain ($\bigtriangleup$) and amplification ($\times$). The "continuous" lines represent the fit of a regression spline after model selection (see Leday et al., 2013).

value for which the following approximation is most accurate.

$$\pi_{\omega_k}(.) \approx \frac{1}{p}\sum_{j=1}^{p} \pi_{\omega_k}(.|y_i) \tag{8}$$

Providing $y_i, i = 1, \cdots, p$ are independent, van de Wiel et al. (2013) showed this is an approximate solution to the likelihood equations that ensure maximization of the marginal likelihood (conventional empirical Bayes). Equation (1) is attractive as it only depends on marginal posteriors. To approximate those we use the Integrated Nested Laplace Approximation (INLA) of Rue et al. (2009).,

The problem of estimating prior parameters is solved iteratively by an EM-

type algorithm, which is briefly sketched as follows:

1. Initiate $\ell = 0$ and $\omega_k^{(0)}$
2. Apply INLA to estimate posteriors $\pi_{\omega_k^{(\ell)}}(\theta|Y_i)$
3. Obtain new estimate $\omega_k^{(\ell+1)}$ by best approximation of parametric prior $\pi_{\omega_k}(\theta)$ to empirical mixture of posteriors.
4. Reiterate from step 2 until convergence.

3 Computational efficiency

We use INLA as a fast alternative to MCMC to approximate marginal posteriors. However, if too many predictors are present, this is (still) computationally prohibitive. To overcome this, we use an *SVD decomposition* of the high-dimensional component in model (1). In our experience, the resulting orthogonality of the components can better accommodate multi-parameter shrinkage than the original setting (faster convergence of the above algorithm). The SVD results can then be back-transformed to the original parameter space (at least in INLA setting with approximately Gaussian posteriors).

4 Conclusion

In all, our model can be seen as a Bayesian graphical ridge that accounts for perturbation effects (DNA copy number). The amount of regularization is learned empirically, and may vary across genes. Sparsity is determined a posteriori through a model selection procedure (Bondell and Reich, 2012).

References

Bondell, Howard D. and Reich, Brian J. (2012) Consistent High-Dimensional Bayesian Variable Selection via Penalized Credible Regions *J Am Stat Assoc*, **107**, 1610 – 1624.

Leday, G. G. R., van der Vaart, A. W., van Wieringen, W. N., and van de Wiel, M. A. (2013). Modeling association between DNA copy number and gene expression with constrained piecewise linear regression splines *Ann Appl Stat*, In press.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *J Roy Stat Soc B*, **71**, 319 – 392.

van de Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., van der Vaart, A. W., and van Wieringen, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors *Biostatistics*. **14**, 113 – 128.

Verhaak, R.G. et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of Glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*, **17**, 98 – 110.

Spatio-temporal seasonal data modelling and forecasting with penalized smooth-ANOVA models

Dae-Jin Lee[1] and María Durbán[2]

[1] CSIRO Mathematics, Informatics and Statistics, Clayton, VIC, Australia
[2] Department of Statistics, Universidad Carlos III de Madrid, Spain.
e-mail: dae-jin.lee@csiro.au and mdurban@est-econ.uc3m.es

E-mail for correspondence: Dae-Jin.Lee@Csiro.au

Abstract: We propose a spatio-temporal seasonal model based on multidimensional P-splines, and the combination of ANOVA-type interaction models of Lee and Durbán (2011) and the smooth modulation model of Eilers (2008). Under the mixed model framework, we also show how to use the model to forecast future observations. We illustrate the methodology with monthly ground-ozone levels taken in 43 monitoring sites in Europe between January 1999 and December 2005.

Keywords: Smooth-ANOVA models, Spatio-temporal smoothing, forecasting, mixed models.

1 Introduction

In recent years, modelling spatio-temporal data has been an area of increasing interest. A particular case is when there is a seasonal trend component, and the seasonality effect might have different characteristics for each location, then, it is necessary to incorporate this effect, not only as an overall effect, but also in the space-time interaction. Lee and Durbán (2011) analyzed monthly averages of air pollution by ground-level ozone (in μg/m^3 units) over Europe. The data were collected in 43 monitoring stations in 15 EU countries from January 1999 to December 2005 (see Figure 1). Given the response vector $y_{i,t}$ of O_3 levels measured at the $i = 1, ..., n$ monitoring stations, at $t = 1, ..., t$ time points. Lee and Durbán (2011) proposed an ANOVA-Type decomposition of the spatio-temporal process given by:

$$y_{ij} = f_s(s_i) + f_t(t_j) + f_{s,t}(s_i, t_j) + \epsilon_{ij}, \tag{1}$$

where s_i is the bivariate vector of spatial geographical coordinates (i.e. $s_i = (lat, lon)$), t_j is the temporal dimension and ϵ_{ij} is a vector of iid uncorrelated errors $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (this assumption can be easily relax within

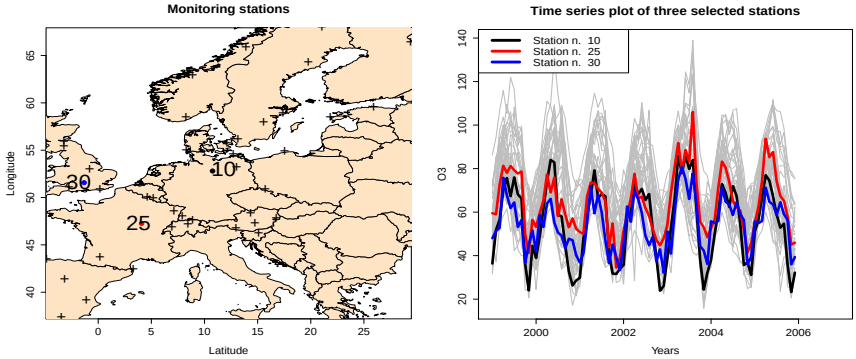


FIGURE 1. O_3 concentration levels data from 01/1999 to 12/2005. Left: Monitoring stations. Right: Raw times series data of three stations (number 10, 25, and 30).

the mixed model framework), and $f_s(\cdot)$, $f_t(\cdot)$, $f_{s,t}(\cdot)$ a set of smooth functions for space, time, and space-time interaction respectively. Each of these smooth functions are modelled with Tensor products of B-splines and penalties (Eilers and Marx, 1996) for each component such that identifiability constraints are imposed.
Lee and Durbán (2011) showed how the mixed model reparameterization of the spatio-temporal model allows for the estimation of the smoothing parameters by restricted maximum likelihood (REML).

$$y_{ij} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \text{ and } \boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{G}), \tag{2}$$

with fixed and random effects matrices

$$\boldsymbol{X} = [\mathbf{1}_{nt} : \boldsymbol{x}_s \otimes \mathbf{1}_t : \mathbf{1}_n \otimes \boldsymbol{x}_t : \boldsymbol{x}_s \otimes \boldsymbol{x}_t] \text{ and} \tag{3}$$

$$\boldsymbol{Z} = [\boldsymbol{x}_s \otimes \boldsymbol{Z}_t : \boldsymbol{Z}_s \otimes \boldsymbol{x}_t : \boldsymbol{Z}_s \otimes \boldsymbol{Z}_t], \tag{4}$$

where $\boldsymbol{x}_s = (lat, lon, lat \cdot lon)$, $\boldsymbol{Z}_s$ is the bivariate spatial random effect matrix, $\boldsymbol{x}_t = 1, ..., t$, and $\boldsymbol{Z}_t$ the temporal random effect matrix. Symbol $\otimes$ denotes the Kronecker product of two matrices. Note that, for notation simplicity, we consider a 2^{nd} order penalty matrix, although higher penalty orders are also possible. It is straight forward to show that the random effects covariance matrix $\boldsymbol{G}$ is a multiple of the identity matrix and depends on a set of variance components for each component of the ANOVA decomposition.
As shown in Figure 1, data presents a seasonal pattern with highest peaks during summer months (June-August) and lowest peaks in winter (December-January). If no seasonal effect is included, then, a large basis for the time component has to be used, and this yields computational problems when the interaction part of the model is estimated. Here, we propose

an extension of the model in Eq. (1), were the time component is decomposed as trend and seasonality using smooth modulation models in Eilers et al. (2008). Our main contribution is the inclusion of the smooth modulation component in the interaction, allowing for a higher level decomposition of the space-time interaction into: space-trend and space-modulation components, and the use of this model for forecasting future observations.

2 Smooth modulation model for spatio-temporal seasonal data

Eilers et. al. (2008) proposed a smooth model for seasonal univariate times series data, where trend and seasonality are modelled as penalized splines (P-splines). Seasonality is accounted for by trigonometric terms based on Fourier series, combined with a varying-coefficients model (Hastie and Tibshirani, 1993). Hence, the main temporal effect in model Eq. (1) can be decomposed as the sum:

$$f_t(x_j) = s(x_i) + \sum_{k=1}^{K} \{g_k(x_j)\cos(k\omega x_j) + h_k(x_i)\sin(k\omega x_j)\}, \tag{5}$$

where for monthly data, $\omega = 2\pi/12$, $s(\cdot)$ accounts for the smooth trend, and $g(\cdot)$ and $h(\cdot)$ are smooth series that describe the local amplitudes of cosine and sine waves. The number of harmonics K required for the seasonal component is usually taken as 1 or 2 to reduce the number of parameters to be estimated. In this paper and for notation simplicity, we use $K = 1$. Modelling and forecasting for univariate seasonal times series as mixed models were shown in Lee and Durbán (2012).

Using this formulation, the mixed model matrices in Eq. (3) include the new terms for the seasonal modulation, i.e. now $\boldsymbol{x}_t$ and $\boldsymbol{Z}_t$ are replaced by:

$$\breve{\boldsymbol{x}}_t = [\boldsymbol{x}_t : \cos(\omega \boldsymbol{x}_t) : \sin(\omega \boldsymbol{x}_t)], \text{ and} \tag{6}$$

$$\breve{\boldsymbol{Z}}_t = [\boldsymbol{Z}_t : \boldsymbol{C}\boldsymbol{Z}_t : \boldsymbol{S}\boldsymbol{Z}_t], \tag{7}$$

where $\boldsymbol{C} = \text{diag}\{\cos(\omega \boldsymbol{x}_t)\}$, and $\boldsymbol{S} = \text{diag}\{\sin(\omega \boldsymbol{x}_t)\}$ are varying-coefficent terms. Hence, the modulation component can be easily included in the space-time interaction, $f_{st}(\cdot)$.

3 Efficient estimation and forecasting

3.1 Estimation

Including a seasonal modulation term result in a considerable increase in the number of parameters, therefore, a fast algorithm is needed to speed up calculations. Lee et al. (2013) proposed a new algorithm for estimating

variance components in bivariate smooth-ANOVA models, their method is based on the results in Schall (1991), and the used of lower rank B-spline basis for the interactions. We use their approach to estimate our model:

$$\begin{aligned}\hat{y}_{ij} = \boldsymbol{X\beta} + f_s(s_i) + & \quad \text{(Fixed effects + smooth space)}\\ + s_{\text{trend}}(t_j) + s_{\text{mod}}(t_j) + & \quad \text{(smooth time trend + modulation)}\\ + f_s(s_i) * s_{\text{trend}}(t_j) + f_s(s_i) * s_{\text{mod}}(t_j), & \quad \text{(space-time interactions)}\end{aligned}$$

where symbol $*$ denotes *"interaction"*. Note that, $\boldsymbol{X\beta}$ is the fixed part, and space and time interaction terms include all possible interactions i.e. linear-by-smooth, and smooth-by-smooth interactions of space and time, i.e. $\boldsymbol{x}_s \otimes \boldsymbol{Z}_t$, $\boldsymbol{x}_s \otimes [\boldsymbol{CZ}_t : \boldsymbol{SZ}_t]$, $\boldsymbol{Z}_s \otimes \boldsymbol{x}_t$, $\boldsymbol{Z}_s \otimes [\cos(\omega \boldsymbol{x}_t) : \sin(\omega \boldsymbol{x}_t)]$, $\boldsymbol{Z}_s \otimes \boldsymbol{Z}_t$, and $\boldsymbol{Z}_s \otimes [\boldsymbol{CZ}_t : \boldsymbol{SZ}_t]$. Using this approach, we have a total of 9 variance components to estimate (a single parameter for the space, two for temporal trend and modulation, and six for all the interaction terms). To further reduce the computational burden (compared to the approach in Lee and Durbán (2011)), we also take advantage of reduced rank nested B-spline basis for computational efficiency. The proposed algorithm is much more efficient than the existing methods in R for this type of models (see Lee et al. (2013) for further details).

3.2 Forecasting

We use the approach of Currie et al. (2004) for fitting and forecasting simultaneously with P-splines models, where extrapolation can be viewed as a missing value problem and future observations are considered as missing data. We adapt this method to extrapolate smooth modulation models by extending the temporal B-spline basis and the penalty (in the main effect and interactions) to include future observations. For the univariate case, in Lee and Durbán (2012) we showed that using a simple diagonal weight matrix $\boldsymbol{W}$ is equivalent to forecasting in the mixed model framework. In the spatio-temporal context, this matrix is the Kronecker product of two weight matrices, $\boldsymbol{W}_n$ for the spatial component, and the weight matrix $\boldsymbol{W}_t$ for the time component, i.e. $\boldsymbol{W} = \boldsymbol{W}_n \otimes \boldsymbol{W}_t$ of dimensions $nt \times nt$, where $\boldsymbol{W}_n$ and $\boldsymbol{W}_t$ are diagonal matrices with diagonal elements $w = 1$ if the data is available and $w = 0$ if the data is missing or to be forecasted.

4 Application

We apply the methodology proposed to the analysis of air pollution by ozone levels in Europe for the period 1/1999-12/2005. We also forecast 12 months (1/2006 to 12/2007). Figure 2 (top) shows the main effects fits for the spatial and temporal components. The time component is decomposed into a trend (for this data the resulting time trend was linear), and

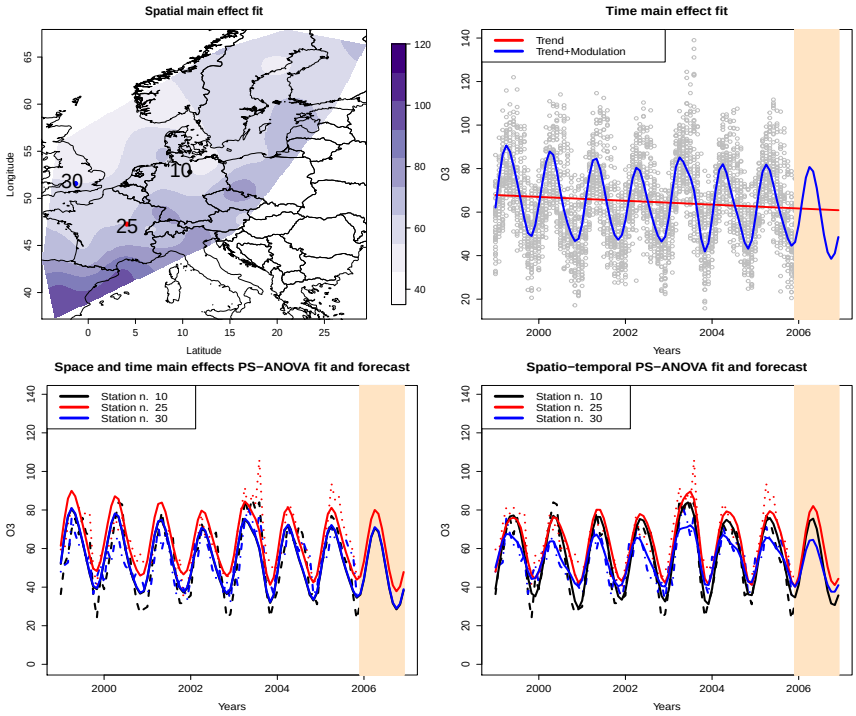


FIGURE 2. Top left: Fitted spatial main effect. Top right: Fitted time main effect (decomposed as trend and modulation). Bottom left: Fitted and forecasted values of selected stations considering the main effects of the Spatio-temporal Smooth-ANOVA model. Bottom right: Fitted and forecasted values of selected stations considering all the main effects and interaction terms of the Spatio-temporal Smooth-ANOVA model.

the seasonal component. The forecasted time main effect is also shown in the shaded area. Figure 2 (bottom left), shows the main effects fits (i.e. only spatial and temporal components), and their forecasts for the next 12 months. It can be noticed that considering only the main effects, the fit is not able to capture the particular characteristics of each station. This is also extended for the extrapolation. Indeed, model fits and forecasts of stations 10 and 30 are undistinguishable, showing the lack of flexibility of ignoring the space-time interaction. Figure 2 (bottom right) shows the complete spatio-temporal model fits and forecasts. Now it can be shown that the space-time interaction terms allows for accounting for the different seasonal patterns across the stations. Hence, we illustrated the need of the space-time modulation, and how the interaction terms are also considered for extrapolation.

5 Concluding remarks and further work

A new Smooth-ANOVA model for spatio-temporal data is proposed in this paper. The model takes advantage of the seasonal pattern of the data, modelling the seasonal component as a varying-coefficient term of sine and cosine waves. This new model reduces the computational cost of the model in Lee and Durbán (2011), where a large basis function for the temporal component was needed to model the temporal main effects and the space-time interaction. The new model also incorporates additional flexibility through the incorporation of meaningful interaction terms, and allowing for different variance components for each term of the space-time decomposition. The main contribution is the extension of the model to forecast new observations. Further work to be considered is the evaluation of the forecast method for a validation set using different criteria for times series forecasting.

Acknowledgements

This research was funded by the Spanish Ministry of Economy and Competitiveness (project MTM2011-28285-C02-02). The research of Dae-Jin Lee was also funded by an NIH grant for the Superfund Metal Mixtures, Biomarkers and Neurodevelopment project 1PA2ES016454-01A2.

References

Currie, I. D., Durbán, M., and Eilers, P. H. C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279298.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science*,**11**,89-121.

Eilers, P. H. C., Gampe, J., Marx, B. D. and Rau, R. (2008). Modulation models for seasonal time series and incidence tables. *Statistics in Medicine*,**27**(17),3430–3441.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient Models (with discussion). *Journal of the Royal Statistical Society, Ser. B*,**60**,271–293.

Lee, D.-J. and Durbán, M. (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling, Vol. 11, Issue 1, Pages 49-69.*

Lee, D.-J. and Durbán, M. (2012). Seasonal modulation smoothing mixed models for times series forecasting. *In proceedings of the 27th IWSM Workshop on Statistical Modelling.* Prague, Czech Republic.

Lee, D.-J., Durbán, M. and Eilers, P.H.C. (2013). Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. *Computational Statistics and Data Analysis*

Decision theory for some elementary statistical problems

Nicholas T. Longford[1]

[1] SNTL and Universitat Pompeu Fabra, Barcelona, Spain

E-mail for correspondence: sntlnick@sntl.co.uk

Abstract: A general alternative to hypothesis testing is described. It is based on decision-theoretical concepts which regard the quantification of the consequences of the two kinds of bad choices and other prior information on par with the data. The approach is applicable in both frequentist and Bayesian paradigms.

Keywords: Decision theory; Equilibrium; Hypothesis testing; Minimum expected loss; Plausible value.

1 Introduction

Hypothesis testing is ubiquitous in contemporary statistical practice. We are aware of its deficiencies, such as the impossibility of finding support for the (null) hypothesis and its inability to reflect the consequences of the two kinds of error that can be committed. We present an alternative in which these two issues are addressed comprehensively, although they require input additional to the problem formulation in the established frequentist setting. Our approach is an application of decision theory; for background, see Berger (1985), Lindley (1985) and DeGroot (2004).

We consider the standard statistical task of deciding whether the value of a parameter θ is smaller or greater than a given value θ_0. In the established approach, the hypothesis A: $\theta < \theta_0$ is tested against the alternative B: $\theta > \theta_0$. We are never justified to continue as if the hypothesis were valid because a test can never confirm it nor support it. Neither can we continue as if the alternative were valid even if there is evidence supporting it, because the power of the test is imperfect. Ignoring such uncertainty is poor practice.

Note that these problems relate not only to hypothesis testing, but also to model selection and model diagnostics where we also wish to choose between two courses of action: to continue the analysis with a more general model or with a submodel and to include all data in the analysis or exclude those highlighted by a given diagnostic procedure.

Key to our approach is the quantification of the consequences of choosing the wrong option. This is a matter for elicitation from the expert or the sponsor of the analysis. An example of the result of such elicitation is that

concluding with option A when in fact $\theta > \theta_0$ costs one unit (a utile), whereas concluding with option B when in fact $\theta < \theta_0$ costs R utiles. We refer to R as the penalty ratio and assume that $R > 1$. There is no loss when the appropriate option is selected. Suppose the choice is based solely on a statistic $\hat{\theta}$. This corresponds to the piecewise constant loss function defined by the values of 0, 1 and R for the various configurations of $\hat{\theta}$ and θ in relation to θ_0 . The choice between A and B is made by comparing the expected losses; smaller expectation is preferred. Details are given in the next section, with some generalisations.

2 Minimum expected loss

For a frequentist, θ is an unknown constant and $\hat{\theta}$ is a random variable, whereas for a Bayesian $\hat{\theta}$ is a constant, having been realised, and θ is a random variable, because it is unknown. We describe the frequentist solution first, but borrow this element of the Bayesian perspective. Suppose $\hat{\theta}$ is an unbiased estimator of θ with a normal distribution with variance σ^2. Having observed $\hat{\theta}$, it is now a constant and $\theta = \hat{\theta} - \sigma X$, where $X \sim \mathcal{N}(0, 1)$. Thus θ is now normally distributed with mean $\hat{\theta}$ and variance σ^2. This switch of status is related to the fiducial argument; see Fisher (1935), Seidenfeld (1992) and Hannig (2009).
Denote by $\phi(x; \mu, \sigma)$ the density of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ evaluated at x. Let $\Phi(x; \mu, \sigma)$ be the corresponding density. We drop the arguments μ and σ when $\mu = 0$ and $\sigma = 1$. Further, let $z_0 = (\hat{\theta} - \theta_0)/\sigma$.
If we choose option A the expectation of the loss we incur is

$$Q_{\mathrm{A}} \;=\; \int_{\theta_0}^{+\infty} \phi\Big(x; \hat{\theta}, \sigma\Big)\, \mathrm{d}x \;=\; \Phi(z_0)\;;$$

if we choose B the expected loss is $Q_{\mathrm{B}} = R(1 - Q_{\mathrm{A}})$. Our choice is based on the balance function $\Delta Q = Q_{\mathrm{A}} - Q_{\mathrm{B}} = (R + 1)Q_{\mathrm{A}} - R$; it is A when $\Delta Q < 0$, that is, when

$$\hat{\theta} \;>\; \theta_0 + \sigma\, \Phi\bigg(\frac{R}{R+1}\bigg)\;,$$

and it is B otherwise. This coincides with hypothesis testing of test size α when $R = 1/\alpha - 1$. Adopting the convention of $R = 19$ ($\alpha = 0.05$) for all problems, as is common practice, may seem singularly restrictive and in some settings unwise.

2.1 Linear and quadratic loss

The consequences of an incorrect decision need not be constant. A more realistic proposal in many settings is that it depends on the magnitude of

the error. Examples of such loss are the piecewise power loss functions, defined as

$$\begin{aligned} L_{\mathrm{A}}\left(\hat{\theta},\theta\right) &= (\theta-\theta_0)^h \\ L_{\mathrm{B}}\left(\hat{\theta},\theta\right) &= R\,(\theta_0-\theta)^h\ , \end{aligned}$$

when respectively $\theta > \theta_0$ and $\theta < \theta_0$, but we choose to the contrary. If we choose B ($\theta > \theta_0$), the expected loss is

$$\begin{aligned} Q_{\mathrm{B}} &= R\int_{-\infty}^{\theta_0}(\theta_0-x)^h\,\phi(x;\hat{\theta},\sigma)\,\mathrm{d}x \\ &= R\sigma^h\int_{-\infty}^{-z_0}(-z-z_0)^h\,\phi(z)\,\mathrm{d}z \qquad (1) \end{aligned}$$

where $z_0 = (\hat{\theta}-\theta_0)/\sigma$. Only powers $h = 1$ and 2 (in addition to $h = 0$) are of any practical importance, defining the respective piecewise linear and quadratic loss functions. The integral in (1) is evaluated by parts. For $h = 1$, we have

$$Q_{\mathrm{B}} \;=\; R\sigma\int_{-\infty}^{-z_0}\Phi(z)\,\mathrm{d}z\,.$$

Denote by Φ_1 the indefinite integral of Φ. It is easy to check that $\Phi_1(x) = x\Phi(x)+\phi(x)$. Thus, $Q_{\mathrm{B}} = R\sigma\Phi_1(-z_0)$ and $Q_{\mathrm{A}} = \sigma\{z_0+\Phi_1(-z_0)\}$. Hence the balance function is

$$\Delta Q \;=\; \sigma\left\{z_0-(R-1)\Phi_1(-z_0)\right\}\ .$$

We choose the option, A or B, that is associated with smaller expected loss, based on the sign of ΔQ. As an alternative, we can find the root of ΔQ, called the equilibrium and denoted by z^*, and compare it with z_0. We apply the Newton-Raphson algorithm, for which we use the identity

$$\frac{\partial\Delta Q}{\partial z_0} \;=\; \sigma\left\{1+(R-1)\Phi(-z_0)\right\}\ .$$

This is positive and since the limits of ΔQ are $\pm\infty$ for $z_0 \to \pm\infty$, ΔQ has a unique root, denoted by z^*. We choose A when $\hat{\theta} < \theta_0+\sigma z^*$ and B otherwise.

For piecewise quadratic loss, we have the identity

$$\Delta Q \;=\; \sigma^2\left\{(R+1)\Phi_2(-z_0)-\left(1+z^2\right)\right\}\ ,$$

where $\Phi_2(x) = (1+x^2)\Phi(x)+x\phi(x)$ is the indefinite integral of $2\Phi_1$. This balance function also has a single root that can be found by the Newton-Raphson algorithm.

2.2 Plausible loss functions

In practice, it is difficult to conclude the elicitation by settling on a single loss function. It is less contentious to declare a range of plausible loss functions, defined by a kernel, such as the linear, and a range of penalty ratios, $(R_{\mathrm{L}}, R_{\mathrm{U}})$. This is called the plausible range, and each value in it is a plausible value of R. We proceed by solving the problem for the limits of this range. If the same option is chosen for both, then it would be chosen for any plausible value of R. In this case, the decision is unequivocal. Otherwise, one option would be preferred for $R \in (R_{\mathrm{L}}, R_{\dagger})$ and the other for $R \in (R_{\dagger}, R_{\mathrm{U}})$. We refer to such a conclusion as impasse. It may be resolved by continuing the elicitation and reducing the range of plausible penalty ratios.
Let z_{L} and z_{U} be the equilibria that correspond to R_{L} and R_{U}, respectively; $z_{\mathrm{U}} > z_{\mathrm{L}}$, because with greater R we are more averse to choosing option B. If $z_0 \in (z_{\mathrm{L}}, z_{\mathrm{U}})$, then we have an impasse.

2.3 Distributions other than normal

The method described in the previous sections has a straightforward generalisation to estimators or statistics with some other distributions for which the calculus in Section 2.1 can be adapted. The key element of this adaptation are tractable expressions for the versions of the integral functions Φ_1 and Φ_2. These are available for t, F, gamma and beta, and for discrete distributions for binomial and Poisson.
For example, the analogues of Φ_1 and Φ_2 for the t distributions with $k > 2$ degrees of freedom are

$$\begin{aligned}\Psi_{1,k}(t) &= t\Psi_k(t) + \frac{1}{\gamma_k}\psi_{k-2}(t\gamma_k)\\ \Psi_{2,k}(t) &= t^2\Psi_k(t) + \frac{1}{\gamma_k^2}\Psi_{k-2}(t\gamma_k) + \frac{t}{\gamma_k}\psi_{k-2}(t\gamma_k)\,,\end{aligned}$$

where ψ_k and Ψ_k are the respective density and distribution function of the t distribution with k degrees of freedom and $\gamma_k = \sqrt{1 - 2/k}$. They are derived by relating $t\psi_k(t)$ to $\psi_{k-2}(t\gamma_k)$; see Longford (2012) for details.
The calculus can be extended further to finite mixtures of distributions, although the proliferation of parameters involved in them makes them suitable only for sizeable datasets.

2.4 Using prior information

Decision making is claimed to be more naturally conducted in the Bayesian paradigm. Prior information is invaluable in many settings, but it can often be converted to prior (hypothetical) data and a frequentist analysis then

conducted with the union of the prior and realised data. In the Bayesian paradigm, the sampling distribution is replaced by the posterior distribution of the relevant estimator ($\hat{\theta}$), and the expected losses with the two options are compared.
Instead of declaring a single (joint) prior for the estimated parameters, we prefer a plausible set of priors, yielding a plausible set of posteriors. One option is preferred unequivocally if it is preferred for every combination of plausible prior parameters, penalty ratio and any other parameter that is specified by its plausible range. While declaring plausible sets of priors is less contentious than declaring a single prior, this advantage over the established Bayesian procedure should be used sparingly, by declaring as narrow ranges as possible, to reduce the chances of an impasse to minimum. See Longford (2010) for an application.

3 Conclusion

Much of statistical practice has accepted the mean squared error (MSE) as the criterion in estimation and the tolerance of 5% error of the first kind in choosing one of two options. Settings in which these criteria are unsuitable are easy to identify; see Longford (2013a) for a generic example. In particular, the symmetric nature of MSE is problematic in many applications, and *ad hoc* adjustments to the formal inferential statements have to be made.
We presented a comprehensive alternative to the established inferential procedures which can be tailored much more closely to the priorities, perspective and further research or business agenda of the sponsor of the analysis. Statistics as a science is impoverished by the effective restriction to a single or a narrow range of criteria for inference, when they should cater for the varied clientele and the entire range of problems we encounter.
The presentation is based on a monograph (Longford, 2013b) to appear in the meantime.

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag.

DeGroot, M.H. (2004). *Optimal Statistical Decisions*. New York:McGraw-Hill.

Fisher, R.A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics* **VI**, 91–98.

Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica* **19**, 491–544.

Lindley, D.V. (1985). *Making Decisions*. Chichester: Wiley.

Longford, N.T. (2010). Bayesian decision making about small binomial rates with uncertainty about the prior. *The American Statistician* **64**, 164–169.

Longford, N.T. (2012). Comparing normal random samples, with uncertainty about the priors and utilities. *Scandinavian Journal of Statistics* **39**, 729–742.

Longford, N.T. (2013a). Assessment of precision with aversity to overstatement. *South African Journal of Statistics* **47**, 49–59.

Longford, N.T. (2013b). *Statistical Applications of Decision Theory*. New York: Springer-Verlag; to appear.

Seidenfeld, T. (1992). R. A. Fisher's fiducial argument and Bayes' theorem. *Statistical Science* **7**, 358–368.

Kernel based dimension reduction and classification of spectroscopy data for authentication of South African wines

Nelmarie Louw[1]

[1] Stellenbosch University, South Africa

E-mail for correspondence: nlouw@sun.ac.za

Abstract: In this paper we compare the performance of several kernel based dimension reduction and classification techniques on mid-infrared spectroscopy spectra of South African young cultivar wines. We also compare the performance to that of partial least squares dimension reduction followed by linear discriminant analysis, which is one of the most popular methods used for the analysis of spectroscopy data. We find that the kernel methods generally yield lower misclassification rates.

Keywords: Kernel methods; Dimension reduction; Classification; Infrared spectroscopy.

1 Introduction

Each of the algorithms are formulated in such a way that the transformed data appear only in the form of inner products and the kernel trick is applied to facilitate calculations.

Infrared spectroscopy has become an important analytical tool in many disciplines, such as the pharmaceutical, agricultural and biomedical fields. It is also widely used in classification and authentication of various food products, such as olive oil, fruit juice and wine. The data considered in this paper consist of mid infrared (MIR) spectra of six different South African young cultivar wines, made from Cabernet, Pinotage, Merlot, Shiraz, Chardonnay and Sauvignon Blanc grapes. The objective is to obtain a model to discriminate between the six cultivars based only on the spectroscopy data. A mid-infrared spectrum of a sample is obtained by modern scanning instruments at hundreds of equally spaced wavelengths in the mid-infrared range, resulting in large numbers of spectral variables. Dimension reduction or variable selection is therefore essential in the analysis of MIR data. Linear discriminant analysis (LDA) is one of the most well known techniques for dimension reduction and classification. There are however two main problems that arise: firstly, linear methods are not always adequate for complex data sets and secondly, the original LDA method is not

applicable when the number of variables is larger than the number of data cases. To deal with the first problem, kernel based techniques, such as kernel Fisher discriminant analysis (KFDA), proposed by Mika et al. (1999) can be applied. To solve the second problem, linear techniques such as regularized discriminant analysis (Friedman, 1989) and linear discriminant analysis by means of generalized singular value decomposition (LDA-GSVD), proposed by Howland and Park (2004), can be used. To deal with nonlinear small sample situations, Shawe-Taylor and Cristianini (2004) developed regularized kernel discriminant analysis (RKDA), while kernel discriminant analysis by means of generalized singular value decomposition (KDA-GSVD) was proposed by Park and Park (2005). In this paper we compare the performance of different dimension reduction and classification methods on MIR data obtained on different young South African single cultivar wines.

2 Dimension reduction

Consider a g-group classification problem where we observe p input variables on $n = \sum_{i=1}^{g} n_i$ cases. Denote the data by $X = \begin{bmatrix} X_1, & \cdots, & X_g \end{bmatrix} \in R^{p \times n}$, where $X_i \in R^{p \times n_i}$ has columns $\mathbf{x_{ij}}, \mathbf{j} = \mathbf{1}, \cdots, \mathbf{n_i}$, containing observations on the p input variables for n_i cases from group i. Denote the mean of group i by $\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x_{ij}}$ and the overall mean by $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{g} n_i \bar{\mathbf{x}}_i$. Consider the between-group scatter matrix $S_B = \sum_{i=1}^{g} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}}$ and the within-group scatter matrix $S_W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i})(\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i})^{\mathrm{T}}$. The aim of LDA is to find a transformation matrix $V \in R^{p \times r}$ to transform the data to an$r$-dimensional space ($r < p$) by maximizing the between-group scatter and minimizing the within-group scatter. If $p < n$, the columns of the matrix V are the eigenvectors corresponding to the $r = min(g-1, p)$ non-zero eigenvalues of $S_W^{-1} S_B$. However, when $p > n$, S_W is singular. One way of solving this problem is by means of regularization (cf. Friedman, 1989). Another approach which makes use of generalized singular value decomposition (GSVD) was suggested by Howland and Park (2004). Define the matrices $H_B = \begin{bmatrix} \sqrt{n_1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}), & \cdots, & \sqrt{n_g}(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}) \end{bmatrix} \in R^{p \times g}$ and $H_W = \begin{bmatrix} X_1 - \bar{\mathbf{x}}_1 \mathbf{u_1^T}, & \cdots, & X_g - \bar{\mathbf{x}}_g \mathbf{u_g^T} \end{bmatrix} \in R^{p \times n}$, where $\mathbf{u_i}$ is a unit vector of length n_i. Then $S_B = H_B H_B^T$ and $S_W = H_W H_W^T$. The GSVD algorithm (Golub and Van der Loan, 1996) is then applied to the matrix $\begin{bmatrix} H_B & H_W \end{bmatrix}$ to find the transformation matrix $V \in R^{p \times r}$ which is used to transform the data to the lower dimensional subspace by the following transformation: $Y = V^T X \in R^{r \times n}$. Classification in this lower dimensional subspace is then done by means of a closest centroid or nearest neighbour classifier. This method is referred to as linear discriminant analysis by means of generalized singular value decomposition (LDA-GSVD).

3 Kernel methods for dimension reduction and classification

Linear dimension reduction and classification methods are not always adequate and kernel methods can be used to obtain non-linear techniques. The basic idea of kernel methods is to transform the data $\mathbf{x_{ij}}$ in the original input space to a higher dimensional feature space by means of a non-linear transformation, $\mathbf{z_{ij}} = \mathbf{\Phi}(\mathbf{x_{ij}})$. A linear technique is then performed in the feature space, but this corresponds to a non-linear technique in the original input space. The dimensionality of the feature space is typically very high and can even be infinite, so performing calculations in this space is difficult or impossible. However, if an algorithm can be expressed in a form which contains the mapped data only as inner products, the kernel trick can be used to circumvent the problem. The kernel trick, which is based on the theory of reproducing kernel Hilbert spaces and specifically Mercer's theorem, entails replacing inner products by a kernel function, $K(\mathbf{x}, \tilde{\mathbf{x}}) = \langle \mathbf{\Phi}(\mathbf{x}), \mathbf{\Phi}(\tilde{\mathbf{x}}) \rangle$. This obviates the need to specify a mapping Φ or to perform calculations in feature space. One of the most frequently used kernels is the Gaussian kernel, $K(\mathbf{x}, \tilde{\mathbf{x}}) = \mathbf{exp}(-\gamma \|\mathbf{x} - \tilde{\mathbf{x}}\|^{\mathbf{2}})$, where γ is a so-called kernel hyperparameter that has to be specified beforehand or determined from the data, typically by means of crossvalidation. Shawe-Taylor and Cristianini proposed regularized kernel discriminant analysis (RKDA), which entails applying linear discriminant analysis in feature space. Park and Park (2005) developed KDA-GSVD by applying generalized singular value decomposition to solve the generalized eigenvalue problem formulated in feature space. Each of the algorithms are formulated in such a way that the transformed data appear only in the form of inner products and the kernel trick is applied to facilitate calculations. Mika et al. (1999) developed kernel Fisher discriminant analysis (KFDA) for two group classification, but this method can also be applied in a multi-group setting by using a pairwise or one-versus-the-rest approach.

For two groups, the KFDA classifier is given by $sign\{b + \sum_{i=1}^{n} \tilde{\alpha}_i K(\mathbf{x_i}, \mathbf{x})\}$. The values of b and $\tilde{\alpha}_1, \tilde{\alpha}_2, \cdots, \tilde{\alpha}_n$ are determined as follows. Evaluating $K(\mathbf{x_i}, \mathbf{x_j})$ for $i, j = 1, 2, \cdots, n$, we are able to construct the so-called Gram matrix, G, with ij^{th} entry $K(\mathbf{x_i}, \mathbf{x_j})$. Let α be an n-vector with elements $\alpha_1, \alpha_2, \cdots, \alpha_n$. Then $\tilde{\alpha}$ maximises the Rayleigh coefficient

$$r(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha}.$$

Here, $M = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$, and $N = GG^T - n_1 \mathbf{m}_1 \mathbf{m}_1^T - n_2 \mathbf{m}_2 \mathbf{m}_2^T$, where the n elements of $\mathbf{m_1}$ are given by $\frac{1}{n_1} \sum_{j=1}^{n_1} K(\mathbf{x_i}, \mathbf{x_j})$, with a similar expression for $\mathbf{m_2}$. The analogy with classical linear discriminant analysis is clear: we may interpret M as the between group scatter matrix, and N as the within group scatter matrix, in both cases taking into account that we

are effectively working in the feature space induced by the kernel function. It is well known that $N^{-1}(\mathbf{m_1}-\mathbf{m_2})$ will maximize the Rayleigh coefficient. There is however one problem: the matrix N is singular and consequently we cannot find $\tilde{\alpha}$ by simply calculating $N^{-1}(\mathbf{m_1}-\mathbf{m_2})$. Mika et al. propose and motivate the use of regularization to overcome this difficulty. In the present context regularization entails replacing N by a matrix $N_\lambda = N + \lambda I$, for some positive scalar λ. This yields a solution $N_\lambda^{-1}(\mathbf{m_1} - \mathbf{m_2})$, depending on λ. Obviously the hyperparameter λ has to be specified, and this is typically done by performing a crossvalidation search along a suitable grid of potential λ-values. The intercept b in the KFD classifier can be specified in different ways. A popular choice, which we will also use, is $b = 0.5(\mathbf{m}_2^T N_\lambda^{-1}\mathbf{m}_2 - \mathbf{m}_1^T N_\lambda^{-1}\mathbf{m}_1) + log(n_1/n_2)$, which is similar to the intercept used in linear discriminant analysis.

4 Analysis of wine cultivar data

The wine spectroscopy data set was obtained from researchers at the Institute for Wine Biotechnology at Stellenbosch University. It consists of $n = 574$ spectra generated on single-varietal young South African wines. The spectra were recorded in the mid infrared range (wavenumber region of 5011 to 929 cm^{-1}. Each spectrum consisted of wavelength dependent absorbance values recorded at $p = 1054$ different wavenumbers. These measurements relate to the chemical composition and internal microstructure of the wines. The aim is to develop a model to discriminate between the wines, with a view to use the model for authentication purposes. We firstly compare the classification performance of the kernel based techniques: KFDA (pairwise, denoted by KFDA1, as well as one-against-the-rest, denoted by KFDA2), KDA-GSVD and RKDA. The data were randomly split into training (70%) and test (30%) data, preserving the group structure in each set. This was repeated 100 times, and the mean and standard error of the misclassification rates were obtained for each of the techniques. The results are reported in Table 1.

TABLE 1. Error rates (and standard errors) for the kernel based techniques.

KFDA1	KFDA2	KDA-GSVD	RKDA
0.1029	0.1113	0.0653	0.0864
(0.0031)	(0.0029)	(0.0019)	(0.0029)

It is clear that the lowest error rate was obtained by the KDA-GSVD method, followed by RKDA.

In the analysis of spectroscopy data, partial least squares is often used for dimension reduction, followed by LDA for classification. We also applied this technique to the wine spectroscopy data. Since this is a linear

technique, we compare its performance to that of LDA-GSVD. The results appear in Table 2.

TABLE 2. Error rates (and standard errors) for the linear techniques

LDA-GSVD	PLS-DA
0.0862	0.1209
(0.0019)	(0.0031)

It is clear that LDA-GSVD performs better on this data set than PLS-LDA. When comparing the results of KDA-GSVD to that of LDA-GSVD, better accuracy was achieved by the nonlinear kernel based method than by the linear method.
Although further investigations on more data sets are needed before general conclusions can be made, it seems as if kernel discriminant analysis by means of generalized singular value decomposition performs well in dimension reduction and classification of spectroscopy data.

References

Friedman, J.H. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association*, **84**, 165 – 175.

Golub, G.H. and Van der Loan, C.F. (1996). *Matrix Computations*. Baltimore: John Hopkins University Press.

Howland, P. and Park, H. (2004). Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 995 – 1006.

Mika, S., Rtsch, G., Weston, J., Schlkopf, B. and Mller, K.-R. (1999). Fisher discriminant analysis with kernels. In: *Neural Networks for Signal Processing, IX*, New York: IEEE Press, 41 – 48.

Park, C.H., and Park, H. (2005). Nonlinear Discriminant Analysis using Kernel Functions and the Generalized Singular Value Decomposition. *Siam Journal on Matrix Analysis and Applications*, **27**, 87 – 102.

Shawe-Taylor, J. and Christianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.

Semiparametric partially nonlinear mixed-effects models with P-splines

Robson J. M. Machado[1], Cibele M. Russo[2]

[1] Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, SP, Brazil

[2] Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brazil

E-mail for correspondence: robsonjmachado@gmail.com

Abstract: The issue of including random effects to nonlinear models is essential to deal with correlated data in repeated measures or longitudinal data with a physical interpretation. However, it usually increases the complexity of the problem since complicated computational methods may be required even for estimation. Moreover, smoothing methods have shown to be interesting techniques to improve the modelling by introducing nonparametric terms to the model and penalizing the likelihood function. We propose semiparametric partially nonlinear mixed effects models with P-splines, where the random effects are included linearly to the model. The proposed method is applied to the famous pharmacokinetic dataset of theophylline concentration.

Keywords: Semiparametric model; nonlinear model; mixed effects model; smoothing; P-splines.

1 Introduction

The theory of linear mixed effects models is well developed for analyzing repeated measures or longitudinal data. For nonlinear data, the inclusion of random effects increases significantly the complexity of the problem, requiring computational methods even for estimation. One alternative to deal with this problem is to include the random effects linearly to the nonlinear model, which has been successfully considered, for example, by Russo et al. (2009). The difficulty increases even more with the inclusion of smoothing terms, but the gain with smoothing is remarkable (see, for instance, Liu & Wu 2008 and Ke & Wang, 2001). Motivated by a pharmacokinetic absortion-elimination data of theophylline concentration, we propose a semiparametric partially nonlinear mixed-effects model with the random effects are included to the model in a linear way, assuming that the random effects and errors jointly follow a multivariate normal distribution.

For adding nonparametric terms, we consider P-splines (Eilers & Marx, 1996).

2 The model

A semiparametric partially nonlinear mixed-effects model for the jth response of the ith subject Y_{ij} may be written as

$$Y_{ij} = \eta(\mathbf{X}_{ij}, \boldsymbol{\beta}) + f(t_{ij}) + \mathbf{Z}_{ij}^{\top}\mathbf{u}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \tag{1}$$

where η is a nonlinear function of $\boldsymbol{\beta}$ and $\mathbf{X}_{ij}$, $\mathbf{X}_{ij}$ is a covariate that can be a scalar or a vector, $\mathbf{Z}_{ij}$ is a known vector related to the random effects, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\top}$ a vector of unknown parameters of major interest, $\mathbf{u}_i = (u_{i1}, \ldots, u_{ir})^{\top}$ the random-effects coefficients for the ith subject, f is a smooth function of time, t_{ij} are the time points. In this work the smooth function is a B-spline of degree 3 with L equidistants knots. Specifically,

$$f(t_{ij}) = \alpha_1 B_1(t_{ij}) + \ldots + \alpha_L B_L(t_{ij}) \tag{2}$$

and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)^{\top}$ is an unknown vector of coefficients related to the B-splines $B_1, \ldots, B_L$. In matrix notation, the model (1) for a vector of longitudinal response variable $\mathbf{Y}_i(m_i \times 1)$ may be expressed as follows

$$\mathbf{Y}_i = \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) + \mathbf{B}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{3}$$

where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^{\top}$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i})^{\top}$, $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{im_i})^{\top}$, $\boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) = (\eta(\mathbf{X}_{i1}, \boldsymbol{\beta}), \ldots, \eta(\mathbf{X}_{im_i}, \boldsymbol{\beta}))^{\top}$ and the elements of matrix $\mathbf{B}_i$ are $b_{jl} = B_l(t_{ij})$ for the ith subject.

It is usual to assume that $\mathbf{u}_i$ and $\boldsymbol{\epsilon}_i$ assumed to be independent and to follow a multivariate normal distribution. We assume that

$$\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{u}_i \end{bmatrix} \sim \mathrm{N}_{m_i+r}\left\{ \begin{pmatrix} \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) + \mathbf{B}_i\boldsymbol{\alpha} \\ \mathbf{0} \end{pmatrix}; \begin{bmatrix} \mathbf{Z}_i\mathbf{Q}\mathbf{Z}_i^{\top} + \sigma^2\mathbf{I}_{m_i} & \mathbf{Z}_i\mathbf{Q} \\ \mathbf{Q}\mathbf{Z}_i^{\top} & \mathbf{Q} \end{bmatrix} \right\}. \tag{4}$$

The matrices $\boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{Q}\mathbf{Z}_i^{\top} + \sigma^2\mathbf{I}_{m_i}$, $\mathbf{Q}$, and $\mathbf{Z}_i\mathbf{Q}$ are the variance-covariance matrices $\mathrm{Var}(\mathbf{Y}_i)$, $\mathrm{Var}(\mathbf{u}_i)$ and $\mathrm{Cov}(\mathbf{Y}_i, \mathbf{u}_i)$. For a parsimonious model, let $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\tau})$ be diagonal with elements of a vector $\boldsymbol{\tau}$, which means that the random effects are uncorrelated. It is easy in this case to work on the marginal model, $\mathbf{Y}_i \sim \mathrm{N}_{m_i}(\boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) + \mathbf{B}_i\boldsymbol{\alpha}; \boldsymbol{\Sigma}_i)$, to preserve the mean of the hierarchical model without requiring numerical integration. The log-likelihood function $L_i = L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_i; \mathbf{Y}_i, \mathbf{X}_i)$ is, apart from a constant, given by

$$L_i = -\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\left[\mathbf{Y}_i - \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) - \mathbf{B}_i\boldsymbol{\alpha}\right]^{\top}\boldsymbol{\Sigma}_i^{-1}\left[\mathbf{Y}_i - \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) - \mathbf{B}_i\boldsymbol{\alpha}\right].$$

This leads to a very rough fit and to make the result less flexible we consider a penalized log-likelihood function based in P-splines

$$L_{p_i}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_i; \mathbf{Y}_i, \mathbf{X}_i) = L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_i; \mathbf{Y}_i, \mathbf{X}_i) - \frac{\lambda}{2n}\boldsymbol{\alpha}^{\top}\mathbf{D}_k^{\top}\mathbf{D}_k\boldsymbol{\alpha}, \tag{5}$$

in which λ is a positive real number to be estimated, $\mathbf{D}_k$ is the matrix representation of the difference operator and k is the order of the differences (see Eilers and Marx, 1996). The complete penalized log-likelihood is given by $L = L_p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}; \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} L_{p_i}$.

3 Estimation method

For a fixed λ, to obtain the penalized maximum likelihood estimates of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^\top = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\tau})^\top$, one possibility is to consider the Fisher scoring algorithm. The penalized score functions may be written as

$$\begin{aligned}
\mathbf{U}_p^{\beta} &= \frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{J}_i{}^\top \boldsymbol{\Sigma}_i^{-1}[\mathbf{Y}_i - \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) - \mathbf{B}_i\boldsymbol{\alpha}], \\
\mathbf{U}_p^{\alpha} &= \frac{\partial L}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n} \mathbf{B}_i^\top \boldsymbol{\Sigma}_i^{-1}[\mathbf{Y}_i - \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) - \mathbf{B}_i\boldsymbol{\alpha}] - \lambda \mathbf{D}_k^\top \mathbf{D}_k \boldsymbol{\alpha} \text{ and} \\
\mathbf{U}_p^{\gamma} &= \frac{\partial L}{\partial \boldsymbol{\gamma}} = (U_{\gamma_1}, \ldots, U_{\gamma_{r+1}})^\top, \text{ with} \\
U_{\gamma_j} &= -\frac{1}{2} \sum_{i=1}^{n} \left\{ \operatorname{tr}\left[\boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i(j)\right] - \mathbf{r}_i^\top \boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i^{-1}(j) \boldsymbol{\Sigma}_i^{-1} \mathbf{r}_i \right\},
\end{aligned}$$

in which $\mathbf{J}_i = \partial \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^\top$, $\mathbf{r}_i = [\mathbf{Y}_i - \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) - \mathbf{B}_i\boldsymbol{\alpha}]$, $\dot{\boldsymbol{\Sigma}}_i(j) = \partial \boldsymbol{\Sigma}_i / \partial \gamma_j$, for $j = 1, \ldots, r+1$, $i = 1, \ldots, n$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{r+1})^\top = (\sigma^2, \boldsymbol{\tau}^\top)^\top$. The penalized Fisher information matrix for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top)^\top$ is given by

$$\mathbf{K}_p^{\theta\theta} = \begin{bmatrix} \mathbf{K}_p^{\beta\beta} & \mathbf{K}_p^{\beta\alpha} & \mathbf{0} \\ \mathbf{K}_p^{\beta\alpha\top} & \mathbf{K}_p^{\alpha\alpha} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{K}_p^{\gamma\gamma} \end{bmatrix}$$

in which

$$\begin{aligned}
\mathbf{K}_p^{\beta\beta} &= \sum_{i=1}^{n} \mathbf{J}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{J}_i, \\
\mathbf{K}_p^{\alpha\alpha} &= \sum_{i=1}^{n} \mathbf{B}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{B}_i + \lambda \mathbf{D}_k^\top \mathbf{D}_k, \\
\mathbf{K}_p^{\beta\alpha} &= \sum_{i=1}^{n} \mathbf{J}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{B}_i \text{ and} \\
\mathbf{K}_p^{\gamma\gamma} &= \sum_{i=1}^{n} K_{i\gamma}, \text{ whose } qs\text{-th element is given by} \\
K_{i\gamma, qs} &= \frac{1}{2} \operatorname{tr}\left[\boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i(r) \boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_i(s)\right],
\end{aligned}$$

An iterative algorithm to obtain the penalized maximum likelihood estimates for $\boldsymbol{\theta}$ using the Fisher scoring method is given by

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix}^{(m+1)} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix}^{(m)} + \begin{bmatrix} \mathbf{K}_p^{\beta\beta} & \mathbf{K}_p^{\beta\alpha} \\ \mathbf{K}_p^{\beta\alpha\top} & \mathbf{K}_p^{\alpha\alpha} \end{bmatrix}^{-1(m)} \begin{bmatrix} \mathbf{U}_p^{\beta} \\ \mathbf{U}_p^{\alpha} \end{bmatrix}^{(m)}$$

$$\widehat{\boldsymbol{\gamma}}^{(m+1)} = \widehat{\boldsymbol{\gamma}}^{(m)} + (\mathbf{K}_p^{\gamma\gamma})^{-1(m)}\mathbf{U}_p^{\gamma(m)},\ m = 0, 1, 2\ldots$$

with $\mathbf{K}_p^{\beta\beta}$,$\mathbf{K}_p^{\alpha\alpha}$, $\mathbf{K}_p^{\beta\alpha}$, $\mathbf{K}_p^{\gamma\gamma}$, $\mathbf{U}_p^{\beta}$, $\mathbf{U}_p^{\alpha}$ and $\mathbf{U}_p^{\gamma}$ as presented previously. The initial values for the algorithm can be the least squares estimates, and the estimates of the random effects can be obtained by using the empirical Bayes method.

An important issue is how to derive the variance-covariance matrix of the penalized maximum likelihood. According to Ibacache-Pulgar et al. (2012), it is possible to derive the variance-covariance matrix by using the inverse of the penalized Fisher information matrix. Thus, the approximate variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be approximated as follows

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\theta}}) \approx [\mathbf{K}_p^{\theta\theta}(\widehat{\boldsymbol{\theta}})]^{-1}.$$

4 Numerical illustration

In an experiment described in Davidian & Giltinan (1995), the anti-ashmatic theophylline substance was administred to 12 subjects and measured in 11 time points. It is usual for this application to consider the nonlinear function, for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$

$$\eta_{ij} = d_i \exp(lK_e + lK_a - lC_l)\frac{[\exp(-e^{lK_e}x_{ij}) - \exp(-e^{lK_a}x_{ij})]}{e^{(lK_a)} - e^{lK_e}},$$

where the parameters lK_a, lK_e and lC_l represent the logarithm of the absortion, elimination and clearance rates and d_i represents the dose administred to the ith individual. In this paper, we consider

$$\mathbf{Z}_i = \left[\frac{\partial\boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{X}_i)}{\partial lK_e}, \frac{\partial\boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{X}_i)}{\partial lK_a}, \frac{\partial\boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{X}_i)}{\partial lC_l}\right]\Bigg|_{\beta=\tilde{\beta}},$$

with $\tilde{\boldsymbol{\beta}}$ being the least squares estimates of $\boldsymbol{\beta} = (lK_a, lK_e, lC_l)^\top$. The smoothing parameter was chosen by Aikake Information Criterion (AIC) method and since the number of parameters are fixed, the AIC method reduce to choose the λ that provides a log-likelihood that minimizes its value. The best model among the fitted was obtained with $\widehat{\lambda} = 0.01$, with a reached log-likelihood of -181.53 and it is represented in Figure 1 and described in Table 1.

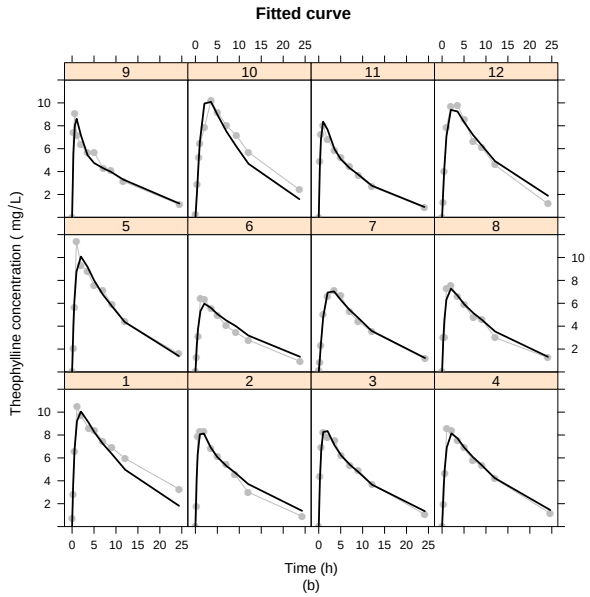


FIGURE 1. Fitted semiparametric model for the theophylline data

TABLE 1. Maximum likelihood estimates of the parametric part and their standard errors for the theophylline application.

Parameter	Estimates	Standard errors
lK_e	-1.9439	0.3703
lK_a	0.41180	0.2286
lC_l	-2.7521	0.3992
σ^2	0.56213	0.0781
τ_1	-0.0069	0.0075
τ_2	0.47219	0.1987
τ_3	0.02617	0.0112

5 Discussion and remarks

Motivated by a nonlinear problem with correlated data, we propose a semiparametric nonlinear mixed-effects models, where the random effects are included linearly to the model. This approach may provide a flexible model, including an interpretable parametric part and smoothing individual profiles.

Acknowledgments: The authors thank to Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, for supporting this research.

References

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data.* Chapman & Hall/CRC, 1995.

Eilers, P. H. C. and Marx, D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* 11, 89-102.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association* 96, 1272–1298.

Ibacache-Pulgar, G., Paula, G. A., and Galea, M. (2012). Influence diagnostics for elliptical semiparametric mixed models. Statistical Modelling, 12(2), 165-193.

Liu, W. and Wu, L. (2008). A semiparametric nonlinear mixed-effects models with non-ignorable missing data and measurement errors for HIV viral data. *Computational Statistics & Data Analysis* 53, 112-122.

Russo, C. M., and Paula, G. A. and Aoki, R. (2009). Influence diagnostics in nonlinear mixed-effects elliptical models. *Computational Statistics & Data Analysis* 53, 4143–4156.

The role of Frailty in survival studies

Gilbert MacKenzie [1] and Il Do Ha[2]

[1] ENSAI, Rennes, France - Daegu Haany, South Korea

E-mail for correspondence: gilbert.mackenzie@ul.ie

Abstract: The focus is mainly, but, not solely, on longitudinal randomised controlled clinical trials. The paper aims to delineate the role of *frailty* in the modern analysis of such trials and also in longitudinal survival studies. Our approach exploits recent developments in statistical modelling and in estimation methods, for example, in non-PH survival modelling, covariance modelling and in h-likelihood inference. We illustrate our approach and findings with examples from the literature.

Keywords: frailty; longitudinal RCTs and studies; MV responses; covariance modelling; PH & non-PH survival distributions, h-likelihood

1 Introduction

The importance of randomized controlled clinical trials in not in dispute. From the perspective of Scientific Method they have the status of experiments. The key feature of such experiments is that their conclusions are protected by randomization. Broadly, non-randomized studies, or observational studies control differences between groups by means of covariate adjustment, using an appropriate statistical model.
In a clinical trial the effect can be attributed to the intervention because, in principle, all other potentially confounding factors are controlled for by the randomization procedure, provided the total sample size, n, is large enough. Of course, in practice the sample size is always finite, and differences exist between the groups being studied and these can be controlled for using statistical modelling methods This is particularly true of trials employing minimization which, typically, only controls for a *pre-selected* subset of factors (Friedman et al., 1998), and hence is considered logically inferior to a conventional, trial.
Thus, today, it is realized that there is much more information to be retrieved about the effect of treatment using a statistical modelling approach and accordingly the era of investing vast sums of money in a clinical trial only to conduct a t-test has gone. Moreover, this has led to a harmonization of methods of analysis in randomized and non-randomized studies, especially as statistical modelling methods have developed.

2 Frailty

2.1 Concept

Consider a survival regression model with failure time density $f(t|\theta, \beta)$, and *basic* hazard function $\lambda(t|.)$ and survivor function $S(t|.)$, where typically θ is a vector valued parameter and β is a regression parameter measuring the influence of p covariates $x' = (x_1, x_2, \ldots, x_p)$. Assume that the basic model is extended to a univariate multiplicative frailty model (Hougaard, 1982) with hazard

$$\lambda(t|u, x) = u\lambda(t|x) \tag{1}$$

where the random variable U, with *mixing* density $g(u|\sigma^2)$, denotes the unobservable individual (i.i.d.) frailties with $E(U) = 1$ and $V(U) = \sigma^2$. The frailties are person specific and may be viewed as allowing for unrecorded covariates. One example is the PH frailty model

$$\lambda(t|u, x) = \lambda_0(t) \exp(x'\beta + v) \tag{2}$$

where $u = \exp(v)$. A multi-component version (Ha, Lee and MacKenzie, 2007) is

$$\lambda(t|u, x) = \lambda_0(t) \exp(x'\beta + Z_1 v_1 + \cdots + Z_q v_q) \tag{3}$$

which may be written as

$$\lambda(t|u^*, x) = u^* \lambda_0(t) \exp(x'\beta) \tag{4}$$

where: $u^* = \prod_{j=1}^{q} \exp(Z_j v_j)$, ie, a given function of $(u_1, u_2, \ldots, u_q)$ and the Z_j are appropriate design matrices, leading naturally to

$$\lambda(t|u^*, x) = \lambda_0(t) \exp(x'\beta + Zv) \tag{5}$$

where Z is a $n \times q$ design matrix and v is a conformable $(q \times 1)$ vector of random effects.

2.2 Model Choice

In the current setting this amounts to a joint choice of a *basic* hazard function $\lambda(t|.)$ and *mixing* density $g(u|\cdot)$. In the multi-component version the latter quantity may be multivariate. There is a wide choice for the *basic* hazard function including: PH (Cox, 1972, 1975), GTDL (MacKenzie, 1996, 1997), XD (Jorgensen, 2011; Burke & MacKenzie, 2011). The latter class covers extreme distributions and is relatively new. For the *mixing* density the choice is usually confined to Gaussian, Log-Normal or Gamma, whence correlation structures may be more easily supported. For a *basic* PH hazard

the resulting marginalized frailty model is not PH (Hougaard, 2000). Moreover, in the univariate case the choice of a PH basic hazard may not always be optimal. In simulation Ha & MacKenzie (2010) report under-estimation of the regression parameter in the PH model with log-Normal frailty, when the data actually follow a GTDL model (non-PH) with log-Normal frailty. As usual, model selection is important. For multi-component models, *focussed* model selection has been developed for selecting the frailty structure best supported by the data for a given mixer (Ha, Lee & MacKenzie, 2007). %subsectionEstimation

3 Model Formulation

We develop methods in the context of time to first recurrence of disease in an EORTC randomized clinical trial of chemotherapy in invasive, non-muscle, bladder cancer patients (Ha et al, 2011). In this multi-centre trial the main interest lies in evaluating *centre* effects and testing for homogeneity across centres. We show how to formulate the associated multi-level frailty models, describe their properties including improved prediction of random effects and perform *focussed* model selection in the h likelihood paradigm.

In general, suppose that data consist of right censored time-to-event observations collected from q centres. Let T_{ij} $(i = 1, \ldots, q,\ j = 1, \ldots, n_i,\ n = \sum_i n_i)$ be the survival time for the jth observation in the ith centre (or cluster) and let C_{ij} be the corresponding censoring time. Then observable data become $y_{ij} = \min\{T_{ij}, C_{ij}\}$ and $\delta_{ij} = I(T_{ij} \leq C_{ij})$, where $I(\cdot)$ is the indicator function.

Denote by v_i a s-dimensional vector of unobserved log-frailties (random effects) associated with the ith cluster. Given v_i, the conditional hazard function of T_{ij} is of the form

$$\lambda_{ij}(t|v_i) = \lambda_0(t) \exp(\eta_{ij}) \tag{6}$$

where $\lambda_0(\cdot)$ is a unknown baseline hazard function, $\eta_{ij} = x_{ij}^T\beta + z_{ij}^T v_i$ is the linear predictor for the hazards, and $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ and $z_{ij} = (z_{ij1}, \ldots, z_{ijs})^T$ are $p \times 1$ and $s \times 1$ covariate vectors corresponding to fixed effects $\beta = (\beta_1, \ldots, \beta_p)^T$ and log-frailties v_i, respectively. Here z_{ij} is often a subset of x_{ij}. In this paper, we assume $v_i \sim N_s(0, \Sigma_i)$, which is useful for modelling multi-component or correlated frailties. Here the covariance matrix $\Sigma_i = \Sigma_i(\theta)$ depends on θ, a vector of unknown parameters.

Let v_{i0} be a *random baseline intercept* and let v_{i1} be a *random slope*. If $z_{ij} = 1$ and $v_i = v_{i0}$ for all i, j, it becomes a random intercept or shared model with $\eta_{ij} = x_{ij}^T\beta + v_{i0}$ where $v_{i0} \sim N(0, \Sigma_i)$ with $\Sigma_i \equiv \sigma_0^2$ for all i. Let β_1 be the effect of primary covariate x_{ij1} such as the main treatment effect and let β_m $(m = 2, \ldots, p)$ be the fixed effects corresponding to the

covariates x_{ijm}. Our two random components lead to a bivariate model with

$$\eta_{ij} = v_{i0} + (\beta_1 + v_{i1})x_{ij1} + \sum_{m=2}^{p} \beta_m x_{ijm} \tag{7}$$

which is easily derived by taking $z_{ij} = (1, x_{ij1})^T$ and $v_i = (v_{i0}, v_{i1})^T$ in (1). Here

$$\begin{pmatrix} v_{i0} \\ v_{i1} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \Sigma_i \equiv \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right\}. \tag{8}$$

allowing a correlation term, $\rho = \sigma_{01}/(\sigma_0\sigma_1)$, between two random effects (v_{i0} and v_{i1}) within a centre thus extending the independent frailty model.

4 Model Interpretation

In order to interpret the fixed and random effects, we consider a nodel with a single binary-treatment indicator, x_{ij}. Then,

$$\lambda_{ij}(t|v_{i0}, v_{i1}; x_{ij}) = \lambda_0(t) \exp\{v_{i0} + (\beta_1 + v_{i1})x_{ij}\}.$$

Now, the time-dependent relative risk for treatment becomes

$$\psi_{ij}(t|x = 1, x = 0) = \frac{\lambda_0(t) \exp\{v_{i0} + (\beta_1 + v_{i1}) \cdot 1\}}{\lambda_0(t) \exp\{v_{i0} + (\beta_1 + v_{i1}) \cdot 0\}} = \exp(\beta_1 + v_{i1}), \tag{9}$$

which is free of time t and holds for all patients in centre i. Here $\exp(\beta_1)$ is the usual expression for the relative risk in a standard PH model. Thus, $\psi_{ij}(t|x = 1, x = 0)$, represents a random multiplicative divergence from the standard relative risk in a PH model which is homogeneous with respect to centres. Note that $\exp(\beta_1 + v_{i1})$ is often called the treatment hazard ratio in the ith centre. We also have that

$$\frac{\exp(\beta_1 + v_{i1})}{\exp(\beta_1)} = \exp(v_{i1}). \tag{10}$$

Thus v_{i1} means the random deviation of the ith centre from the overall treatment effect. Similarly, in order to interpret v_{i0} we consider the model without the covariate x_{ij} $\lambda_{ij}(t|v_{i0}) = \lambda_0(t) \exp(v_{i0})$ whence, $\phi_{ij}(t) = \frac{\lambda_0(t) \exp(v_{i0})}{\lambda_0(t) \exp(0)} = \exp(v_{i0})$ which is free of time t and holds for all patients in centre i, and v_{i0} represents the random deviation of the ith centre from the overall underlying baseline risk.

TABLE 1. Results for fitting the four models to the bladder cancer data

Model	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\sigma}_0^2$ (SE)	$\hat{\sigma}_1^2$ (SE)	$\hat{\sigma}_{01}$ (SE)	$[\hat{\rho}]$
M1 (Cox)	-0.667	0.509	–	–	–	
	(0.170)	(0.144)				
M2 (Indep)	-0.695	0.544	0.070	3×10^{-12}	–	
	(0.175)	(0.149)	(0.058)	(1×10^{-4})		
M3 (Corr)	-0.757	0.532	0.161	0.036	-0.068	[-0.893]
	(0.191)	(0.150)	(0.178)	(0.170)	(0.149)	
M4 (B)	-0.695	0.544	0.070	–	–	
	(0.175)	(0.149)	(0.058)			

M1: Cox model without frailties; M2: independent frailty model with $\rho = 0$;
M3: correlated frailty model with $\rho \neq 0$; M4: shared frailty model with random baseline risk (B) only; β_1 and β_2, effects of treatment and tumor status, respectively;
σ_0^2 and σ_1^2, the variances of random baseline risk and random treatment effect, respectively;
σ_{01} and ρ, the corresponding covariance and correlation with $\rho = \sigma_{01}/(\sigma_0\sigma_1)$;
SE, the estimated standard error for parameters.

5 Analysis of EORTC Trial Data

The duration of the Disease Free Interval (DFI) in non muscle invasive bladder cancer patients, treated in various centres in Europe, is analysed. The DFI is defined as the time from randomization to the date of the first recurrence. Patients without recurrence at the end of the follow-up period were censored at their last date of follow-up. For simplicity of analysis, we consider only 410 patients from 21 centres included in EORTC trial 30791. The two covariates of interest are: CHEMO x_{ij1} (0=No, 1=Yes) and TUSTAT x_{ij2} (0=Primary, 1=Recurrent). Notice that x_{ij1} is the main treatment covariate. The numbers of patients per centre varied from 3 to 78, with mean 19.5 and median 15. Of the 410 patients, 204 patients (49.8 per cent) without recurrence were censored at the date of last follow up.
For the purpose of analysis, we consider the three submodels of (3):, M1 (Cox): Cox model without frailties (basic hazard),, M2 (Indep): Cox models, with two independent frailty terms ($\rho = 0$),, M3 (Corr): Cox models, with two correlated frailty terms ($\rho \neq 0$).

Models M2 and M3 contain the random baseline risk v_{i0} and the random treatment-by-centre interaction term, $v_{i1}x_{ij1}$. The models were fitted using SAS/IML. The results are summarized in Table 1. In all three models the two fixed effects ($\beta_j, j = 1, 2$) are significant. In particular, the use of chemotherapy (CHEMO = 1) significantly prolongs the time to first recur-

rence as compared to patients who do not receive chemotherapy (CHEMO = 0). The two nested models (M1 and M2) ignoring random components or their correlation show similar results for β_j $(j = 1, 2)$. However, the absolute magnitude and SE of the estimate for the main treatment effect β_1 in M1 and M2 are smaller than those for the correlated model (M3). In M2 and M3, the variances (σ_0^2 and σ_1^2) indicate the amount of variation between centres in the baseline risk and in the treatment effect, respectively. Here, the estimate of σ_0^2 is relatively larger than that of σ_1^2. This does not seem surprising since differences in outcome according to treatment effect are typically smaller than differences due to patient characteristics which often vary across centres. However, care may be necessary in comparing the two variances because these two values should not be interpreted on the same scale.
Moreover, the correlated model M3 explains the degree of dependency between the two random components (i.e. the random centre effect v_0 and the random treatment-by-centre interaction v_1). The estimate of ρ ($\hat{\rho} = -0.893$) gives a large negative value, indicating that the two predicted random components ($\hat{v}_0$ and $\hat{v}_1$) have a strong negative correlation.

6 Discussion

The methods developed lead to an interesting analysis However, the modelling scheme described above needs to be extended in a number of important ways for use in routine biostatistical analysis. These issues which involve including individual level frailties and utilising focussed model selection will be discussed in the presentation.

Acknowledgments: This work was supported by the SFI's (www.sfi.ie) BIO-SI (www.ul.ie/bio-si) research programme, grant number, **07MI012**.

References

Cox, D. R. (1972) Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187-220.

Ha ID, Lee Y, & MacKenzie G (2007). Model selection for multi-component frailty models. Statistics in Medicine, 26, 4790-4807.

Ha ID, Sylvester R, Legrand C, & MacKenzie G. (2011). Frailty modelling for survival data from multi-centre clinical trials. Statist. Med. 2011, 30 2144-2159.

MacKenzie, G. (1996) Regression models for survival data. *J. R. Statist. Soc.* D **45**, 1, 21-34.

Model-based source estimation during foodborne disease outbreaks

Juliane Manitz[1], Thomas Kneib[1]

[1] Department of Statistics and Econometrics, University of Göttingen, Göttingen, Germany

E-mail for correspondence: jmanitz@uni-goettingen.de

Abstract: The 2011 E. coli outbreak in Germany exposed the lack of timely and efficient source detection as an integral part of mitigation strategies during foodborne disease outbreaks. Conventional public health source detection procedures use case-control studies and tracings along the food shipping chain. Such methods are typically very time-consuming and suffer from problems associated with data collected from patient interviews, such as bias. We introduce a new network theoretic method to estimate the spatial source of food-borne disease outbreaks and similar dynamical contagion phenomena. Our method requires only infection reports regularly collected by public health institutes and knowledge of the underlying food shipment network topology. We fit a hierarchical Bayesian spatio-temporal model to the infection counts, which uses the shortest path tree of the contaminated food shipping network as neighbourhood definition. Using Bayesian model comparison criteria, our method assignes an epicenter plausibility to each outbreak source candidate. We test our method in a spatial dynamic simulation model for foodborne diseases and specifically validate our approach for the German E. coli outbreak in 2011.

Keywords: Complex network; Spatio-temporal model; foodborne disease.

1 Statement of the problem

Recently observed frequent foodborne disease outbreaks acutely demonstrated the need for timely and efficient source detection in the case of a foodborne disease outbreak to public health bodies, risk assessment authorities, food industry, and society. Only efficient epidemic source detection and timely removal of the contaminated product can prevent further disease spread and impact on the population and economy.
Only in 66% of observed outbreaks, public health investigations were able to find an evidence for the infection source (O'Brien et al., 2006). One reason is the uncertain association between aetiology and food vehicle. Furthermore, the multi-disciplinary nature of a foodborne disease outbreak investigation task passes another major difficulty. It usually requires information from many sources including interview-based data from case-control and cohort

studies as well as tracings along the food shipping chain. We eliminate this source of bias by basing our new source estimation method only on the topology of the underlying food shipping network and infection reports. Our methodological developments were motivated during the 2011 EHEC/-HUS outbreak in Germany, which has been the largest worldwide reported *E. coli* outbreak regarding the number of severe HUS cases. The vast majority of the infections occurred in northern Germany, while other cases were travel related (see Figure 1). The source in the Lower Saxonian district Uelzen as well as outbreak trend and sprouts as transmission vehicle were hard to predict, because the outbreak was caused by a rare serotype O104:H4 (Frank et al., 2011).

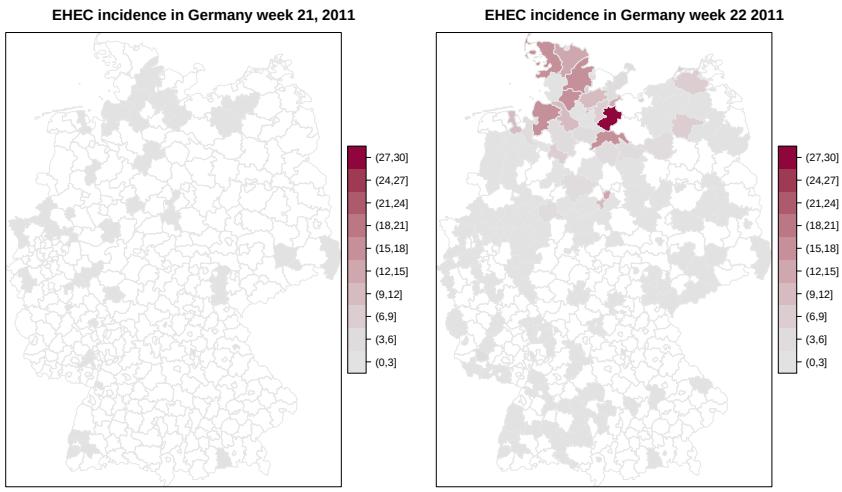


FIGURE 1. EHEC incidence in Germany in 21st and 22nd calendar week, 2011

2 Deterministic source detection

First, we introduce the deterministic groundwork method for source detection during foodborne disease outbreaks. On the basis of these ideas, we will develop the model-based source detection method in the next section. We define a food shipping network $G = (\mathcal{K}, \mathcal{L})$ as a collection of nodes $k \in \mathcal{K}$ connected by links $\mathcal{L}$, where $\mathcal{K} \neq \emptyset$ and $\mathcal{L}$ is a set of unordered pairs of elements of $\mathcal{K}$. In our context, nodes represent German districts while links refer to their trade connections. Thus, the network by definition includes all districts suspected to be the source of the outbreak, i.e. $k_0 \in \mathcal{K}_0 \subseteq \mathcal{K}$. For simplicity, we construct the network using the gravity law of trade (Anderson, 1979). Then, link weights are proportional to the connected node population and inversely proportional to their Euclidian distance.

For outbreak source tracing, we define the effective distance, which combines the deterministic distance, measured as shortest path length, and the corresponding path probability. In this way, we are able to reorganize the spatial pattern of infection counts to a circular tree representation for all source candidates $k_0 \in \mathcal{K}_0$ as root. We minimize the effective network distance of all potential sources to the median centre of reported epidemic mass. Looking at the circular tree representation with the true source as root node, it is easy to spot a circle-like pattern of infected nodes (see Figure 2).

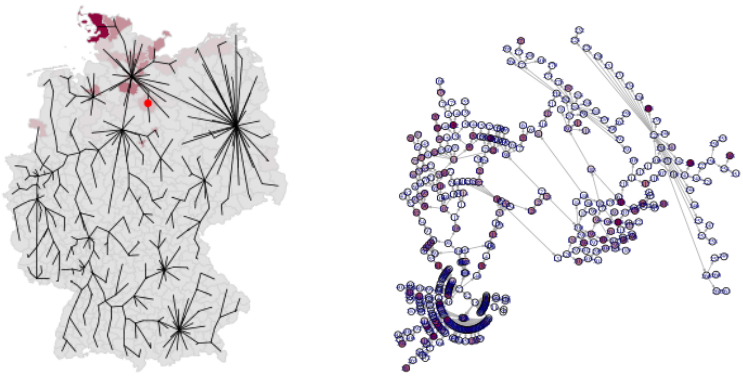


FIGURE 2. Deterministic source detection with application to EHEC/HUS data. (A) German map with shortest path tree. (B) Circular shortest path tree representation with root in Uelzen. Color-coding corresponds to the aggregated EHEC incidence during the first three outbreak weeks (19th-22nd calendar week 2011)

In extensive simulation studies, we used a novel spatial dynamic model for foodborne diseases and showed high detection probabilities in a variety of scenarios. These simulated scenarios were also used to assess the uncertainty of the estimated outbreak source through prediction with a generalized linear model (McCullagh and Nelder, 1989).

3 Model-based source detection

A statistical method for source detection should be able to deal with different types of uncertainty. The shipping network is defined under uncertainty, because its structures are highly adaptive to varying demand. During the outbreak, public health institute investigations gain knowledge from risk-oriented sampling. Moreover, the infectious disease counts show trend and seasonality and suffer from under-reporting and reporting delay.
We assess this problem by a hierarchical Bayesian spatio-temporal model, which distinguishes the regular background trend of sporadic cases and the

epidemic part given a specific source candidate using the network structure as neighbourhood definition. We fit this model for all potential source candidates and derive a posterior plausibility for being the source causing the given outbreak pattern.
We assume the number of infected y_{kt} in district k at time t to be negative Binomial distributed, i.e.

$$\begin{aligned} y_{kt}|\mu_{kt},\nu &\sim \mathrm{NB}(\mu_{kt},\nu), \\ \log(\mu_{kt}) = \eta_{kt} &= \log(E_k) + \beta_t + s_{k|k_0} + x_k, \quad t = 1,\ldots,T, k \in \mathcal{K}, \end{aligned}$$

where $\boldsymbol{\theta}_{kt} = (\beta_1,\ldots,\beta_T, s_{1|k_0},\ldots,s_{K|k_0}, x_1,\ldots,x_K,\nu)$. Thereby, β_t describes the epidemic time trend, $s_{k|k_0}$ the dispersal of the contaminated infection vehicle alog the food shipping chain, and x_k the local dispersal of sproradic cases. Furthermore, the model is rescaled by offset E_k, the population in district k.
The typical epidemic curve can be adapted by a time trend prior, which follows a random walk of order one, i.e.

$$\beta_t|\beta_{t-1},\tau_\beta \sim \mathrm{N}(\beta_{t-1},\tau_\beta^{-1})$$

Furthermore, we model the dispersal of the contaminated food item on a shortest path tree with source candidate $k_0 \in \mathcal{K}_0$ by

$$s_{k|k_0}|s_l, k \neq l, \mathbf{w}, k_0, \tau_s \sim \mathrm{N}\left(\sum_{k\sim l} \frac{w_{kl}}{w_{k+}} s_l, \frac{1}{w_{k+}\tau_s}\right),$$

where $w_{k+} = \sum_k w_{kl}$. The weights w_{kl} represent the link weights in the shortest path tree with root k_0.
For the regular background trend of sporadic cases, we assume local spatial dispersal. Thus, we define a standard Besag model prior for x_k

$$x_k|x_l, k \neq l, \tau_x \sim \mathrm{N}\left(\frac{1}{n_k}\sum_{k\sim l} x_l, \frac{1}{n_k\tau_x}\right),$$

where $k \sim l$ represent links in the local neighborhood structure.
On another level, we plan to incorporate uncertainty about the food shipping network structure by assigning priors to the shortest path tree weights w_{kl}

$$w_{kl} \sim \mathrm{Ga}(\nu/2, \nu/2).$$

The estimation of the model becomes very complex, but can be solved elegantly using a strategy suggested by Brezger et al. (2007).
We fit this Bayesian model with epidemic dispersal along differing shortest path trees. They represent the efficient food shipping network of a contaminated infection vehicle with root in a potential source candidates $k_0 \in \mathcal{K}_0$. Comparing these model fits for different sources, we can assign a plausibility to each tested potential source candidate for being the true source of the observed outbreak.

4 Conclusion

Altogether, the introduced model-based source detection method for foodborne disease outbreaks uses minimal data basis and introduces a network-theoretic approach for detection. Alternative approaches are time-consuming and usually based on potentially biased patient interview data. The method is designed to work for various indirectly transmitted infectious diseases, but has as well the potential to detect the origin of various other propagtion patterns, such as the spread of technical innovations in agriculture or delays in a railway system.

Acknowledgments: This research has been financially supported by grants from the German Science Foundation (DFG), Research Training Group 1644 'Scaling problems in Statistics'.

References

Anderson, J. (1979). A Theoretical Foundation for the Gravity Equation. *American Economic Review*, **69**, 106 – 116.

Brezger, A., Fahrmeir, L. and Hennerfeind, A. (2007). Adaptive Gaussian Markov random fields with applications in human brain mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **56**, 327 – 345.

Frank, C., et al. (2011). Epidemic Profile of Shiga-Toxin Producing Escherichia coli O104:H4 Outbreak in Germany. *New England Journal of Medicine*, **365(19)**, 1771 – 1780.

Manitz, J., Kneib, T., Schlather, M., and Brockmann, D. (2013). Network-based Source Detection for Foodborne Disease Epidemics Applied to the German 2011 E. coli outbreak. *Working paper.*

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models.* London: Chapman & Hall.

O'Brien, S.J., Gillespie, I.A., Sivanesan, M.A., Elson, R., Hughes, C., and Adak, G.K. (2006). Publication bias in foodborne outbreaks of infectious intestinal disease and its implications for evidence-based food policy. England and Wales 19922003. *Epidemiology and Infection*, **134**, 667 – 674.

Boosting sonographic birth weight estimation

Andreas Mayr[1], Florian Faschingbauer[2], Matthias Schmid[1]

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] Department of Obstetrics and Gynecology, FAU University Hospital, Erlangen, Germany

E-mail for correspondence: andreas.mayr@fau.de

Abstract: Sonographic measurements of the fetus are often used to predict birth weight as it is the most important indicator for possible complications during delivery. Statistical challenges in the modelling of these kind of prediction models include multicollinearity and variable selection. The aim of our investigation is to analyze if modern variable selection and regularization tools like component-wise boosting algorithms, which can cope with both of those issues, can improve existing prediction formulas. We therefore consider boosting generalized additive models (GAM) as well as the recently proposed, more flexible gamboostLSS algorithm for boosting generalized additive models for location, scale and shape (GAMLSS). As distribution-free competitor we applied additive quantile regression boosting.

Keywords: Gradient boosting; GAMLSS, quantile regression

1 Background

Ultrasound measurements during pregnancy are often used to estimate the weight of the fetus and to compare the resulting values with standardized growth charts. In the last few days before delivery, the focus shifts towards predicting the birth weight (BW, see Figure 1), which has a high clinical relevance as both very low as well as very high birth weight are associated with a greater risk of complications during labor and the first days after birth. There exists a broad range of prediction formulas for birth weight based on different anthropometric measurements of the fetus by ultrasound. One of the most popular formulas, used in clinics around the world, is the Hadlock III formula which will serve as benchmark in our analysis. It was the best-performing formula among eleven others in a recent large-scale comparison (Siemer et al., 2008). Our analysis is based on a large naturalistic study containing 10281 singleton pregnancies in Germany between 2003 and March 2013. The main statistical challenges in the modelling of these kind of data are *multicollinearity* and *variable selection*. Both issues can be addressed by component-wise boosting algorithms.

FIGURE 1. Example of standard birth weight prediction. Sonographic image on the left and output containing resulting prediction (3710g ± 542g) on the right.

2 Methods

All presented regression models are estimated by component-wise boosting algorithms, which have their roots in the field of machine learning. The basic idea is to iteratively apply simple regression tools (base-learners) and aggregate them to a final additive model. In our case, we apply P-splines as base-learners to account for possible non-linear effects of the predictors. The boosting algorithm is then stepwise descending the empirical risk by fitting the gradient of the loss function to the base-learners. Fitting is carried out component-wise: Each base-learner typically corresponds to one single predictor and the base-learners are fitted one at a time. This procedure effectively overcomes the problem of multicollinearity.
Furthermore, in every iteration, boosting algorithms only include the effect of the best-performing base-learner in the resulting additive model. If the algorithm is stopped before each base-learner was included at least once, the corresponding left-out predictors are excluded from the final model which hence leads to an automated and fully data-driven variable selection.
By adapting the gradient, various loss functions can be optimized, leading to various regression settings ranging from modelling count data to time-to-event analysis (Bühlmann and Hothorn, 2007). In our analysis, we first consider the L_2 loss leading to a common GAM, where $f_1(\cdot), ..., f_p(\cdot)$ represent unspecified functional forms for the effects of components $X_1, ..., X_p$ on the expected mean of the conditional distribution:

$$\mathrm{E}(Y|X) = \mu(X) = \beta_0 + f_1(X_1) + ... + f_p(X_p)$$

A far more flexible model class is GAMLSS (Rigby and Stasinopoulos, 2005) which goes beyond mean regression by modelling not only the location via the expected value, but also other distribution parameters like scale and shape parameters. In order to fit GAMLSS via boosting, the recently proposed gamboostLSS algorithm (Mayr et al., 2012) extended the common

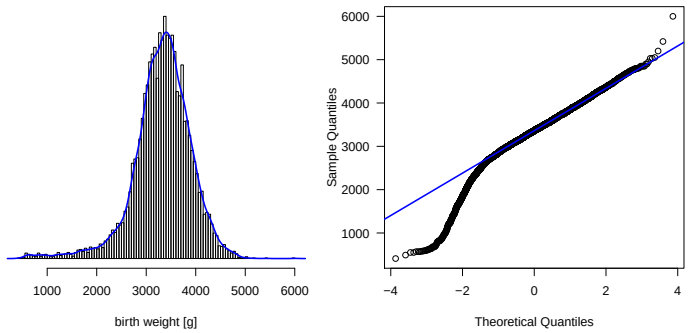


FIGURE 2. Distribution of birth weight in our sample (left) and a QQ-plot (right) comparing it to a standard normal.

component-wise boosting approach to multiple parameter dimensions. In every boosting iteration, the algorithm circles between the different gradients corresponding to different parameters. Via boosting GAMLSS, we can hence account also for possible effects on the variance of the conditional distribution while carrying out variable selection for two additive models.

$$\begin{aligned} \mathrm{E}(Y|X) &= \mu(X) = \beta_{0\mu} + f_{1\mu}(X_1) + ... + f_{p\mu}(X_p) \\ \mathrm{Var}(Y|X) &= \sigma(X) = \exp(\beta_{0\sigma} + f_{1\sigma}(X_1) + ... + f_{p\sigma}(X_p)) \end{aligned}$$

A completely distribution-free approach is quantile regression, which can also be successfully estimated via boosting by considering the check-function as corresponding loss (Fenske et al., 2011). The fundamental idea is not to focus on parameters of an assumed distribution but to directly model the τ-quantiles of the conditional distribution.

$$\mathrm{Q}_\tau(Y|X) = \beta_{0\tau} + f_{1\tau}(X_1) + ... + f_{p\tau}(X_p)$$

For our application to predict birth weight, we specified $\tau = 0.5$, leading to robust median boosting via the L_1 loss.

3 Results

Due to the naturalistic study design, our sample contains also various extreme observations (see Figure 2): 191 babies (1.9%) have a very low birth weight $\leq$ 1600g and 137 babies (1.3%) are diagnosed with extreme fetal macrosomia (BW $\geq$ 4500g). The median birth weight in our data set is 3360g (range: 410g – 6000g). On the predictor side, the data set contains

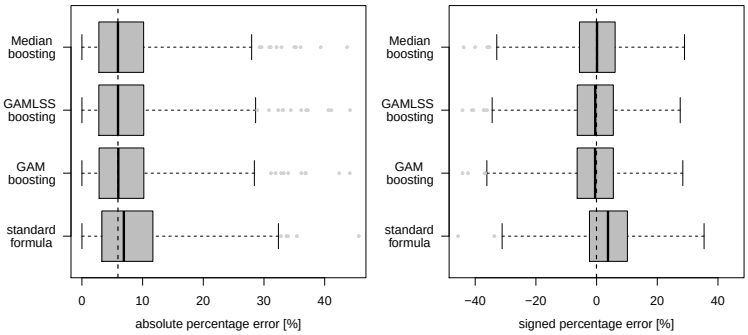


FIGURE 3. APE (left) and SPE (right) on the test data for different model classes, the dashed line represents the best performing quantile regression.

seven different anthropometric measurements of the fetus (e.g., head circumference, femur length). The correlation between the measurements is high, ranging from 0.46 to 0.91 which could lead to severe multicollinearity problems, at least for standard fitting algorithms. Additionally to the main effects, we considered all pairwise interactions, leading to a total of 28 potential predictor variables which are all represented by separate P-spline base-learners.

To assess the prediction accuracy, we split the data in 6000 training observations which are used to fit the different formulas and 4281 test observations. The tuning of the boosting algorithms via their stopping iteration (which controls variable selection and shrinkage of effect estimates) is carried out via 25-fold bootstrapping on the training observations. In case of gamboostLSS, in order to account for different complexities of the mean and variance models, we searched a two-dimensional grid of possible combinations of stopping iterations.

In the literature on sonographic weight estimation, the performance of prediction formulas is typically compared regarding two measures: The absolute percentage error ($\text{APE} = 100 \cdot \frac{|\text{BW}-\widehat{\text{BW}}|}{\text{BW}}$) as well as the signed percentage error ($\text{SPE} = 100 \cdot \frac{\text{BW}-\widehat{\text{BW}}}{\text{BW}}$). The resulting APE and SPE values are presented in Table 1 (see also Figure 3). The best performing model class, at least regarding these two measures, is the non-parametric quantile boosting approach (in this case, median boosting) followed by the predictions of GAMLSS boosting. It has to be noted, that all boosting methods consistently outperform the best performing standard formula, regarding prediction accuracy. However, this effect diminishes to some extent if the standard Hadlock formula is re-fitted on our training data (see Table 1).

TABLE 1. Percentage errors of the different prediction schemes on test data.

	absolute percentage error			signed percentage error		
method	median	mean	sd	median	mean	sd
median boosting	5.86	7.07	5.58	-0.19	-0.29	9.01
GAMLSS boosting	5.87	7.12	5.63	-0.71	-0.83	9.04
GAM boosting	5.88	7.14	5.61	-0.77	-0.84	9.04
standard formula	6.64	7.92	6.11	3.67	3.58	9.35
re-fitted	6.17	7.54	6.32	-0.77	-0.85	9.81

4 Discussion

Our data analysis on boosting sonographic birth weight estimation demonstrates the ability of component-wise gradient boosting algorithms to yield sparse and accurate prediction formulas based on various loss functions. All three considered model classes, the common GAM approach, the more flexible GAMLSS and also quantile regression outperformed the highly popular and accurate Hadlock III formula which is used in clinics around the world. The boosting algorithms selected sparse models in a data-driven manner, choosing from 28 highly correlated predictors. The quantile regression approach lead to the best results, which might be explained by the higher robustness towards outliers due to the L_1 loss.

Another advantage of quantile regression, but also of GAMLSS, is their greater flexibility when it comes to prediction intervals. While the size of common prediction intervals (which are based on GAMs or other standard mean regression approaches) are fixed, in case of GAMLSS or quantile regression they may also depend on the information of the predictors. This leads to subject-specific instead of population-specific interval sizes (Figure 4). As a result, the size of the individual prediction interval differs from fetus to fetus, although they could have the same point-prediction. The intervals can therefore report a greater level of uncertainty for some fetuses instead of giving a general level of certainty for the complete sample.

However, one has to acknowledge that the benefit of modern boosting algorithms in combination with P-splines, regarding the accuracy of the point predictions is relatively small. Standard birth weight prediction formulas, developed in the 1980s with classical linear models, might on average underestimate the birth weight in this large sample of German fetuses from 2003 to 2013. However, they are a strong competitor if they are re-fitted on the new training data and therefore adjusted for a possible trend towards higher birth weight or regional differences.

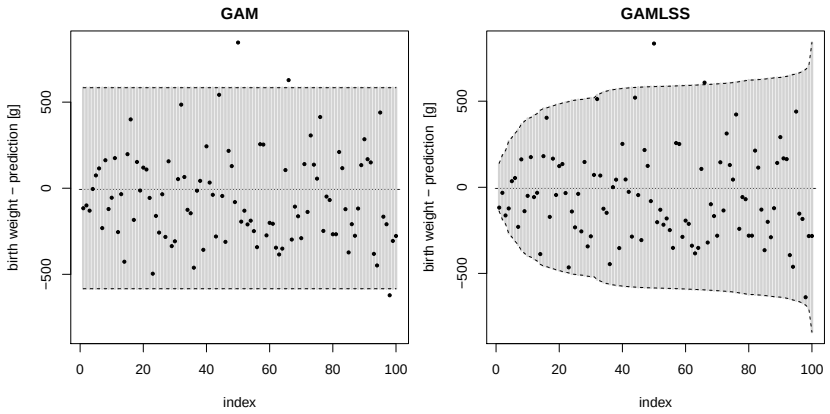


FIGURE 4. Comparing the prediction intervals from GAM boosting and GAMLSS boosting on 100 randomly selected fetuses from the test-data set. The plots are centered around the point prediction; the dark points are the observed birth weights; the intervals are shaded gray. The 100 observations are ordered by the size of the GAMLSS interval, the x-axis refers to this rank.

Acknowledgments: The work of Andreas Mayr and Matthias Schmid was supported by Deutsche Forschungsgemeinschaft (DFG) (www.dfg.de), grant SCHM 2966/1-1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477-522.

Fenske, N., Kneib, T., Hothorn, T (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, **106**(494): 494-510.

Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012). Generalized additive models for location, scale and shape for high-dimensional data – a flexible approach based on boosting. *Applied Statistics*, **61**(3), 403-427.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**, 507-554.

Siemer, J., Egger, N. , Hart, N., et al. (2008). Fetal weight estimation by ultrasound: comparison of 11 different formulae and examiners with differing skill levels. *European Journal of Ultrasound*, **18**, 159-164.

Bayesian P-splines with a multiplicative term in EMG trace data

James P. McKeone[1], Anthony N. Pettitt[1]

[1] Queensland University of Technology, Australia

E-mail for correspondence: james.mckeone@qut.edu.au

Abstract: A method is proposed to describe force or compound muscle action potential (CMAP) trace data collected in an electromyography study for motor unit number estimation (MUNE). Experimental data was collected using incremental stimulation at multiple durations. However, stimulus information, vital for alternate MUNE methods, is not comparable for multiple duration data and therefore previous methods of MUNE (Ridall et al., 2006, 2007) cannot be used with any reliability. Hypothesised firing combinations of motor units are modelled using a multiplicative factor and Bayesian P-spline formulation. The model describes the process for force and CMAP in a meaningful way.

Keywords: Bayesian P-splines; MUNE; Multiple duration data

1 Introduction

Reliable motor unit number estimates (MUNE) are of key interest in clinical and experimental neurophysiology. MUNE has been a popular area of research since the seminal work of McComas et al. (1971) who pioneered the method of incremental stimulation in electromyography (EMG). MUNE is efficacious in assessing the severity or tracking the progression of diseases characterised by muscle weakness resulting from the death or inaction of motor units (Baumann et al., 2012). An exciting application of MUNE is being pursued by the Miami project to cure paralysis, who research methods to replace dead nervous system cells and promote and guide the regrowth of axons with the aims of muscle re-innervation and ultimately motor unit function for patients suffering spinal injury or damage to the central nervous system (Casella et al., 2010). MUNE may be used to investigate the success of a cell transplant, specifically, whether a treatment has produced new, active motor units.

MUNE techniques involve analysis of compound muscle action potential (CMAP) or force data resulting from electromyography (EMG) studies. In a clinical setting, a surface EMG study involves applying an electrical stimulus at the nerve for a fixed duration to observe a CMAP or force

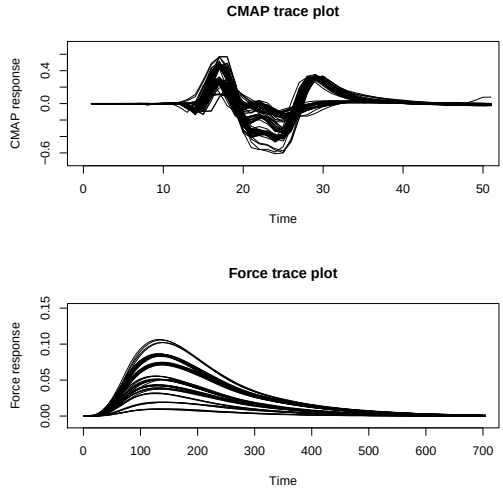


FIGURE 1. Electromyography (EMG) responses to stumulus collected on an anesthetised rat with stimuli applied at three durations, 10, 20 and 50 μs. **(Top)** Compound muscle action potential (CMAP) trace data. **(Bottom)** Force trace data.

response through electrodes or transducers attached to the skin. The technique of incremental stimulation involves setting the stimulus intensity such that no units are firing initially, and gradually incrementing the stimulus intensity, invoking more motor units until all are believed to be firing. Clinicians tend to adopt a relatively long duration (usually $50\mu s$) and therefore smaller stimulus intensities to maintain a level of comfort for the patient. The trace responses for each stimulus intensity can be plotted against time for CMAP and force, see Figure 1.

In an experimental setting, specifically the case considered here of anethetised rats, it is more common to apply the incremental stimulation technique at multiple durations. It is believed that this allows the process of alternation and therefore individual motor units to be investigated more thoroughly (Casella et al., 2010). However, it is not immediately obvious how existing MUNE techniques, all of which rely on stimulus intensity information as input data, may be adjusted to account for multiple durations. Suppose an experienced clinician applied the incremental stimulation technique and identified the combinations of motor units thought to be active at each particular stimulus. Such information is labelled as a firing pattern. A potential firing pattern for investigation could be the result of a naïve approach as the former suggestion or a statistically based approach, such as MUNE techniques using Bayesian hierarchical modelling (Ridall et al., 2006) combined with reversible jump Markov chain Monte Carlo sampling (Ridall et al., 2007). These MUNE techniques are henceforth referred to as Bayesian MUNE. An example of a firing pattern for an assumed 9 unit model is presented in Figure 2.

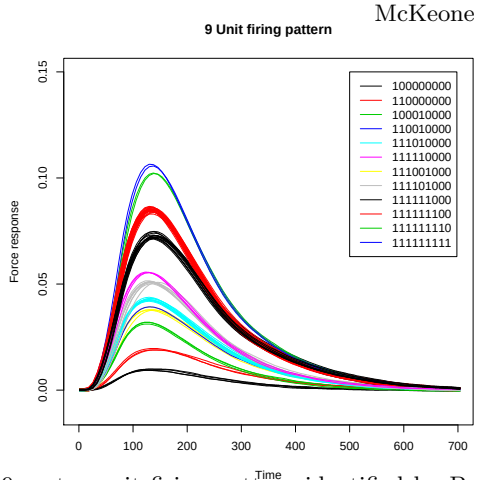


FIGURE 2. A 9 motor unit firing pattern identified by Bayesian MUNE. The set of units active within each group are identified in the legend, 1 is active, 0 is inactive for each particular motor unit.

The contribution of this work is to investigate the plausibility of different firing patterns − and usually different numbers of total motor units − in multiple duration EMG data, without using stimulus information to inform the model.

Consider a set of plausible firing patterns objectively identified by Bayesian MUNE. The stimulus data is adjusted using an approximation based on the strength-duration relationship (Hill, 1936) that assumes the stimulus multiplied by duration is constant before the algorithm is implemented. The output from Bayesian MUNE is a relatively small number of plausible firing patterns corresponding to values of the total number of units. The posterior model probabilities from Bayesian MUNE may not be relied upon due to the uncertainty in making the varying stimulus duration correction. Here, trace data and each potential firing pattern are modelled using Bayesian P-splines with a multiplicative scale factor. The scale term shifts the spline to the mean of each pre-determined group.

2 Bayesian P-splines and Multiple Duration Data

Let i be a location on curve j in group k. A group is a set of motor units active at a particular stimulus intensity. Bayesian MUNE suggests seven potential firing patterns for subsequent investigation. Consider the model

$$y_{ijk} = \phi_k g(x_i) + \epsilon_{ijk} \tag{1}$$

that allows the trace data y_{ijk} to be represented as a spline $g(x_i)$, that is scaled to each cluster by a coefficient ϕ_k and error ϵ_{ijk}. The spline may be written as $g(\mathbf{x}) = X\beta$ for $\mathbf{x} = (x_1, \ldots, x_m)'$, with design matrix X and a

vector of coefficients, β. The spline is fitted as a reference shape across all curves.

A multivariate normal likelihood, $\mathbf{y}_{j,k} \sim N(\phi_k X\beta, \epsilon^2 I)$, is assumed for these data . The model variance parameter ϵ^2 is assigned the prior distribution $\epsilon^2 \sim$ Inverse-Gamma$(\alpha_\epsilon, \beta_\epsilon)$ and assumed fixed across all curves and groups.

The spline $X\beta$ is formulated as a Bayesian P-spline with a Bézier basis in the method set out by Eilers and Marx (1996) and extended by Lang and Brezger (2004). A random walk difference prior is placed on the coefficients β such that,

$$p(\beta|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{\frac{\text{rank}(K)}{2}} \exp\left(-\frac{1}{2\tau^2}\beta^{\mathrm{T}} K\beta\right) , \tag{2}$$

for the first-order difference matrix K, with the variance parameter defined as $\tau^2 \sim$ Inverse-Gamma$(\alpha_\tau, \beta_\tau)$ a priori. It is important that the prior on τ^2 is not too diffuse, the values $\alpha_\tau = 1$, $\beta_\tau = 0.005$ were adequate, but ultimately the prior was found to have little effect on the posterior. It remains to describe how the scale parameters ϕ_k will be dealt with, two prior formulations are considered.

In keeping with the assumptions applied in Bayesian MUNE (Ridall et al. 2006), the largest observed group of firing motor units is assumed to be the sum of all units in the firing pattern. The result of this assumption is a data dependent prior on units, μ_n, $n = 1, \ldots, N$, parameters to represent individual unit size. Each scale coefficient, ϕ_k, is assumed to be defined by the unique sum $\phi_k = \sum \mu_m$ for units identified by the firing pattern. A Dirichlet prior is a natural formulation for coefficients μ_n. With N the total number of units and K the group with the largest members in terms of amplitude, $\sum_{n=1}^{N} \mu_n = \phi_K = 1$ is satisfied under a Dirichlet prior specification. Therefore for individual units write, $\mu_n \sim$ Dirichlet(α_μ) for some concentration parameter α_μ. The implicit constraint $\sum_{n=1}^{N} \mu_n = 1$ is sufficient for identifiability of the spline and the scale coefficients.

2.1 Posterior Distribution and Algorithm

The full-conditional distributions for parameters ϵ^2, β and τ^2 may be found after some algebra. The posterior distribution for the μ_n has no closed form density and so we use a Metropolis-Hastings within Gibbs approach to sampling from the posterior distribution in tandem with a systematic approach to sampling and updating parameters from the full-conditional distributions.

3 Results

Figure 3 presents the data and fitted models for each group of curves according to an assumed eleven unit firing pattern. The model describes the

smooth force curves very well though produces what may seem like a less convincing fit for the CMAP traces. For CMAP data, the region of the curve thought to be most important in action potential investigations is the first peak and while CMAP curves of different groups have heterogeneous shape, the model does indeed fit the first peak quite well in each group.

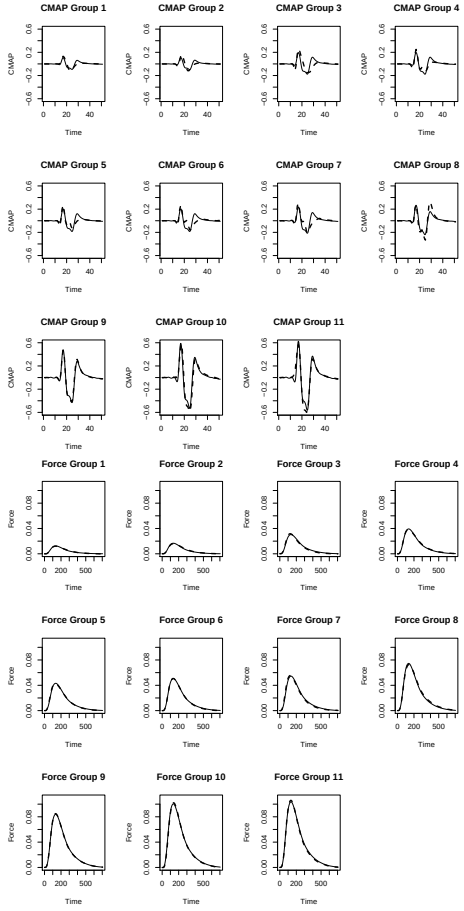


FIGURE 3. Model fit for an eleven unit firing pattern. **(Top)** CMAP response. **(Bottom)** Force response. Individual group data (solid line) and the fitted multiplicative spline (dashed line).

4 Discussion

The force trace data are used as illustration for the model, as seen in Figure 1. However, CMAP data is far more prevalent in clinical and experimental

investigations involving MUNE. A spline based approach was adopted so as not to exclude the possibility of analysing the more erratic CMAP traces. The model considered is a method to describe different firing patterns for the force or CMAP trace data with stimulus applied at the nerve at multiple durations. Future work involves deciding between different suggested firing patterns in an automatic way. Model fit calculations such as DIC and BIC perform poorly so we propose a model choice approach to target the marginal likelihood or statistical evidence, though such a presentation is beyond the scope of this short paper.

Acknowledgments: The authors would like to thank Christine Thomas, Gareth Ridall and Chris Drovandi for assistance. James McKeone is grateful for the support of an Australian Postgraduate Award (APA). Tony Pettitt is supported by an Australian Research Council Discovery Grant.

References

Baumann, F., Henderson, R.D.. Ridall, P.G.. Pettitt, A.N. and McCombe, P.A. (2012). Use of Bayesian MUNE to show differing rate of loss of motor units in subgroups of ALS. *Clinical Neurophysiology*, **123**, 2446 – 2453.

Casella, G.T., Almeida, V.W., Grumbles, R.M., Liu, Y. and Thomas, C.K. (2010). Neurotrophic factors improve muscle reinnervation from embryonic neurons. *Muscle and Nerves*, **42(5)**, 788 – 797.

Eilers, P.H. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11(2)**, 89 – 121.

Hill, A.V. (1936). Excitation and accommodation in nerve. *Proceedings of the Royal Society of London, Serial B*, **116**, 305 – 355.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13(1)**, 183 – 212.

McComas, A., Fawcett, P., Campbell, M. and Sica, R. (1971). Motor unit number estimation - a Bayesian approach. *Journal of Neurology, Neurosurgery, and Psychiatry*, **34**, 121 – 131.

Ridall, P.G., Pettitt, A.N., Henderson, R.D. and McCombe, P.A. (2006). Motor unit number estimation in human neurological diseases and animal models. *Biometrics*, **62**, 1235 – 1250.

Ridall, P.G., Pettitt, A.N., Friel, N., McCombe, P.A. and Henderson, R.D. (2007). Motor unit number estimation in human neurological diseases and animal models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **56(3)**, 235 – 269.

Testing of Growth Curves using Cubic Smoothing Splines: A Multivariate Approach

Nicholas Mesue [1], Tapio Nummi [2]

[1] School of Health Sciences, FIN–33014, University of Tampere, Finalnd.
[2] School of Health Sciences, FIN–33014, University of Tampere, Finalnd.

E-mail for correspondence: Nicholas.Mesue@uta.fi

Abstract: The primary objective of this paper is to present a new method for testing growth curves. The method is based on spline approximation and on F-test. This method also applies under a certain type of correlation structures that are especially important in the analysis of repeated measures and growth data. In this paper, it is shown how the basic spline model and the test can be extended to more general multiple response situation. The methods are illustrated by real data sets. The new method proved to be a very powerful modeling and testing tool especially in situations where the growth curve may not be easily approximated using simple parametric models.

Keywords: Balanced data; Correlated observations; F–test; Longitudinal data.

1 Spline Growth Model

Estimation has been the main focus of smoothing spline methods whereas hypothesis testing has not received a considerable attention. The idea here is to test if the progression in time is equal over the set of correlated observations for groups of data for a single measured response variable and then briefly introduce its extension to a multivariate situation. The method presented here was initially introduced by Nummi and Mesue (2013). The growth curve model for complete and balanced data (Generalized multivariate analysis of variance, GMANOVA) model was introduced by Potthoff and Roy (1964). This model can be written as

$$\mathbf{Y} = \mathbf{TBA}^{\mathrm{T}} + \mathbf{E}, \tag{1}$$

where $\mathbf{Y} = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n})$ is the matrix of n mutually independent response vectors, $\mathbf{T}$ is a $q \times p$ within-individual design matrix, $\mathbf{A}$ is an $n \times m$ between-individual design matrix, $\mathbf{B}$ is an unknown $p \times m$ parameter matrix to be estimated and $\mathbf{E}$ is a $q \times n$ matrix of random errors. It is assumed that the columns $\mathbf{e_1}, \ldots, \mathbf{e_n}$ of $\mathbf{E}$ are independently normally distributed i.e., $\mathbf{e_i} \sim N(\mathbf{0}, \mathbf{\Sigma}),\ i = 1, \ldots, n$; where $\mathbf{\Sigma}$ is assumed to be unstructured.

Using smooth curves, the model (1) can be written in a more general form as

$$\mathbf{Y} = \mathbf{G}\mathbf{A}^{\mathrm{T}} + \mathbf{E}, \tag{2}$$

where $\mathbf{G} = (\mathbf{g_1}, \ldots, \mathbf{g_m})$ is the matrix of smooth mean growth curves in time points $t_1, t_2, \ldots, t_q$. We assume that the covariance matrix $\mathbf{\Sigma}$ takes certain type of parsimonious structure $\mathbf{\Sigma} = \sigma^2\mathbf{R}(\theta)$ with covariance parameters θ. Model (2) is now referred to as the ***Spline Growth Model*** (SGM). The growth curve model of Potthoff and Roy (1964) is now the special case $\mathbf{G} = \mathbf{TB}$.

2 Model Estimation

The smooth solution for $\mathbf{G}$ can be obtained by minimizing the penalized least squares (PLS) criterion

$$Q = \mathrm{tr}[(\mathbf{Y} - \dot{\mathbf{G}})^{\mathrm{T}}\mathbf{H}(\mathbf{Y} - \dot{\mathbf{G}}) + \alpha\dot{\mathbf{G}}\mathbf{K}\dot{\mathbf{G}}], \tag{3}$$

where $\alpha > 0$ is a fixed smoothing parameter, $\dot{\mathbf{G}} = \mathbf{G}\mathbf{A}^{\mathrm{T}}$, $\mathbf{H} = \mathbf{R}^{-\mathbf{1}}$ and the roughness matrix $\mathbf{K}$, (from the roughness penalty, $RP = \int g''^2$) is defined as

$$\mathbf{K} = \nabla\mathbf{\Delta}^{-\mathbf{1}}\nabla^{\mathrm{T}}, \tag{4}$$

where the non-zero elements of banded $q \times (q-2)$ and $(q-2) \times (q-2)$ matrices ∇ and $\mathbf{\Delta}$ are defined as

$$\nabla_{k,k} = \frac{1}{h_k},\ \nabla_{k+1,k} = -(\frac{1}{h_k} + \frac{1}{h_{k+1}}),\ \nabla \tag{5}$$

and

$$\mathbf{\Delta}_{k,k+1} = \mathbf{\Delta}_{k+1,k} = \frac{h_{k+1}}{6},\ \mathbf{\Delta}_{k,k} = \frac{h_k + h_{k+1}}{3}, \tag{6}$$

where $h_j = x_{j+1} - x_j, j = 1, 2, ..., (q-1)$ and $k = 1, 2, ..., (q-2)$.
For a given α and $\mathbf{H}$, the spline estimator becomes as

$$\tilde{\mathbf{G}} = (\mathbf{H} + \alpha\mathbf{K})^{-\mathbf{1}}\mathbf{H}\mathbf{Y}\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-\mathbf{1}}. \tag{7}$$

For more details, see Nummi and Mesue (2013) and Green and Silverman (1994). However, the covariance matrix $\mathbf{H}$ may not be known and therefore the estimator (7) would be difficult to use in practical situations. Fortunately, it can be shown that in certain important special cases the general spline estimator (7) simplifies to simple linear functions of the original observations $\mathbf{Y}$. One obvious condition for such kind of simplification is $\mathbf{KR} = \mathbf{K}$. If this condition holds, we get $\mathbf{K} = \mathbf{KH}$ and therefore, the spline estimator $\tilde{\mathbf{G}}$ simplifies to

$$\hat{\mathbf{G}} = (\mathbf{I} + \alpha\mathbf{K})^{-\mathbf{1}}\mathbf{Y}\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-\mathbf{1}} = \mathbf{S}\mathbf{Y}\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-\mathbf{1}}, \tag{8}$$

where the smoother matrix is $\mathbf{S} = (\mathbf{I} + \alpha\mathbf{K})^{-\mathbf{1}}$.
Some important special cases for growth data are $\mathbf{R} = \mathbf{I}$ (Independence), $\mathbf{R} = \mathbf{I} + \sigma^{\mathbf{2}}_{\mathbf{d}}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ (Uniform or compound symmetry), $\mathbf{R} = \mathbf{I} + \sigma^{\mathbf{2}}_{\mathbf{d}'}\mathbf{X}\mathbf{X}^{\mathrm{T}}$ (Linear with constant variance) and $\mathbf{R} = \mathbf{I} + \mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}$ (Linear), where $\mathbf{X} = (\mathbf{1}, \mathbf{x})$ and $\mathbf{x}$ is a vector of q measuring times.

3 Spline Approximation

In general, the smoother matrix § is not a projection matrix and therefore certain results, e.g. in testing, developed for general linear models are not directly applicable. Our approach is to utilize an approximation for the smoother matrix § with the properties of a projection matrix. Clearly, one obvious approximation of the spline fit (8) is the spline estimator

$$\bar{\mathbf{G}} = \mathbf{P}_{\mathbf{m}}\mathbf{Y}\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-\mathbf{1}}, \tag{9}$$

where $\mathbf{P}_{\mathbf{m}} = \mathbf{M}_*\mathbf{M}_*^{\mathrm{T}}$ and $\mathbf{M}_*$ contains the $c(\leq q)$ first eigenvectors of $\mathbf{M}$. This arises from the eigenvalue decomposition of the smoother

$$\mathbf{S} = \mathbf{M}(\mathbf{I} + \alpha\mathbf{\Lambda})^{-\mathbf{1}}\mathbf{M}^{\mathrm{T}}, \tag{10}$$

where $\mathbf{M}$ is the matrix of q orthogonal eigenvectors of $\mathbf{K}$ and **Lambda** is a diagonal matrix of corresponding q eigenvalues. Note that the matrices $\mathbf{K}$ and $\mathbf{S}$ share the same set of eigenvectors $\mathbf{m_1}, \mathbf{m_2}, \ldots, \mathbf{m_q}$. Here, the eigenvectors are ordered by eigenvalues of $\mathbf{S}$. It is known that the sequence of eigenvectors appears to increase in complexity like a sequence of orthogonal polynomials. The first two eigenvalues of $\mathbf{S}$ are always 1. The first two eigenvectors $\mathbf{m}_1$ and $\mathbf{m_2}$ span the subspace corresponding to the straight line model. In the mixed model formulation of the spline solution e.g. Verbyla et al., (1999), this corresponds to the fixed part of the model. The smoother matrix $\mathbf{S}$ and the smoothing parameter need not be computed here. However, the number of eigenvectors c from $\mathbf{K}$ used in the approximation need to be calculated. This can be done by using a modified Generalized Cross-Validation criterion. See Nummi and Mesue (2013) for details.

4 Hypothesis Testing

Consider the set of fitted spline curves $\hat{\mathbf{Y}} = \hat{\mathbf{G}}\mathbf{A}^{\mathrm{T}}$ From model (7), we may use the approximation

$$\hat{\mathbf{Y}} = \bar{\mathbf{G}}\mathbf{A}^{\mathrm{T}} = \mathbf{M}_*\hat{\mathbf{\Omega}}\mathbf{A}^{\mathrm{T}}, \tag{11}$$

where we denoted $\hat{\Omega} = \mathbf{M}_*^{\mathrm{T}}\mathbf{Y}\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-\mathbf{1}}$. All the relevant information for ***testing mean profiles*** is now in the matrix $\hat{\Omega}$, which can now be considered to be an unbiased estimate of the unknown parameter matrix of

the statistical model $E(\mathbf{Y}) = \mathbf{M}_*\mathbf{\Omega}\mathbf{A}^{\mathbf{T}}$. Therefore in sequel, we confine in testing linear hypothesis of the form

$$H_0 : \mathbf{C\Omega D} = \mathbf{0},$$

where $\mathbf{C}$ and $\mathbf{D}$ are known $\nu \times c$ and $m \times g$ matrices with ranks ν and g, respectively. Testing can be based on the ratio of sum of squares matrices. First, dropping the first eigenvector $\mathbf{m}_1$ corresponding to the constant term in the approximation model, we can take $\mathbf{C} = [\mathbf{0}, \mathbf{I}]$. If we also assume the uniform covariance model $\mathbf{R} = d^2\mathbf{1}\mathbf{1}^{\mathrm{T}} + \mathbf{I}$, the sum of squares needed for testing becomes as

$$Q_* = \mathrm{tr}\{[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]\{[\mathbf{D}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-\mathbf{1}}\mathbf{D}]^{-\mathbf{1}}[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]^{\mathrm{T}}\}. \tag{12}$$

If σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{1}{n(q-c)}\mathrm{tr}\{\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{m}})\mathbf{Y}\}, \tag{13}$$

then testing can be based on

$$F = \frac{Q_*/\nu g}{\hat{\sigma}^2} \sim F[\nu g, n(q-c)]. \tag{14}$$

Note that Q_* does not contain unknown parameters of the covariance matrix and therefore for this special case the distribution of the $F-$statistic is exact. This is an important result since the uniform covariance model is quite common and a good approximation in many situations.

5 Computational example

The cattle data (Kenward, 1987), consist of 60 cattles that were randomly assigned to two treatment groups, A and B say, of size 30. The bodyweight (kg) of each cattle was measured 11 times (2 weeks interval) over a 133 day period. The 11^{th} measurement was taken after a week. We want to test if the mean progression in bodyweight of cattles in time is the same for both treatment groups.

To set up a Spline Growth Model, the between-individual design matrix $\mathbf{A}$ was defined as follows. For cattles on treatment A, the rows of $\mathbf{A}$ would be $(1, 0)$, $i = 1, \cdots, 30$ and for those on treatment B, the rows of $\mathbf{A}$ would be $(0, 1)$, $i = 1, \cdots, 30$. The minimum value of $GCV = 579.4783$ is obtained at $\alpha = 9870120$. The effective degrees of freedom $edf_* = 6.247323$. The fitted mean spline curves (**bold lines**) are shown in Figure 1. To test if the mean progression in time of bodyweights of cattles is the same for the two treatment groups, the dimension of $c = 3$ was used in the spline approximation.

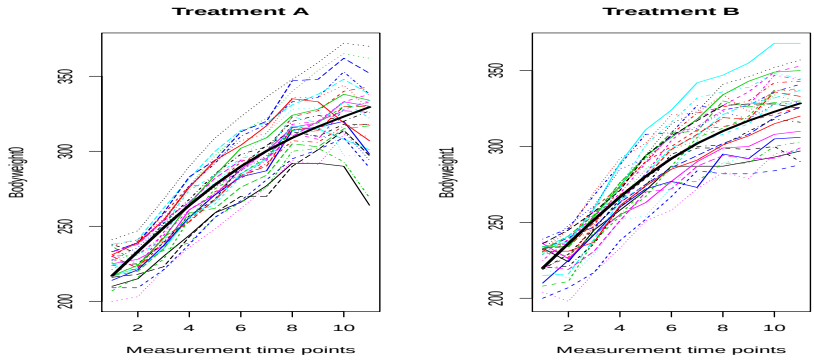


FIGURE 1. Bodyweight of cattles taken over 11 time points for two treatment groups, A and B.

To test the null hypothesis of equal progression, we took $\mathbf{C} = [\mathbf{0}, \mathbf{I_2}]$ and $\mathbf{D} = [1, -1]^{\mathrm{T}}$. Next, we estimate $\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}$. This yields $\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D} = (-4.090355,\ 1.654439)^{\mathrm{T}}$ and the residual variance estimate is $\hat{\sigma}^2 = 45.06687$. For the covariance matrix $\mathbf{R}$, we assumed the uniform correlation model and therefore the exact version of the test statistic can be used. The value of Q_* is given as

$$Q_* = \mathrm{tr}\{[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]\{[\mathbf{D}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-\mathbf{1}}\mathbf{D}]^{-\mathbf{1}}[\mathbf{C}\hat{\mathbf{\Omega}}\mathbf{D}]^{\mathrm{T}}\} = 292.0225$$

and the value of the test statistic is then

$$F = \frac{Q_*/\nu g}{\hat{\sigma}^2} = \frac{292.0225/2}{45.06687} = 3.239881$$

Comparing the obtained value of F to the critical value $F_{0.95}(2, 480) = 3.00$, the null hypothesis of equal progression in time of mean bodyweight of cattles on treatments A and B is clearly rejected. This means that treatments are significantly different in improving the growth of cattles by controlling the levels of their intestinal parasites that would retard growth/bodyweight gain.

6 Extension: Multivariate Spline Growth Curve Model

In this section, we briefly introduce how the spline growth model can be extended to a multivariate situation. The multivariate spline growth curve model can be written as

$$\mathbf{Y} = \mathbf{G}\mathbf{A}^{\mathrm{T}} \tag{15}$$

where

$$\mathbf{Y} = (\mathbf{y_1}, \ldots, \mathbf{y_n}), \quad \mathbf{y_i} = (\mathbf{y_{i1}^T}, \ldots, \mathbf{y_{is}^T})^T, i = 1, \ldots, n$$

and

$$\mathbf{G} = (\mathbf{g_1}, \ldots, \mathbf{g_m}), \quad \mathbf{g_j} = (\mathbf{g_{j1}^T}, \ldots, \mathbf{g_{js}^T})^T, j = 1, \ldots, m$$

Here, $\mathbf{y_i} = (\mathbf{y_{i1}^T}, \ldots, \mathbf{y_{is}^T})^T$ is a vector of measurements of s responses and $\mathbf{g_j} = (\mathbf{g_{j1}^T}, \ldots, \mathbf{g_{js}^T})^T$ is the corresponding vector of smooth mean curves. For the covariance matrix $\mathbf{R}$, we can take, for example

$$\mathbf{R} = (\mathbf{I_s} \otimes \mathbf{X_c})\mathbf{D}(\mathbf{I_s} \otimes \mathbf{X_c}) + \mathbf{I},$$

where $\mathbf{X_c}$ consists of c(=0,1,2) columns of $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$. If $\mathbf{X_c} = \mathbf{1}$ and $s = 2$, (bivariate uniform structure) for example,

$$\mathbf{R} = \left(\mathbf{I}_2 \otimes \mathbf{1}\right) \begin{pmatrix} d_1^2 & d_{12} \\ d_{21} & d_2^2 \end{pmatrix} \left(\mathbf{I}_2 \otimes \mathbf{1}^T\right) + \mathbf{I} = \begin{pmatrix} d_1^2\mathbf{1}\mathbf{1}^T + \mathbf{I} & d_{12}\mathbf{1}\mathbf{1}^T \\ d_{21}\mathbf{1}\mathbf{1}^T & d_2^2\mathbf{1}\mathbf{1}^T + \mathbf{I} \end{pmatrix}.$$

If we now define

$$\mathbf{K_s} = \mathbf{W} \otimes \mathbf{K},$$

where $\mathbf{W}$ is a diagonal matrix of smoothing parameters $\lambda_1, \ldots, \lambda_s$. Then we have $\mathbf{R}\mathbf{K_s} = \mathbf{K_s}$ and the unweighted estimator becomes as

$$\hat{\mathbf{G}} = (\mathbf{I} + \mathbf{W} \otimes \mathbf{K})^{-1}\mathbf{Y}\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}$$

For hypothesis testing, we may follow lines similar to those in section 4.

Conclusions: This kind of testing can be applied in a complex modeling situation where more traditional parametric growth curves are not applicable. The approach is also easily extended to a multivariate situation.

References

Diggle, P. J., Liang, K-Y., and Zeger, S. L. (1994). Analysis of Longitudinal Data. Oxford: Clarendon Press.

Green, P. J. and Silverman, B. W. (1994). Nonparametric Regression and Generalized Linear Models. London: Chapman & Hall.

Hastie, T. (1996). Pseudosplines. *Jounal of the Royal Statistical Society, Series B,* 58: 379 – 396.

Kenward, M. G. (1987). A method of comparing pofiles of repeated measurements, *Appl. Statist.,* 36: 296 – 308.

Nummi, T. and Mesue, N. K. (2013). Testing of Growth Curves Using Cubic Smoothing Splines. Advances in Growth Curve Models: Topics from the Indian Statistical Institute. Proceedings from Giridih conference, Springer.

Nummi, T., Jianxin, P., Siren, T., and Liu, K. (2011). Testing for Cubic Smoothing Splines under Dependent Data. *Biometrics,* 65(3): pp. 871 – 875.

Nummi, T. and Mottonen, J. (2000). On the analysis of multivariate growth curves. *Metrika,* 52: pp. 77 – 89.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika,* 5: 313 – 326.

Pan, J. and Fang, K. (2002). Growth curve models and statistical diagnostics. New York: Springer Series in Statistics.

Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using cubic smoothing splines (with discussions). *Journal of the Royal Statistical Scociety Series C,* 48: 269 – 311.

Wu, L. and Zhang, J. T. (2006). Nonparametric Regression Methods for Longitudinal Data Analysis. John Wiley & Sons, New Jersey.

Posterior approximations for Gaussian models with "non-detect" data

Daniel A. Molinari[1], Ludger Evers[1], Adrian W. Bowman[1]

[1] School of Mathematics and Statistics, University of Glasgow, UK

E-mail for correspondence: d.molinari.1@research.gla.ac.uk

Abstract: In many applications one has to deal with censored data, i.e. data for which it is only known that their value is below or above a certain threshold. We propose a posterior approximation similar to the Laplace approximation for partly censored data with Gaussian response. The approximation can be computed very efficiently and is typically close to the full Bayesian solution, which requires the use of sampling strategies such as MCMC. The proposed method is illustrated using an example from environmental modelling.

Keywords: Non-detects; Censored; Laplace; Bayesian.

1 Background

In many environmental applications data are gathered by monitors which cannot record measurements which are below (or above) a detection limit. For observations outside the detection range it is only known that they are below a lower detection limit or above an upper detection limit. A naïve approach, still used by many practitioners, is to replace the non-detected observations by a deterministic function of the detection limit (e.g. half the detection limit). This approach underestimates the uncertainty of the estimated mean parameters and can introduce a substantial bias. More formal approaches to this problem include the EM algorithm or methods derived from techniques used for time-to-event data. Some of these methods are implemented in the R package *NADA* (Helsel, 2012).

We propose a posterior approximation to the Bayesian solution similar to the Laplace approximation. Using a Bayesian framework has the advantage of being able to use Bayesian methods for the determination of hyperparameters such as the penalisation parameter in spline models, for which Bayesian methods perform better than their frequentist or heuristic counterparts (see e.g. Wood, 2011, or Molinari, 2012).

2 Model

We will consider a Bayesian linear model, i.e. $\mathbf{Y}|\boldsymbol{\alpha}, \sigma^2 \sim \mathcal{N}_n\left(\boldsymbol{B}\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I}_n\right)$ with prior distribution $\boldsymbol{\alpha}, \sigma^2 \sim \mathcal{NIG}_m\Big(\mathbf{0},\, (\lambda \boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1},\, a,\, b\Big)$ and a suitable prior on λ. A P-Spline regression model is a special case of the above with $\boldsymbol{B}$ being a matrix of basis functions and $\boldsymbol{D}$ a suitable difference matrix.
For simplicity of presentation, we will only consider the case of a lower detection limit. We assume that n_u observations have been observed directly; these will be denoted by $\boldsymbol{Y^u}$. For the remaining n_c observations the response $\boldsymbol{Y^c}$ could not be observed directly; it is only known that Y_i^c is less than $d_i\,(i = 1, \ldots, n_c)$. Splitting $\boldsymbol{B}$ in two sub-matrices $\boldsymbol{B^u}$ and $\boldsymbol{B^c}$ gives $\boldsymbol{Y^u}|\boldsymbol{\alpha}, \sigma^2 \sim \mathcal{N}_{n_u}\left(\boldsymbol{B^u}\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I_{n_u}}\right)$ and $\boldsymbol{Y^c}|\boldsymbol{\alpha}, \sigma^2 \sim \mathcal{N}_{n_c}\left(\boldsymbol{B^c}\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I_{n_c}}\right)$ with $\boldsymbol{B^u} \in \mathbb{R}^{n_u \times m}$, $\boldsymbol{B^c} \in \mathbb{R}^{n_c \times m}$ and $\boldsymbol{\alpha} \in \mathbb{R}^m$. This model is known as *Tobit regression model* in econometrics.

3 Approximation to the log-likelihood

As the only part of the log-posterior which is not a quadratic function of $\boldsymbol{\alpha}$ is the likelihood contribution from the censored observations, only this component has to be approximated using the Laplace approximation,

$$\begin{aligned} L(\boldsymbol{\alpha}) &\propto \prod_{i=1}^{n_u} \frac{1}{\sigma}\varphi\Big(\frac{y_i^u - \boldsymbol{B}_i^{u\mathrm{T}}\boldsymbol{\alpha}}{\sigma}\Big) \prod_{i=1}^{n_c} \Phi\Big(\frac{d_i - \boldsymbol{B}_i^{c\mathrm{T}}\boldsymbol{\alpha}}{\sigma}\Big) \\ &\approx L(\hat{\boldsymbol{\alpha}}) \exp\big\{ -\tfrac{1}{2}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^{\mathrm{T}} \boldsymbol{Q}\, (\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})\big\} \end{aligned}$$

with $\hat{\boldsymbol{\alpha}}$ corresponding to the mode of $L(\boldsymbol{\alpha})$ and $\boldsymbol{Q} = -\frac{\partial^2 \ell(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^2} = \frac{\boldsymbol{B}^{u\mathrm{T}}\boldsymbol{B}^u - \boldsymbol{B}^{c\mathrm{T}}\boldsymbol{W}^c\boldsymbol{B}^c}{\sigma^2}$ where $\ell(\boldsymbol{\alpha})$ represents the log-likelihood and $\boldsymbol{W}^c$ the diagonal matrix defined by $\boldsymbol{W}^c_{ii} = \frac{\varphi(t_i)^2 - \varphi(t_i)^{\mathrm{T}}\Phi(t_i)}{\Phi(t_i)^2}$ with $t_i = \frac{d_i - \boldsymbol{B}_i^{c\mathrm{T}}\boldsymbol{\alpha}}{\sigma}$. $\boldsymbol{Q}$ and $\boldsymbol{W}^c$ are evaluated at $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$, which are computed iteratively using a Newton-Raphson algorithm which also takes the priors into account. If the proportion of censored observations is small one can exploit the Woodbury formula to perform the necessary matrix inversions more efficiently.

4 Interpretation in terms of imputed values

It can be shown that the Newton-Raphson method for computing $\hat{\boldsymbol{\alpha}}$ can be rewritten as $\hat{\boldsymbol{\alpha}} = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{B} + \lambda \boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{W}\tilde{\boldsymbol{y}}$ with $\tilde{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y}^u \\ \tilde{\boldsymbol{y}}^c \end{pmatrix}$, $\tilde{\boldsymbol{y}}^c = \boldsymbol{B}^c\hat{\boldsymbol{\alpha}} - \hat{\sigma}\boldsymbol{W}^{c-1}\boldsymbol{v}^c$, $\boldsymbol{W} = \begin{pmatrix} \boldsymbol{I}_{n_u} & 0 \\ 0 & \boldsymbol{W}^c \end{pmatrix}$, and $\boldsymbol{v}_i^c = \frac{\varphi}{\Phi}\Big|_{t=t_i}$. Thus, the iterative method for finding the mode of $\boldsymbol{\alpha}$ can be viewed as a weighted regression model where the response is made up of the true observed responses and

imputed values. It follows, for left-censored data, that these imputed values are always smaller than the corresponding fitted ones. The weight function $w(t) = \frac{\varphi(t)^2 - \varphi(t)^{\mathrm{T}}\Phi(t)}{\Phi(t)^2}$ is evaluated at $t_i = \frac{d_i - \boldsymbol{B}_i^{c\mathrm{T}}\hat{\boldsymbol{\alpha}}}{\hat{\sigma}}$ i.e., as a function of the difference between the detection limit and the fitted value. The weight function is decreasing and thus, the larger the difference the smaller the weight given to the corresponding imputed value.
Using an EM algorithm corresponds to $\boldsymbol{W}_{EM}^c = \boldsymbol{I}_n$ and $\tilde{\boldsymbol{y}}_{EM}^c = \boldsymbol{B}^c\hat{\boldsymbol{\alpha}} - \hat{\sigma}\boldsymbol{v}^c$. Thus, the EM algorithm gives the values imputed for censored observations the same weight as for uncensored observations. On the other hand, the values imputed by the EM algorithm are closer to the fitted values (see also Figure 1). The reason for this difference is that the likelihood contribution coming from the censored observations can be highly skewed and thus the Laplace approximation has to move into the tails of the Gaussian to obtain a similar curvature. The EM algorithm typically needs many more iterations than the proposed Laplace approximation.

5 Approximation to the posterior distribution of $\boldsymbol{\sigma^2}$

It is clear from the derivations above that the posterior distribution of $\boldsymbol{\alpha}$ can be approximated by a Gaussian distribution. It is also clear that a Gaussian distribution will not be a suitable approximation for the variance parameter. The joint distribution of $\boldsymbol{\alpha}$ and σ^2 is thus approximated by a Normal Inverse Gamma distribution where the parameters are found by matching the first and second derivatives with respect to σ^2 of the exact log-posterior and the approximating NIG distribution. This approximation gives exact results if no censored data is present.
Laplace approximations for the Tobit model have already been studied by Chib (1992), who only uses a proper Laplace approximation (without the Inverse-Gamma part).

6 Example

We will consider a time series of concentrations of a groundwater contaminant recorded over 1379 days at a well at an industrial site. 75 observations have been recorded, 49 (65%) of which are below the detection threshold. A B-spline basis with 25 basis functions is used as design matrix.
The data, together with the predicted mean function and 95% prediction intervals, are shown in Figure 1 employing: Laplace approximation, full Bayesian solution using MCMC, EM algorithm and the result obtained when replacing the censored values with one-half the detection limit. For the Laplace approximation and for the model using half the detection limit the MAP (maximum a posteriori) estimate of the penalty parameter λ was used. For the full Bayesian solution λ was integrated out using MCMC. It

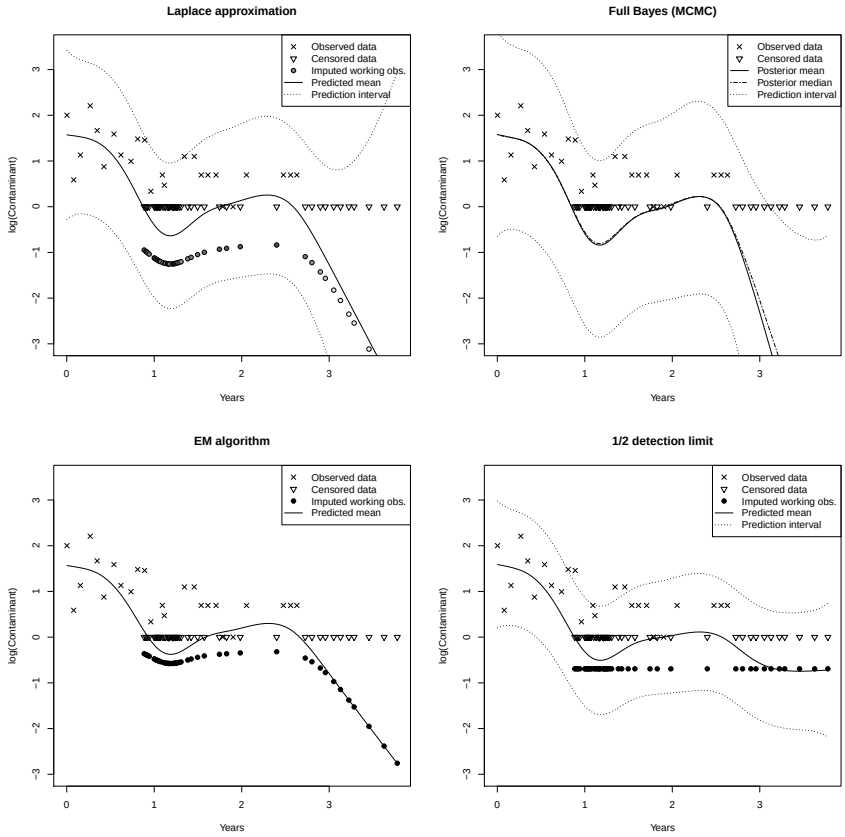


FIGURE 1. Predicted mean function and 95% prediction intervals for the contamination data obtained from the different methods. For the Laplace approximation the intensity of the imputed data indicates their weights.

can be seen that, except at the very end of the series, the Laplace approximation is close to the full Bayesian solution. Using half the detection limit underestimates the uncertainty and yields too narrow prediction bands. Figure 2 shows posterior densities of the prediction $\mathbf{B}(1.2)\boldsymbol{\alpha}$ at time $t = 1.2$ and the variance σ^2 for the different methods. It shows that the Laplace approximation underestimates the variance slightly. This figure also includes the results obtained when using a full Bayesian model with the smoothing parameter λ set to the MAP estimate obtained for the Laplace approximation. It suggests that some of the difference between the full Bayesian solution and the Laplace approximation is due to the two different ways of handling the penalty parameter λ. Again one can also see that using half the detection limit yields both biased and overconfident results.

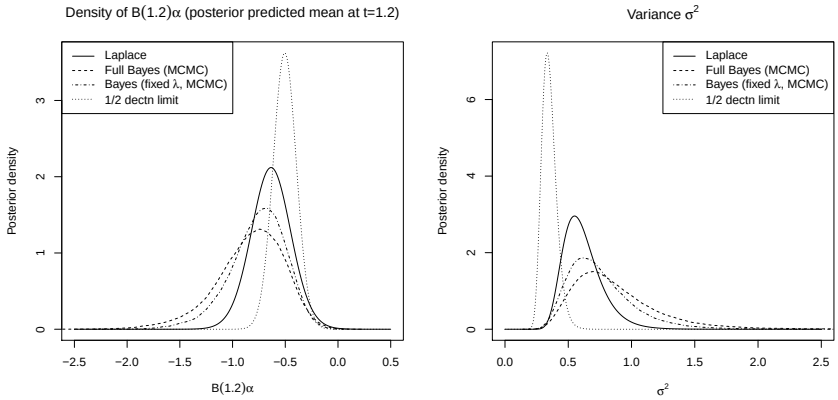


FIGURE 2. Comparison of the posterior densities of the prediction $\mathbf{B}(1.2)\boldsymbol{\alpha}$ at time $t = 1.2$ and the variance σ^2 for the different methods.

7 Conclusion and future work

The proposed Laplace-type approximation to the posterior distribution of models for censored Gaussian data, allows for an intuitive interpretation in terms of imputed observations. Though the approximation is quite accurate, future research will investigate alternative approximations such as the Expectation Propagation (EP) method (Minka, 2001).

References

Chib, S. (1992). Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, **51**, 79 – 99.

Helsel, D. (2012). *Statistics for Censored Environmental Data using Minitab and R.* Wiley.

Minka, T. (2001). Expectation Propagation for Approximate Bayesian Inference. *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 362 – 369.

Molinari, D. (2012). Smoothing parameter selection for spatiotemporal models with application to the analysis of contaminants in groundwater. *Proceedings of the 27th International Workshop on Statistical Modelling*, **2**, 637 – 642.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation on semiparametric generalized models. *Journal of Royal Statistical Society: Series B*, **73**, 3 – 36.

Characterisation and Mixed Effects Models for EEG Signals

Kathakali Ghosh Mukherjee[1], Claire Miller[1], Adrian W Bowman[1], Gregor Thut[2]

[1] School of Mathematics and Statistics, University of Glasgow, Glasgow, UK
[2] Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK

E-mail for correspondence: k.mukherjee.1@research.gla.ac.uk

Abstract: Artefact corrected and smoothed brain signals from TMS-EEG experiments may be characterised into functional forms of phase/frequency and amplitude to reduce the dimension of the data. In this paper, linear mixed effects models are fitted to repeated measurements from such functional frequency data in order to assess the changes in functional mean and standard deviation curves over various experimental conditions.

Keywords: Repeated Measures; Mixed Model; TMS; EEG; Entrainment.

1 Introduction

Data obtained from functional neuroimaging techniques such as electroencephalography (EEG) consist of high dimensional spatiotemporal signals which have particularly high resolution in time. It is therefore useful to characterise such datasets in terms of fewer parameters before model fitting is carried out. Further, in situations where the brain is stimulated by externally controlled Transcranial Magnetic Stimuli (TMS), statistical evidence of entrainment of the brain signal to a specific α band frequency (8-12 Hz) is of interest. In this paper, we propose characterising the EEG signals, which have been preprocessed using smoothing, in terms of parameters such as phase, instantaneous frequency and amplitude. The experimental TMS setup then leads to investigating evidence of entrainment in the signals in a repeated measures mixed model framework. The differences in the degree of variation in the frequency curves are studied, and a model for standard deviation of the frequency curves is the focus of this paper.

1.1 Data Preprocessing

The motivating dataset comes from a TMS-EEG experiment described by Thut et al. (2011) designed to assess the evidence of entrainment. A subset of the data that are used in this paper comprise EEG recordings from 6

subjects under 4 conditions, each having 54 trials. Data are recorded at S = 60 channels connected to an EEG cap and T = 5500 time points spanning a time of 1.1 seconds (-0.1 to 1.0 s with approximately 0-0.4 s being the duration of stimulation). In the main experimental condition, 5 TMS pulses are administered at regular intervals in an orientation perpendicular to the gyrus (a portion of the parietal lobe of the brain) with the exact inherent subject specific alpha frequency ('Main'). The EEG signal is recorded for several replicates in the subjects at (1) the main condition and three control conditions: (2) when the orientation of the TMS equipment has been rotated by 90° ('TMS-90'), (3) when the TMS pulses are applied asymmetrically ('Arrythmic') and (4) a sham condition where the pulses are only sound beeps ('Sound clicks').
A preprocessing step is applied to the data to remove artefacts and estimate the 'brain' signal of interest. For this, an additive model with a mean response μ and 4 components is fitted to model the artefacts appropriately and to simultaneously estimate the signals (Mukherjee, et al. (2012)). The additive model is given by:

$$y(s,t) = \mu + m_1(.,t') + m_2(.,t'') + m_3(.,t) + m_4(s,t) + \varepsilon \quad \forall\, t = 1, ..., T \quad (1)$$

where s is the spatial coordinate of the channel and t is the time of the recordings. Here μ is the overall mean, m_1 represents the temporally smooth trace of the TMS pulses, m_2 is a temporally smooth function defining the underlying cyclic 50 Hz. component, m_3 is a temporally smooth long term nonlinear trend component and m_4 is the spatiotemporally smooth signal of interest. Estimated signals m_4 from model (1) are treated as 'brain' signals of interest for the remainder of this paper.

2 Methods

Smooth 'brain' signal estimates at each channel may be described using characteristic parameters such as phase or frequency and amplitude. In order to compute the phase and amplitude curves from the signal, certain concepts from the complex logarithm model paradigm discussed in Eilers (2010) are used. Frequency is estimated as a function of time, given by $\phi'(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt}$, where ϕ is the phase function of the signal at each channel. The zero crossings of an oscillatory signal trace (when the magnitude of the signal is zero) determine the cumulative phase in multiples of 2π. The zero crossings are plotted against their corresponding times of occurrence to obtain a phase function. Instantaneous frequency is obtained by differentiating the smooth phase function.
Mean frequency curves for each channel and intervals based on standard deviations of the frequency curves at each channel are produced to assess the behaviour of the signals in terms of their average frequency. Estimates of mean and standard deviation are computed for discrete time epochs from

these curves. A natural epoch to consider would be the time interval between two successive TMS pulses. This provides four repeated observations for any characteristic or summary measure between pulses 1 and 2, 2 and 3, 3 and 4 and 4 and 5.

2.1 Linear Mixed Effects Models with Repeated Measures

The nature of the experiment produces multiple replicates within each subject as a random effect and subject itself as a random effect. When considering the mean signal for frequency, both random effects are of interest. When considering the standard deviation of the frequency and the corresponding interval, only the random effect of subjects is relevant. The latter will be the focus of this paper denoted by model 2. The repeated measures mixed effects model is fitted as:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon \tag{2}$$

where $\mathbf{Y}$ is the $n \times 1$ vector of n response observations, $\mathbf{X}$ is the $n \times p$ model matrix for p fixed effects, β is a $p \times 1$ vector of fixed effect coefficients, $\mathbf{Z}$ denotes the $n \times q$ model matrix of q random effects and γ is the $q \times 1$ vector of coefficients of random effects where $\gamma \sim \mathbf{N_q}(\mathbf{0}, \mathbf{\Psi})$. $\mathbf{\Psi}$ denotes the $q \times q$ covariance matrix for the random effects. $\epsilon \sim \mathbf{N_n}(\mathbf{0}, \sigma^\mathbf{2}\mathbf{\Lambda})$ is the error term where $\mathbf{\Lambda}$ is the $n \times n$ covariance matrix for errors., With the aim of assessing statistical evidence of entrainment of the brain signal to the α band frequency (8-12 Hz), repeated measures ANOVA is applied on the width of the interval 2 standard deviations on either side of the mean frequency for each of the 6 subjects across all conditions. The two fixed effects for this model are (i) the time indices at the repeated measurements of the average width of the interval over all replicates within a subject, and (ii) the experimental conditions; γ denotes the random effect of subjects.

3 Model Fitting

Signal estimates of interest are simultaneously obtained for each replicate in 6 subjects for all 4 conditions using the additive model (1). Fig. 1 displays the signal at channel CP4 (the channel of principal interest) from one replicate under the main experimental condition in subject 2, and its corresponding frequency curve, as an example. The functional curves of all replicates at channel CP4 across the 4 conditions are summarized as functions for mean $\pm$ 2 functional standard deviations (Fig. 2). The functional mean appears to always lie in the α band frequency. However, the functional variability appears to change depending on condition. In some subjects, it can be informally ascertained that the variability around the mean frequency tends to decrease over time during the stimulus period

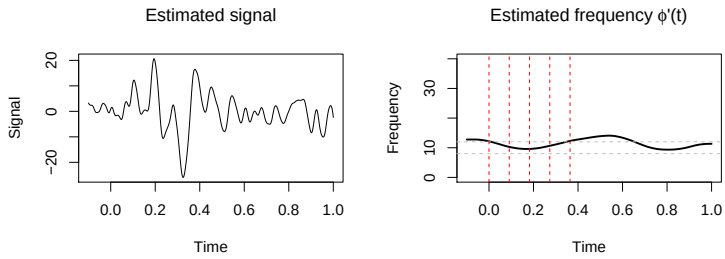


FIGURE 1. Estimated signal (left) and corresponding frequency function (right) for 1 replicate of **subject 02** at channel CP4 in the main condition. Vertical lines show TMS pulse locations and horizontal lines show α band.

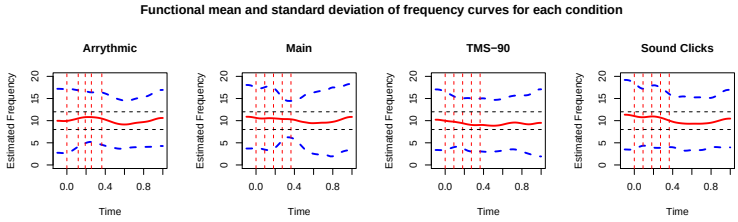


FIGURE 2. Estimated mean frequency curve (solid) and intervals based on 2 functional standard deviations (dashed) for **subject 02** at all 4 conditions at channel CP4. Vertical lines show TMS pulse locations and horizontal lines show α band.

(when the TMS pulses are administered), specifically in the main condition as illustrated by subject 2 in Fig. 2. In order to formalise this finding, a model of the repeated measures of the standard deviations of the estimated frequency over all subjects and conditions is fitted with subject as a random effect. The model with two main effects, namely time indices at the repeated measures of the standard deviation over replicates in each subject and a condition effect, is fitted. An interaction term of time indices at repeated measures $\times$ conditions is also included in the model. The interaction term is not significant (marginal ANOVA F statistic $= 1.634$, p-value $= 0.104$) although this could be due to relatively small sample size. Since it is close to 10% level of significance, this term has been retained in the model to allow the standard deviations to show different patterns across conditions. The estimates of the standard deviations of the average frequency across conditions is given in Fig. 3.

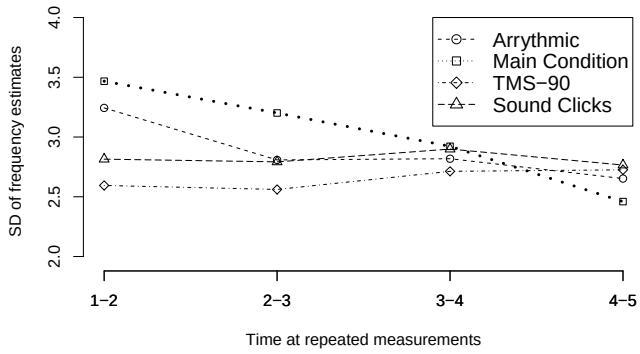


FIGURE 3. Estimates from mixed effects model for standard deviation of frequency between successive TMS pulses over replicates for **all subjects** (with subject as random effect) for each condition, and interaction of the repeated measures with condition.

4 Discussion

It can be seen from Fig. 2 that the standard deviation of the estimated frequencies tend to become smaller over time during the stimulus period, especially between the third and fifth TMS pulse, for the main condition in subject 2. The model indicates that the standard deviation around a mean frequency of 10 Hz (in the α band) decreases over time for the main condition over all subjects (Fig. 3). Similar repeated measures models may be applied to amplitude and phase functions (results not shown here) to characterise and describe the 'brain' signals in terms of fewer functional parameters.

References

Thut, G., Veniero, D., Romei, V., Miniussi, C., Schyns, P.G. and Gross, J. (2011). Rhythmic TMS Causes Local Entrainment of Natural Oscillatory Signatures. *Current Biology*, **21**:14, 1176 – 1185.

Mukherjee, K.G., Miller, C., Bowman, A.W., Thut, G. (2012). A Flexible Regression Framework for TMS-EEG Signals. *Proceedings of the 27th International Workshop on Statistical Modelling*, **Vol.2**, 127 – 132.

Eilers, P (2010). The smooth complex logarithm and quasi-periodic models. *Statistical Methods and Regression Structures*, Springer-Verlag.

A comparison between landmarking and joint modeling for producing predictions using longitudinal outcomes

Magdalena Murawska[1,*], Dimitris Rizopoulos[1], Emmanuel Lesaffre[1,2]

[1] Department of Biostatistics, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands,*e-mail: m.murawska@erasmusmc.nl
[2] I-Biostat, Catholic University of Leuven, Belgium

E-mail for correspondence: m.murawska@erasmusmc.nl

Abstract: In medical studies often longitudinal measurements reflecting the health of the patient are collected during the follow-up time. In this setting it is often of primary interest to assess whether the available history of the patient can be used for predicting patient survival. In this work we compare two approaches to dynamic predictions of survival probability namely, landmarking and joint modeling of the longitudinal and survival processes.

Keywords: Dynamic predictions; Joint models; Landmarking.

1 Introduction

Motivated by current trends towards personalized medicine, there is great interest nowadays to develop prognostic models. Examples are numerous and come from a wide spectrum of diseases, including prognostic models applied in cancer research, risk scores for cardiovascular diseases, such as the Framingham score, and prognostic models for HIV infected patients. The common characteristic for all these diseases is that the rate of progression is not only different from patient to patient but also dynamically changes in time for the same patient. Hence, it is medically relevant to investigate whether repeated measurements of specific biomarkers can ultimately provide a better understanding of disease progression.
In this work we compare two approaches for producing dynamic predictions of survival probabilities using the recorded longitudinal information, namely landmarking (van Houwelingen and Putter, 2011) and joint modeling (Rizopoulos, 2012). Landmarking requires adjusting the risk set at the landmatk point and fitting a Cox model, whereas joint modeling explicitly takes into account all the longitudinal history of a subject to produce predictions. We show that because the subject-specific longitudinal trajectories can be quite complex (e.g. nonlinear, plateaus) different features of

these trajectories may be more predictive for the event of interest. To this end we study how the aforementioned approaches perform under different functional relationship between the longitudinal and event time outcomes. Our proposals are exemplified in the primary biliary cirrhosis (PBC) study conducted by the Mayo Clinic between 1974 and 1984. For patients with PBC serum bilirubin is known to be a good marker of progression. In our analysis aim to find which characteristics of the serum bilirubin profile are most predictive for death.

2 Methodology

2.1 Dynamic Prediction of Survival

Let $Y_i(u)$ denote the longitudinal measurement for individual i ($i = 1, \ldots, n$) at time u, and let u_{ij} ($j = 1, \ldots, m_i$) denote points at which measurements are taken for subject i. In addition let T_i^* denote the true failure times for individual i. Since the failure times are right censored we observe only $T_i = \min(T_i^*, C_i)$, where C_i is the censoring time with the binary failure indicator Δ_i which equals 0 if subject was censored and 1 otherwise. We construct dynamic predictions by calculation of the predicted survival functions S_k, for a new subject k for whom we have a set of longitudinal measurements $Y_k(t) = \{Y_k(s); 0 \leq s \leq t\}$ available. We are therefore interested in the conditional probability of surviving time $u > t$ given that the subject has survived up to t:

$$S_k(u \mid t) = \Pr(T_k^* > u \mid T_k^* > t, Y_k(t)). \tag{1}$$

2.2 Landmark Approach

The landmark method simplifies the longitudinal history $Y_k(t)$ in (1) to the last value $y_k(t)$. Dynamic predictions are obtained by suitably adjusting the risk set and refitting the Cox model. In particular, for a given time t_L a corresponding landmark data set $\mathcal{L}_L$ is constructed, by selecting individuals that are at risk at time t_L. Then for this set of individuals a simple Cox model is fitted with the current value of the longitudinal marker $y_i(t_L)$ at time t_L:

$$\lambda_i(t) = \lambda_0(t) \exp\{\gamma^{\mathrm{T}} v_i + \alpha y(t_L)\}, \qquad i \in \mathcal{L}_L,$$

where v_i is the vector of baseline covariates, and $\lambda_0(t)$ is the baseline hazard that can be modeled parametrically or left unspecified. From the fitted model (2) the survival probabilities $\Pr(T^* > u \mid T^* > t_L, y(t_L))$ can be computed.

2.3 Joint Model Approach

In the joint modeling approach we postulate two submodels for the longitudinal and survival processes. For continuous longitudinal markers usually a linear mixed model is postulated:

$$y_i(t) = m_i(t) + \epsilon_i(t) = x_i^{\mathrm{T}}(t)\beta + z_i^{\mathrm{T}}(t) + \epsilon_i(t), \tag{2}$$

where $m_i(t)$ denotes the true value of the longitudinal marker at time t, β denotes the vector of the fixed-effects parameters, $b_i \sim N(0, D)$ is the vector of random effects, $x_i(t)$ and $z_i(t)$ denote the design matrices for the fixed and random effects, respectively, and $\epsilon_i(t)$ is the measurement error term, $\epsilon_i(t) \sim N(0, \sigma^2)$. To model a nonlinear behaviour of the longitudinal marker one may include splines in the design matrices $X_i(t)$ and $Z_i(t)$ or a nonlinear mixed model might be posited instead of (2).
For the survival process a standard relative risk model is usually assumed that shares some common, possible time-dependent term $f(t, b_i, \alpha)$, with the longitudinal mixed effects model:

$$\lambda_i(t) = \lambda_0(t) \exp(\alpha^{\mathrm{T}} f(t, b_i, \alpha) + \gamma^{\mathrm{T}} v_i), \tag{3}$$

where v_i denotes again a vector of baseline covariates as in (2) and γ is a vector of associated coefficients. Parameters α measure the strength of the association between longitudinal and survival processes. The joint model consisting (2) and (3) can be estimated using maximum likelihood or Bayesian methods. Here we opt for the latter. Based on the fitted model the dynamic predictions can be constructed. Let θ denote the vector of parameters from the joint model and $\mathcal{S}_n$ - a sample of size n on which the joint model was fitted. Then $S_k(u \mid t)$ from (1) can be estimated as a Bayesian posterior expectation:

$$S_k(u \mid t) = \int \Pr\left(T_k^* > u \mid T_k^* > t, Y_k(t), \mathcal{S}_n; \theta\right) p(\theta \mid \mathcal{S}_n) d\theta. \tag{4}$$

The first part of the integrant in (4) can be written as:

$$\begin{aligned} &\Pr\left(T_k^* > u \mid T_k^* > t, Y_k(t), \mathcal{S}_n; \theta\right) \\ &= \int \Pr\left(T_k < u \mid T_k^* > t, b_k; \theta\right) \times p\left(b_k \mid T_k^* > t, Y_k(t), \theta\right) db_k. \end{aligned} \tag{5}$$

The survival function $S_k(u \mid t)$ has a dynamic nature because when new information is recorded for a patient k at time $t' > t$, we can update these predictions and obtain $S_k(u \mid t')$ for $u > t'$. Therefore when combining (4) and (5) a Monte Carlo approach can be used to compute $S_k(u \mid t)$ for each patient and $S_k(u \mid t')$ can be updated for every time point $t' > t$. Following Rizopoulos (2012), we used a similar sampling scheme to obtain the dynamic subject-specific predictions based on the fitted joint model. The advantage in the joint modeling approach to dynamic predictions is the possibility of defining different association structure between the longitudinal and survival processes. Some examples of alternative functional forms for the survival submodel (3) are:

$$\begin{aligned}
\text{I} \quad \lambda_i(t) &= \lambda_0(t)\exp\{\gamma^{\mathrm{T}} v_i + \alpha_1 m_i(t)\} \\
\text{II} \quad \lambda_i(t) &= \lambda_0(t)\exp\{\gamma^{\mathrm{T}} v_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\} \\
\text{III} \quad \lambda_i(t) &= \lambda_0(t)\exp\left\{\gamma^T v_i + \alpha_1 \int_0^t m_i(s)ds\right\} \\
\text{IV} \quad \lambda_i(t) &= \lambda_0(t)\exp\{\gamma^{\mathrm{T}} v_i + \alpha^{\mathrm{T}} b_i\}.
\end{aligned} \tag{6}$$

In parametrization I we assume that hazard of death at time t is associated with the true current value of the marker $m_i(t)$ at time t and α measures the strength of this association. In parametrization II we extend model I by including also the slope of the trajectory at time t, $m_i'(t)$. In parametrization III we postulate that hazard for death at time t is associated with area under the trajectory up to t. Finally in parametrization IV we only assume that the two submodels for the longitudinal and survival process are linked through the vector of the shared random terms b_i.

3 Application

We have applied the described methodology to the PBC data set. To model longitudinal serum bilirubin level $Y_i(u)$ we postulate a mixed effects model with natural cubic splines to account for nonlinear character of the marker evolution. We included interaction terms between B-spline basis and the treatment group to model different trajectories for the two treatment groups. For the survival process we used a standard relative risk model for which we considered different forms of the association structure from (6). In every survival submodel we included treatment as a baseline covariate. The baseline hazard $\lambda_0(t)$ was modeled parametrically using the Weibull distribution, i.e: $\lambda_0(t) = \phi t^{\phi-1}$. We have compared the results with the landmark approach adjusting for treatment effect also in the landmark Cox model (2). Substantial differences between the prediction from the joint models with different parametrization and the landmark approach were observed. Different joint models were compared using DIC criterion.

4 Simulation Results

Additionally we have performed a series of simulations to evaluate both methods of constructing dynamic predictions. We simulated data using the joint model corresponding to the model fitted for the PBC data. In particluar, for the longitudinal response we used mixed effects model with natural cubic splines described in previous section and for the survival submodel we considered all scenarios I-IV from (6). Baseline hazard was simulated from the Weibull distribution. All the parameters of the longitudinal and survival part were taken from the models fitted for the PBC data. In each

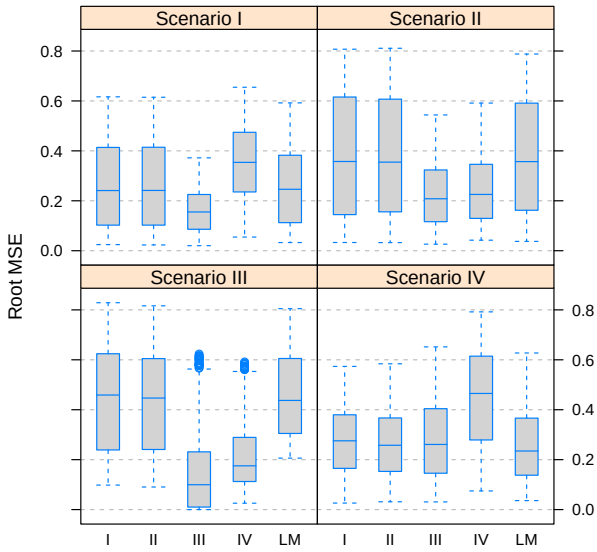


FIGURE 1. *RMSE of the distance between the prediction from the gold standard model and the four fitted joint models (I,II,III,IV) together with the landmark model (LM) for the different simulated scenarios.*

scenario we excluded randomly ten patients. For the remaining patients we fitted the joint models with the same longitudinal submodel parametrization and different survival submodel parametrization I-IV. We constructed the dynamic predictions of survival probability for the excluded patients based on the fitted four joint models as well as the landmark approach. The predictions from both approaches were compared with the predictions obtained from the gold standard model, namely the model with the true parametrization and true values of the fixed and the random effects. The obtained results are illustrated on Figure 1. As can be observed for some simulated scenarios the predictions from joint modeling and landmarking are substantially different. Also within the joint modeling approach the results are influenced by the choice of the model parametrization.

References

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-To-Event Data: With Applications in R.* Boca Raton : CRC Press.

van Houwelingen, H.C., Putter, H. (2011). *Dynamic Prediction in Clinical Survival Analysis.* Boca Raton : CRC Press.

Model averaging quantiles for censored data

Ruth Nysen[1], Marc Aerts[1], Christel Faes[1]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Belgium

E-mail for correspondence: ruth.nysen@uhasselt.be

Abstract: Quantiles are of interest in food safety data dealing with a limit of detection. The limit of detection introduces a lot of uncertainty in the left tail of the underlying distribution, making quantile estimation for this part of the distribution difficult. Therefore we fit a model to the data and derive the model-based estimate for the quantile. Since the true distribution is unknown, model averaging is used to combine information from a set of models. In this paper we discuss two approaches to use model averaging for quantiles. The methods are applied to a data example and compared in a simulation study. The effect of an increasing percentage of censoring on the estimates is explored.

Keywords: Censoring; Model averaging; Quantiles.

1 Introduction

We are interested in the lower quantiles of the distribution: e.g. which concentration is the cut-off for the 10% lowest concentrations? The most common estimate is the nonparametric or empirical quantile. However, censoring occurs in particular in the left tail of the distribution and introduces a lot of uncertainty about the quantiles. Therefore we fit a model to the data and derive the model-based estimate for the quantile. First we fit several parametric models to the data. All models are related to the log-normal distribution, a distribution that is regularly used in food safety data. Next we consider a semi-nonparametric family of distributions that consists of extensions of the log-normal distribution. Gallant and Nychka (1987) and Fenton and Gallant (1996) studied this family of distributions. Nysen, Aerts and Faes (2012) based a goodness-of-fit test for censored data on this semi-nonparametric family of distributions.
We can select the best model from the set of parametric ($\mathcal{M}_P$) and semi-nonparametric models ($\mathcal{M}_S$), based on a model selection criterion, e.g. Akaike's information criterion (AIC). This model is considered as the best approximating model and the cumulative distribution is used to estimate the quantile. However, if we would have a second and similar data set, it is possible that a different model is selected by the model selection criterion. This might result in a quite different estimate of the quantile. By

selecting one single model, we ignore the model uncertainty. The idea of model averaging is to start from a large set of plausible models and combine information from all models. In Section 2 we discuss two approaches to apply model averaging for quantiles. The approaches are applied to a data example in Section 3 and compared in a simulation study in Section 4. Although we focus on left censored data in this paper, the estimation can deal with other types of censoring, like right and interval censoring.

2 Model averaging of quantiles: two approaches

Let $M_i, i = 1 \ldots, K$ be a rich set of candidate models, with F_i the corresponding cumulative distribution function. The natural parameters θ_i are estimated by maximum likelihood theory and their variance-covariance matrix is denoted by $\text{Var}(\hat{\theta}_i)$. Suppose we want to estimate the p-quantile of the data ξ_p, where p is a fixed number between 0 and 1. There are several approaches in model averaging to obtain an estimate for the quantiles.
In the first approach, the quantile is estimated for each candidate model and the model averaged estimate is a weighted average of the K estimates. Based on model M_i, the quantile is obtained by $\xi_{p,i} = F_i^{-1}(p; \theta_i)$. The variance of the quantile can be approximated by the delta method:

$$\text{Var}(\hat{\xi}_{p,i}) \approx \nabla F_i^{-1}(p; \theta_i)^T \text{Var}(\hat{\theta}_i) \nabla F_i^{-1}(p; \theta_i). \tag{1}$$

In (1) the gradient $\nabla F_i^{-1}(p; \theta_i)$ is with respect to θ_i and can be estimated by $\nabla F_i^{-1}(p; \hat{\theta}_i)$.
Burnham and Anderson (1998) calculate the weight for each model, based on the difference of its AIC with the smallest AIC of all candidate models:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{j=1}^{K} \exp\left(-\frac{1}{2}\Delta_j\right)},$$

where $\Delta_i = \text{AIC}_i - \text{AIC}_{\min}$.
The model averaged value of the quantile $\xi_{p,MA1}$ is given by the weighted average of the estimates of all candidate models:

$$\hat{\xi}_{p,MA1} = \sum_{i=1}^{K} w_i \hat{\xi}_{p,i} \tag{2}$$

with estimated variance

$$\widehat{\text{Var}}(\hat{\xi}_{p,MA1}) = \left[\sum_{i=1}^{K} w_i \sqrt{\widehat{\text{Var}}(\hat{\xi}_i) + \left(\hat{\xi}_i - \hat{\xi}_{p,MA1}\right)^2} \right]^2.$$

The variance estimator is the sum of two components. The first component is the conditional variance, given model M_i. The second component reflects the variation in the estimates across the K models.

A second approach applies model averaging at a different level. The distribution of each candidate model is estimated and combined in a model averaged cumulative distribution function. The quantile of the combined distribution is the model averaged quantile estimate. Let x be a real number in the domain of the candidate distribution function. The averaged distribution function is given by

$$\hat{F}_{MA}(x;\theta) = \sum_{i=1}^{K} w_i \hat{F}_i(x;\theta_i)$$

with θ the vector of the natural parameters of all candidate models. The estimated variance of $\hat{F}_{MA}(x;\theta)$ is

$$\widehat{\mathrm{Var}}(\hat{F}_{MA}(x)) = \left[\sum_{i=1}^{K} w_i \sqrt{\widehat{\mathrm{Var}}(\hat{F}_i(x;\theta_i)) + \left(\hat{F}_i(x;\theta_i) - \hat{F}_{MA}(x;\theta)\right)^2}\right]^2. \tag{3}$$

The quantile can be estimated by

$$\hat{\xi}_{p,MA2} = F^{-1}_{MA}(p;\theta). \tag{4}$$

Based on the implicit function theorem, the variance of $\hat{\xi}_{p,MA2}$ can be approximated by

$$\mathrm{Var}(\hat{\xi}_{p,MA2}) = \frac{\mathrm{Var}(\hat{F}_{MA}(\hat{\xi}_{p,MA2}))}{\hat{f}^2_{MA}(\xi)},$$

where $\mathrm{Var}(\hat{F}_{MA}(\hat{\xi}_{p,MA2}))$ can be estimated by (3). Because the true value ξ is unknown, we need to estimate $\hat{f}_{MA}(\xi)$ by $\hat{f}_{MA}(\hat{\xi}_{p,MA2})$.
In general, the estimates (2) and (4) result in different estimates, because quantile estimation is not a linear functional. When estimating for instance the cumulative distribution function, the two approaches would be equivalent. We compare the performance of the two approaches in a simulation study, but first we illustrate the approaches in a real data analysis. We will consider a set of parametric and semi-nonparametric models $\mathcal{M}_P \cup \mathcal{M}_S$.

3 Data analysis

The motivating data for this study, is a sample of measurements of cadmium level. The data set consists of almost 100 observations, but 37% of the measurements are censored by the limit of detection (LOD). The LODs are small and lie in between 0.001 and 0.01. A visual representation of the data is given in Figure 1, where a kernel density of the logarithm of the concentrations is shown.

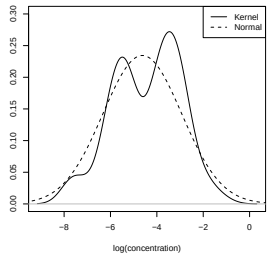


FIGURE 1. Cadmium data. Kernel density function of the concentrations where LODs are replaced by LOD/2 (solid line), transformed to the log scale. Normal fit (dashed line) based on likelihood for censored data.

Figure 1 shows that the (log-)normal distribution is a reasonable fit, but other distribution might fit better. Table 1 lists the AIC for several parametric models and for the semiparametric extensions of the log-normal distribution. The best global fit is given by the log-skew-t distribution and the distribution with two extra parameters from the semi-nonparametric family of distributions. For each model, the 10% and 25% quantiles are given in Table 1. For comparison, we also computed the nonparametric quantile estimates, where we substituted the LOD in the data with LOD/2 to cope with the left censoring. We see that the parametric models have smaller estimates for the quantiles. This is due to the left censoring, because the parametric models put weight on the left of the distribution, while the raw data only provide information on the LOD.
The quantile estimates based on the two model averaging approaches from Section 2 are close together. If we average the distribution function before computing the quantile (approach 2), the 10% quantile is slightly smaller. The estimated standard error for the estimate of the second approach is also smaller.

4 Simulation study

The performance of both methods is demonstrated in a simulation study. Data are simulated from 3 kinds of distributions, i.e. the log-normal distribution, the gamma distribution with same mean and variance as the log-normal distribution, and a mixture of 2 log-normal distributions. For each distribution, 500 samples are drawn. The censoring scheme is defined by limits of detection that correspond to five quantiles (1%, 5%, 10%, 20%, 25%) of the original log-normal distribution, resulting in 12% censoring on average when sampling from the log-normal distribution.
On each sample we fit 13 models: seven parametric (log-normal, log-skew-normal, log-t, log-skew-t, gamma, Weibull), denoted by $\mathcal{M}_P$, and seven

TABLE 1. Cadmium data. AIC and model averaging weights. For each parametric and semi-nonparametric model, the quantile was estimated (standard error).

	AIC	weight	$\hat{\xi}_{0.1}$	$\hat{\xi}_{0.25}$
Non-parametric			0.00250	0.00500
Log-normal	-135.7	0.000	0.00110 (0.00036)	0.00310 (0.00076)
Log-skew-n	-149.3	0.023	0.00045 (0.00026)	0.00226 (0.00086)
Log-t	-133.4	0.000	0.00111 (0.00036)	0.00314 (0.00078)
Log-skew-t	-154.2	0.262	0.00000 (0.00000)	0.00109 (0.00122)
Weibull	-149.1	0.021	0.00061 (0.00019)	0.00291 (0.00058)
Gamma	-151.8	0.081	0.00029 (0.00021)	0.00222 (0.00093)
GenGam	-150.0	0.033	0.00023 (0.00021)	0.00208 (0.00097)
SemiNP1	-151.5	0.068	0.00060 (0.00013)	0.00128 (0.00028)
SemiNP2	-154.3	0.275	0.00019 (0.00006)	0.00137 (0.00133)
SemiNP3	-152.6	0.120	0.00022 (0.00054)	0.00149 (0.00118)
SemiNP4	-150.7	0.047	0.00023 (0.00111)	0.00156 (0.00109)
SemiNP5	-148.8	0.018	0.00021 (0.00372)	0.00160 (0.00105)
SemiNP6	-150.3	0.038	0.00025 (0.00293)	0.00152 (0.00122)
SemiNP7	-148.3	0.014	0.00015 (0.00581)	0.00154 (0.00119)
$\mathcal{M}_P \cup \mathcal{M}_S$ 1 *			0.00020 (0.00052)	0.00147 (0.00120)
$\mathcal{M}_P \cup \mathcal{M}_S$ 2 **			0.00015 (0.00050)	0.00148 (0.00107)

* quantiles are estimated for each distribution and then averaged (MA1)
** cumulative distribution function is averaged (MA2)

extensions of the log-normal defined by the semi-nonparametric family of distributions, denoted by $\mathcal{M}_S$. The generalized gamma distribution was not included in the simulation study due to convergence issues. We provide here the preliminary results for data simulated from the log-normal distribution, estimating the 1% quantile. Table 2 shows the variance and the squared bias, together with the sign of the bias. The mean squared error (mse) is obtained by adding these two quantities.

A first observation is that the mean squared error is larger when the data are censored, which is to be expected because censoring introduces more uncertainty in the data and the estimation process. From the family of parametric models, the log-t and the true log-normal distribution provide the best fits. In the semi-nonparametric family of distributions, the models perform worse, because more parameters are added. Indeed, since we simulate from the log-normal distribution, the extra parameters are redundant to describe the data. If the data are not censored, there is no difference between the two modeling approaches, regarding the mean squared error. On the contrary, the bias is smaller for the second approach, while the variance is smaller for the first approach. In the censoring case, the first approach does result in a smaller mean squared error. The variance is higher in the

TABLE 2. Simulation study. Data simulated from log-normal distribution with sample size 100. ($\times 10^{-5}$)

Censoring	No			Yes		
	$\text{bias}^{2\ (\text{sign})}$	var	mse	$\text{bias}^{2\ (\text{sign})}$	var	mse
Log-normal	0.258 $^{(+)}$	3.700	3.957	0.283 $^{(+)}$	4.381	4.664
Log-skew-n	0.338 $^{(+)}$	7.184	7.522	2.266 $^{(+)}$	10.909	13.175
Log-t	0.000 $^{(+)}$	3.467	3.467	0.051 $^{(-)}$	4.482	4.533
Log-skew-t	0.006 $^{(+)}$	7.179	7.185	0.357 $^{(-)}$	15.422	15.780
Weibull	40.546 $^{(-)}$	0.252	40.799	43.228 $^{(-)}$	0.211	43.439
Gamma	39.419 $^{(-)}$	2.136	41.555	45.539 $^{(-)}$	1.064	46.603
SemiNP1	0.494 $^{(+)}$	6.076	6.569	0.020 $^{(-)}$	12.088	12.108
SemiNP2	1.208 $^{(+)}$	6.999	8.207	0.100 $^{(+)}$	16.869	16.969
SemiNP3	1.467 $^{(+)}$	8.543	10.010	0.601 $^{(+)}$	19.938	20.540
SemiNP4	1.641 $^{(+)}$	8.811	10.452	0.944 $^{(+)}$	22.621	23.565
SemiNP5	1.673 $^{(+)}$	9.383	11.056	1.394 $^{(+)}$	23.612	25.006
SemiNP6	1.880 $^{(+)}$	9.594	11.474	2.222 $^{(+)}$	24.477	26.699
SemiNP7	2.053 $^{(+)}$	10.417	12.471	3.280 $^{(+)}$	25.546	28.826
$\mathcal{M}_P \cup \mathcal{M}_S$ 1 *	0.351 $^{(+)}$	6.018	6.369	0.056 $^{(+)}$	10.916	10.973
$\mathcal{M}_P \cup \mathcal{M}_S$ 2 **	0.248 $^{(+)}$	6.121	6.369	0.199 $^{(-)}$	13.016	13.215

* quantiles are estimated for each distribution and then averaged (MA1)
** cumulative distribution function is averaged (MA2)

model averaging compared to the single-model inference, because the model selection uncertainty is now incorporated.
In future research we will compare the model averaging approaches for the other distributions, sample sizes and different quantiles. We will also consider a different family of distributions, e.g. restricted to the parametric models $\mathcal{M}_P$ or to the semi-nonparametric family of distributions $\mathcal{M}_S$.

References

Burnham, K.P. and Anderson, R.A. (1998). *Model selection and inference: A practical information-theoretic approach.* New York: Springer-Verlag.

Fenton, V.M. and Gallant, A.R. (1996). Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics*, **74**, 77 – 118.

Gallant, A.R. and Nychka, D.W. (1987) Semi-nonparametric maximum likelihood estimation. *Econometrica*, **55**(2), 363 – 390.

Nysen, R., Aerts, M. and Faes, C. (2012), Testing goodness of fit of parametric models for censored data. *Statistics in Medicine*, **31**, 2374-2385.

CircSiZer for exploring circular data

María Oliveira, Rosa María Crujeiras, Alberto Rodríguez–Casal[1]

[1] Department of Statistics and Operations Research, University of Santiago de Compostela (Spain)

E-mail for correspondence: maria.oliveira@usc.es

Abstract: SiZer (SIgnificant ZERo crossing of the derivatives) is a visualization method for exploring significant underlying structures in data samples based on nonparametric smoothers, originally designed for kernel estimators. An extension of SiZer to circular data, namely CircSiZer, is introduced for the regression setting, when dealing with a circular covariate and a scalar response. CircSiZer presents a graphical device to assess which observed features are statistically significant. With the same purpose, the CircSiZer based on smoothing splines, is also presented. The proposed tool is used for analyzing the influence of the wind direction over wind speed in the atlantic coast of Galicia (NW Spain). This analysis is particularly interesting given the heavy marine traffic in this area.

Keywords: circular data; CircSiZer; data smoothers; wind pattern.

1 Introduction

Coastal and marine ecosystems suffer from a variety of threats due to human and industrial activity, being these ecosystems specially vulnerable to oil spills and toxic dumping. Specifically, the atlantic coast of Galicia (NW Spain) has suffered two major ship accidents which caused serious environmental and ecological damages: the burning of a cargo ship named Casón in 1987, and the oil spill of the Prestige tanker, in 2002. In both accidents, the strong winds played a decisive role.

A buoy anchored in the area provides hourly collected wind speed and wind direction, being the measurements of this latter variable a set of circular or periodic data. Thus, in order to describe the relation of the wind pattern with wind speed in the Galician coast during winter season, a new exploratory tool based on nonparametric smoothers (kernel and smoothing splines) is introduced, taking into account the circular nature of the measurements for wind direction.

In any nonparametric procedure, a smoothing parameter should be chosen, controlling the global aspect of the estimator and its dependence on the sample. The SiZer method, developed by Chaudhuri and Marron (1999) for

linear data, provides a means of circumventing the smoothing parameter selection and, at the same time, allows for the assessment of statistically significant features (peaks and valleys) in the data structure by finding the regions of significant gradient (zero crossings of the derivative). The SiZer ideas can be fitted to the circular setting yielding the CircSiZer.
This paper is organized as follows. Section 2 provides a brief overview on kernel regression estimation when the explanatory variable is circular and the response is linear. Section 3 is devoted to the introduction of the CircSiZer plot in the regression setting, detailing its interpretation. The idea of CircSiZer with smoothing splines is briefly introduced. Finally, the performance of the new CircSiZer is illustrated with a real data set in Section 4.

2 Nonparametric circular-linear regression

Let $\{(\Theta_i, Y_i),\ i = 1, \ldots, n\}$ be a random sample from (Θ, Y) a circular and a linear random variables, respectively. The relation between these variables can be modeled by

$$Y_i = f(\Theta_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where, f denotes the regression function and ε_i are real–valued random variables with zero mean and variance σ^2.
The regression function f can be estimated by using kernel smoothers. The local linear regression estimate for $f(\theta)$ and $f'(\theta)$ at an angle θ are given by $\hat{f}(\theta;\nu) = \hat{a}$ and $\hat{f}'(\theta;\nu) = \hat{b}$, where

$$(\hat{a}, \hat{b}) = \arg\min_{(a,b)} \sum_{i=1}^{n} K_\nu(\theta - \Theta_i)\left[Y_i - (a + b\sin(\theta - \Theta_i))\right]^2 \tag{2}$$

(see Di Marzio et al., 2009 for details). In equation (2), ν is the smoothing parameter and K_ν is a circular kernel function, e.g., the von Mises kernel with concentration parameter ν. Large values of ν lead to undersmoothed estimations of the regression curve, exaggerating the local features in the sample and tending to an interpolation of the data. On the other hand, small values of ν result in a global averaging, oversmoothing the local characteristics in the data (see Figure 1).

3 CircSiZer: SiZer map for circular data

As noticed in the Introduction, smoothing parameter selection is a critical issue. Apart from the lack of a uniformly superior rule for that purpose, from a practical point of view, the exploration of the estimators at different smoothing degrees (for a range of reasonable bandwidth values, between

oversmoothing and undersmoothing levels) will provide more in–depth information about the available data. However, significant features in the underlying data structure should be effectively disentangled from sampling artifacts. Features like peaks and valleys of a smooth curve can be characterized in terms of zero crossings of derivatives. Hence, the significance of such features can be judged from statistical significance of zero crossings or equivalently the sign changes of derivatives. This idea has been sucessfully exploited by Chaudhuri and Marron (1999) in developing a simple yet effective tool called SiZer for exploring significant structures in density and regression curves.

In the usual inferential approach in the statistical literature, the spotlight is placed on the true underlying curve f and doing inference on it, in particular, based on confidence bands. A crucial problem in nonparametric estimation is that $f(\theta;\nu) = \mathbb{E}(\hat{f}(\theta;\nu))$ is not necessarily equal to $f(\theta)$, involving an inherent bias specially for small values of ν (see Figure 1, left). The bias can be reduced by taking large values of ν, but in this case the estimator is highly variable, depending strongly on the data sample (see Figure 1 right). Chaudhuri and Marron (1999) avoid the bias-variance trade off problem by adopting the scale–space ideas which naturally lead to make inference on the smoothed curve $f(\cdot;\nu)$ rather than on the curve f. It should be noted that, for small values of ν, the smoothed curve $f(\cdot;\nu)$ can be very different from f. However if ν is within a reasonable range, $f(\cdot;\nu)$, which can be thought as the curve at a resolution level ν, shows the same valley-peaks structure as f (see Figure 1, center).

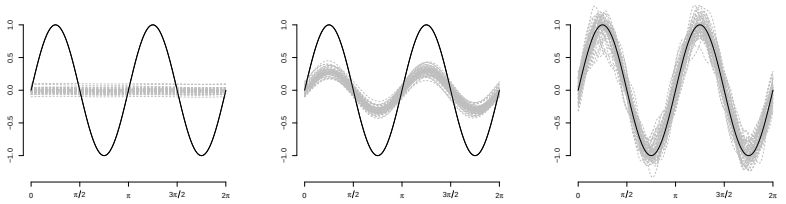


FIGURE 1. Nonparametric regression estimators (gray curves) from 50 random samples of size 250 from model (1) with $f(\theta) = \sin(2\theta)$ (solid line) and normally distributed errors with variance $\sigma^2 = 0.5$ and with $\nu = 0.2$ (left), $\nu = 2$ (center) and $\nu = 50$ (right).

Thus, in order to assess the significance of features such as peaks and valleys, instead of constructing confidence intervals for $f'(\theta)$, CircSiZer seeks confidence intervals for the scale–space version $f'(\theta;\nu) \equiv \mathbb{E}(\hat{f}'(\theta;\nu))$. The confidence intervals are constructed as detailed in Oliveira et al. (2012). So, for a given pair (θ,ν), the curve at a smoothing level ν is significantly increasing (decreasing) if the confidence interval is above (below) 0 and

if the confidence interval contains 0, the curve at the smoothing level ν and at the point θ does not have a statistically significant slope. This information can be displayed in a circular color map in such a way that, at a given ν, the performance of the estimated curve is represented by a color ring where differents colors will allow to indentify peaks and valleys. Blue (black, for black and white versions) color indicates locations where the curve is significantly increasing; red (dark gray) color shows where it is significantly decreasing and purple (gray) indicates where it is not significantly different from zero. Regions where there is not enough data to make statements about significance are gray (light gray) coloured. Thus, at a given bandwidth, a significant peak can be identified when a region of significant positive gradient is followed by a region of significant negative gradient (i.e. blue–red pattern), and a significant trough by the reverse (red–blue pattern), taking clockwise as the positive sense of rotation. Values of the smoothing parameter ν, which are transformed to $-log_{10}$ scale, are indicated along the radius.

In order to construct the smoother for f, kernel methods are not the only alternative. As in Marron and Zhang (2005), the CircSiZer can also be adapted to smoothing splines, by replacing the kernel estimator by a smoothing spline conveniently adapted to the circular nature of the explanatory variable. In this case, the regression function is estimated by finding the smooth function $\hat{f}_\lambda$ that minimizes the penalized least squares criterion

$$S(g) = \sum_{i=1}^{n} \left[Y_i - g(\Theta_i)\right]^2 + \lambda \int_0^T \left[g''(\theta)\right]^2 d\theta, \tag{3}$$

over the class of twice continuously differentiable periodic functions with period $T = 2\pi$. The parameter λ plays the role of the smoothing parameter. When λ is large, a premium is being placed on smoothness and potential estimators with large second derivatives are penalized. Conversely, a small value of λ corresponds to more emphasis on goodness of fit with $\lambda = 0$ giving an estimator that interpolates the data. It can be shown that $\hat{f}_\lambda$ is neccesarily a periodic cubic spline on $[\Theta_1, \Theta_{n+1}]$ with knots at the points Θ_i, $i = 1, \ldots, n+1$, where $\Theta_{n+1} = \Theta_1 + T$.

4 Real data analysis

The practical usefulness of the proposed CircSiZer map is illustrated by the analysis of a real dataset concerning wind direction and speed in the atlantic coast of Galicia (NW–Spain). The dataset consists of hourly observations of wind direction (in degrees) and wind speed (in m/s) in winter season (from November to February), from 2003 until 2012. In order to avoid the dependence present between consecutive measurements in the time series, observations were taken with a lag period of 95 hours.

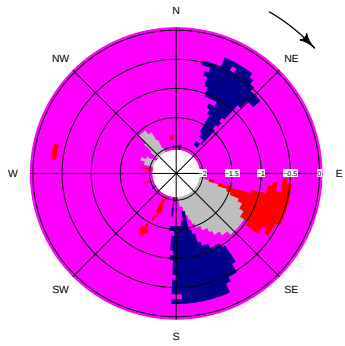


FIGURE 2. CircSiZer map for wind speed (m/s) with respect to wind direction.

Figure 2 shows the CircSiZer map for regression, applied for exploring the relation between wind speed as a response and wind direction as a covariate. It can be seen that wind speed increases when wind direction comes from NE and S. Winds from SE are not frequent at all during winter period, being this fact reflected by the gray shaded area.

Acknowledgments: This work has been supported by Project MTM2008–03010 from the Spanish Ministry of Science and Innovation, and by the IAP network StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences), from Belgian Science Policy. We also acknowledge the advice of José A. Crujeiras, an experienced skipper working in the Galician coast.

References

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves, *Journal of the American Statistical Association*, **94**, 807 – 823.

Di Marzio, M., Panzera A. and Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statistics & Probability Letters*, **79**, 2066 – 2075.

Marron, J. S. and Zhang, J. T. (2005). SiZer for smoothing splines, *Computational Statistics*, **20**, 481 – 502.

Oliveira, M., Crujeiras, R. M. and Rodríguez–Casal, A. CircSiZer: an exploratory tool for circular data. *Journal of Environmental and Ecological Statistics*. DOI 10.1007/s10651-013-0249-0.

Variable Selection and Shrinkage of Varying to Fixed Effects in Finite Mixtures of Generalized Linear Models

Wolfgang Pößnecker[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-University Munich, Germany

E-mail for correspondence: Wolfgang.Poessnecker@stat.uni-muenchen.de

Abstract: Standard regression models like GLMs implicitly assume a homogeneous influence of the predictor variables on the response across the entire sample. If the underlying population exhibits latent heterogeneity, this can adequately be accounted for by Finite Mixtures of GLMs (FMGLMs). Because FMGLMs use a GLM of its own for each subpopulation of the heterogeneous overall population, they are inherently of a higher dimensionality than ordinary GLMs. In order to retain stability and interpretability, regularization and variable selection are required. Due to the special structure of FMGLMs, variable selection is not directly induced by lasso penalties. We therefore suggest a group lasso approach to variable selection in FMGLMs and additionally introduce a novel penalty that is able to shrink varying to fixed effects, which means that mixing only affects the predictors for which a heterogeneous effect is justified. The rest of the predictors is incorporated with a more parsimonious fixed effect. We show how the corresponding estimator can be computed using a penalized EM algorithm and demonstrate the usefulness of our approach by an application to forensic data.

Keywords: Finite mixture model; lasso; group lasso; regularization; penalization.

1 Introduction

The goal of regression analysis is to model the dependence of a response variable Y on a vector of (potential) predictor variables $\boldsymbol{x} = (x_1, \dots, x_p)^{\mathrm{T}}$. Typically, data pairs $\{y_i, \boldsymbol{x}_i\}$, $i = 1, \dots, n$ are observed and used to fit the regression. Generalized Linear Models (GLMs) (McCullagh & Nelder, 1989) are among the most important classes of regression models. In GLMs, it is assumed that the conditional density $f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = f(y_i|\mu(\boldsymbol{x}_i))$ belongs to the exponential family and is solely influenced by the predictors through the conditional mean, which is given by

$$\mu(\boldsymbol{x}_i) = \mathrm{E}(y_i|\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \qquad i = 1, \dots, n,$$

where g denotes a link function and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^{\mathrm{T}}$ is the coefficient vector. (Note that some GLM densities, like the Gaussian, also depend on

a dispersion parameter. For notational simplicity, we will omit dispersion parameters in this paper.) It is seen from the formula above that GLMs implicitly assume the effect of all predictors x_j on Y to be constant across all observations. If the considered population is heterogeneous, this assumption may not hold, resulting in a poor and misleading fit and low prediction accuracy. If the heterogeneity is latent, Finite Mixture of Regression models are required to deal with these issues. A comprehensive reference on Finite Mixture Models is McLachlan & Peel (2000). In this paper, we focus on Finite Mixtures of GLMs (FMGLMs), which assume that the considered population consists of K different subpopulations in which the relationship between Y and $\boldsymbol{x}$ is described by a GLM. Since the class membership is latent, the conditional density of Y is a convex combination of GLM densities $f_k(\cdot)$:

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \sum_{k=1}^{K} \pi_k f_k(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_k). \tag{1}$$

The $f_k(\cdot)$ are called mixture components and are densities from the simple exponential family with mean

$$\mu_{ik} = \mu_k(\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_k) = g^{-1}(\beta_{k0} + x_{i1}\beta_{k1} + \ldots + x_{ip}\beta_{kp}).$$

The coefficient β_{kj} represents the effect of predictor x_j on the response of observations belonging to the k-th mixture component. The component weights π_k represent an observation's prior probability of belonging to the different classes and must satisfy $\sum_{k=1}^{K} \pi_k = 1,\ \pi_k > 0\ \forall k$. Additionally, let $l(\boldsymbol{\beta})$ denote the log-likelihood corresponding to (1). The overall coefficient vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$ is of length $K \cdot (p+1)$, so that FMGLMs require K times more coefficients than ordinary GLMs. This inherently higher dimensionality complicates the interpretation and threatens the stability of FMGLMs. Therefore, it seems natural to regularize the model. In this paper, we pursue a penalty approach, in which the estimator $\hat{\boldsymbol{\beta}}$ is obtained by maximizing a penalized log-likelihood $l_{\mathrm{pen}}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda P(\boldsymbol{\beta})$, where λ is a tuning parameter controlling the amount of penalization and the penalty term P is a functional that penalizes the coefficient vector $\boldsymbol{\beta}$. The specific choice of P determines the structure and properties of the resulting estimator $\hat{\boldsymbol{\beta}}$ and is discussed in the next section. Afterwards, we quickly describe how the penalized estimator can be computed using a penalized EM algorithm, followed by an application of our regularized FMGLM method to forensic data.

2 Choice of the penalty term

2.1 A penalty for structured variable selection

One possible way to reduce the aforementioned complexity of FMGLMs is to set some coefficients β_{kj} to zero, so that the corresponding predictor

does not influence Y in the k-th subpopulation. For this purpose, it was suggested in previous work on regularization in Finite Mixtures of Regressions (Khalili & Chen, 2007; Städler, Bühlmann, van de Geer, 2010) to use a penalty term whose core idea, apart from some technical details like weighting schemes, is given by the lasso (Tibshirani, 1996):

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{k=1}^{K} |\beta_{kj}|. \tag{2}$$

While this ordinary lasso approach is able to shrink single coefficients β_{kj} to zero, it does not directly induce variable selection in FMGLMs since in this model class, the influence of a predictor x_j is captured by a vector of coefficients: $\boldsymbol{\beta}_{\bullet j} = (\beta_{1j}, \ldots, \beta_{Kj})^{\mathrm{T}}$. Thus, the lasso approach only eliminates x_j from the model if K coefficients are shrunk to zero simultaneously while the penalty term does not encourage such a simultaneous shrinkage. To achieve a structured selection behavior, we suggest to instead apply the group lasso (Yuan & Lin, 2006; Meier, van de Geer, Bühlmann, 2008) to the vectors $\boldsymbol{\beta}_{\bullet j}$:

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} ||\boldsymbol{\beta}_{\bullet j}||_2, \tag{3}$$

with $||\boldsymbol{u}||_2 = \sqrt{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}}$ denoting the L_2-norm.

2.2 Shrinkage of varying to fixed effects

Another possibility to reduce the complexity of FMGLMs is to restrict the use of heterogeneous effects β_{kj} that are varying across mixture components to a subset of the potential predictors and to use a homogeneous effect that is fixed across all components for the rest. Formally, we say that predictor x_j has a fixed effect if $\beta_{1j} = \beta_{2j} = \ldots = \beta_{Kj}$. Shrinking a varying to a fixed effect removes $K - 1$ parameters from the model which greatly helps interpretability and stability. Since the concept of fixed effects is defined within the parameter vectors $\boldsymbol{\beta}_{\bullet j}$, we can structure the overall penalty term in the following way:

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \psi P_1(\boldsymbol{\beta}_{\bullet j}) \; + \; (1 - \psi) P_2(\boldsymbol{\beta}_{\bullet j}), \tag{4}$$

where $P_1(\boldsymbol{\beta}_{\bullet j}) = ||\boldsymbol{\beta}_{\bullet j}||_2$ is a penalty term that induces variable selection and P_2 has to be chosen so that it encourages the shrinkage of varying to fixed effects. The parameter $\psi \in [0, 1]$ distributes the overall penalty level λ to the two penalties. The last remaining step is to choose a suitable penalty P_2. To relate this goal to existing penalization approaches, note that if x_j has a fixed effect, all pairwise differences are zero, that is, $|\beta_{rj} - \beta_{sj}| =$

$0\ \forall r \neq s$. Therefore, we suggest the following penalty for the reduction of varying to fixed effects in FMGLMs:

$$P_2(\boldsymbol{\beta}) = \sum_{j=1}^{p} P_2(\boldsymbol{\beta}_{\bullet j}) = \sum_{j=1}^{p} \sqrt{\sum_{r=1}^{K} \sum_{s>r} (\beta_{rj} - \beta_{sj})^2}. \tag{5}$$

This penalty term combines the idea of the group lasso and the fused lasso (Tibshirani, 2005). Note that we penalize all pairwise differences, compared to the penalization of only adjacent coefficients in Tibshirani (2005), in order to prevent label switching from affecting our penalty. Following Tibshirani (2005), who called the combination of the lasso and the L_1-norm of a fusion term "fused lasso", we denote our overall penalty term "group fused lasso" since it is the combination of the group lasso and a grouped fusion term:

$$\begin{aligned} P(\boldsymbol{\beta}) &= \sum_{j=1}^{p} P(\boldsymbol{\beta}_{\bullet j}) \\ &= \sum_{j=1}^{p} \left(\psi \sqrt{\sum_{k=1}^{K} \beta_{kj}^2} \; + \; (1-\psi) \sqrt{\sum_{r=1}^{K} \sum_{s>r} (\beta_{rj} - \beta_{sj})^2} \right) \end{aligned} \tag{6}$$

3 Estimation

The standard method for the estimation of a Finite Mixture Model is the EM algorithm of Dempster et al. (1977). When applying the EM algorithm to the problem of maximizing a penalized log-likelihood, the E-step of the algorithm remains unaffected and the penalty term is passed to the M-step:

Choose starting values $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\pi}^{(0)}$. For $m = 0, 1, 2, \ldots$ until convergence, iterate:

E-step: For $i = 1, \ldots, n;\ k = 1, \ldots, K$ compute the weights τ_{ik}:

$$\tau_{ik} = \frac{\pi_k^{(m)} f_k(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}_k^{(m)})}{\sum_{s=1}^{K} \pi_s^{(m)} f_s(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}_s^{(m)})}.$$

M-step: Given the weights τ_{ik}, compute:

$$\pi_k^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}, \quad k = 1, \ldots, K$$

$$\boldsymbol{\beta}^{(m+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\tau_{ik} \log \left(f_k(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}_k) \right) \right) - \lambda P(\boldsymbol{\beta}).$$

Application of this penalized EM algorithm is straight-forward except for the maximization of the penalized, weighted GLM log-likelihood in the penalized M-step. To solve this task, we use the Fast Iterative Thresholding Algorithm (Fista) of Beck & Teboulle (2009). Further technical details can be obtained from the authors.

4 Application to the scapula data

We consider data from a dissertation (Feistl, 2004) at the Institute of Forensic Medicine of the LMU Munich which examined if the state of the scapula (shoulder) bone can be related to the age of dead bodies. Determining the age of a corpse is one of the most important tasks in forensics because it is pivotal in identifying the person.
The data set consists of 154 observations for which the age as well as height, weight and sex are known. Additionally, 15 physical features of the scapula bone were measured. The goal of the study was to find out which of these features carry information about the age of the body. Because the sex of a corpse is frequently indeterminable in practice, a particular aim of the study was to investigate if at least some of the scapula bone features are worthwhile predictors of age irrespective of sex. Since there are many potential sources of heterogeneity among these features, say healthy versus unhealthy lifestyles, we use a mixture of Gaussians with repsonse age and the 15 scapula features plus height and weight as predictors. Because we excluded the variable sex from the model, any heterogeneity between sexes is captured by the mixture. If the effect of a predictor is estimated to be fixed, we can thus conclude that the connection between this predictor and age is ambisexual. The model was fit using our group fused lasso approach, with K selected via BIC, which yielded $K = 3$ as the optimal choice, and λ chosen by 10-fold crossvalidation. The hyperparameter ψ was set to 0.5 and not tuned. The estimated parameters of the best model found by our regularized approach are summarized in the following two tables:

TABLE 1. Parameters of the regularized mixture model for the scapula data.

Comp.	π_k	Interc.	height	weight	bgi	agi	il	bpc	cyh	cde
#1	0.64	26.99	-0.05	0	0	1.05	0	0	0	0
#2	0.23	46.11	-0.89	0	0	4.83	0	0	0	0
#3	0.13	46.98	-0.23	0	0	6.04	0	0	0	0

It is seen from tables 1 and 2 that our approach was able to remove 6 of the 17 variables from the model and to shrink 6 of the remaining 11 variables to a fixed effect.

TABLE 2. Parameters of the mixture model for the scapula data, part II.

Comp.	y3c	y6a	y6b	y10	y13b	y13d	y14a	y14b	y15
#1	1.27	0.81	0.77	2.52	3.15	2.00	2.11	4.76	4.92
#2	1.27	0.81	-0.06	2.52	3.15	-5.81	2.11	1.77	4.92
#3	1.27	0.81	0.99	2.52	3.15	2.28	2.11	8.79	4.92

References

Beck, A., Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **2**, 183 – 202.

Dempster, A., Laird, N., Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1 – 38.

Feistl, W. (2004). *Lebensaltersschätzung am menschlichen Schulterblatt.* Dissertation, Ludwig-Maximilians-Universität München.

McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models.* London: Chapman & Hall.

McLachlan, G.J., Peel, G. (2000). *Finite Mixture Models.* New York: Wiley & Sons.

Meier, L., van de Geer, S., Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, **70**, 53 – 71.

Khalili, A., Chen, J. (2007). Variable Selection in Finite Mixture of Regression Models. *Journal of the American Statistical Association*, **102**, 1025 – 1038.

Städler, N., Bühlmann, P., van de Geer, S. (2010). L1-penalization for mixture regression models. *Test*, **19**, 209 – 256.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267 – 288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, **67**, 91 – 108.

Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49 – 67.

Cholesky decomposition for multivariate volatilities

Xanthi Pedeli[1], Konstantinos Fokianos[1], Mohsen Pourahmadi[2]

[1] Department of Mathematics & Statistics, University of Cyprus, Cyprus
[2] Department of Statistics, Texas A&M University, USA

E-mail for correspondence: pedeli.xanthi@ucy.ac.cy

Abstract: Parsimonious estimation of high–dimensional covariance matrices is of fundamental importance in multivariate statistics. Typical examples occur in finance, where the instantaneous dependence between several asset returns has to be taken into account. Multivariate GARCH processes have been established as a standard approach for modelling such data. However, estimation in this case involves several computational complexities. As an alternative, we put forward the idea of Cholesky decomposition, initially proposed in the context of longitudinal studies. We revisit the Cholesky decomposition approach in a time series framework where individual volatilities are modeled through the application of log-GARCH models. Simulation results show that the reparameterized covariance matrix estimate is of the same order of accuracy as the estimates obtained by traditional stochastic volatility models, such as the CCC- and DCC-GARCH.

Keywords: Cholesky decomposition; covariance matrix; GARCH.

1 Introduction

During the last three decades, parsimonious modelling of covariance structures has received increased interest (see for example Tsay, 2010) because of two important challenges that need to be addressed: the covariance matrix $\mathbf{\Sigma}$ has to be positive definite and might be high-dimensional. The complexity of the problem increases sharply with the number of correlated series or groups under study. A typical example is the case of multivariate volatility in finance where the number of covariance matrices to be estimated is the same as the number of observations.

Several multivariate extensions of the univariate GARCH models have been proposed in the literature. Early variants of multivariate GARCH models heavily depend upon a growing number of free parameters. More recent variants (see for example Francq and Zakoian, 2010, Chap.11), are either based on strong assumptions, that may not be realistic, or require restrictions that are often not met in practice.

Such problems are resolved simultaneously by employing the idea of Cholesky decomposition of a covariance matrix. More specifically, the

Cholesky decomposition provides statistically meaningful and unconstrained parameterizations and it also guarantees the positive definiteness of the estimated covariance matrix (Pourahmadi, 1999). For the case of multivariate time series, see Tsay (2010, Chap.10), Dellaportas and Pourahmadi (2012) and Lopes et al. (2012).
We combine the ideas of Cholesky decomposition

$$\mathbf{T}_t \mathbf{\Sigma}_t \mathbf{T}_t^{'} = \mathbf{D}_t \quad \text{or} \quad \mathbf{\Sigma}_t = \mathbf{T}_t^{-1} \mathbf{D}_t (\mathbf{T}_t^{-1})^{'} \tag{1}$$

and asymmetric log-GARCH models (Francq et al. 2012) for the estimation of time-varying variances and covariances of a vector of correlated time series. Modelling the diagonal entries of $\mathbf{D}_t$ using log-GARCH models comes as a natural choice since working with $\log \mathbf{D}_t$ achieves the positive definite constraint of $\mathbf{\Sigma}_t$. Equation (1) will be explained in more detail below.
Section 2 summarizes the basic idea of Cholesky decomposition. In Section 3 we briefly describe the asymmetric log-GARCH(p,q) model. Section 4 introduces the suggested modelling approach. Some preliminary simulations are reported in Section 5.

2 The Cholesky decomposition

The Cholesky decomposition of a symmetric matrix $\mathbf{\Sigma}$ is based on equation (1); see Pourahmadi (1999). We use the fact that a symmetric matrix $\mathbf{\Sigma}$ is positive definite if and only if there exists a unique lower triangular matrix $\mathbf{T}$, with 1's as diagonal entries, and a unique diagonal matrix $\mathbf{D}$ with positive diagonal entries such that

$$\mathbf{T} \mathbf{\Sigma} \mathbf{T}^{'} = \mathbf{D}. \tag{2}$$

In the time series framework, a different covariance matrix $\mathbf{\Sigma}_t$ has to be estimated at each time point $t = 1, \ldots, n$ and hence n equations of type (2) should be formed. For reparameterization of $\mathbf{\Sigma}_t$ using (2) and interpretation of the entries of the matrices $\mathbf{T}_t$ and $\mathbf{D}_t$, the idea of regression is the basic tool (Pourahmadi, 1999; Tsay, 2010, Chap.10). More specifically, for a given time point t, let $\mathbf{Y}_t = (Y_{1;t}, \ldots, Y_{m;t})$ be a generic random vector with mean zero and positive–definite covariance matrix $\mathbf{\Sigma}_t$, where m stands for the dimensionality of the multivariate series. The linear least squares regression of $Y_{j;t}$, $j = 1, \ldots, m$ on its predecessors $Y_{1;t}, \ldots, Y_{j-1;t}$, is defined as

$$Y_{j;t} = \sum_{k=1}^{j-1} \phi_{jk;t} Y_{k;t} + \epsilon_{j;t}, \quad j = 1, \ldots, m, \tag{3}$$

where the regression coefficients $\phi_{jk;t}$ determined from $\mathbf{\Sigma}_t$ are unconstrained. The prediction error $\epsilon_{j;t}$ has variance $\mathrm{Var}(\epsilon_{jt}) = d_{j;t}^2$. Successive prediction errors are uncorrelated so that the covariance matrix of

$\boldsymbol{\epsilon}_t = (\epsilon_{1;t}, \ldots, \epsilon_{m;t})'$ is given by $\mathbf{D}_t = \text{cov}(\boldsymbol{\epsilon}_t) = \text{diag}(d^2_{1;t}, \ldots, d^2_{m;t})$. Hence, (3) can be written in matrix form as $\boldsymbol{\epsilon}_t = \mathbf{T}_t \mathbf{Y}_t$, where $\mathbf{T}_t$ is a unit lower triangular matrix with $-\phi_{jk;t}$ in the (j,k)th position, for $j = 2, \ldots, m$ and $k = 1, \ldots, j-1$. Then it follows that the unit lower triangular matrix $\mathbf{T}_t$ diagonalizes $\boldsymbol{\Sigma}_t$ as in (1).

3 The log-GARCH model

The (asymmetric) log-GARCH(p,q) model is defined as

$$\begin{aligned} \epsilon_t &= \sigma_t \eta_t \qquad (4) \\ \log \sigma_t^2 &= \omega + \sum_{i=1}^{q} (\alpha_{i+} 1_{\{\epsilon_{t-i}>0\}} + \alpha_{i-} 1_{\{\epsilon_{t-i}<0\}}) \log \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \log \sigma_{t-j}^2 \end{aligned}$$

where $\sigma_t > 0$ and η_t is a sequence of independent and identically distributed (i.i.d.) variables with $E\eta_0 = 0$ and $E\eta_0^2 = 1$. When $\boldsymbol{\alpha}_+ = \boldsymbol{\alpha}_-$ with $\boldsymbol{\alpha}_+ = (\alpha_{1+}, \ldots, \alpha_{q+})$ and $\boldsymbol{\alpha}_- = (\alpha_{1-}, \ldots, \alpha_{q-})$, the usual summetric log-GARCH is obtained. This formulation allows for asymmetric effects between positive and negative asset returns (leverage effect), and it does not impose any positivity restrictions on the volatility coefficients (Francq et al., 2012).

4 Modelling covariances via log-GARCH

Assume that $\mathbf{Y}_t = (Y_{1;t}, \ldots, Y_{m;t}) \sim N(0, \boldsymbol{\Sigma}_t)$ and $\mathbf{T}_t \boldsymbol{\Sigma}_t \mathbf{T}_t' = \mathbf{D}_t$, or equivalently $\boldsymbol{\Sigma}_t = \mathbf{T}_t^{-1} \mathbf{D}_t (\mathbf{T}_t^{-1})'$, where $\mathbf{T}_t = (-\phi_{jk})$ is a unit lower triangular matrix and $\mathbf{D}_t = \text{diag}(d^2_{1;t}, \ldots, d^2_{m;t})$. Then, $\mathbf{T}_t^{-1} \mathbf{D}_t^{1/2}$ is the lower triangular Cholesky decomposition of $\boldsymbol{\Sigma}_t$ such that $\mathbf{T}_t \mathbf{Y}_t \equiv \boldsymbol{\epsilon}_t \sim N(0, \mathbf{D}_t)$. It follows that the joint normal distribution of $\mathbf{Y}_t$ given its past, can be expressed as a set of m recursive conditional regressions where

$$Y_{1;t} \sim N(0, d^2_{1;t}), \quad Y_{j;t} \sim N(\sum_{k=1}^{j-1} \phi_{jk;t} Y_{k;t}, d^2_{j;t}), \; j = 2, \ldots, m \qquad (5)$$

To allow for a time-varying $\boldsymbol{\Sigma}_t$ without any restrictions, we define $\log d^2_{j;t}$, $j = 1, \ldots, m$ as a log-GARCH(1,1) model, i.e.

$$\log d^2_{j;t} = \omega_j + \{\alpha_{j+} 1_{\{\epsilon_{j;t-1}>0\}} + \alpha_{j-} 1_{\{\epsilon_{j;t-1}<0\}}\} \log \epsilon^2_{j;t-1} + \beta_j \log d^2_{j;t-1} \qquad (6)$$

Equations (5) and (6) define our stochastic volatility model since availability of $\phi_{jk;t} \equiv \phi_{jk}$'s and $d^2_{j;t}$'s implies availability of $\mathbf{T}_t$ and $\mathbf{D}_t$ and hence $\boldsymbol{\Sigma}_t$ can be reconstructed.

Estimates of the log-GARCH parameters are obtained by Quasi Maximum Likelihood (QMLE). In particular, a QMLE of the vector of unknown

parameters $\boldsymbol{\theta}$ is defined as any solution $\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \tilde{Q}_n(\boldsymbol{\theta})$ where $\tilde{Q}_n(\boldsymbol{\theta}) = \sum_{t=r_0+1}^{n} \tilde{\ell}_t(\boldsymbol{\theta})/n$, $\tilde{\ell}_t(\boldsymbol{\theta}) = \epsilon_t^2/\tilde{d}_t^2(\boldsymbol{\theta}) + \log \tilde{d}_t^2(\boldsymbol{\theta})$, r_0 is a fixed integer and $\log \tilde{d}_t^2(\boldsymbol{\theta})$ is recursively defined by (6). Details on the choice of initial values, strong consistency and asymptotic normality of the QMLE can be found in Francq et al. (2012).

5 Simulations

The forecasting performance of the suggested approach was assessed through a small set of simulation experiments. Bivariate series of length $n = 500$ were simulated as follows: Let $\boldsymbol{\Sigma}_0$ be the identity and $\boldsymbol{\Sigma}_1$ a (2×2) covariance matrix. We define $\boldsymbol{\Sigma}_t = (1-w_t)\boldsymbol{\Sigma}_0 + w_t\boldsymbol{\Sigma}_1$ where w_t are weights of the form $w_t = (1 - c^{t/n})/(1 - c)$, $t = 1, 2, \ldots, n$, and at each t we draw $\mathbf{Y}_t = (Y_{1;t}, Y_{2;t}) \sim N(0, \boldsymbol{\Sigma}_t)$. The constant c involved in the computation of w_t can be any real number and we have chosen $c = 0.5$.

Apart from the proposed Cholesky-log-GARCH model, the CCC- and DCC-GARCH models of order (1,1), as well as a moving blocks approach (see Lopes et al., 2012) were used for the purposes of comparison. Regarding the moving blocks approach, at each time t, we compute the sample covariance of q observations centered at t. We used $q = 11, 31$ and 75. At both the left and right end of the data range, the block size is truncated when it exceeds the observed time span. The sample covariance matrices obtained in this way, served as estimates of the original $\boldsymbol{\Sigma}_t$'s but they were also used for the calculation of starting values for the GARCH-type models.

To measure the accuracy of a covariance matrix estimate $\hat{\boldsymbol{\Sigma}}_t$, we used the entropy loss and Kullback-Leibler loss, $\Delta_{1t} = \operatorname{tr}\left(\boldsymbol{\Sigma}_t^{-1}\hat{\boldsymbol{\Sigma}}_t\right) - \log\left|\boldsymbol{\Sigma}_t^{-1}\hat{\boldsymbol{\Sigma}}_t\right| - m$ and $\Delta_{2t} = \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\Sigma}_t\right) - \log\left|\hat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\Sigma}_t\right| - m$ respectively. In each setting we conducted 200 simulations. To compare the Cholesky-log-GARCH model with alternative modelling approaches in each simulation we defined $\delta_i^{(\mathrm{CHOL}-\cdot)} = \sum_{t=1}^{n}\left(\Delta_{it}^{(\mathrm{CHOL})} \leq \Delta_{it}^{(\cdot)}\right)/n$, where $i = 1, 2$, $(\cdot) = \mathrm{MB}, \mathrm{CCC}, \mathrm{DCC}$ and 'CHOL', 'CCC', 'DCC'and 'MB' stand for the Cholesky-log-GARCH, CCC-GARCH, DCC-GARCH and moving-blocks approaches respectively. Following, we averaged over the total number of simulations to get the final measures for the comparison of the forecasting performances.

Results are summarized in Table 1. Almost all δ_i's are greater than 50% implying that the Cholesky-log-GARCH model performs better than the CCC-, DCC-GARCH and the moving blocks approach. As expected, $\delta_i^{(\text{CHOL-MB})}$ is affected by the block's size. In particular, increasing q improves the forecasting performance of the moving blocks approach. However, even for large q's the Cholesky-log-GARCH model still has a better predictive ability. These results can be seen as an early indication that the

suggested Cholesky-log-GARCH model is promising in terms of parsimonious modelling of conditional covariances without violating the positive definite constraint.

TABLE 1. Simulation results.

q	Σ_1		$\delta_1^{(\text{CHOL-})}$ (s.e.)			$\delta_2^{(\text{CHOL-})}$ (s.e.)		
			MB	CCC	DCC	MB	CCC	DCC
11	4	1.8	0.920 (0.037)	0.682 (0.143)	0.590 (0.141)	0.924 (0.033)	0.666 (0.147)	0.570 (0.144)
	1.8	2						
11	16	8	0.927 (0.036)	0.628 (0.123)	0.634 (0.121)	0.928 (0.034)	0.618 (0.126)	0.625 (0.125)
	8	25						
31	4	1.8	0.757 (0.081)	0.673 (0.135)	0.558 (0.144)	0.764 (0.079)	0.659 (0.140)	0.539 (0.150)
	1.8	2						
31	16	8	0.777 (0.097)	0.611 (0.125)	0.615 (0.129)	0.779 (0.095)	0.600 (0.130)	0.606 (0.134)
	8	25						
75	4	1.8	0.529 (0.119)	0.674 (0.149)	0.561 (0.163)	0.534 (0.120)	0.663 (0.151)	0.544 (0.167)
	8	25						
75	16	8	0.593 (0.122)	0.619 (0.111)	0.616 (0.113)	0.593 (0.123)	0.610 (0.114)	0.608 (0.117)
	8	25						

Acknowledgments: K. Fokianos and X. Pedeli are supported by a Leventis Foundation grant. The authors would like to thank Christian Francq for kindly providing part of the algorithm for the estimation of the log-GARCH model.

References

Dellaportas, P., and Pourahmadi, M. (2012). Cholesky-GARCH models with applications to finance. *Statistics and Computing*, **22**, 849 – 855.

Francq, C., Wintenberger, O. and Zakoian, J. (2012). GARCH models without positivity constraints: exponential or log GARCH? *MPRA Paper 41373, University Library of Munich, Germany.*

Francq, C., and Zakoian, J. (2010). *GARCH Models: Structure, Statistical Inference and Financial Applications.* John Wiley, New York.

Lopes, H., McCullogh, R., and Tsay, R. (2012). Cholesky Stochastic Volatility Models for High-Dimensional Time Series. *Technical Report, Chicago Booth.*

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterization. *Biometrika*, **86**, 677 – 690.

Tsay, R. (2010). *Analysis of Financial Time Series (THIRD Edition).* John Wiley, New York.

A unifying framework for specifying generalized linear models for categorical data

Jean Peyhardi[1] [2], Catherine Trottier[1], Yann Guédon[2]

[1] Université Montpellier 2, I3M, Montpellier, France
[2] CIRAD, UMR AGAP and Inria, Virtual Plants, Montpellier, France

E-mail for correspondence: jean.peyhardi@math.univ-montp2.fr

Abstract: In the context of categorical data analysis, the case of nominal and ordinal data has been investigated in depth while the case of partially ordered data has been comparatively neglected. We first propose a new specification of generalized linear models (GLMs) for categorical response variables which encompasses all the classical models such as multinomial logit, odds proportional or continuation ratio models but also led us to identify new GLMs. This unifying framework makes the different GLMs easier to compare and combine. We then define the more general class of partitioned conditional GLMs for categorical response variables. This new class enables to take into account the case of partially ordered data by combining nominal and ordinal GLMs.

Keywords: categorical data analysis; generalized linear model; partitioned conditional model; recursively partitioned categories.

1 Specification of generalized linear models for categorical response variables

Let Y denote the response variable with J categories ($J > 1$) and $X = (X_1, \ldots, X_p)$ be a vector of explanatory variables in a general form (a categorical variable being represented by an indicator vector). The definition of a GLM includes the specification of a link function g which is a C^1-diffeomorphism from $M = \{(\pi_1, ..., \pi_{J-1}) \in]0,1[^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\}$ to an open subset of $\mathbb{R}^{J-1}$, between the expectation $\pi = E[Y|X{=}x] = (\pi_1, ..., \pi_{J-1})^T$ and the linear predictor $\eta = (\eta_1, ..., \eta_{J-1})^T$. All the classical link functions $g = (g_1, \ldots, g_{J-1})$, described in the literature -see Agresti (2002) and Fahrmeir and Tutz (2001)- share the same structure which we propose to write as

$$g_j = F^{-1} \circ r_j \ , \quad j = 1, \ldots, J-1,$$

where F is a continuous and strictly increasing cumulative density function (cdf) and $r = (r_1, \ldots, r_{J-1})^T$ is a C^1-diffeomorphism from M to an open

subset of $]0,1[^{J-1}$. Thus we have

$$r_j(\pi) = F(\eta_j)\ , \qquad j = 1, \ldots, J-1.$$

In the following we describe in more details the components r, F and η.

Ratio r: The linear predictor η is not directly related to the expectation π but to a particular transformation r of the vector π which we call the ratio. In the following we will consider four particular C^1-diffeomorphism. The *adjacent*, *sequential* and *cumulative* ratios are respectively defined by $\pi_j/(\pi_j+\pi_{j+1})$, $\pi_j/(\pi_j+\ldots+\pi_J)$ and $\pi_1+\ldots+\pi_j$ for $j = 1, \ldots, J-1$, assume order among categories but with different interpretations. The *reference* ratio, defined by $\pi_j/(\pi_j + \pi_J)$ for $j = 1, \ldots, J-1$, is mainly useful for nominal response variables.

Latent variable cdf F: The most commonly used symmetric distributions are the *logistic* and *Gaussian* distributions but the *Laplace* and *Student* distributions may also be useful. The most commonly used asymmetric distributions are the *Gumbel max* and *Gumbel min* distributions. Playing on the symmetrical or asymmetrical character and the more or less heavy tails may markedly improve the model fit. In applications the Student(d) distribution will be approximated by a Gaussian distribution when $d > 30$.

Linear predictor η: It can be written as the product of the design matrix Z and the vector of parameters β (Fahrmeir and Tutz, 2001). Each explanatory variable can have its own design effect. For example, if X_1 has a *global* effect, X_2 a *local* effect, ... and X_p a *global* effect, the corresponding design matrix, with $J-1$ rows, is

$$Z = \begin{pmatrix} 1 & & & x_1^T & x_2^T & & & & x_p^T \\ & 1 & & x_1^T & & x_2^T & & & x_p^T \\ & & \ddots & \vdots & & & \ddots & \cdots & \vdots \\ & & & 1 \quad x_1^T & & & & x_2^T & x_p^T \end{pmatrix}.$$

This design will be denoted by the tuple (global, local,..., global) and a single word global or local will denote the same design for all the explanatory variables $X_1, \ldots, X_p$.

Finally, we propose to specify a particular GLM for categorical response variables by the (r, F, Z) triplet with

$$r(\pi) = \mathbf{F}(Z\beta),$$

where $\mathbf{F}(\eta) = (F(\eta_1), \ldots, F(\eta_{J-1}))^T$.

This specification eases the comparison of GLMs for categorical response variables; see examples in Table 1. Moreover, it enables to define an enlarged

TABLE 1. (r, F, Z) specification of some classical GLMs for categorical response variables.

Multinomial logit model $P(Y=j) = \frac{\exp(\alpha_j + x^T\delta_j)}{1+\sum_{k=1}^{J-1}\exp(\alpha_k + x^T\delta_k)}$	(reference, logistic, local)
Odds proportional logit model $\log\left\{\frac{P(Y\leq j)}{1-P(Y\leq j)}\right\} = \alpha_j + x^T\delta$	(cumulative, logistic, global)
Proportional hazard model (Grouped Cox Model) $\log\{-\log P(Y>j)\} = \alpha_j + x^T\delta$	(cumulative, Gumbel min, global)
Adjacent logit model $\log\left\{\frac{P(Y=j)}{P(Y=j+1)}\right\} = \alpha_j + x^T\delta_j$	(adjacent, logistic, local)
Continuation ratio logit model $\log\left\{\frac{P(Y=j)}{P(Y>j)}\right\} = \alpha_j + x^T\delta_j$	(sequential, logistic, local)

set of GLMs for nominal response variables by $\{(\text{reference}, F, Z)\}$ triplets, which includes the multinomial logit model. GLMs for nominal and ordinal response variables are usually defined with different design matrices Z; see the first two rows in Table 1. Fixing the desing matrix Z may ease the comparison of GLMs for nominal and ordinal response variables.

Finally, a single estimation procedure based on Fisher scoring algorithm can be applied to all the GLMs specified by (r, F, Z) triplets. Using the chain rule, the score function can be separated into two parts where the first depends on the triplet (r, F, Z), whereas the second does not.

$$\frac{\partial l}{\partial \beta} = \underbrace{Z^T \frac{\partial \mathbf{F}}{\partial \eta}\frac{\partial \pi}{\partial r}}_{(r,F,Z)\text{ dependant part}} \underbrace{\text{Cov}(Y|X=x)^{-1}\,[y-\pi]}_{(r,F,Z)\text{ independent part}}.$$

2 Partitioned conditional GLMs for categorical response variables

The main idea is to recursively partition the J categories and then to specify a GLM for each partition. Such combinations of GLMs have already been proposed such as the two-step model of Morawitz and Tutz (1990), that combines sequential and cumulative models, or the partitioned conditional model for partially ordered set (POS-PCM) of Zhang and Ip (2012) that combines multinomial logit and odds proportional logit models. Our proposal can be seen as a generalization of POS-PCMs that benefits from the genericity of the (r, F, Z) specification. In particular, our objective was not only to propose GLMs for partially-ordered response variables but also to differentiate the role of explanatory variables for each partition of categories using for instance different design matrices.

Definition: Let $J \geq 2$ and $1 \leq k \leq J-1$. A ***k*-partitioned conditional GLM** for categories $1, \ldots, J$ is defined by:

- A **partition tree $\boldsymbol{\mathcal{T}}$** of $\{1, \ldots, J\}$ with $\mathcal{V}^*$, the set of non terminal nodes of cardinal $\boldsymbol{k}$. Let Ω_j^V be the children of node $V \in \mathcal{V}^*$.
- A **collection $\{(\boldsymbol{r^V}, \boldsymbol{F^V}, \boldsymbol{Z^V}(\boldsymbol{x^V})) \mid \boldsymbol{V} \in \boldsymbol{\mathcal{V}^*}\}$** of GLM(s) for each conditional probability vector $\pi^V = (\pi_1^V, \ldots, \pi_{J_V-1}^V)$, where $\pi_j^V = P(Y \in \Omega_j^V | Y \in V, X^V = x^V)$ for $j = 1, \ldots, J_V$.

Model estimation: It can be shown that the log-likelihood of partitioned conditional GLMs can be decomposed into components such that each component can be maximised individually because GLMs attached to each partition of categories do not share common regression coefficients (Zhang and Ip, 2012). Each component corresponds to the partition of a parent node $V \in \mathcal{V}^*$, and therefore, each GLM $(r^V, F^V, Z^V(x^V))$ can be estimated separately using the procedure described in Section 1.

3 Application to back pain prognosis

Doran and Newell (1975) describe a back pain study with 101 patients. The response variable y was the assessment of back pain after three weeks of treatment using the six ordered categories: *worse, same, slight improvement, moderate improvement, marked improvement, complete relief.* The three selected explanatory variables observed at the beginning of the treatment period were x_1 = *length of previous attack* (1=short, 2=long), x_2 = *pain change* (1=getting better, 2=same, 3=worse) and x_3 = *lordosis* (1=absent/decreasing, 2=present/increasing).
The best model we obtained for this data set was a 2-partitioned conditional GLM (log-likelihood of -151.36 with 9 parameters); see figure 1.

Anderson (1984) obtained a log-likelihood of −154.39 with 9 parameters for the stereotype model. This gain is mainly due to the modularity of partitioned conditional GLMs (change of ratio r and design matrix Z between the two partitions).

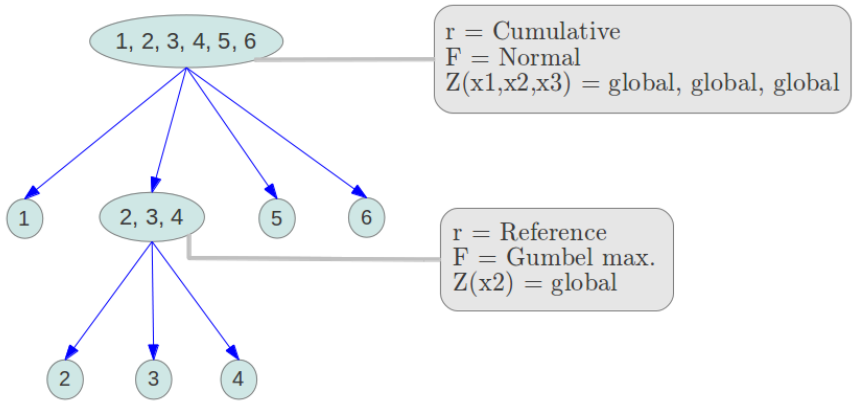


FIGURE 1. Representation of a 2-partitioned conditional GLM (partition tree $\mathcal{T}$ of six response categories and two associated GLMs for categorical response variables)

References

Agresti, A. (2002). Categorical Data Analysis. *John Wiley and Sons.*

Anderson, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, **46**, 1 – 30.

Doran, D.M.L. and Nowell, D.J. (1975). Manipulation in treatment of low back pain: a multicentre study. *British medical journal*, **2**, 161 – 164.

Fahrmeir, L. and Tutz, G. (2001). Multivariate Statistical Modelling Based on Generalized Linear Models. *Springer Verlag.*

Morawitz, B. and Tutz, G. (1990). Alternative parameterizations in business tendency surveys. *Mathematical Methods of Operations Research, Springer*, **34**, 143 – 156.

Zhang, Q. and Ip, E.H. (2012). Generalized linear model for partially ordered data. *Statistics in Medicine.*

Bayesian Sparse Factor Regression Approach to Genomic Data Integration

Veronika Ročková[1], Emmanuel Lesaffre[12]

[1] Department of Biostatistics Erasmus MC Rotterdam, The Netherlands
[2] L-BioStat, KU Leuven, Belgium

E-mail for correspondence: v.rockova@erasmusmc.nl

Abstract: We consider a sparse factor regression model for interpretable partition of variation in multiple responses. The variability is separated into three components: the one (a) attributable to shared latent variables, (b) explained by a set of explanatory variables, and (c) idiosyncratic to each response. The model augments the exchangeable regressions approach by adding a latent factor structure, which allows for dependent patterns of marginal covariance between the responses. In order to enable identification of a parsimonious structure, we impose spike and slab priors on the individual entries in both the factor loading and regression matrices. The computation is carried out by an EM algorithm, which typically converges much faster than sampling schemes. The model is applied to a problem of integrating two genomic datasets, where expression of microRNA's is related to expression of genes with underlying connectivity pathway network. The model allows simultaneous identification of likely pathway groupings as well as functional gene-microRNA interactions.

Keywords: EM algorithm; Factor Analysis; Sparsity; Spike and Slab.

1 Factor Regression Model Structure

The data setup under consideration consists of a $n \times G$ matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]'$ containing n observations on G related genes and a $n \times p$ predictor matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]'$. The columns in $\boldsymbol{X}$ and $\boldsymbol{Y}$ have been centered and $\boldsymbol{X}$ further standardized to have unit column-wise variances. We assume throughout that $\boldsymbol{y}_i$'s arise as independent realizations from a latent factor regression model with a linear mapping representation using observed explanatory variables as well as unobserved (latent) factors. Given $\boldsymbol{\omega}_i$, a $(d \times 1)$ vector of latent variables for the case i, we assume

$$f(\boldsymbol{y}_i \mid \boldsymbol{\omega}_i, \boldsymbol{A}, \boldsymbol{B}, \Sigma) = \mathrm{N}_G(\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{B}\boldsymbol{\omega}_i, \Sigma), \quad 1 \leq i \leq n, \tag{1}$$

where the $G \times p$ matrix $\boldsymbol{A}$ consists of unknown regression coefficients, $\Sigma = \mathsf{diag}\{\sigma_j^2\}_{j=1}^G$ is a diagonal matrix of unknown positive scalars and the $G \times d$ matrix $\boldsymbol{B}$ contains factor loadings weighting the contributions

of individual factors in gene expression. Following the standard assumption, the latent vectors are considered to arise through random sampling from Gaussian distribution $\mathrm{N}_d(\mathbf{0}, \sigma_\omega^2 \mathrm{I}_d)$. The equation (1) induces a corresponding Gaussian distribution for observations $f(\boldsymbol{y}_i \mid \boldsymbol{A}, \boldsymbol{B}, \Sigma) = \mathrm{N}_G(\boldsymbol{A}\boldsymbol{x}_i, \boldsymbol{B}\boldsymbol{B}' + \Sigma)$, $1 \le i \le n$, which permits dependent patterns of covariance among $\boldsymbol{y}_i$ attributable to the common latent factors. The factor model (1) is not identifiable without further constraints. Following a common convention, we assume that $\boldsymbol{B}$ has a full-rank lower triangular structure with unit elements on the diagonal and $\sigma_\omega^2 = 1$.

2 Sparsity Modeling with Spike and Slab Priors

The Bayesian approach to defining sparse factor and regression structures uses priors on the elements in $B = \{b_{jl}\}_{j,l=1}^{G,d}$ and $A = \{a_{jl}\}_{j,l=1}^{G,p}$ that induce either zeroes or values close to zero with high-probability. We take the latter approach, exploiting the continuous relaxation of a point-mass mixture prior (George and McCulloch (1993)). Denote $[\boldsymbol{A}, \boldsymbol{B}] = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G]'$, where

$$\boldsymbol{\beta}_j = (\boldsymbol{a}_j', \boldsymbol{b}_j')' = (\underbrace{a_{j1}, \ldots\ldots\ldots, a_{jp}}_{\text{regression coefficients}}, \underbrace{b_{j1}, \ldots, b_{jd}}_{\text{factor loadings}})', \quad 1 \le j \le G.$$

Then each $\boldsymbol{\beta}_j$ is assigned a conjugate Gaussian mixture prior

$$\pi(\boldsymbol{\beta}_j) \sim \mathrm{N}_{p+d}(\mathbf{0}, \sigma_j^2 \boldsymbol{D}_j)$$

where $\boldsymbol{D}_j = \mathsf{diag}\{(1-\gamma_{jl})v_{0l} + \gamma_{jl}v_{1l}\}_{l=1}^{p+d}$ and v_{0l} (resp. v_{1l}) takes two different values v_{0a} and v_{0b} (resp. v_{1a} and v_{1b}) depending whether $1 \le l \le p$ or $p+1 \le l \le p+d$. The hyper-parameters $0 < v_{0a} < v_{1a}$ (resp. $0 < v_{0b} < v_{1b}$) are set to be small and large to distinguish the β_{jl} values which warrant a functional relationship between j-th response and l-th predictor (resp. factor). Here

$$\boldsymbol{\gamma}_j = (\underbrace{\gamma_{j1}, \ldots\ldots\ldots, \gamma_{jp}}_{\text{predictor indicators}}, \underbrace{\gamma_{jp+1}, \ldots, \gamma_{jp+d}}_{\text{factor indicators}})', \quad 1 \le j \le G,$$

denotes a vector of inclusion indicators, which characterizes the binary selection status of each predictor/factor. To define the joint prior distribution on $\Gamma = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_G)'$, we use a hierarchical prior which allows different occurrence probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p+d})'$ of non-zero elements in the different columns of $[\boldsymbol{A}, \boldsymbol{B}]$ and assumes that all indicators in Γ are column-wise exchangeable. In particular,

$$\pi(\Gamma) = \prod_{l=1}^{p+d} \theta_l^{\sum_{j=1}^G \gamma_{jl}} (1-\theta_l)^{G - \sum_{j=1}^G \gamma_{jl}}.$$

To complete the specification, θ_l's are assigned Beta distribution $\mathcal{B}(a, b)$ and for σ_j^2 we assume independent Inverse Gamma priors $\mathrm{IG}(\eta/2, \eta\lambda/2)$.

3 EM Algorithm for Sparse Bayesian Factor Regression

As an alternative to stochastic search, we pursue a deterministic approach to finding modes of the posterior distribution $\pi(\boldsymbol{A}, \boldsymbol{B}, \Sigma, \boldsymbol{\theta}|\boldsymbol{Y})$ using an EM algorithm, treating both $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n]'$ and Γ as "missing data". We combine the EM algorithm for estimation in probabilistic principal components (Tipping and Bishop (1999)) together with EM-based Bayesian variable selection method in linear regression using the mixture priors (Rockova and George (2012)). The EM algorithm locates posterior modes by iteratively maximizing the objective function:

$$Q\left(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\theta}, \Sigma\right) = \mathsf{E}_{\boldsymbol{\Gamma}, \boldsymbol{\Omega}|\cdot} \left[\log \pi \left(\underbrace{\boldsymbol{A}, \boldsymbol{B}, \Sigma, \boldsymbol{\theta}}_{\text{unknown parameters}} , \overbrace{\boldsymbol{\Gamma}, \boldsymbol{\Omega}}^{\text{missing data}} \mid \underbrace{\boldsymbol{Y}}_{\text{observed data}} \right) \right],$$

where $\mathsf{E}_{\boldsymbol{\Gamma}, \boldsymbol{\Omega}|\cdot}(\cdot)$ denotes the conditional expectation given observed data and current parameter estimates at the k-th iteration.

3.1 Closed Form E-step

The E-step entails computation of a conditional mean and covariance of the latent data in $\boldsymbol{\Omega}$ and the conditional expectation of the matrix of indicators Γ. These can be evaluated according to:

$$\begin{aligned}
\langle \boldsymbol{\omega}_i \rangle &= M^{(k)} B^{(k)'} \Sigma^{(k)-1} \left(\boldsymbol{y}_i - A^{(k)} \boldsymbol{x}_i \right), \\
M^{(k)} &= \left(B^{(k)'} \Sigma^{(k)-1} B^{(k)} + \mathrm{I}_d \right)^{-1}, \\
\langle \gamma_{jl} \rangle &= \frac{\mathrm{N}(\beta_{jl}^{(k)}; 0, \sigma_j^{(k)2} v_{1l}) \theta_l^{(k)}}{\mathrm{N}(\beta_{jl}^{(k)}; 0, \sigma_j^{(k)2} v_{1l}) \theta_l^{(k)} + \mathrm{N}(\beta_{jl}^{(k)}; 0, \sigma_j^{(k)2} v_{0l})(1 - \theta_l^{(k)})}.
\end{aligned}$$

3.2 Closed Form M-step

Denote $\boldsymbol{y}^j$ the j-th column in $\boldsymbol{Y}$, $||\cdot||$ the l^2 norm, $\langle \Omega \rangle = [\langle \boldsymbol{\omega}_1 \rangle, \ldots, \langle \boldsymbol{\omega}_n \rangle]'$ and $\boldsymbol{D}_j^\star = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & M^{(k)} \end{pmatrix} + \mathrm{diag}\{\langle \gamma_{jl} \rangle / v_{1l} + (1 - \langle \gamma_{jl} \rangle) / v_{0l}\}_{l=1}^{p+d}$. The M-step consists of updates (for $j > d$):

$$\begin{aligned}
\boldsymbol{\beta}_j^{(k+1)} &= \left([\boldsymbol{X}, \langle \Omega \rangle]' [\boldsymbol{X}, \langle \Omega \rangle] + \boldsymbol{D}_j^\star \right)^{-1} [\boldsymbol{X}, \langle \Omega \rangle]' \boldsymbol{y}^j, \\
\sigma_j^{(k+1)} &= \sqrt{\frac{||\boldsymbol{y}^j - [\boldsymbol{X}, \langle \Omega \rangle] \boldsymbol{\beta}_j^{(k+1)}||^2 + ||\boldsymbol{D}_j^{\star 1/2} \boldsymbol{\beta}_j^{(k+1)}||^2 + \nu\lambda}{n + p + d + \nu}}, \\
\theta_l^{(k+1)} &= \frac{\sum_{j=1}^G \langle \gamma_{jl} \rangle + a - 1}{a + b + G - 2},
\end{aligned}$$

For $j \leq d$, each vector $\boldsymbol{\beta}_j^{(k+1)}$ is confined by the lower triangular structure in the factor loading matrix and can be obtained again as a ridge regression solution, only with a restricted predictor matrix and a modified response. The updates $\sigma_j^{(k+1)}$ change correspondingly.

4 Recovering Sparsity

Sparsity in the matrices $\widehat{A}$ and $\widehat{B}$ can be recovered by thresholding estimates that are small in magnitude. This is equivalent to screening out coefficients $\widehat{\beta_{jl}}$ with a small conditional posterior inclusion probability, e.g. according to a local median probability model rule $\mathsf{P}(\gamma_{jl} = 1 \mid \cdot) < 0.5$. These conditional probabilities are easily obtained as a by-product of the EM algorithm.

5 Simple Simulated Example

In this simulated example, $n = 100$ observations were drawn from a $G = 200$ dimensional factor regression model (1) with $p = 10$ predictors and $d = 5$ factors. The rows of the regression matrix $\boldsymbol{X}$ were drawn independently from $\mathrm{N}_{10}(\mathbf{0}, \mathrm{I})$. We set σ_j's to one. The matrix $\boldsymbol{B}$ equals $\mathrm{I}_d \otimes \mathbf{1}_{40}$ apart from the diagonal elements b_{kk}, which are set to one. The columns of $\boldsymbol{A}$ were drawn independently from a Bernoulli distribution with success probabilities $\{(j-1) \times 0.1\}_{j=1}^{10}$. We consider unit starting values for elements in $\boldsymbol{A}^{(0)}, \boldsymbol{B}^{(0)}$ and $\sigma_j^{(0)}$'s. We set $v_{0a} = v_{0b} = 0.1, v_{1a} = v_{1b} = 100$ and $\eta = \lambda = a = b = 1$. We run the EM algorithm assuming $d = 3, 5, 7$. The heatmap of $[\boldsymbol{A}, \boldsymbol{B}]$ and the posterior inclusion probability matrix $\widehat{\Gamma}$ for the true number of factors $d = 5$ is given in the upper left and right panels in Figure 1. The 0.5 thresholding rule yields 14 false negatives in determining the true pattern of sparsity in $[\boldsymbol{A}, \boldsymbol{B}]$. We further observe that (a) under-determining the factor dimension (assuming $d = 3$) leads to more false negatives in the matrix $\boldsymbol{A}$ (lower left panel in Figure 1), (b) over-determining the factor dimension (assuming $d = 7$) leads the same number of false negatives in $\boldsymbol{A}$, where the estimates $\widehat{\boldsymbol{\theta}}$ indicate the redundancy of the last two factors (lower right panel in Figure 1).

6 Data Analysis

We analyze data collected at the Department of Hematology at Erasmus MC Rotterdam on patients with acute myeloid leukemia (AML). The 212 cases were analyzed for (1) the expression of $M = 177$ microRNA's using quantitative PCR, (2) gene expression using Affymetrix Human Genome Gene-Chips. Only $G = 2\,087$ genes that were involved in common pathways (according to the KEGG database) were analyzed. Setting

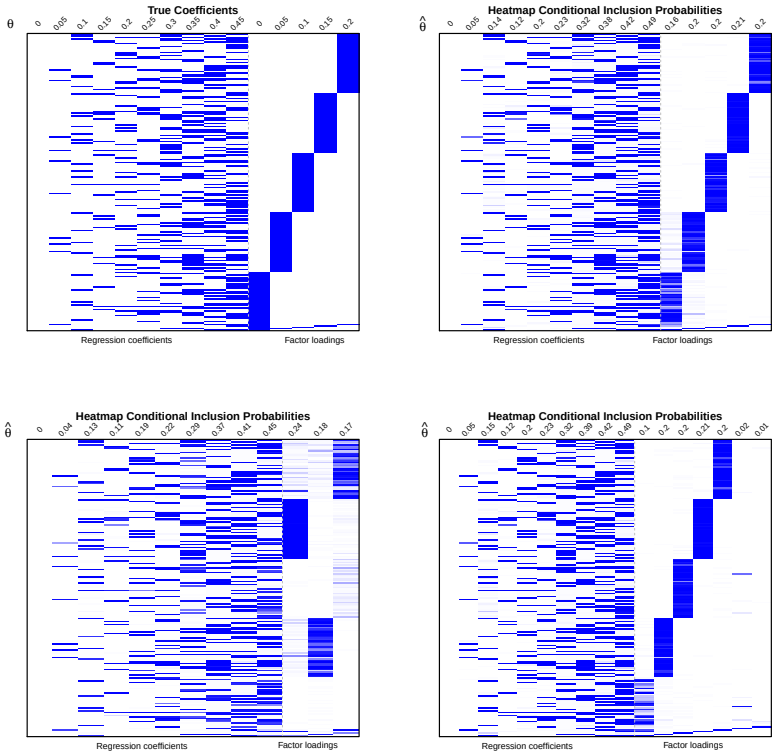


FIGURE 1. Heatmap of true regression coefficients (upper left) and conditional posterior inclusion probabilities (upper right $d = 5$, lower left $d = 3$, lower right $d = 7$), $\boldsymbol{\theta}$ or $\widehat{\boldsymbol{\theta}}$ above each column

$d = 20, v_{0a} = 1, v_{0b} = 0.5$ and $v_{1a} = v_{1b} = 100$ we obtain a sparse network consisting of 494 gene-microRNA and 162 gene-factor associations involving 33 microRNA's, 495 genes and 3 factors. The association matrix (obtained after thresholding $\widehat{\Gamma}$ based on the 0.5 threshold) is depicted in Figure 2. MicroRNA's known to be associated with either a clinical outcome in AML (miR-181 family) or specific AML subtypes (miR-10, miR-26 and miR-196 families) are marked in red.

7 Discussion

We considered a sparse factor regression model to discern putative associations between microRNAs and genes, while simultaneously unraveling likely pathway structure. The model differs from similar proposals (Carvalho et al. (2008)) in the choice of the continuous variable selection prior and the fast EM-based deterministic strategy for the computation.

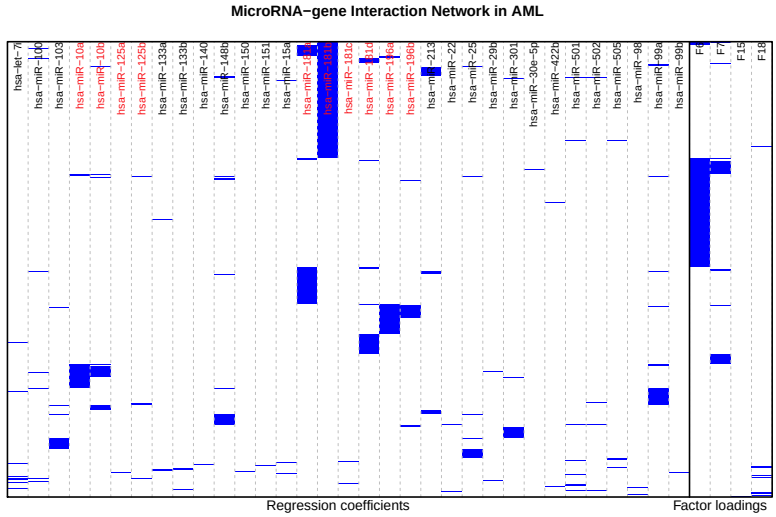


FIGURE 2. Results of the factor regression model applied on the AML data

References

Carvalho, C., Chang, J., Lucas, J., Nevins, J.R., Wang, Q. and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics *Journal of the American Statistical Association*, **103**, 1438 – 1456.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling *Journal of the American Statistical Association*, **88**, 881 – 889.

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis *Journal of the Royal Statistical Society, Series B*, **61**, 611 – 622.

Rockova, V. and George, E. (2012). EMVS: The EM approach to Bayesian variable selection *Submitted manuscript*

Fast algorithm for smoothing parameter selection in multidimensional P-splines

María Xosé Rodríguez - Álvarez[1], Dae - Jin Lee[2], Thomas Kneib[3], María Durbán[4], Paul H.C. Eilers[5]

[1] Unit of Clinical Epidemiology and Biostatistics, Complexo Hospitalario Universitario de Santiago de Compostela, Spain
[2] CSIRO Mathematics, Informatics and Statistics, Clayton, VIC, Australia
[3] Department of Economics, Georg August University Göttingen, Germany
[4] Department of Statistics, Universidad Carlos III de Madrid, Spain
[5] Erasmus Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: mariajose.rodriguez.alvarez@usc.es

Abstract: In this work a new computational algorithm for estimating a multivariate P-spline model with anisotropic penalizations is presented. Our proposal is based on the mixed model representation of a multivariate P-spline, in which the smoothing parameter for each covariate becomes a ratio between variances. On the basis of the restricted maximum likelihood (REML), we obtain closed-form expressions for the variance components estimators. This formulation leads to an efficient implementation that can considerably reduce the computational load. The proposed algorithm can be seen as a generalization of the algorithm by Schall (1991) - for variance components estimation - to deal with non-standard structures of the covariance matrix of the random effects. Finally, we illustrate our proposal with historical records of monthly precipitation data.

Keywords: Smoothing; P-splines; Anisotropic penalization; Mixed Models

1 Introduction

In this paper, for the sake of illustration we focus our attention on a bivariate P-spline with B-splines basis and difference penalties (Eilers and Marx, 1996). However, the proposed algorithm can be easily extended to the multivariate case, and any basis and quadratic penalty combination can be also accomodated.
Consider a bivariate regression problem

$$y_i = f\left(x_{i1}, x_{i2}\right) + \varepsilon_i \sim N\left(0, \sigma^2\right), i = 1, \ldots, n. \tag{1}$$

where f is a smooth and unknown function. Within the P-spline framework of Eilers and Marx (1996, 2003), the unknown surface $f(x_1, x_2)$ can be approximated by the tensor product of two univariate B-splines, i.e.,

$f(x_1, x_2) = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \theta_{jk} B_j^1(x_1) B_k^2(x_2)$. In matrix notation, model (1) is then expressed as

$$\mathbf{y} = \mathbf{B}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{B} = \mathbf{B_2} \square \mathbf{B_1}$ (with $\square$ denoting the 'row-wise' kronecker product), and $\boldsymbol{\theta} = (\theta_{11}, \cdots, \theta_{d_1 1}, \cdots, \theta_{d_1 d_2})'$. By assuming an anisotropic penalization (i.e. a different amount of smoothing for x_1 and x_2), the penalty matrix is given by

$$\breve{\mathbf{P}} = \lambda_1 \mathbf{I}_{d_2} \otimes \mathbf{D}_1' \mathbf{D}_1 + \lambda_2 \mathbf{D}_2' \mathbf{D}_2 \otimes \mathbf{I}_{d_2},$$

where $\mathbf{D}_i$ is a matrix that forms differences of order q_i, $(i = 1, 2)$ and $\otimes$ denotes the kronecker product.

Under this representation, the P-spline approach is based on minimizing the penalized least-squares function $S(\boldsymbol{\theta}; \mathbf{y}, \lambda_1, \lambda_2) = \|\mathbf{y} - \mathbf{B}\boldsymbol{\theta}\| + \boldsymbol{\theta}' \breve{\mathbf{P}} \boldsymbol{\theta}$, with the smoothing parameters being selected (usually within a 2D - grid) using some criteria such as (generalized) cross-validation, the Akaike information criterion or the Bayesian information criterion. However, this search can be very expensive to compute, specially for large datasets.

A different approach for estimating model (2) is to use the equivalence between P-splines and mixed models (Currie et al., 2006). However, under the mixed-model formulation of a bivariate P-spline with anisotropic penalizations, the covariance matrix of the corresponding random effects has a non - standard form, with a block involving both the smoothing parameters λ_1 and λ_2. This feature, makes model estimation unfeasible using standard mixed modelling software. Although estimation can be done by numerical maximization of the (restricted) log - likelihood (ML or REML), it has also the drawback of being computationally demanding.

In this work, we present a new computational algorithm for estimating a bivariate P-spline with anisotropic penalizations on the basis of the mixed model formulation. Following the ideas presented in Harville (1977) and Schall (1991), we derive closed-form expressions for the variance components estimators, based on REML. The algorithm is, therefore, straightforward to implement in practice. Moreover, some characteristics of the derived expressions can be used to improve even further the computational time, thus rendering very good computing times. Finally, our approach allows to obtain the total effective dimension of the fitted bivariate function f in (1) as the sum of components related to each covariate x_1 and x_2.

2 Computational algorithm

Let us consider the mixed model representation of model (2):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\alpha} \sim N(0, \boldsymbol{G}) \text{ and } \boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 \boldsymbol{I}_n\right).$$

Using the 2D reparameterization of model (2) based on the Singular Value Decomposition (SVD) of the marginal penalties $\mathbf{D}_i' \mathbf{D}_i$ $(i = 1, 2)$ (see Lee and Durbán, 2011), the random effects covariance matrix $\mathbf{G}$ has inverse

$$\mathbf{G}^{-1} = \begin{pmatrix} \frac{1}{\tau_2^2}\tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{q_1} & & \\ & \frac{1}{\tau_1^2}\mathbf{I}_{q_2} \otimes \tilde{\mathbf{\Sigma}}_1 & \\ & & \frac{1}{\tau_2^2}\tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{d_1-q_1} + \frac{1}{\tau_1^2}\mathbf{I}_{d_2-q_2} \otimes \tilde{\mathbf{\Sigma}}_1 \end{pmatrix},$$

where $\tilde{\mathbf{\Sigma}}_i$ are the non-zero eigenvalues of $\mathbf{D}_i'\mathbf{D}_i$ $(i = 1, 2)$, $\tau_1^2 = \frac{\lambda_1}{\sigma^2}$ and $\tau_2^2 = \frac{\lambda_2}{\sigma^2}$. It is important to notice that using the SVD, $\mathbf{G}^{-1}$ is a diagonal matrix, then

$$\mathbf{G} = \text{diag}\left(\frac{1}{\{\mathbf{G}^{-1}\}_{ll}}\right),$$

where $\{\cdot\}_{ll}$ denotes the lth diagonal element of $\mathbf{G}^{-1}$.

2.1 Fixed and random effects parameters estimation

By *Theorem 2* in Harville (1977), $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ are estimated as:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \\ \hat{\boldsymbol{\alpha}} &= \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{y}, \end{aligned} \tag{3}$$

where

$$\mathbf{V} = \sigma^2\mathbf{I}_n + \mathbf{Z}\mathbf{G}\mathbf{Z}',$$

and

$$\mathbf{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}.$$

2.2 Variance components estimation

The variance components estimators are obtained by maximizing the restricted log-likelihood

$$l^* = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{V}^{-1}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Given that $\boldsymbol{G}$ is a diagonal matrix, we find that the partial derivatives of l^* with respect to the variance components τ_i^2 $(i = 1, 2)$ are

$$2\frac{\partial l^*}{\partial \tau_i^2} = -\frac{1}{\tau_i^2}trace\left(\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\mathbf{G}\frac{\mathbf{\Lambda}_i}{\tau_i^2}\mathbf{G}\right) + \frac{1}{\tau_i^4}\hat{\boldsymbol{\alpha}}'\mathbf{\Lambda}_i\hat{\boldsymbol{\alpha}}, \tag{4}$$

where

$$\mathbf{\Lambda}_2 = \begin{pmatrix} \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{q_1} & & \\ & \vec{0}_{q_2(d_1-q_1)} & \\ & & \tilde{\mathbf{\Sigma}}_2 \otimes \mathbf{I}_{d_1-q_1} \end{pmatrix},$$

$$\mathbf{\Lambda}_1 = \begin{pmatrix} \vec{0}_{q_1(d_2-q_2)} & & \\ & \mathbf{I}_{q_1} \otimes \tilde{\mathbf{\Sigma}}_1 & \\ & & \mathbf{I}_{d_2-q_2} \otimes \tilde{\mathbf{\Sigma}}_1 \end{pmatrix}.$$

By equating expression (4) to zero, the estimators of the variance components are then obtained:

$$\hat{\tau}_i^2 = \frac{\hat{\boldsymbol{\alpha}}' \boldsymbol{\Lambda}_i \hat{\boldsymbol{\alpha}}}{\mathrm{ed}_i}, \tag{5}$$

where

$$\mathrm{ed}_i = trace\left(\boldsymbol{Z'PZG}\frac{\boldsymbol{\Lambda}_i}{\tau_i^2}\mathbf{G}\right),$$

and

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\boldsymbol{\alpha}}\|^2}{n - \sum_{i=1}^{2} \mathrm{ed}_i - p}, \text{ with } p = \mathrm{ncol}(\mathbf{X}).$$

2.3 Remarks

It should be noted that the sum of the effective dimensions involved in the estimation of the variance component τ_1^2 and τ_2^2 (see (5)) corresponds to the effective dimension of the penalized part (or random part) of the fitted surface:

$$\begin{aligned}
\mathrm{ed}_1 + \mathrm{ed}_2 &= trace\left(\boldsymbol{Z'PZG}\frac{\boldsymbol{\Lambda}_1}{\tau_1^2}\mathbf{G}\right) + trace\left(\boldsymbol{Z'PZG}\frac{\boldsymbol{\Lambda}_2}{\tau_2^2}\mathbf{G}\right) \\
&= trace\left(\boldsymbol{Z'PZG}\right) \\
&= trace\left(\boldsymbol{ZGZ'P}\right) \\
&= trace\left(H_{Random}\right)
\end{aligned}$$

where H_{Random} denotes the hat matrix of the random part (see (3)).

3 Application to precipitation data

We have applied the proposed computational algorithm to a dataset containing weather observation records compiled in the United States of America (USA). The data came from the National Climatic Data Center (NCDC) of the USA, and contain monthly total precipitation (in millimeters) from January 1895 to December 1997. For illustration purposes, we focus our analysis on estimating the spatial pattern of precipitation for April 1948 in the USA. Specifically, the dataset comprises a total of 11918 records.
For each record, the longitude-latitude position of monitoring stations is provided, jointly with the monthly total precipitation in millimeters and a standardization of this raw observation, called *anomaly* (see Johns et al. 2003). From these 11918 records, only 5906 correspond to stations where monthly total precipitation values were observed, and the remainder 6012 correspond to missing station precipitation values, that have been filled in using spatial statistics (Johns et al. 2003). We therefore restricted our analysis to the 5906 true records. Figure 1 shows the raw data of the monthly

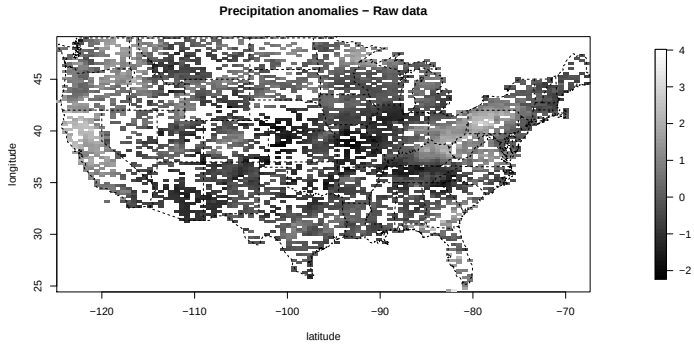


FIGURE 1. Raw data the monthly precipitation anomalies in USA for April 1948.

precipitation anomalies in USA for April 1948. We fitted model (1) using our aproach, with second order penalties and 40 inner knots for each marginal cubic B-spline basis. The fitted surface is shown in Figure 2. The effective dimension for longitude and latitude was 287.297 and 397.043 respectively. As regards the computing time, the algorithm took 357.57 seconds with a 2.40GHz Intel Core i5 processor and 4GB of RAM.

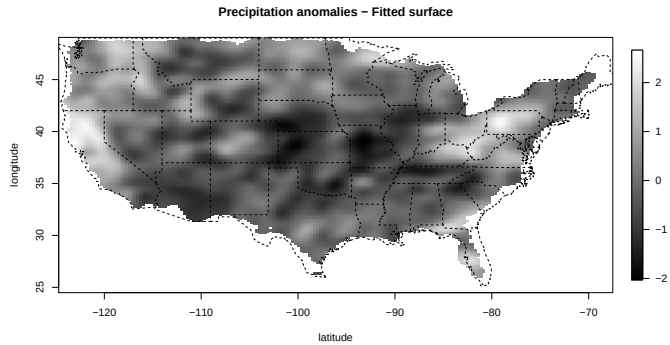


FIGURE 2. Estimated spatial pattern of the monthly precipitation anomalies in USA for April 1948.

For comparison purposes, we also analised this dataset using the `bam` function in `R`-package `mgcv` (version 1.7-22)(Wood, 2006; R Development Core Team, 2013). `bam` function has been specially designed to deal with very large datasets. As before, we chosen second order penalties and 40 inner knots for each marginal cubic B-spline basis, and the REML criterion (`method = "fREML"`) was used for the automatic selection of the smoothing parameters (Wood, 2011). The fitted model had an effective dimension

of 744.6, and the computing time achieved by this approach was 1152.13 seconds, about 3.2 times more than with using our algorithm.

Acknowledgments: The authors would like to express their gratitude for the support received in the form of the Spanish Ministry of Economy and Competitiveness grants MTM2011-28285-C02-01 and MTM2011-28285-C02-02. Work of María Xosé Rodríguez - Álvarez was supported by grant CA09/0053 from the Instituto de Salud Carlos III. The research of Dae-Jin Lee was funded by an NIH grant for the Superfund Metal Mixtures, Biomarkers and Neurodevelopment project 1PA2ES016454-01A2.

References

Johns, C., Nychka, D., Kittel, T. and Daly, C. (2003). Infilling sparse records of spatial fields. *Journal of the American Statistical Association*, **98**, 796 – 806.

Currie, I., Durban. M. and Eilers, P. H. C. (2006) Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, **68**, 259 – 280.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89 - 121

Eilers, P. H. C. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*, **66**, 159 - 174.

Harville, D. A. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, **72**, 320 - 338

Lee, D.-J. and Durbán, M. (2011) P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, **11 (1)**, 49 - 69

R Development Core Team (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719 - 721.

Wood, S. N. (2006). *Generalized Additive Models. An introduction with R.* Chapman & Hall/CRC.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B*, **73**, 3 – 36.

Approximate Bayesian inference based on modified log-likelihood ratios

Erlis Ruli[1], Laura Ventura[1], Walter Racugno[2]

[1] Department of Statistical Sciences, University of Padova, Italy
[2] Department of Mathematics and Informatics, University of Cagliari, Italy

E-mail for correspondence: ruli@stat.unipd.it

Abstract: The aim of this contribution is to discuss recent advances in approximate Bayesian computation based on the asymptotic theory of modified log-likelihood ratios. Some new results for a vector parameter of interest are presented. These approximations may routinely be applied for Bayesian inference, since little more than standard likelihood quantities is required for their implementation, and hence they may be available at little additional computational cost over simple first-order approximations. In particular, they can be used to define accurate Bayesian credible sets with good frequentist properties. The method is illustrated by an example.

Keywords: Asymptotic expansion; Bayesian credible set; Laplace approximation; Tail area probability.

1 Introduction

The aim of this contribution is to discuss recent advances in approximate Bayesian computation based on the asymptotic theory of modified log-likelihood ratios for a vector of parameters. We show that this theory provides asymptotic formulae for posterior Bayesian credible sets with accurate frequentist coverage.

Higher-order approximations for posterior distributions based on modifications of likelihood roots have been widely discussed in the Bayesian literature; see, among others, Reid (2003), Sweeting (1996), Ventura *et al.* (2013), and references therein. One appealing feature of these approximations is that they may routinely be applied for Bayesian inference, since they require little more than standard likelihood quantities for their implementation, and hence they may be available at little additional computational cost over simple first-order approximations.

In this paper, a new result for multivariate posterior distributions based on modifications of the likelihood ratio is presented and its use for Bayesian computation of credible sets is illustrated. Paralleling approximations for univariate posterior distributions, the proposed results are based on the

asymptotic theory of modified log-likelihood ratios and only routine maximization output is required for its implementation.
The paper is organized as follows. In Section 2, after a brief review of higher-order Bayesian approximations for a scalar parameter, the proposed extension for the multiparameter case is developed. In Section 3 the method is illustrated by a numerical example. Concluding remarks are given in Section 4.

2 An asymptotic formula for vector parameters

Consider a sampling model $p(y;\theta)$, with $\theta \in \Theta \subseteq \mathcal{R}^d$, and let $L(\theta) = L(\theta;y) = \exp\{\ell(\theta)\}$ be the likelihood function based on data y. Given a prior $\pi(\theta)$ for θ, Bayesian inference is based on the posterior distribution $\pi(\theta|y) \propto \pi(\theta)L(\theta)$.
Let us start from the simplest case $d = 1$. In many applications an approximation of the posterior tail area

$$\int_{-\infty}^{\theta_0} \pi(\theta|y)\, d\theta \;=\; \Pr(\theta \le \theta_0|y) \tag{1}$$

is required. A higher order approximation of (1) can be derived as follows. The first step is to consider the Laplace expansion of $\pi(\theta|y)$ in (1), given by

$$\pi(\theta|y) \doteq \frac{1}{\sqrt{2\pi}} j(\hat\theta)^{1/2} \exp\{\ell(\theta) - \ell(\hat\theta)\} \frac{\pi(\theta)}{\pi(\hat\theta)}, \tag{2}$$

where $\hat\theta$ is the maximum likelihood estimator (MLE) of θ, $j(\theta)$ is the observed information and the equality holds with relative error of order of $O(n^{-1})$. Therefore, we obtain

$$\int_{-\infty}^{\theta_0} \pi(\theta|y)\, d\theta \;\doteq\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\theta_0} j(\hat\theta)^{1/2} \exp\left\{-\frac{r(\theta)^2}{2}\right\} \frac{\pi(\theta)}{\pi(\hat\theta)}\, d\theta, \tag{3}$$

where $r(\theta) = \mathrm{sign}(\hat\theta - \theta) w(\theta)^{1/2}$ is the likelihood root, with $w(\theta) = 2(\ell(\hat\theta) - \ell(\theta))$ the log-likelihood ratio statistic.
The second step is to change the variable of integration from θ to $r(\theta)$. The motivation for this change of variable is that, in terms of $r(\theta)$ the posterior distribution is approximately equal to the standard normal density times a suitable function of r. After the change of variable, with Jacobian $d\theta/dr = -r/\ell'(\theta)$, with $\ell'(\theta) = d\ell(\theta)/d\theta$, we obtain

$$\int_{-\infty}^{\theta_0} \pi(\theta|y)\, d\theta \;\doteq\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r_0} \exp\left\{-\frac{r^2}{2} - 2\log b(r)\right\} dr, \tag{4}$$

where $b(r) = j(\hat{\theta})\frac{r\pi(\theta)}{\ell'(\theta)\pi(\hat{\theta})}$ and $r_0 = r(\theta_0)$. The last step is a further change of variable from r to $r^\star = r - r^{-1}\log b(r)$, so that $-(r^\star)^2 = -r^2 + 2\log b(r) - (r^{-1}\log b(r))^2$. The Jacobian of this transformation and the term $(r^{-1}\log b(r))^2$ contribute only to the error of (4) and it can be shown that

$$\int_{-\infty}^{\theta_0} \pi(\theta|y)\, d\theta \stackrel{\cdot\cdot}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r_0^\star} \exp\left\{-\frac{(r^\star)^2}{2}\right\} dr^\star = \Phi\{r_0^\star\}, \tag{5}$$

where $\Phi(\cdot)$ is the standard normal distribution, $r_0^\star = r_0 + r_0^{-1}\log b(r_0)^{-1}$ and the symbol "$\stackrel{\cdot\cdot}{=}$" indicates that the equality holds with accuracy of order $O(n^{-3/2})$.

Form (5) an accurate $(1-\alpha)$ equi-tailed credible interval can be computed as

$$CI_{1-\alpha} = \{\theta : |r^\star(\theta)| \leq z_{1-\alpha/2}\}, \tag{6}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal density. In general, the solution of $r^\star(\theta) = z_\alpha$ gives the approximate α-quantile of the posterior distribution and, in particular, for $\alpha = 1/2$ an approximation of the posterior median is obtained.

A possible limitation of (5) is that it does not allow to compute moment-based posterior summaries. Nevertheless, it can be used to simulate values from the approximate marginal posterior; for more details see Ruli *et al.* (2012).

In this contribution we are concerned on the extension of (6) in the presence of many parameters ($d > 1$). Following the same steps of the scalar case, consider first the Laplace approximation of $\pi(\theta|y)$. Second, a change of variable form θ to a suitable statistic $r = r(\theta)$, such that $w(\theta) = 2(\ell(\hat{\theta}) - \ell(\theta)) = r^T r$, is considered. Third, a change of variable from r to $r^\star = r - \delta(r)$, with $\delta(r)$ chosen to satisfy $r^T\delta(r) = \log b(r)$ for a suitably defined $b(r)$ is considered. We obtain

$$(r-\delta)^T(r-\delta) = r^T r - 2\log b(r) + O(n^{-2}),$$

which is asymptotically χ^2_d in large deviation regions., To compute the second step we propose to use the signed root log-likelihood ratio transformation defined in Sweeting (1996). In particular let $\theta = (\theta_1, \ldots, \theta_d) = (\theta^i, \theta^{(i+1)})$, with $\theta^i = (\theta_1, \ldots, \theta_i)$ and $\theta^{(i+1)} = (\theta_{i+1}, \ldots, \theta_d)$. Let $\hat{\theta}^{(i+1)}_{\theta^i}$ be the partial MLE of $\theta^{(i+1)}$ given θ^i, and let $\hat{\theta}_{j,\theta^i}$ be the jth component of $(\theta^i, \hat{\theta}^{(i+1)}_{\theta^i})$, for $j > i$. The signed root log-likelihood ratio transformation is thus given by

$$r(\theta) = (r_1(\theta), \ldots, r_d(\theta)), \tag{7}$$

with

$$r_i = \text{sign}(\theta_i - \hat{\theta}_{i,\theta^{i-1}})\sqrt{2[\ell(\theta^{i-1}, \hat{\theta}^{(i)}_{\theta^{i-1}}) - \ell(\theta^i, \hat{\theta}^{(i+1)}_{\theta^i})]}, \tag{8}$$

for $i = 1, \ldots, d$. From (7) and (8), it follows that r_i is a function of θ^i and (7) is a one-to-one data-dependent transformation of θ, such that $\exp\{-r^T r/2\} = L(\theta)/L(\hat{\theta})$ (Sweeting, 1996). Moreover, for the Jacobian of the transformation it holds

$$\left|\frac{dr}{d\theta}\right| = \prod_{i=1}^{d} \left|\frac{\ell_i(\theta^i, \hat{\theta}^{(i+1)}_{\theta^i})}{r_i}\right|, \tag{9}$$

where $\ell_i(\theta)$ is the ith component of the score vector $\partial\ell/\partial\theta$, for $i = 1, \ldots, d$. Finally, after reparametrizing in terms of r in the second step and changing the variable from r to $r^\star$ in the third step, we obtain

$$w^\star = w^\star(\theta) = r^T r - 2\log b(r), \tag{10}$$

with

$$b(r) = |j(\hat{\theta})|^{1/2} \frac{\pi(\theta)}{\pi(\hat{\theta})} \left[\prod_{i=1}^{d} \left|\frac{\ell_i(\theta^i, \hat{\theta}^{(i+1)}_{\theta^i})}{r_i}\right|\right]^{-1}.$$

Asymptotically, we have that $w^\star \sim \chi^2_d$ to order $O(n^{-1})$ in large deviation regions. To obtain a statistic which generalizes the scalar version $r^\star(\theta)$, Skovgaard (2001) suggests the asymptotically equivalent approximation

$$w^{\star\star} = w^{\star\star}(\theta) = r^T r \left(1 - \frac{\log b(r)}{r^T r}\right)^2, \tag{11}$$

since, for $d = 1$, it holds $w^{\star\star}(\theta) = (r - r^{-1}\log b(r))^2 = (r^\star)^2$.
Extending (6) to the multivariate case, a $(1-\alpha)$ credible region (CR) from $w^{\star\star}$ (or $w^\star$) may be defined as $\{\theta : w^{\star\star} \leq \chi^2_{d,1-\alpha}\}$, where $\chi^2_{d,1-\alpha}$ is the $(1-\alpha)$-th quantile of the χ^2_d distribution.

3 An example: the gamma model

Let y be a sample of n observations from the gamma density Gamma(a, b), with $a = b = 1$. For simplicity, the parameters are taken in logarithmic scale, i.e. $\theta = (\log a, \log b)$, and two different prior specifications are considered: the flat prior $\pi(\theta) \propto 1$ and two independent normal priors $\text{N}(\mu, 5) \times \text{N}(\mu, 5)$, with $\mu = \{0, 2, 5\}$.
As a first illustration, a sample of $n = 5$ observations is taken, and the improper uniform prior for θ is assumed. In this case, the 0.95 CR computed with $w^\star$ and $w^{\star\star}$ are compared with a Wald-type credible set $CR_W = \{\theta : (\theta - \tilde{\theta})^T j(\tilde{\theta})(\theta - \tilde{\theta}) \leq \chi^2_{2,0.95}\}$, an asymptotic deviance-type

credible set $CR_D = \{\theta : -2\log\frac{\pi(\theta|y)}{\pi(\tilde{\theta}|y)} \leq \chi^2_{2,0.95}\}$, and an exact deviance-type credible set $CR_E = \{\theta : -2\log\frac{\pi(\theta|y)}{\pi(\tilde{\theta}|y)} \leq c\}$, with c computed by simulation, with $\tilde{\theta}$ posterior mode.

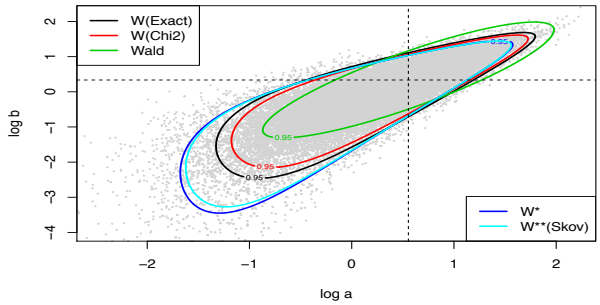


FIGURE 1. Gamma model with the flat prior.

These comparisons are shown in Figure 1, where the grey dots represents posterior samples taken with importance sampling. Clearly, the Wald-type CR is very misleading since the shape of the posterior is far from quadratic. The $w^{\star\star}$ has similar shape to the exact and approximate deviance-type CRs, although little stretched from the bottom-left side.

The effective posterior probability content of the various CRs, computed over a large sample of posterior simulations, is `0.837` for Wald-type, `0.927` for the approximate deviance-type, `0.955` for $w^{\star}$ and `0.953` for $w^{\star\star}$.

Lastly, a small simulation study is performed, through a sample of 10^4 Monte Carlo trials from the gamma model with $a = b = 1$, sample size $n = \{5, 10\}$, normal priors with $\mu = \{0, 2, 5\}$, as well as the flat prior. The aim is to compare the empirical coverage of the various CRs as well as their posterior probability contents over repeated sampling from the fixed model. The posterior probability content is computed over a subset of 500 Monte Carlo trials, randomly selected from the full set of 10^4 samples.

The results are summarized in Table 1, from which it can be deduced that Wald-type CR_W perform quite poorly, both in terms of empirical coverages and posterior probability contents. Deviance-type CR_D perform better in terms of posterior probabilities than coverages, whereas the credible sets based on $w^{\star\star}$ are superior to both. An other striking feature the credible sets based on $w^{\star\star}$ is their robustness with respect to μ. In fact, although the empirical coverage tends to vary somehow with μ, the posterior probability remains very close to the nominal value 0.95.

4 Concluding remarks

In this contribution we discuss a procedure to obtain approximate Bayesian credible sets with the right posterior probability content and with accurate

TABLE 1. Empirical coverage probabilities and empirical posterior probabilities of the 0.95 $w^{\star\star}$, Wald-type (CR_W) and deviance-type (CR_D) credible regions over 10^4 and 500 Monte Carlo trials, respectively.

Sample	Prior	$w^{\star\star}$	CR_W	CR_D
5	Flat	0.950 (0.952)	0.799 (0.831)	0.908 (0.923)
	$\mu = 0$	0.961 (0.955)	0.829 (0.853)	0.926 (0.931)
	$\mu = 2$	0.934 (0.952)	0.732 (0.862)	0.866 (0.933)
	$\mu = 5$	0.900 (0.951)	0.668 (0.873)	0.812 (0.935)
10	Flat	0.952 (0.951)	0.878 (0.894)	0.934 (0.939)
	$\mu = 0$	0.956 (0.952)	0.887 (0.899)	0.938 (0.940)
	$\mu = 2$	0.942 (0.951)	0.838 (0.902)	0.911 (0.941)
	$\mu = 5$	0.927 (0.950)	0.804 (0.906)	0.887 (0.942)

frequentist properties. These properties seem to hold regardless of the prior used, although some priors may be more preferable than others. Other possible applications are currently under examination.

Acknowledgments: This work was supported by a grant from the University of Padua (Progetti di Ricerca di Ateneo 2011) and by the Cariparo Foundation Excellence-grant 2011/2012.

References

Reid, N. (2003). The 2000 Wald memorial lectures: asymptotics and the theory of inference. *Annals of Statistics*, **31**, 1695 – 1731.

Ruli, E., Sartori, N., Ventura, L. (2012). A note on marginal posterior simulation via higher-order tail area approximations. http://arxiv.org/pdf/1212.1038.pdf (**submitted**).

Skovgaard, I.M. (2001). Likelihood asymptotics. *Scandinavian Journal of Statistics*, **28**, 3 – 32.

Sweeting, T.J. (1996). Approximate Bayesian computation based on signed roots of log-density ratios (with Discussion). In: *Bayesian Statistics*, 5, Ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, 427 – 444. Oxford University Press.

Ventura, L., Sartori, N., Racugno, W. (2013). Objective Bayesian higher-order asymptotics in models with nuisance parameters. *Computational Statistics & Data Analysis*, **60**, 90 – 96.

Joint modeling of longitudinal and time-to-event data with application to the prediction of prostate cancer recurrence

Mbéry Séne[1], Carine A. Bellera[2], Cécile Proust-Lima[1]

[1] INSERM, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France,
Univ. Bordeaux, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France

[2] INSERM, ISPED, CIC-EC7, F-33000 Bordeaux, France,
Department of Clinical Epidemiology and Clinical Research, Institut Bergonié, Regional Comprehensive Cancer Center, Bordeaux, France.

E-mail for correspondence: Mbery.Sene@isped.u-bordeaux2.fr

Abstract: In the last decade, the joint modeling has rapidly developed in the field of biostatistics and medical research to simultaneously study a longitudinal marker and a correlated time-to-event. Among joint models, the shared random-effects models, that define a mixed model for the longitudinal marker and a survival model for the time-to-event in which characteristics of the mixed model are included as covariates, received the main interest. Indeed, they extend naturally the survival models with time-dependent covariates and offer a flexible framework to explore the link between a longitudinal biomarker and a risk of event. The objective of this work is to present the shared random-effect model methodology, and illustrate their implementation and evaluation through a real example from the study of prostate cancer progression after a radiation therapy. In particular, different specifications of the dependency between the longitudinal biomarker, the prostate specific antigen (PSA), and the risk of clinical recurrence are investigated to better understand the link between these two processes. These different joint models are compared in terms of goodness-of-fit and adequation to the joint model assumptions but also in terms of predictive accuracy using the expected prognostic cross-entropy. Indeed, in addition to better understand the link between the PSA dynamics and the risk of clinical recurrence, the perspective in prostate cancer studies is to provide dynamic prognostic tools of clinical recurrence based on the biomarker history.

Keywords: Dynamic predictions; Joint models; Predictive accuracy; Prognostic cross-entropy; Prostate cancer; Shared random-effect models.

1 Introduction

In longitudinal studies, repeated measurements of biomarkers and time-to-event data are often collected. To assess the relationship between a longitudinal biomarker and a time-to-event, joint models have been developed. They allowed to eliminate the different sources of bias which are inherent when using a classic Cox model with the biomarker included as a time-dependent covariate. These models consist in simultaneously modeling the longitudinal biomarker and the time-to-event processes and in characterizing their relationship. Therefore, they provide an interesting framework to (i) assess the longitudinal trajectory of the biomarker and its association with covariates without the bias introduced by the informative dropout, (ii) assess the risk of event and its association with covariates, including the repeated measures of the biomarker and (iii) directly explore the association between the longitudinal and survival processes.
The principle of joint models is to model the repeated measurements of the biomarker using a mixed model, to model the risk of event using a survival model and to link the two sub-models through a common latent structure (Wulfsohn and Tsiatis, (1997)). This common latent structure captures the association between the two processes, therefore conditionally to the latent structure, the two processes are independent. There exist two types of joint models for longitudinal and time-to-event data: shared random effect models (continuous latent structure, the random effects) and joint latent class models (discrete latent structure, latent classes).
In addition to evaluating the association between the longitudinal and survival processes, joint models also allowed the development of new types of prediction tools that incorporate repeated measurements of biomarkers to predict the risk of an event (Proust-Lima et al., (2012)). These prognostic tools are dynamic and have the advantage of being updated at each new available measurement of the biomarker.
In this context, the aim of this work is to present the shared random effects models and the dynamic prognostic tools that can be derived from them. We considered different shared random-effect models which differ in the form of the dependence between the longitudinal biomarker and event risk and compared them in terms of goodness-of-fit and adequation to the joint model assumptions but also in terms of predictive accuracy using the expected prognostic cross-entropy (Commenges et al., (2012)). We illustrated this work on real data of progression of localized prostate cancer after radiation therapy.

2 Shared random-effect models

2.1 Longitudinal submodel

Repeated measures of the biomarker are analyzed by a linear mixed model. Specifically, we assume that the repeated measures $Y_i(t_{ij})$ are noisy mea-

sures of $Y_i^*(t_{ij})$ the true unobserved biomarker value for $j = 1, ...n_i$ at time t_{ij}. We model the mean change over time of $Y_i^*(t_{ij})$ by taking into account the correlation within the biomarker repeated measures of a same subject:

$$\begin{aligned} Y_i(t_{ij}) &= Y_i^*(t_{ij}) + \epsilon_i(t_{ij}) \\ &= X_{li}(t_{ij})^T \beta + Z_i(t_{ij})^T b_i + \epsilon_i(t_{ij}) \end{aligned} \quad (1)$$

where $X_{li}(t_{ij})$ and $Z_i(t_{ij})$ are vectors of time-dependent covariates associated respectively with the vector of fixed effects β and the vector of Gaussian random-effects b_i with mean 0 and variance-covariance matrix D. The vectors of errors of measurement $\epsilon_i = (\epsilon_i(t_{i1}), \cdots, \epsilon_i(t_{in_i}))^T$ are assumed to follow independently a multivariate Gaussian distribution with mean $\mathbf{0}$ and diagonal variance-covariance matrix $\Sigma_i = \sigma^2 I_{n_i}$; ϵ_i and b_i are independent.

2.2 Survival submodel

The risk of event could be modelled using any survival model but in practice, proportional hazard models are mostly considered and defined as follows:

$$\lambda_i\left(t|X_{ei}, b_i\right) = \lambda_0\left(t\right) e^{X_{ei}^T \gamma + h(b_i,t)^T \eta} \quad (2)$$

where $\lambda_0\left(t\right)$ is the baseline hazard function, γ is the vector of coefficients defining the association between the vector of covariates X_{ei} and the survival time, and $h(b_i, t)$ is a multivariate function of the random-effects b_i defined in (1) and associated with the vector of parameters η. The coefficients η measure the association between the longitudinal and survival processes while $h(b_i, t)$ defines the nature of the dependence between the two processes. In this work, we explore and compare different functions $h(b_i, t)$.

2.3 Estimation of joint model

Shared random-effect models can be estimated within the maximum likelihood framework. Let θ be the whole vector of parameters defined in (1) and (2). Using the assumption of independence between the longitudinal and the survival processes conditionally to the random effects, the joint log-likelihood of the observed data is:

$$\begin{aligned} l\left(\theta\right) &= \log\left[\prod_{i=1}^{N}\left(\int_{b_i} f_Y\left(Y_i|b_i;\theta\right) f_T\left(T_i|b_i;\theta\right) f_b\left(b_i;\theta\right) db_i\right)\right] \\ &= \sum_{i=1}^{N} \log\left(\int_{b_i} f_Y\left(Y_i|b_i;\theta\right) \lambda_i\left(T_i|b_i;\theta\right)^{E_i} S_i\left(T_i|b_i;\theta\right) f_b\left(b_i;\theta\right) db_i\right) \end{aligned}$$

where f_b and f_Y are multivariate Gaussian density functions of b and Y; $\lambda_i(T_i|b_i;\theta)$ is the hazard function defined in (2) and taken at the observed time T_i

($T_i = \min(T_i^*, C_i)$, with T_i^* the actual time-to-event and C_i the censoring time, and $E_i = \mathbb{1}_{\{T_i^* \leq C_i\}}$ the indicator of event); $S_i(T_i|b_i;\theta) = e^{-\int_0^{T_i} \lambda_i(t|b_i;\theta)dt}$ is the derived survival function. This joint likelihood involves two integrals that are usually approximated by Gauss-Hermite and Gauss-Kronrod quadratures (Rizopoulos, (2010)).

3 Dynamic predictions

3.1 Dynamic prognostic tools

Individual dynamic predictions of the event can be derived from a joint model. They consist in the individual predicted probability of event between times s and $s+t$ given the biomarker data $Y_i^{(s)} = \{Y_i(u), u \leq s\}$ collected until the time s of prediction (Rizopoulos, (2011); Proust-Lima and Taylor, (2009); Proust-Lima et al., (2012)):

$$\begin{aligned} p_i(s,t;\theta) &= \mathbb{P}\left(T_i \leq s+t | T_i \geq s, Y_i^{(s)}, X_i; \theta\right) \\ &= 1 - \frac{\int_{b_i} f_{Y^{(s)}}\left(Y_i^{(s)}|b_i, X_i;\theta\right) S_i\left(s+t|b_i, X_i;\theta\right) f_b\left(b_i;\theta\right) db_i}{\int_{b_i} f_{Y^{(s)}}\left(Y_i^{(s)}|b_i, X_i;\theta\right) S_i\left(s|b_i, X_i;\theta\right) f_b\left(b_i;\theta\right) db_i} \end{aligned}$$

From this, dynamic prognostic tools can be constructed: at a time s of prediction, for a new subject i with biomarker history $Y_i^{(s)}$ and covariates X_i, a dynamic prognostic tool can be computed as the predicted probability of event defined in (3) and computed with the parameter estimates $\hat{\theta}$ obtained on a large dataset. An alternative is to use a Monte-Carlo method to approximate the distribution of the predicted probability of event. In other words, a large set of $\theta^{(b)}$ $(b = 1, ..., B)$ can be generated from the asymptotic distribution of the parameter estimates $\mathcal{N}\left(\hat{\theta}, \widehat{\mathrm{V}(\hat{\theta})}\right)$, with $\widehat{\mathrm{V}(\hat{\theta})}$ the variance of the parameter estimates; and $p_i(s,t;\theta^{(b)})$ can be computed. The median of $p_i(s,t;\theta^{(b)})$ gives a point estimate over the B draws while the 2.5% and 97.5% percentiles give the 95% confidence bands (Proust-Lima et al., (2012)).

3.2 Evaluation of the predictive accuracy

To evaluate the predictive accuracy of dynamic prognostic tools, we used a measure from the information theory, the expectation prognosis observed cross-entropy (EPOCE) (Commenges et al., (2012); Proust-Lima et al., (2012)). It is defined as $E\left[-\log\left(f_{T|Y^{(s)},T^*\geq s}(T)\right)|T^* \geq s\right]$ where $f_{T|Y^{(s)},T^*\geq s}$ is the conditional density of the time-to-event given the history of the marker until the time of prediction s.
The EPOCE can be estimated on external data as well as the data used to estimate the model thanks to an approximated cross-validated estimator.

4 Application

The application aimed at illustrating how shared random-effects models could be implemented and evaluated in practice on a real dataset. We used the data from a cohort of patients followed-up after a localized prostate cancer treated by radiation therapy (N=459 subjects with 74 clinical recurrences). We considered a series of joint models to predict the risk of clinical recurrence according to the standard prognostic factors known at diagnosis and the dynamics of the prostate Specific Antigen (PSA) which is the biomarker of progression of prostate cancer. The joint models only differed in the way the dependence between the dynamics of PSA and the risk of clinical recurrence was defined through $h(b_i, t)$. We considered alternately functions including only random effects of the mixed model, or the current level of PSA and/or the current slope of PSA, or even non-linear functions of the current level of PSA to investigate the non-linearity of its effect on the risk of clinical recurrence. These models were compared in terms of goodness-of-fit and adequation to the joint model assumptions (especially the loglinearity assumption in the survival model) but also in terms of predictive accuracy using the expected prognostic cross-entropy to assess their ability to predict the risk of clinical recurrence.

Acknowledgments: Special Thanks to the French National Institute of Cancer INCa for financing the projet PREDYC and Jeremy M.G. Taylor for providing the dataset.

References

Commenges, D., Liquet, B., and Proust-Lima, C. (2012). *Choice of prognostic estimators in joint models by estimating differences of expected conditional Kullback-Leibler risks.* Biometrics, **68(2)**, 380 – 7.

Proust-Lima, C., Séne, M., Taylor, J.M.G., and Jacqmin-Gadda, H. (2012). *Joint latent class models for longitudinal and time-to-event data: A review.* Statistical methods in medical research, **0(0)**, 1 – 17.

Proust-Lima, C., and Taylor, J.M.G. (2009). *Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach.* Biostatistics (Oxford, England), **10(3)**, 535 – 49.

Proust-Lima, C., Taylor, J.M.G., Scott, W., Ankerst, D., Liu, N., Kestin, L., Bae K., and Sandler, H. (2008). *Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts.* International Journal of Radiation Oncology Biology Physics.

Rizopoulos, D. (2010). *JM : An R Package for the Joint Modelling of longitudinal and time-to-event data.* Journal Of Statistical Software, **35(9)**.

Rizopoulos, D. (2011). *Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data.* Biometrics, **67(3)**, 819 – 29.

Wulfsohn, M.S., and Tsiatis, A.A. (1997). *A joint model of survival and longitudinal data measured with error.* Biometrics, **53**, 330 – 339.

Estimating prediction error in mixed models

Benjamin Saefken[1], Sonja Greven[2], Thomas Kneib[1]

[1] Georg-August-University Goettingen, Germany
[2] Ludwig-Maximilians University Munich, Germany

E-mail for correspondence: bsaefke@uni-goettingen.de

Abstract: In mixed models, there are two possible perspectives on the prediction error. The prediction may either be based on data from new unobserved clusters or on data from already observed clusters. This corresponds to the use of the marginal and the conditional likelihood for the prediction error measurement respectively. Especially when the focus is on the link between mixed models and semiparametric regression, the conditional prediction error is the appropriate approach. When choosing among different models this leads to the conditional Akaike information criterion (cAIC). For Gaussian responses the resulting criterion is observable and analytically accessible. For exponential family distributions, we derive similar asymptotic measures. They allow for cAIC based model choice and variable selection in generalized linear mixed models. We also give an intuitive explanation of the model choice behaviour of the resulting criterion.

Keywords: Conditional AIC; Deviance error; Variable selection; Covariance penalties; Random effect

1 Marginal and conditional perspective on generalized mixed models

A generalized linear mixed model

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + z_i'\boldsymbol{u} \tag{1}$$

for responses $y_1, \ldots, y_n$ from some exponential family distribution with fixed effects $\boldsymbol{\beta}$, random effects $\boldsymbol{u}$ and link-function $g()$ is considered. The random effects $\boldsymbol{u}$ are normally distributed with mean zero and covariance matrix $\boldsymbol{G}(\boldsymbol{\tau}^2)$. In this setting, the focus may either be on the marginal distribution of the responses $\boldsymbol{y}$ or on the distribution conditioned on the random effects $\boldsymbol{u}$, respectively.

For clustered data the model (1) can be specified as

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i$$

with independent random effects $\boldsymbol{u}_i \sim \mathcal{N}(0, \boldsymbol{G}_i(\boldsymbol{\tau}_i^2))$ for each i, where $i = 1, \ldots, m$ indicates the i-th cluster, $\boldsymbol{y}_i$ is the vector of n_i responses from the

i-th cluster and $\boldsymbol{\beta}$ are the fixed effects. If the main interest is in the overall effects $\boldsymbol{\beta}$ and the random effects model the within cluster correlation of the responses, the marginal perspective is appropriate. Then the marginal log-likelihood

$$\log f(y_i|\boldsymbol{\beta}) = \log \int f(y_i|\boldsymbol{\beta}, \boldsymbol{u}) p(\boldsymbol{u}) d\boldsymbol{u}$$

is obtained by integrating out the random effects.

If on the other hand the focus is on the random effects, an approach based on the log-likelihood conditioned on the random effects, i.e. the conditional log-likelihood

$$\log f(y_i|\boldsymbol{\beta}, \boldsymbol{u})$$

is suitable. In this case the random effects act as normal parameters with regularized estimation due to a penalty term induced by the covariance structure of the random effects. For example in penalized regression the random effects are used as a tool to model the penalized parameters, see for example Wood (2006).

2 Marginal and conditional error measurement

In generalized regression based on maximum likelihood estimation, the apparent error is measured by the deviance error

$$err = -2 \sum \log f(y_i|\hat{\boldsymbol{\beta}}(y_i)) + C$$

with a constant C not depending on y_i. The constant C is the log-likelihood of the saturated model and stays the same for different models and is therefore ignored when choosing among competing models. This error term is too optimistic to predict future values. Therefore the quantity of interest is the expected prediction error to new observations

$$Err = -2\mathbb{E}_z \left(\sum \log f(z|\hat{\boldsymbol{\beta}}(y_i)) \right).$$

Efron (2004) showed that for any exponential family with corresponding natural parameter θ

$$\mathbb{E}\,(Err) = \mathbb{E}\left[err + 2\mathrm{Cov}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{y}) \right]. \tag{2}$$

In generalized linear models, the approximation $\mathrm{Cov}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{y}) \approx p$ is used with p being the number of parameters in the model. The approximation is exact in case of normal responses. The resulting criterion is Akaike's information criterion, see Efron (2004).

In mixed models, the main question that arises in prediction error measurement is, if the prediction should be based on the marginal or conditional

likelihood. Greven and Kneib (2010) show that equation (2) does not hold in linear mixed models with Gaussian responses if the marginal log-likelihood is used and $\text{Cov}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{y}) = p + q + 1$ is assumed, with p the number of fixed parameters and q the number of variance parameters of the random effects. That criterion is called marginal Akaike information criterion.

Vaida and Blanchard (2005) introduced the conditional Akaike information criterion for Gaussian responses using the conditional likelihood in equation (2). Since $\text{Cov}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{y})$ is not observable, they propose to use the trace of the hat matrix $\boldsymbol{H}$ mapping the data vector $\boldsymbol{y}$ on the fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}}$ as covariance penalty. In case of known random effects variance parameters $\boldsymbol{\tau}^2$, the mapping is linear and

$$\text{Cov}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{y}) = \text{tr}(\boldsymbol{H}).$$

In most applications though $\boldsymbol{\tau}^2$ will not be known and plugging in a consistent estimator $\hat{\boldsymbol{\tau}}^2$ induces a bias that does not disappear asymptotically, see Greven and Kneib (2010). Stein (1981) showed that

$$\text{Cov}(\hat{y}(y), y) = \sigma^2 \mathbb{E}\left(\frac{\partial \hat{y}}{\partial y}\right), \tag{3}$$

for normal y with variance σ^2. This was applied to the conditional Akaike information criterion by Liang et. al (2008) resulting in

$$cAIC = -2 \log f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}(\boldsymbol{y}), \hat{\boldsymbol{u}}(\boldsymbol{y})) + 2 \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i}. \tag{4}$$

Thus the log-likelihood measures the fit of the model to the data while the covariance penalty, i.e. the sum over the sensitivity of the fitted values $\hat{y}_i(\boldsymbol{y})$ with respect to small changes in the response values y_i, measures the stability of the fitting process. This formulation of the covariance penalty has been made analytically accessible avoiding high computational costs and imprecise numerical approximations by Greven and Kneib (2010).

3 Covariance penalties in generalized mixed models

For Gaussian responses, the conditional expected prediction error (2), i.e. conditioned on the random effects, can be evaluated by use of the Stein formula (3) leading to the conditional Akaike information (4). Response variables from some other exponential family distribution would need formulas similar to the Steinian (3) in order to evaluate the covariance penalty. If the response variable is Poisson distributed then the Chen-Stein formula, due to Chen (1975),

$$\text{Cov}(\hat{\theta}(y), y) = \mathbb{E}\left(y(\hat{\theta}(y) - \hat{\theta}(y-1))\right),$$

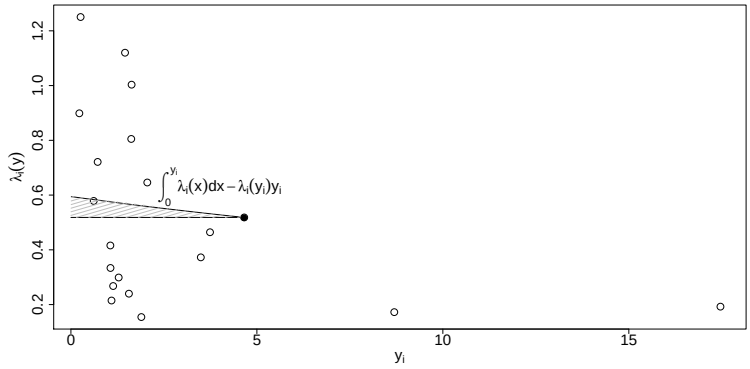


FIGURE 1. i-th covariance of exponential data with $\lambda_i = -\hat{\theta}_i$.

with $y\hat{\theta}(y-1) = 0$ if $y = 0$ by convention, where $\hat{\theta}$ is the estimated natural parameter of the exponential family, gives an observable quantity that allows to evaluate the conditional expected prediction error (2). Lian (2012) proposed the resulting criterion

$$cAIC = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}(\boldsymbol{y}), \hat{\boldsymbol{u}}(\boldsymbol{y})) + 2\sum_{i=1}^{n} y_i(\hat{\theta}_i(y_i) - \hat{\theta}_i(y_i - 1))$$

as the conditional Akaike information criterion for Poisson responses. Similar to the covariance penalty term in the Gaussian criterion (4), the covariance penalty measures the weighted sensitivity of the estimated natural parameter to a change in the response value.

For exponentially distributed responses, the covariance penalty in (2) can also take an observable form

$$\mathrm{Cov}(\hat{\theta}(y), y) = \mathbb{E}\left(y\hat{\theta}(y) - \int_0^y \hat{\theta}(x)dx\right).$$

So both of the covariance penalties, for Poisson and exponential data are also measures of the stability of the fitting process, since they are derived by calculating the change of the estimated natural parameter if other response values would have been observed. In case of exponential data this can be seen in Figure 1.
Therefore it is also possible to write an observable cAIC for exponentially distributed responses as

$$cAIC = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}(\boldsymbol{y}), \hat{u}(y)) + 2\sum_{i=1}^{n}\left(\int_0^{y_i} \hat{\lambda}_i(x)dx - y_i\hat{\lambda}_i(y_i)\right),$$

where $\hat{\lambda} = -\hat{\theta}_i$ is the estimated rate.

Since there is no way to rewrite the covariance penalties in (2) such that they are observable for all exponential family distributions other measures need to be used for the remaining exponential family distributions. In case of Bernoulli responses, i.e. binary data, the covariance penalty may be rewritten as

$$\mathrm{Cov}(\hat{\theta}(y), y) = \mathbb{E}\left(\mu(1-\mu)(\hat{\theta}(1) - \hat{\theta}(0))\right).$$

Since the true value of μ is not available it can be replaced by a consistent estimator $\hat{\mu}$. With this covariance penalty and the help of equation (2) a conditional Akaike information criterion for binomial responses can be given by

$$cAIC = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}(\boldsymbol{y}), \hat{\boldsymbol{u}}(\boldsymbol{y})) + 2\sum_{i=1}^{n} \hat{\mu}_i(1-\hat{\mu}_i)(\hat{\theta}_i(y_i) - \hat{\theta}_i(y_i - 1)). \quad (5)$$

Similarly the conditional Akaike information criterion for gamma distributed responses can be approximated by

$$cAIC = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}(\boldsymbol{y}), \hat{\boldsymbol{u}}(\boldsymbol{y})) + 2\sum_{i=1}^{n} \frac{\partial \hat{\mu}_i(y_i)}{\partial y_i}.$$

4 Model choice behaviour

When choosing the model with the lowest estimated prediction error an important question that arises in mixed models is whether or not the model should include random effects. Greven and Kneib (2010) show that the marginal AIC is biased when choosing between a model including random effects and a model without random effects. Therefore we investigate the behaviour of different criteria in the selection of random effects in a simulation study. 1000 datasets were generated from a random intercept model with binary responses $y_{ij} \sim \mathcal{B}(1, \pi_{ij})$ with

$$\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 x_j + u_i)}{1 + \exp(\beta_0 + \beta_1 x_j + u_i)};\ i = 1, \ldots, m;\ j = 1, \ldots, n_i,$$

where $u_i \sim \mathcal{N}(0, \tau^2)$, $\beta_0 = 0.1$, $\beta_1 = 0.2$, $x_j = j$ and number of clusters $m = 5, 10$ and the cluster sizes are $n_i = 5, 10$. For $\tau^2 = 0,\ 0.1, \ldots,\ 1.8$ the different AIC values: the proposed conditional AIC (5), the marginal AIC, a asymptotic cAIC proposed by Yu and Yau (2012) and an asymptotically true cAIC, are compared with the value of the AIC without any random effects , i.e. with covariance penalty assumed to be $\mathrm{Cov}(\hat{\boldsymbol{\theta}}(\boldsymbol{y}), \boldsymbol{y}) = p$. The proportion of sets where the complex model is favored are plotted in Figure 2. The marginal AIC includes random effects only if the random effects variance is high. The proposed cAIC (5) offers better model choice behaviour when comparing a model including random effects ($\tau^2 > 0$) and a model without random effects ($\tau^2 = 0$).

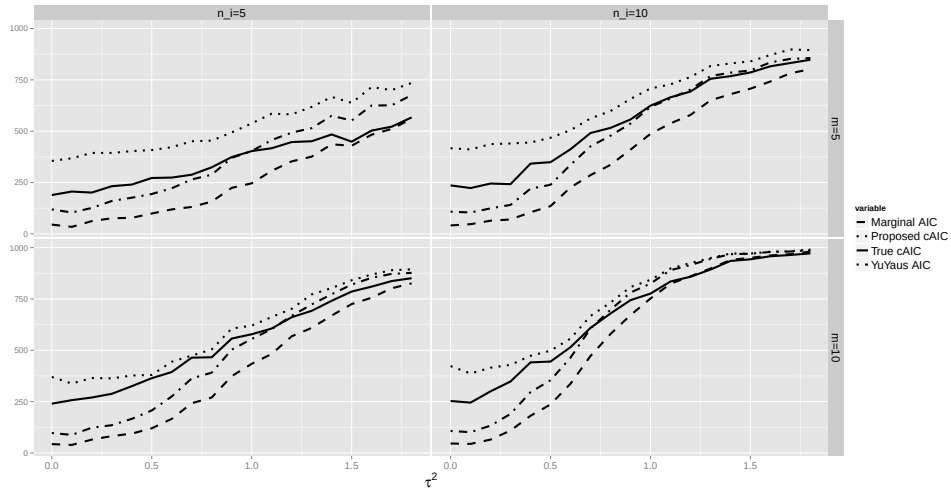


FIGURE 2. Proportion of simulation replications where the more complex model was favored by the AIC.

References

Chen, L.H.Y. (1975). Poisson approximation for dependent trials. *Annals of Probability* , **3**, 534 – 545.

Efron, B. (2004). The Estimation of Prediction error Covariance Penalties and Cross-Validation. *Journal of the American Statistical Association*, **99**, 619 – 642.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, **97**, 773 – 789.

Lian, H. (2012). A note on conditional akaike information for poisson regression with random effects. *Electronic Journal of Statistics*, **6**, 1 – 9.

Liang, H., Wu, H. and Zou, G. (2008). A note on conditional aic for linear mixed-effects models. *Biometrika*, **95**, 773 – 778.

Vaida F. and Blanchard, S. (2005). Conditional Akaike information in mixed effects models. *Biometrika*, **92**, 351 – 370.

Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. *Chapman & Hall/CRC*

Yu D. and Yau, K.K.W. (2012). Conditional Akaike information criterion for generalized linear mixed models. *Computational Statistics & Data Analysis*, **56**, 629 – 644.

A two-part model using quantile regression under a Bayesian perspective

Bruno Santos[1], Heleno Bolfarine[1]

[1] Instituto de Matemática e Estatística - Universidade de São Paulo, Brazil

E-mail for correspondence: bramos@ime.usp.br

Abstract: We develop an extension of the two-part model proposed by Cragg (1971) considering the asymmetric Laplace distribution for the continuous density, proposing a quantile regression analysis in the process, within a Bayesian approach. We also consider the case where there could be a zero inflation process while estimating a Bayesian tobit quantile regression, and by the imputation of the latent variable indicating whether a zero observation belongs to a point mass or the continuous distribution, we are able to obtain a generalization of our two-part model. We illustrate our method in a known data set in the field of econometrics.

Keywords: Bayesian quantile regression; Two-part model; MCMC.

1 Introduction

Quantile regression, introduced by Koenker (1978), is now widely recognized as a major tool to address the relationship between the response variable, Y, and the explanatory variables, X, not only in a central location, which can be studied with the median for symmetric distributions, but in others positions of the conditional distribution of Y given X. This model was first addressed in a Bayesian setting with Yu and Moyeed (2001) and later this idea was improved by Kozumi and Kobayashi (2011), who developed an effective Gibbs sampler using the formulation of the asymmetric Laplace distribution as a mixture of normal and exponential distributions and proposed an extension of their method to the Bayesian tobit quantile regression (BTQR) model. In this article, we introduce the idea of the zero inflated Bayesian tobit quantile regression (ZIBTQR) as an extension of the BTQR combined with the two-part model proposed by Cragg (1971), when there is a point mass density at zero beyond the zeros coming from the censoring of the asymmetric Laplace distribution.
This paper is organized as follows. We adapt the two-part model to the Bayesian quantile regression framework in Section 2. In Section 3, we formulate the ZIBTQR, as a extension of the BTQR model, using the two-part model approach. In the Section 4, we illustrate our method with data about

work supply of married women in the United Stated, first studied by Mroz (1971), using the tobit model. We finish with our final remarks in Section 5.

2 Two-part model

We consider the following linear model

$$y_i = x_i^{\mathrm{T}}\beta_\tau + \epsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

where ϵ_i is distributed according to the asymmetric Laplace distribution (Yu and Zhang, 2005), with its τ-th quantile equal to zero.
The implementation of the Gibbs sampler to conduct the Bayesian inference in quantile regression is simplified by the the representation of the asymmetric Laplace distribution as a mixture of the normal and the exponential distributions (Kotz et al., 2001). Considering a scale parameter σ for the model, we can rewrite the model in (1) as

$$y_i = x_i^{\mathrm{T}}\beta_\tau + \theta v_i + \psi\sqrt{\sigma v_i}u_i, \tag{2}$$

where $\theta = (1-2\tau)/(\tau(1-\tau))$, $\psi = 2/(\tau(1-\tau))$, v_i and u_i are mutually independent and distributed as exponential with mean σ and standard normal, respectively.
If we take into consideration the possibility of y_i being modeled by a mixture of two distributions, e.g., an asymmetric Laplace distribution and a point mass distribution at zero, we can use the two-part model introduced by Cragg (1971). So we can write the density of y_i as

$$g(y_i) = p_i I_i + (1-p_i)f(y_i), \quad i = 1, \ldots, n, \tag{3}$$

where $I_i = 1$ if $y_i = 0$ and zero otherwise, $p_i = P[y_i = 0]$ and $f(y_i)$ is the asymmetric Laplace density function.
We can also incorporate covariates to explain p_i by making $p_i = H(z_i^{\mathrm{T}}\gamma_\tau)$, where $H(.)$ is a link function, that could be, for instance, the normal cumulative distribution function (cdf), producing the probit model or the logistic cdf, producing the logistic model. The set of variables in this case, Z, can be either different or the same as that used to infer about the continuous density.
Let C denote the set of censored observations and D denote the set of uncensored observations. If we consider the model in (2) for the continuous part, we have that $Y_i \mid v_i \sim N(x_i^{\mathrm{T}}\beta_\tau + \theta v_i, \psi^2\sigma v_i)$ and $v_i \sim \mathrm{Exp}(\sigma)$. So we can write the likelihood function in the following way:

$$f(y, v \mid \beta_\tau, \gamma, \sigma) = \prod_{i\in C} H(z_i^{\mathrm{T}}\gamma_\tau) \prod_{i\in D} (1 - H(z_i^{\mathrm{T}}\gamma_\tau))\, f(y_i \mid v_i) f(v_i),$$

where

$$f(y_i \mid v_i, \beta_\tau, \gamma, \sigma) \propto \left((v_i\sigma)^{-1/2}\right) \exp\left\{-\frac{(y_i - x_i^{\mathrm{T}}\beta_\tau - \theta v_i)^2}{2\psi^2\sigma v_i}\right\}$$

and $f(v_i \mid \beta_\tau, \gamma, \sigma) = \sigma^{-1}\exp\{-v_i/\sigma\}$.
Next, we use the Gibbs sampler developed in Kozumi and Kobaysahi (2011), with a Metropolis-Hastings step similar to the algorithm considered in Luo et al. (2012), used to develop a Bayesian method for quantile regression model for longitudinal data. Completing the model specification, we assume the priors for $\beta_\tau \sim N(b_0, B_0)$, $\gamma_\tau \sim N(g_0, G_0)$ and $\sigma \sim IG(n_0/2, s_0/2)$, where $IG(g_1, g_2)$ denotes an inverse Gamma distribution with parameters equal to g_1 and g_2. We also assume that all hyperparameters are known. We can write the posterior distribution of the parameters $(\beta_\tau, \gamma_\tau, \sigma)$ as

$$\begin{aligned}\pi(\beta_\tau, \gamma_\tau, \sigma \mid y, v) &\propto f(y, v \mid \beta_\tau, \gamma)\pi(\beta_\tau)\pi(\gamma_\tau)\pi(\sigma)\\ &\propto f(y \mid v, \beta_\tau, \gamma)f(v \mid \beta_\tau, \gamma)\pi(\beta_\tau)\pi(\gamma_\tau)\pi(\sigma).\end{aligned}$$

We obtain that the full conditional posteriors of all parameters and latent variables are given by

$$\pi(\gamma_\tau \mid y, v, \beta_\tau, \sigma) \propto \prod_{i\in C} H(z_i^{\mathrm{T}}\gamma_\tau) \prod_{i\in D} (1 - H(z_i^{\mathrm{T}}\gamma_\tau)) \exp\left\{-\frac{1}{2}(\gamma_\tau - g_0)^{\mathrm{T}} G_0^{-1}(\gamma_\tau - g_0)\right\}$$

$$\begin{aligned}\beta_\tau \mid y, v, \gamma_\tau, \sigma &\sim N(b_1, B_1)\\ v_i \mid y, \beta_\tau, \gamma_\tau, \sigma &\sim GIG(1/2, \hat{\delta}_i, \hat{\xi}_i)\\ \sigma \mid y, v, \beta_\tau, \gamma_\tau &\sim IG(\tilde{n}/2, \tilde{s}/2)\end{aligned}$$

where $b_1 = B_1(X^{\mathrm{T}}W(y - \theta v) + B_0^{-1}b_0)$, $B_1 = \left(X^{\mathrm{T}}WX + B_0^{-1}\right)^{-1}$, W is a diagonal matrix with entries $v_i/(\psi^2\sigma)$, $i \in D$, $\hat{\delta}_i = (y_i - x_i^{\mathrm{T}}\beta_\tau)^2/\psi^2\sigma$, $\hat{\xi}_i = 2/\sigma + \theta^2/\psi^2\sigma$, $\tilde{n} = n_0 + 3n$, $\tilde{s} = s_0 + 2\sum v_i + \sum\left[(y_i - x_i^{\mathrm{T}}\beta_\tau - \theta v_i)^2/\psi^2\sigma\right]$, with the sums defined in D and n is the total number of observations in D. So we can define a Metropolis-Hastings within Gibbs sampler algorithm as follows:

1. Define initial values to $\gamma_\tau^{(0)}, \beta_\tau^{(0)}, \sigma^{(0)}$
2. Using a multivariate normal as the proposal density for a random walk Metropolis-Hastings algorithm, at the ith step of the algorithm, draw $\gamma_\tau^{(i)}$ from $N(\gamma_\tau^{(i-1)}, \sigma_\gamma\Omega_\gamma)$ and accept $\gamma_\tau^{(i)}$ with probability

$$\min\left\{1, \frac{\pi(\gamma_\tau^{(i)} \mid y, v, \beta_\tau, \sigma)}{\pi(\gamma_\tau^{(i-1)} \mid y, v, \beta_\tau, \sigma)}\right\}$$

3. Update all others parameters and latent variables according to theirs conditional posteriors presented before.

It is important to tune σ_γ to reach an overall acceptance probability between 0.15 and 0.50. Following the approach by Luo et al. (2012), we take Ω_γ to be the identity matrix. We note, however, that the validity of this assumption should be verified with some simulation studies in a later work. If we consider $H(.)$ to be the cdf, $\Phi(.)$, of the standard normal distribution, then it is possible to define a complete Gibbs sampler, without the need of a Metropolis-Hastings step, using the ideas of Albert and Chib (1993).

3 Zero inflated Bayesian tobit quantile regression model

Moulton and Halsey (1995) proposed an extension of Cragg's model, considering that an observed zero, or a lower detection limit (could be different from zero), can be either from the point mass distribution or from the continuous distribution, being in the latter case a censored observation. So we should rewrite the density in (3) as

$$g(y_i) = (p_i + (1 - p_i)F(T))I_i + (1 - p_i)f(y_i),$$

where T is the lower detection limit, e.g., zero, and $F(.)$ is a cdf of the continuous part. For quantile regression models, we use the cdf of the asymmetric Laplace distribution.

For the censored observations, one possible approach is the one suggested by Chao (1998), where the author working with censored observations in a survival analysis setup and with no possibility to observe whether one person on the study was cured or not, i.e., cure was a latent variable, decided to impute this variable by the complementary probability of the person being not cured given the explanatory variables of the problem. The same can be done in our model using the probability p_i and imputing an auxiliary variable, so there is a two-part model and we have a problem like the one stated in Section 2, in which the observations come from the point mass density or the continuous density.

There is no reason to explicitly describe the algorithm for this method, except for the statement that it should be added a step in the suggested algorithm in the previous section. A binary auxiliary variable should be imputed based on the probability p_i to divide the censored values between observations of the point mass distributions and the continuous distribution. After this imputation, the algorithm should continue almost exactly as explained in Section 2, with the only difference being that sampling of the censored observations should consider a truncated normal distribution, as described by Kozumi and Kobayashi (2011). About the imputation proposal, it should be noted that at every step, this process is repeated so it becomes more precise with the updates of the model of probabilities p_i's.

One important remark about this model is that it is a generalization of the famous tobit model, proposed by Tobin (1958), widely known in the field

of econometrics, but in its Bayesian quantile regression version. This new model can be seen as a way of checking if there is a zero inflation process in the censored observations used to estimate the Bayesian tobit quantile regression model.

4 Application

In order to illustrate our method, we examine the data set from Mroz (1987), about female labor supply. Consider hours worked during the year as a variable that measures labor supply, this data set has 753 married women, from which 325 worked zero hours in 1975, the year that the survey was conducted. This variable is then considered left censored and an usual tobit regression analysis is considered. Kozumi and Kobayashi (2011) performed a Bayesian tobit quantile regression analysis with this data set. In our application, we defined as explanatory variables for both parts of the model: years of education, income which is not due to the wife, age, number of children under 6 years old and number of children over 6 years old.

In our example, we tuned σ_τ so we could have an acceptance rate for the Metropolis-Hastings algorithm between 0.40 and 0.50. Our inference is based in a MCMC sample of 10.000 observations obtained after taking 5.000 observations as burn-in.

We found that the hypothesis of a point mass density in the zero should not be discarded, since the corresponding parameter of age is significant to model the probability. In this case, we believe that there is a zero inflation process that should be taken into account in the analysis of the Bayesian tobit quantile regression model. We also should add that the analysis of the continuous part still provides interesting results, but we decided not to give further details here.

5 Final remarks

In this paper, we develop a new two-part model using quantile regression, in a Bayesian setting. We propose a Metropolis-Hastings within Gibbs sampling algorithm to estimate the parameters of interest. We also extended our model to consider the situation where there is a zero inflation process in the Bayesian tobit quantile regression model. Using a imputation step in the algorithm, we are able to transform this situation into a two-part model. We illustrate our method with a known data set in the field of econometrics about female labor supply.

The code to implement this two-part model was implemented in the statistical software R and is available upon request from the first author. These programs are intended to be released in the form of a package in www.cran.r-project.org.

Acknowledgments: Special thanks to FAPESP for financial support of the first author and CNPq for financial support of the second author.

References

Albert, J.H. and Chib, S. (1993). Bayesian analysis ob bynary and polychtomous response data. *Journal of the American Statistical Association*, **88**, 669 – 679.

Chao, E.C. (1998). Gibbs Sampling for Long-Term Survival Data with Competing Risks. *Biometrics*, **54**, 350 – 366.

Cragg, J. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, **39**, 829 – 844.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33 – 50.

Kotz, S., Kozubowski, T.J. and Podgórski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Boston: Birkhauser.

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, **81**, 1565 – 1578.

Luo, Y., Lian, H. and Tian, M. (2012). Bayesian quantile regression for longitudinal data models. *Journal of Statistical Computation and Simulation*, **82**, 1635 – 1649.

Moulton, L. and Halsey, N.A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, **51**, 1570 – 1578.

Mroz, T. (1987). The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. *Econometrica*, **55**, 765 – 799.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24 – 36.

Yu, K. and Moyeed, J. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, **54**, 434 – 447.

Yu, K. and Zhang, J. (2005). A Three-Parameter Asymmetric Laplace Distribution and Its Extension. *Communications in Statistics - Theory and Methods*, **34**, 1867–1879.

DIF-LASSO: Differential Item Functioning in Rasch Models

Gunther Schauberger[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-University Munich

E-mail for correspondence: gunther.schauberger@stat.uni-muenchen.de

Abstract: The Rasch model is the most widely used model in assessment tests. It assumes that the probability of solving an item is determined by the latent ability of the person and the difficulty of the item. A problem in assessment tests is differential item functioning (DIF), which means that the probability of solving an item may depend on the membership to an ethnic, racial or gender group. A general model is proposed that is able to model DIF depending on a vector of covariates. Regularized estimation methods are proposed that solve the high dimensional estimation problem.

Keywords: Rasch model, differential item functioning, DIF, Group Lasso, DIF-lasso

1 Differential Item Functioning Model

The Rasch Model (Rasch, 1960) is a commonly used model for item response data. For binary test scores, it models the probability for a participant to score on an item by estimating both a parameter for the person ability and a parameter for the item difficulty. In the case of P persons and I items, the Rasch Model has the form

$$P(Y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} \quad p = 1, \dots, P \quad , i = 1, \dots, I, \qquad (1)$$

where $Y_{pi} \in \{0, 1\}$ represents the score of person p on item i. Thus, person parameters $\theta_p, \quad p = 1, \dots, P$, and item parameters $\beta_i, \quad i = 1, \dots, I$, have to be estimated.
An alternative representation of model (1) uses logits and is given by

$$\log\left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)}\right) = \theta_p - \beta_i. \qquad (2)$$

As model (2) is not identifiable, a restriction on the parameters has to be applied. For simplicity we use $\theta_P = 0$.

In item response models, DIF appears if an item has different difficulties depending on which person tries to solve the item. Therefore, DIF changes the item difficulty depending on covariates of the participants. This idea can be formalized by the so-called Differential Item Functioning Model (DIF Model)

$$\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) = \theta_p - (\beta_i + \boldsymbol{x}_p^T\boldsymbol{\gamma}_i), \tag{3}$$

where $\boldsymbol{x}_p = (x_{p1}, \ldots, x_{pm})$ denotes a person-specific covariate vector and m denotes the number of covariates. It is an extension of the Rasch Model (2) that allows for person-specific item difficulties $\beta_i + \boldsymbol{x}_p^T\boldsymbol{\gamma}_i$. The item-specific parameters $\boldsymbol{\gamma}_i^T = (\gamma_{i1}, \ldots, \gamma_{im})$ represent the change of the item difficulty depending on the covariate values. Therefore, $m \cdot I$ parameters additional to the regular Rasch Model have to be estimated.

Since the model contains a large number of parameters, maximum likelihood (ML) estimates will only exist in situations where both the number of covariates and the number of items are very small. To overcome th e estimation problem we propose to use the Group Lasso penalty (Yuan and Lin, 2006). The general assumption for our model is, that only a part of the items induces DIF and only for these items item-specific parameters $\boldsymbol{\gamma}_i$ have to be estimated. Therefore, the objective is to perform variable selection on the item-specific parameters. The advantage of the approach is that it automatically detects which items induce DIF.

2 Estimation Procedure

2.1 The DIF Model as a Generalized Linear Model

The Rasch Model (1) and also the DIF Model (3) can be embedded into the framework of Generalized Linear Models (GLMs).
Let the data be given by $(Y_{pi}, \boldsymbol{x}_p)$, $p = 1, \ldots, P$, $i = 1, \ldots, I$. Additionally, we use the notation $\mathbf{1}_{P(p)}^T = (0, \ldots, 0, 1, 0, \ldots, 0)$ and $\mathbf{1}_{I(i)}^T = (0, \ldots, 0, 1, 0, \ldots, 0)$, where $\mathbf{1}_{P(p)}$ and $\mathbf{1}_{I(i)}$ have lengths $P-1$ and I and have the value 1 at positions p and i, respectively. Thus, model (3) can be represented as

$$\begin{aligned}\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) &= \theta_p - \beta_i - \boldsymbol{x}_p^T\boldsymbol{\gamma}_i \\ &= \mathbf{1}_{P(p)}^T\boldsymbol{\theta} - \mathbf{1}_{I(i)}^T\boldsymbol{\beta} - \boldsymbol{x}_p^T\boldsymbol{\gamma}_i = \boldsymbol{z}_{pi}^T\boldsymbol{\alpha}.\end{aligned} \tag{4}$$

Here, $\boldsymbol{\alpha}^T = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_I^T)$, $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_{P-1})$ and $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_I)$, denotes the complete parameter vector and

$$\boldsymbol{z}_{pi}^T = (\mathbf{1}_{P(p)}^T, -\mathbf{1}_{I(i)}^T, 0, \ldots, 0, -\boldsymbol{x}_p^T, 0, \ldots, 0)$$

denotes the design vector for the p-th person and the i-th item. In $\boldsymbol{z}_{pi}$, the component $-\boldsymbol{x}_p$ corresponds to the parameter $\boldsymbol{\gamma}_i$ in $\boldsymbol{\alpha}$. Therefore, (4) represents the structural component of a GLM for binary response with logit link.
Of course, also the regular Rasch Model can be represented in the GLM framework by

$$\begin{aligned} \log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) &= \theta_p - \beta_i \\ &= \mathbf{1}_{P(p)}^T\boldsymbol{\theta} - \mathbf{1}_{I(i)}^T\boldsymbol{\beta}, \end{aligned} \tag{5}$$

where the design vector and the parameter vector reduce to $(\mathbf{1}_{P(p)}, -\mathbf{1}_{I(i)})$ and $(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)$, respectively.

2.2 Penalized Estimation

In the following we show how model (3) can be estimated by maximizing the penalized likelihood

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}),$$

where $l(.)$ is the common log-likelihood of the model and $J(\boldsymbol{\alpha})$ is an appropriate penalty term. The *group lasso penalty for item differential functioning* (DIF-LASSO) that is propsed has the form

$$J(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_I^T) = \sum_{i=1}^{I} \|\boldsymbol{\gamma}_i\|,$$

where $\|\boldsymbol{\gamma}_i\| = (\gamma_{i1}^2 + \cdots + \gamma_{im}^2)^{1/2}$ is the L_2-norm of the parameters of the ith item with m denoting the length of the covariate vector. The penalty encourages sparsity in the sense that either $\hat{\boldsymbol{\gamma}}_i = \mathbf{0}$ or $\gamma_{is} \neq 0$ for $s = 1, \ldots, m$. Thus, the whole group of parameters collected in $\boldsymbol{\gamma}_i$ is shrunk simultaneously toward zero. The effect is that in a typical application only some of the parameters get estimates $\hat{\boldsymbol{\gamma}}_i \neq \mathbf{0}$. These correspond to items that show DIF.

3 Application

In the following, we present some results for applications of our method on simulated and real data sets. First, we consider a simulated data set for $P = 100$ persons, $I = 10$ items and $m = 3$ covariates. Figure 1 shows the path of the $\boldsymbol{\gamma}$-parameters for one simulated data set.
The dashed paths correspond to the three items containing DIF. They are in the model at the BIC-optimal path point whereas the non-DIF items

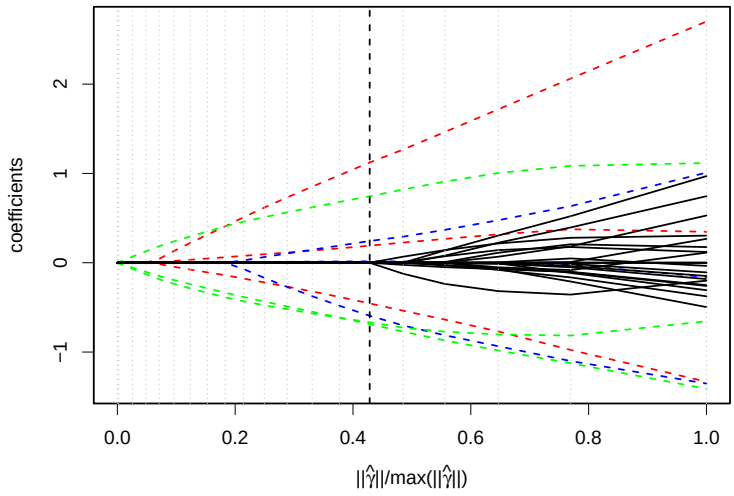


FIGURE 1. Coefficient paths for simulated data set; dashed paths represent DIF items; the bold vertical line represents the BIC-optimal path point

are excluded from the model. Therefore, the selection in this scenario is perfect.

We apply the method to an example that has first been considered by Strobl et al. (2010). It uses data from an online quiz for testing one's general knowledge conducted by the weekly German news magazine SPIEGEL. The 45 test questions were from five topics, politics, history, economy, culture, and natural sciences. We use the same sub sample as Strobl et al. (2010) consisting of 1075 university students from Bavaria, who had all been assigned a particular set of questions. The covariates that we included as potentially inducing DIF are gender, age, semester of university enrollment, an indicator for whether the student's university received elite status by the German excellence initiative (elite), and the frequency of accessing SPIEGEL's online magazine (spon).

Out of the 45 items, according to our analysis 16 showed DIF. Figure 2 shows the profile plot for the coefficient estimates $\hat{\gamma}_i$ of the items with DIF.

The highlighted paths represent the items with the highest DIF which are all items from the topic economy. The items with the highest DIF were:

- Zetsche: "Who is this?" (a picture of Dieter Zetsche, the CEO of the Daimler AG, maker of Mercedes cars, is shown).

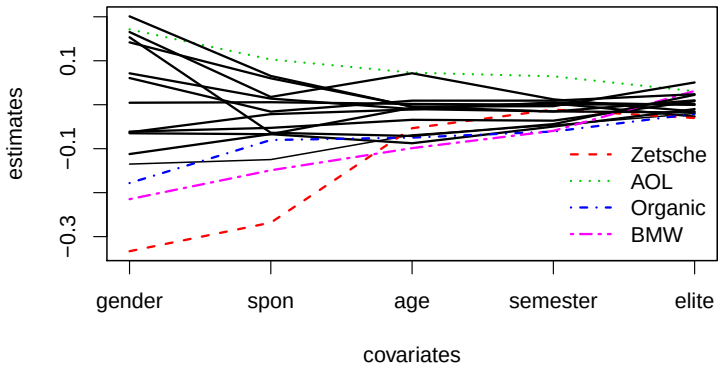


FIGURE 2. Profile plot for coefficient estimates of items with DIF

- AOL: "Which internet company took over the media group Time Warner?"
- Organic: "What is the meaning of the hexagonal organic logo?" (Synthetic pesticides are prohibited)
- BMW: "Which German company took over the British automobile manufacturers Rolls-Royce?"

It can be seen, that most of the DIF is induced by the covariates gender and spon.

References

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*, Danish Institute for Educational Research

Strobl, C. and Kopf, J. and Zeileis, A. (2010). A new method for detecting differential item functioning in the Rasch model, *Technical Report 92, Department of Statistics, LMU Munich.*

Tutz, G. and Schauberger, G. (2012). A Penalty Approach to Differential Item Functioning in Rasch Models, *Technical Report 134, Department of Statistics, LMU Munich.*

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, **68**, 49 – 67.

Modelling plant height data with scaled and shifted prototype curves

Sabine K. Schnabel[1], Paul H.C. Eilers[1,2], Fred A. van Eeuwijk[1]

[1] Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands
[2] Erasmus Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: sabine.schnabel@wur.nl

Abstract: In agricultural research phenotypic data are mainly collected through field and greenhouse experiments. Often a whole population of plants is monitored at different time points during the growing season. Here we are analyzing time series of plant height data in potato. The plant-specific data is modelled inspired by a model that was originally developed for a study of growth of children. We are using P-splines for a smooth curve and introduce a vertical as well as a horizontal shift per plant.

Keywords: growth modelling; smooth curve; shift; scale; P-splines

1 Introduction

Measurements on growing plants often show a characteristic shape: a monotonically increasing curve. In agricultural trials many (up to hundreds) of such curves are being collected for one genetic population. Data are collected either by hand or completely automatically. To combine such measurements with genomic data, meaningful summaries have to be developed. A number of methods have already been proposed to estimate good characteristics of plant growth curves. They include classical parametric approaches based on the logistic curve (Malosetti et al., 2006), a semi-parametric approach based on splines (Hurtado et al., 2012) and a survival analysis approach for phenotypic data series on an ordinal scale (Schnabel et al., 2010).
In this contribution we study models that assume that there is one prototype curve, which has been stretched and shifted on both the horizontal (time) and the vertical axes. The transformation parameters as well as the prototype curve itself are to be estimated from the data. In our example the response variable has been log-transformed. Therefore a shift along the vertical axis translates into a rescaling on the original response scale.
The inspiration for our model comes from a publication by Cole et al. (2010) based on work by Beath (2007). Cole and co-authors call their model

SITAR, which stands for *SuperImposition by Translation And Rotation*. This acronym is not obvious, as the rotation property is not immediately clear from the model. However, it shows an excellent fit to growth data of children as demonstrated in an example for height. It is relatively parsimonious, because the curve for any individual is summarized with only three parameters in addition to the spline coefficients common to all curves.
The model has one parameter for shifting along the vertical scale. Given the prototype curve it can be estimated by linear regression. However, the two other parameters occur in the argument of the curve and lead to a non-linear problem see (1). Beath used natural B-splines to model the prototype curve and non-linear mixed models to estimate the parameters. After some experiments with the software provided in his paper, we decided to start from scratch. We model the curves with P-splines. In addition because of the rather precise and detailed data, we drop the mixed model approach. A fixed model is sufficient. We also experiment with simplifications using less parameters and initially include only two parameters.
In the next section we present briefly the original model and explain our approach and its estimation procedure. Finally we apply it to plant height measurements from a potato field experiment.

2 Method

Epidemiological studies often deal with longitudinal data for developmental characteristics of the cohort under study. Cole et al. (2010) presented the SITAR model based on an earlier paper (Beath, 2007). Both publications propose a shape invariant model with a single fitted curve. There are three different mechanisms that drive the shape of an individual curve in relation to the mean curve for the whole population: a curve can be shifted up or down, left or right, or the scale on the x-axis can be shrunk or stretched. The original SITAR model uses three subject-specific random effects for the characterization of a response y_{ij} for subject i at age j:

$$y_{ij} = \alpha_i + h\left(\frac{t - \beta_i}{\exp(-\gamma_i)}\right) \tag{1}$$

with α a random intercept adjusting for height, β a random shift along the x-axis and γ a random scaling factor. The three parameters are termed *size*, *tempo* and *velocity* respectively by Cole et al.. In the original notation $h(\cdot)$ is a natural cubic spline of the response over age t. This model formulation has the advantage that the parameters are directly interpretable in a biological context.
Inspired by this model we propose a simplification and adaptation of it. Instead of the mixed model we are sticking to a fixed model using P-splines for the functional form (Eilers and Marx, 1996). The response is log-transformed in our context.

As an initial step we formulate the curves for genotype i at time j as

$$\log(y_{ij}) = \alpha_i + f(t_j) \tag{2}$$

including a subject-specific vertical shift α (as intercept of the model or *size* in Cole's terminology). In our case we use P-splines for the functional form f, therefore:

$$\log(y_{ij}) = \alpha_i + \sum_k b_{jk}\beta_k \tag{3}$$

with $B = [b_{jk}]$ a B-spline basis with a generous number of splines and β the associated coefficients. In the implementation two parameters λ –for a difference penalty– and κ –for a ridge penalty– are included to ensure smoothness as well as numerical stability.
In order to correct for the horizontal shifts that the different curves undergo, we introduce a transformation of the horizontal axis. To this end we rewrite and substitute in (3):

$$\begin{aligned} b_{jk} &= B_k(t_j) \\ b_{ijk} &= B_k(t_j + \delta_i) \approx B_k(t_j) + \delta_i \dot{B}_k(t_j) \end{aligned} \tag{4}$$

where $\dot{B}_k(t_j)$ is the first derivative of the kth spline evaluated at time t_j. δ is the so-called *tempo* effect inducing a non-linear transformation of the horizontal scale.
In the model above we assume that for all subjects i the measurements are taken at the same time points t_j. However, this might not be the case in all applications. This can be easily included by using subject-specific time points t_{ij}.

3 Application

We analyze data from a field experiment with a diploid potato mapping population with more than 150 different genotypes. Over the course of the growing season different characteristics are measured. Plant height (in cm) has been assessed at nine time points over the course of three months. Figure 1(a) shows the log-transformed data of this heterogeneous potato population. We apply the model including a vertical and horizontal shift as explained above. The fitted curves and its residuals are depicted in Figures 1(b) and 1(d). Figure 1(c) shows a scatterplot of the results on the shifted scales $\log(y_{ij}) - \alpha_i$ versus $t_j + \delta_i$. The colors are ordered according to the order of the individual *tempo* effects δ_i.

4 Conclusion and Discussion

Our model, inspired by Cole et al. (2010), was successfully applied to longitudinal phenotypic data generated in a field experiment. We can estimate

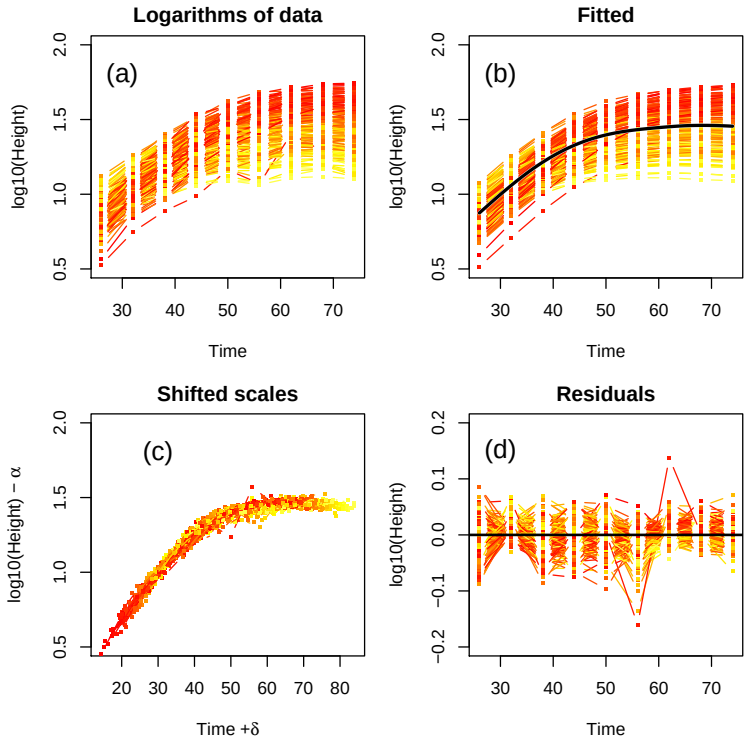


FIGURE 1. (a) Log-transformed data for the whole population, (b) fitted curves using vertical and horizontal shifts with the mean curve, (c) scatterplot on shifted scales, (d) residuals .

a mean curve that provides general information about the growth of the potato plants in the population as a whole. More importantly we also determine characteristics for the individual genotypes that can be used in further genetic analyses. A scatterplot of the vertical and horizontal shifts per genotype can be found in Figure 2.
In future work we plan to extend the model in different ways. At the moment our model includes a so-called *size* as well as a *tempo* effect. In a next step we plan to extend it with a *velocity* effect γ:

$$b_{ijk} = B_k(\gamma_i t_j + \delta_i). \tag{5}$$

A preliminary analysis with the example data this extension did not seem to improve the estimation results, but it is important for future applications. Although the data presented in this manuscript are typical for plant breeding trials, they are still a simplification of the real data situation. To complicate matters data are often measured for several replications of the same

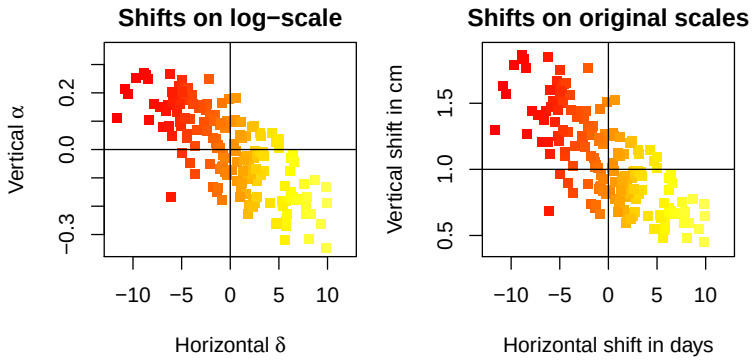


FIGURE 2. Right: horizontal shift (δ) versus vertical shift (α) on log scale for the fitted curves. Left: horizontal shifts (in days) versus vertical shifts (in cm). Colour codes in increasing size of the tempo effect δ.

genetically identical plant. Additionally in some trials the data might include a mixture of cross-sectional and longitudinal data due to intermediate harvest of parts of the experimental field. For replicates a direct solution is at hand by extending the model to accommodate replicates as a random effect within the genotype. Mixture of data through different collection methods need a more theoretical treatment before this situation can be integrated in the current context. Last but not least the estimated individual characteristics of the genotypes will be used in further genetic analysis to associate these with regions on the chromosomes. In order to offer more powerful tools for the plant research community these topics will be treated in future work and reported elsewhere.

Acknowledgments: The data set used in the application has been collected at the Holetta Agricultural Research Center in Holetta, Ethiopia, by Biructawit Bekele Tessema who is financed by a grant from NUFFIC and is part of the Laboratory of Plant Breeding at Wageningen University and Research Centre.

References

Beath, K.J. (2007). Infant growth modelling using a shape invariant model with random effects. *Statistics in Medicine*, **26**, 2547 – 2564.

Cole, T.J., Donaldson, M.D.C., and Y. Ben-Shlomo (2010). SITAR — a useful instrument for growth curve analysis. *International Journal of Epidemiology*, **39**, 1558 – 1566.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science*, **11**, 89 – 121.

Hurtado, P., Schnabel, S., Zaban, A., Veteläinen, M., Virtainen, E. , Eilers, P., van Eeuwijk, F., Visser, R., and Maliepaard, C. (2010). Dynamics of senescence-related QTL in potato. *Euphytica*, **183**, 289 – 302.

Malosetti, M., Visser, R.G.F., Celis-Gamboa, C., and van Eeuwijk, F.A. (2006). QTL methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. *Theoretical Applied Genetics*, **113**, 288 – 300.

Schnabel, S.K., Eilers, P.H.C., Hurtado López, P., Visser, R.G.F., and van Eeuwijk, F.A. (2010). Haulm senescence in potatoes and semi-parametric survival models. In: *Proceedings of the 25th International Workshop on Statistical Modelling*, Glasgow, UK, pp. 489 – 494.

Model interpretation from the additive elements of the PWRSS in GLMMs

Mariangela Sciandra[1], Gianfranco Lovison[1]

[1] Dipartimento di Scienze Economiche, Aziendali e Statistiche, Palermo, Italy

E-mail for correspondence: mariangela.sciandra@unipa.it

Abstract: Generalized Linear Mixed models (GLMMs) have rapidly become a widely used tool for modelling clustered and longitudinal data with non-Normal responses. Although a large amount of work has been done in the literature on likelihood-based inference on GLMMs, little seems to have been done on the decomposition of the total variability associated to the different components of a mixed model. In this work we try to generalize the idea of *likelihood additive elements* (Whittaker, 1984), proposed in the context of GLMs, to the case of GLMMs by using the Penalized Weighted Residual Sum of Squares (PWRSS). The proposal is illustrated by means of a real application.

Keywords: Additive elements; GLMMs; Penalized Weighted Residual Sum of Squares.

1 Introduction

An extensive methodology exists in the literature for investigating the explanatory power of each covariate in regression-type, like linear and generalized linear, models. It is well known that with exact orthogonality of the covariates and Normal errors distribution, the contribution to the global goodness of fit of each explanatory variable added to the linear predictor can be uniquely evaluated by the consequent reduction in the residual sum of squares. When there is a substantial non-orthogonality such contribution can be highly dependent on the presence or absence of other specific variables in the model.
A somewhat neglected contribution to deal with such situations in the classical linear model setting, is due to Newton and Spurrel (1967), who proposed a way to partition the residual sum of squares of the null model containing only the intercept into additive components, the *regression elements*, which can be attributed to each variable alone, to pairs of variables, to triples of variables and so on. Whittaker (1984) extended Newton and Spurrel's idea to more general settings, at the same time giving it a more rigorous formalization. In particular, Whittaker showed the potential for model interpretation provided by the *additive elements of the like-*

lihood function, obtained partitioning the maximized log-likelihood ratio test statistic into such additive elements.
So far, Whittacker's approach has been applied to interpret classical linear or generalized linear models; some difficulties rise when trying to extend it to mixed models. Mixed models assume that some of the regression parameters are fixed whereas others are random. It follows that, while in fixed-effects models the concept of explained variation refers necessarily to the reduction in variation due to the fixed covariates, in mixed models we have to distinguish between variation explained by the random effects and variation explained by the fixed effects (Chen and Dunson, 2003). This poses a serious problem, since in general the contributions of these two components cannot be easily separated.
In this paper, we generalize Whittaker's approach to GLMMs using the Penalized Weighted Residual Sum of Squares (PWRSS) proposed by Bates (2013). Whittaker's approach to model interpretation, a brief introduction to GLMMs and definition of PWRSS are presented in Section 2. The proposed generalization and the real data example are given in Section 3.

2 Model interpretation using Whittaker's additive elements of the likelihood function

Define a set of variables indexed by the first k integers; denote the index set $\{1, 2, \ldots, k\}$ by K and its power set by L:

$$L = \{\emptyset, 1, 2, \ldots, k, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots, \{1, 2, \ldots, k\}\}$$

L is a binary lattice denoting which variables are and which are not included in the fitted model. Then a function $s()$ can be defined on L, with values in the real line, which associates to each element a in L a quantity summarizing the model goodness of fit. To define the additive elements, additivity of $s()$ is postulated and the elements are evaluated using the following recursion: in a first step all the elements $G(: a)$ are computed as

$$G(: a) = s(a) \quad \forall a \in L \tag{1}$$

and then the *additive elements* are recursively obtained as

$$G(ia : b) = G(a : b) - G(a : ib) \quad \forall a, b \in L \quad and \quad i \in K \tag{2}$$

where ia is the subset containing i and the integers in a.
The additive elements result invariant to permuting the integers within the subsets a and b; moreover, repeated use of the relation (2) leads to a full decomposition of $G(:)$ as

$$G(:) = \sum_{a,b \in L} G(a : b)$$

Similarly, the amount of s attributable to the variable i can be obtained as

$$G(i:) = \sum_{a,b \in L} G(ia:b) \tag{3}$$

So, in a formal way, Whittaker defines *additive elements* of s in L the set $\{G(a:b)\}$, where $\{a,b\}$ is a partition of the index set K.
Generally,in reporting results additive elements are properly distinguished in residual, primary, secondary and higher orders elements, to which different interpretations are attached. In particular, primary elements measure the unique contribution of each explanatory variable and for this reason they cannot assume negative values. Secondary elements can assume several meanings; between these, the most interesting is that the secondary elements are a measure of proxy between the two explanatory variables, i.e. the amount of variation neither uniquely attributable to one of the two regressors nor explained by both. They can assume negative values: if a secondary element is positive than the two regressors involved have a *competitive* explanatory role; on the contrary, a negative value of a secondary element means that the two regressors are *complementary* in explaining the variability in the response variable. Higher order elements will be interpreted in terms of partial and marginal elements, while the term with the highest order will be interpreted in terms of residual variation. The definition of Whittaker's additive elements component is based on a properly chosen function $s()$. As emphasized in previous sections, in classical linear models $s()$ corresponds to the residual sum of squares, and the additive elements then are the regression elements of Newton and Spurrel. In the Generalized Linear Model setting, Whittaker(1984) proposed to use the deviance as function $s()$ to be decomposed. Aim of next section will be to further generalize the function $s()$ in order to define additive components for the class of GLMMs.

2.1 Generalized Linear Mixed Models and the PWRSS

Let Y_{ij} be the jth measurement for cluster i, $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$ and let $\mathbf{Y}_i$ be the n_i dimensional vector of all observations in cluster i. A GLMM assumes that, conditionally on q-dimensional random effects $\mathbf{b}_i$, the elements Y_{ij} of $\mathbf{Y}_i$ are independent and follow a distribution in the exponential family:

$$f\{y_{ij}|\mathbf{b}_i, \beta, \phi\} = exp\{\phi^{-1}[y_{ij}\vartheta_{ij} - \psi(\vartheta_{ij})] + c(y_{ij}, \phi)\} \tag{4}$$

with $\mu_{ij} = E(Y_{ij}|\mathbf{b}_i) = \psi'(\vartheta_{ij})$ and $Var(Y_{ij}|\mathbf{b}_i) = \psi''(\vartheta_{ij})\phi$, where ϑ_{ij} is the canonical parameter; finally, a known link function $g(\cdot)$ relates the linear predictor to the transformed mean response as

$$g(\mu_{ij}) = g[E(Y_{ij}|\mathbf{b}_i)] = \mathbf{x}_{ij}^{\mathrm{T}}\beta + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b}_i \tag{5}$$

with $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ the corresponding $(p \times 1)$ and $(q \times 1)$ vectors of covariates associated with the fixed effects and random effects, respectively, and ϕ a dispersion parameter. Finally, in addition to the distributional assumption (4), the random effects are assumed to be drawn independently from $\mathcal{N}(\mathbf{0}, \mathbf{D}(\tau))$. The presence of random parameters prevents the use of a standard likelihood function and hence a remarkable amount of work in the literature on likelihood-based inference for GLMMs has been devoted to the search of extended versions of the likelihood function and the deviance measure of discrepancy used in standard GLMs (for a review, see Molenberghs and Verbecke, 2005). Since, in order to generalize Whittaker's approach, we need a function $s()$ which: (i) depends on both the fixed parameters β, τ and on the random parameters $\mathbf{b}_i, i = 1, \ldots, m$; (ii) has the property of being additive over the lattice indexing all possible models; (iii) is computationally efficient, we opted for the **penalized weighted residual sum of squares (PWRSS)** introduced by Bates(2013) and used by the `R` function `glmer`:

$$PWRSS = [\mathbf{y} - \mathbf{g}^1(\mathbf{Z}\mathbf{\Lambda}(\tau)\mathbf{u} + \mathbf{X}\beta)]^{\mathrm{T}}\mathbf{W}[\mathbf{y} - \mathbf{g}^1(\mathbf{Z}\mathbf{\Lambda}(\tau)\mathbf{u} + \mathbf{X}\beta)] + \parallel \mathbf{u} \parallel \quad (6)$$

where: $\mathbf{y} = [\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_m^{\mathrm{T}}]^{\mathrm{T}}$, $\mathbf{X} = [\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_m^{\mathrm{T}}]^{\mathrm{T}}$, $\mathbf{Z} = \bigoplus_{i=1}^{m} \mathbf{Z}_i$, $\mathbf{W} = \bigoplus_{i=1}^{m} \mathbf{W}_i$, with $\mathbf{W}_i$ the usual weight matrix of the conditional GLM in the i-th cluster, $\mathbf{b}_i = \mathbf{\Lambda}(\tau)\mathbf{u}_i, i = 1, \ldots, m$, and $\mathbf{u} = [\mathbf{u}_1^{\mathrm{T}}, \ldots, \mathbf{u}_m^{\mathrm{T}}]^{\mathrm{T}}$, with $\mathbf{\Lambda}(\tau)$ a transformation matrix such that $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

3 Generalization of Whittaker's approach to GLMMs

Denote by $K = \{1, 2, \ldots, k\}$ the index set referring to the variables in the fixed part of the model, i.e. in matrix $\mathbf{X}$, and by $Q = \{1, 2, \ldots, q\}$ the index set referring to the variables in the random part of the model, i.e. in matrix $\mathbf{Z}$. Denote by F (Fixed) the power set of K and by R (Random) the power set of Q. Let L be the lattice generated from the Cartesian product of the two binary lattices F and R, $L = F \times R$, so that:

$$L = \{(\emptyset; \emptyset), \ldots, (\{1, 2, \ldots, k\}; \emptyset), (1; 1), \ldots, (\{1, 2, \ldots, k\}; \{1, 2, \ldots, q\})\}$$

Now, a scalar function $s()$, $s : L \mapsto \Re^+$, can be defined on L such that applied to a generic element $(a; \alpha)$ in L it summarizes the corresponding GLMM goodness of fit. In particular, we use as function $s()$ the PWRSS defined in equation (6). Unlike in (fixed) linear model, in which the reference null model is uniquely determined, here we assume as "null" model the one which contains only the intercept in both the fixed and random parts. Once defined the "null" model, definition of the additive elements requires a recursive procedure consisting of a first step in which all the elements $G(: a; : \alpha)$ are computed as

$$G(: a; : \alpha) = s(a; \alpha) \quad \forall (a; \alpha) \in L \quad (7)$$

and a following step in which additive elements are obtained as

$$G(ia : b; j\alpha : \gamma) = G(a : b; \alpha : \gamma) - G(a : ib; \alpha : j\gamma) \tag{8}$$

where $a, b \in F, \alpha, \gamma \in R, i \in K$ and $j \in Q$ and ia is the subset of F containing i and the integers in a; jα is the subset of R containing j and the integers in α. Applying recursively the relation (8) it can be shown that the PWRSS associated to the "null" model satisfies:

$$G(:;:) = \sum_{(a,b;\alpha,\gamma)\in L} G(a : b; \alpha : \gamma) \tag{9}$$

3.1 Example. *High school and beyond (Hsb) data*

In the following example a sequence of Poisson GLMMs were fitted in order to model the number of awards received by students from High school and beyond. The set of regressors used were $F = \{Gender, SES\}$ and $R = \{Prog, SES\}$ representing gender, social economical status and the type of school program of students. Students are clustered within schools.

TABLE 1. Additive elements for the *High school and beyond* data.

Order	Element	PWRSS	%	Sub totals
Residual	*:12 ; :12*	145.270	92.235	
	:12 ; 1:2	1.327	0.843	
	:12 ; 12:	0.992	0.630	
	:12 ; 2:1	0.805	0.511	94.219
Primaries	*1:2 ; :12*	1.953	1.240	
	1:2 ; 1:2	1.953	1.240	
	1:2 ; 12:	5.045	3.203	
	1:2 ; 2:1	-0.416	-0.264	
	2:1 ; :12	-4.419	-2.806	
	2:1 ; 1:2	0.289	0.183	
	2:1 ; 12:	4.931	3.131	
	2:1 ; 2:1	-1.360	-0.863	5.064
Secondaries	*12: ; :12*	5.737	3.643	
	12: ; 1:2	-3.508	-2.227	
	12: ; 12:	3.386	2.150	
	12: ; 2:1	-4.485	-2.848	0.718
Total	*:;:*	157.500	100.00	100.00

Graphical inspection of the elements in Table 1 helps to identify variables with a high explanatory power, both in the fixed and the random part. As the value of the residual element $G(: 12; : 12)$ is extremely high with respect to all other additive elements, it has been eliminated from the graph in Fig.1. Since $G(: 12; 12 :)$ is only 0.992 implies that the gain from a model with only random intercept to a model with both regressors in the random part is negligible. On the other hand, the fact that the signs of two among the four primary elements with the same fixed additive elements $(2 : 1)$ become negative when the regressor SES is added to the random part, suggests that if one regressor has to be added as a random effect, besides the intercept, it should be PROG.

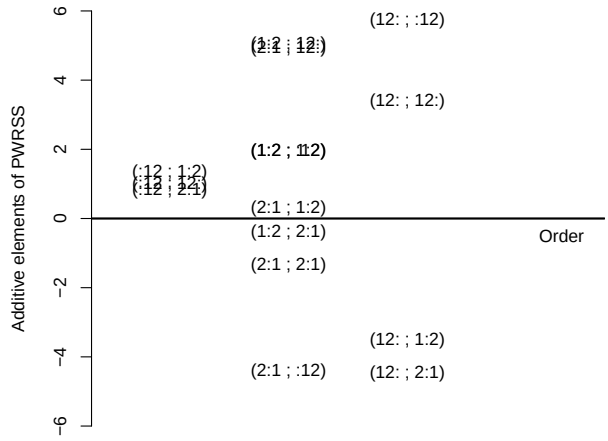


FIGURE 1. Penalized weighted residual sum of squares plot for the Hsb data.

References

Bates, D. (2013). Computational methods for mixed models. *http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf*.

Chen, Z. and Dunson, D.B. (2003). Random Effects Selection in Linear Mixed Models. *Biometrics*, **59, 4**, 762 – 769.

Molenberghs, G. and Verbeke, G (2005). *Models for discrete longitudinal data*. Springer, New York.

Newton, R.G. and Spurrell, D.J. (1967). Examples of the Use of Elements for Clarifying Regression Analyses. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **16, 2**, 165 – 172.

Whittaker, J. (1984). Model Interpretation from the Additive Elements of the Likelihood Function. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*,**33, 1**, 52 – 64.

Fractile Boosting: a novel approach to mode regression

Fabian Sobotka [1], Andreas Mayr[2], Thomas Kneib[1]

[1] Chair of Statistics, Georg August University Goettingen, GERMANY
[2] IMBE, Friedrich Alexander University Erlangen-Nuremberg, GERMANY

E-mail for correspondence: fabian.sobotka@wiwi.uni-goettingen.de

Abstract: Mode regression would be very interesting in practice, if the estimation methods were easy to apply and could include semiparametric models. In the framework of boosting we are able to generalise quantile and expectile regression to fractiles, which minimise a loss function derived from error terms taken to flexible powers. This regression problem can also be modeled as a generalised model for location, scale and shape using an auxiliary likelihood. We combine both approaches to perform an approximation to and generalisation of mode regression, by inserting a low power to the losses and asymmetric weights. This approach is able to handle bimodal, highly skewed or truncated distributions and semiparametric models.

Keywords: mode regression; asymmetric loss regression; semiparametric models; componentwise functional gradient descent boosting; GAMLSS.

1 Introduction

Quantile and expectile regression (see Koenker, 2005, or Sobotka and Kneib, 2010, for example) are becoming increasingly popular. They provide easy ways to estimate more than the conditional location of a response variable. They also do so without many assumptions to the distribution of the aforementioned response. While expectiles allow for more flexible, semiparametric models, fast, easy and efficient estimation methods are nowadays available for both quantiles and expectiles. They aim on solving the following estimation problem

$$\hat{\beta}_\tau = \arg\min_\beta \sum_{i=1}^{n} w_i(\tau)|y_i - x_i'\beta|^k \tag{1}$$

with

$$w_i(\tau) = \begin{cases} \tau & \text{if } y_i > x_i'\beta \\ 1-\tau & \text{if } y_i \leq x_i'\beta \end{cases}, \ \tau \in (0,1)$$

for a response y_i, a covariate or design vector x_i, asymmetry $\tau \in (0,1)$ and power $k > 0$. The estimate is computed for a power of $k = 1$ using linear

programming techniques and provides regression quantiles, while $k = 2$ allows for a direct calculation of the estimate, resulting in expectiles. Setting $\tau = 0.5$ reduces the problem to classical median and mean regression, respectively.

In theory, a power of $k = 0$ would result in mode regression and the inclusion of weights $w_i(\tau)$ in its generalisation. However, the practical execution is futile for finite samples with metric responses. Instead, mode regression was introduced as a form of kernel regression (Lee, 1989) and generalised to additive and nonlinear models without a penalty term by Kemp and Santos Silva (2010), but only for unimodal and to some extent only for symmetric distributions. In addition, semiparametric models with smooth spatial or random effects cannot be included here.

In order to offer a solution to this scenario, we propose the introduction of k-fractiles as solutions of the minimisation problem (1). Hence, quantiles and expectiles are also possible fractiles. However, for $k \searrow 0$, fractiles can approximate and generalise mode regression while incorporating semiparametric models and smoothing. Due to the distribution-free formulation of the estimate, skewed and bimodal distributions for the response are also allowed.

Since a direct search of the minimum (1) might in general be a hard task, we offer two estimation algorithms based on boosting, a flexible framework for regression models and also provide some first examples.

2 Boosting

Component-wise gradient boosting as introduced by Bühlmann and Hothorn (2007) is a very flexible framework that allows for penalised semiparametric modelling. Smooth, spatial and random effects, e.g., are included by simple (least squares-like) base-learning procedures $f_1(\cdot), ..., f_p(\cdot)$ independent from the applied loss function. It divides the minimisation problem in a fixed, large number of small steps where only the model part with the steepest gradient is included. Hence, automatic variable selection takes place during the estimation. The following estimation algorithm can be executed with the R-package **mboost** (Hothorn et al., 2013) for a pre-fixed number of iterations m_{stop}.

1. **Initialization:** $m = 0$. Initialize the additive predictor $\eta_i^{[0]} = 0$ for $i = 1, ..., n$. Specify a set of base-learners $f_1(\cdot), ..., f_p(\cdot)$, one for each covariate.

2. **Negative gradient:** $m = m + 1$. Compute the negative gradient vector $u^{[m]}$:

$$u_i^{[m]} = \begin{cases} k * \tau * |y_i - \eta_{\tau i}^{[m-1]}|^{(k-1)} & (y_i - \eta_{\tau i}^{[m-1]}) \geq 0 \\ k * (1-\tau) * |y_i - \eta_{\tau i}^{[m-1]}|^{(k-1)} & (y_i - \eta_{\tau i}^{[m-1]}) < 0. \end{cases}$$

3. **Component-wise estimation:** Use the base-learners to fit the negative gradient vector $u^{[m]}$ to every possible covariate $x_1, ..., x_p$ separately
$$u^{[m]} \overset{\text{base-leaner}}{\longrightarrow} \hat{f}_j(x_j) \quad \text{for} \quad j = 1, ..., p$$

4. **Update one component:** Select the component j^* that best fits the negative gradient vector and update the additive predictor:
$$\hat{\eta}^{[m]} = \hat{\eta}^{[m-1]} + \text{sl} \cdot f_{j^*}(x_j) \ ,$$
where sl is a small step-length ($0 < \text{sl} \ll 1$). Therefore, only the best-performing base-learner (and hence the best-performing covariate) contributes to the update.

5. **Iteration:** Iterate steps 2 to 4 until $m = m_{\text{stop}}$. ,

As penalised least squares regression is highly dependent on the optimal choice of the smoothing parameters, in this algorithm the optimal stopping iteration is also computed via cross-validation. This allows for the regulation of smoothness in our estimates.
The only requirement is the existence of a gradient, which is available for almost all $k > 0$. So, boosting theoretically allows for the fit of semiparametric models to fractiles, and especially also for $k \ll 1$.

3 GAMLSS

An alternative approach can be constructed by the combination of generalised additive models for location scale and shape (Rigby and Stasinopoulos, 2005) with boosting by Mayr et al. (2012). Instead of solving the minimisation problem (1) we choose a skew exponential power distribution

$$\begin{aligned} f_{\mu,\sigma,w,k}(y) = \quad & \frac{c}{\sigma} \left\{ I(y < \mu) \exp\left[-\frac{1}{2} \left| w \frac{y-\mu}{\sigma} \right|^k \right] \right. \\ & \left. + I(y \geq \mu) \exp\left[-\frac{1}{2} \left| \frac{1}{w} \frac{y-\mu}{\sigma} \right|^k \right] \right\} \end{aligned}$$

with $c = \frac{wk}{(1+w^2)2^{1/k}\Gamma(\frac{1}{k})}$ and $w = \left(\sqrt{\frac{1-\tau}{\tau}} \right)^{1/k}$ as auxiliary likelihood. In the R-package **gamboostLSS** (Hofner et al., 2011) the boosting algorithm from the previous section is adapted to maximise such likelihoods for multiple parameters. While we fix w and k in order to rewrite our loss function (1), the algorithm iterates between a fit for μ and σ. Hence, we get a good fit of the distribution to our data and escpecially an estimate for the location μ. By this reformulation we also end up with a maximisation problem instead of a minimisation problem which appears to be more stable for very small powers k and extreme asymmetries $\tau \to 1$, $\tau \to 0$.

4 Examples

In order to test our procedures we generate simple, bivariate data with $n = 500$ observations of the form $y = f(x) + \varepsilon$ with $f(x) = x^2$ or $f(x) = \exp(-x^2)$ and $\varepsilon \sim Exp(0.5)$ or $\varepsilon \sim 0.55N(0, 0.6^2) + 0.45N(3, 0.6^2)$. In these cases the conditional mode would be either at the lowest ends of the response's values or in the lower half of the data with an additional, weaker mode in the top half. Both scenarios could not be estimated sensibly by kernel mode regression as the properties of the estimate are only known for symmetric and unbounded kernels.
However, as Figure 1 shows, boosting is possible down to $k = 0.15$, which seems to be a good approximation of the mode while still using a lot of information from the data. In comparison to the mean, median and (for fun) the 50% 4-fractile we can also see that the mode approximation is a much better measure of location for the data at hand. Especially with bimodal errors (on the right) the mean represents the data rather poorly.

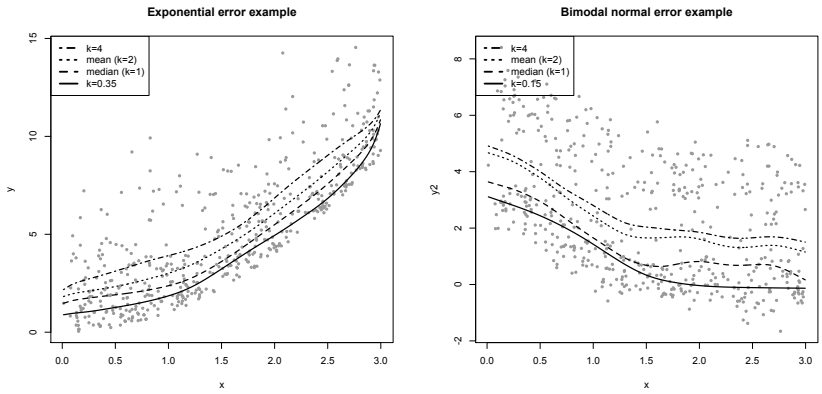


FIGURE 1. example analyses of generated data with exponential and bimodal normal errors for fractiles with $k = 0.15/0.35, 1, 2, 4$ and a symmetric ($\tau = 0.5$) loss function

The addition of asymmetric weights as shown in Figure 2 also seems to work as it helps to uncover the second mode that is present in the data.
While the first experiments were successful, we also uncovered an instability of the procedure for very small powers and extreme asymmetries. The combination might likely lead to artefacts in the estimate. The possibilities as well as the limits of fractiles are yet to be figured out. Next we perform a simulation study to assess the qualities of fractiles as replacement of mode regression and a comparison with GAMLSS by Rigby and Stasinopoulos (2005).

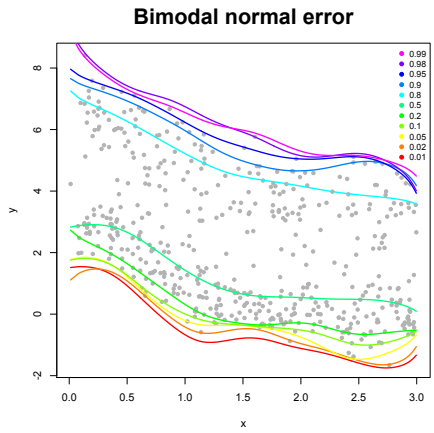


FIGURE 2. example analysis of generated data with bimodal normal errors for fractiles with $k = 0.35$ and asymmetries between 0.01 and 0.99

5 Simulations

The estimators are compared for nonlinear models of the form $y = f(x) + \varepsilon$ with $X \sim U(0,3)$, $f(x) = 5\exp(-x^2)$ or $f(x) = 5sin(2x)$ and $\varepsilon \sim N(0, 0.5^2), Exp(0.5), LN(0,1)$. Hence, we have a symmetric, truncated and a skewed error distribution. The data is generated for $n = 100, 500$ and replicated 100 times.

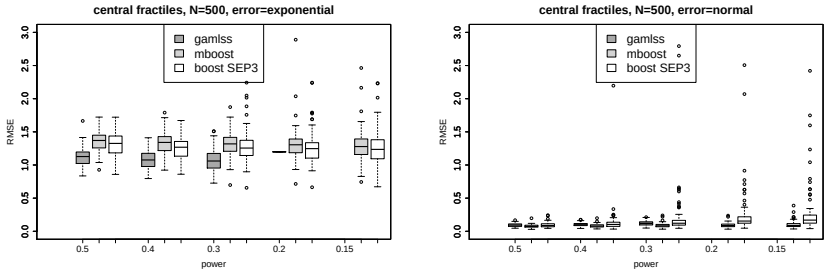


FIGURE 3. Comparison of the three possible estimation methods: root mean squared errors for the modes of an exponential and a normal error in a non-linear regression model. Estimated fractiles for $k = 0.5, 0.4, 0.3, 0.2, 0.15$ and a symmetric ($\tau = 0.5$) loss function.

In Figure 3 we can see the results in terms of RMSE for fractile exponents of $k = 0.5, 0.4, 0.3, 0.2, 0.15$ and no asymmetry ($\tau = 0.5$). An improvement

of the estimation for lower k is generally visible. However, the frequentist GAMLSS method would not work for $k < 0.3$. Hence, the missing boxplots. Surprisingly, boosting the gradient of the loss function proves to be slightly better than with the specification of the SEP3 distribution. In a symmetric distribution the error is also overall smaller than in the truncated case where the mode is directly at the edge of the data and therefore hardly accessible by estimation. We also find that higher values of k lead to less variable estimates. However, the results do net yet lead to a specific decision for one algorithm and one k, if we want to estimate a mode regression. For the different scenarios the quality of the available methods changes relative to each other.

Acknowledgments: Financial support from the German Research Foundation (DFG) grant KN 922/4-1 is gratefully acknowledged.

References

Bühlmann, P., Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science* 22(4), 477-505.

Hofner, B., Mayr, A., Fenske, N., Schmid, M. (2011). gamboostLSS: Boosting Methods for GAMLSS Models. R package version 1.0-3.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., Hofner, B. (2013). mboost: Model-Based Boosting. R package version 2.2-1.

Kemp, G. C. R., Santos Silva, J. M. C. (2010). Regression towards the mode. *Economics Discussion Papers, University of Essex, Department of Economics* http://EconPapers.repec.org/RePEc:esx:essedp:686.

Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press.

Lee, M.J. (1989). Mode Regression. *Journal of Econometrics* 42, 337-349.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., Schmid, M. (2012). Gamlss for high-dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society, Series C* 61, 403–427.

Rigby, R. A., Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Applied Statistics* 54, 507-554.

Sobotka, F. and T. Kneib (2010). Geoadditive expectile regression. *Computational Statistics and Data Analysis*, doi: 10.1016/j.csda.2010.11.015.

Do not use a cannon to kill a mosquito: a comparison of supervised classification algorithms in the context of MS lesion segmentation in structural MRI

Elizabeth M. Sweeney [1] [2], Russel T. Shinohara [3], Joshua T. Vogelstein [4], Daniel S Reich[2], Ciprian M. Crainiceanu[1]

[1] Department of Biostatistics, The Johns Hopkins University, United States
[2] Translational Neuroradiology Unit, Neuroimmunology Branch, National Institute of Neurological Disease and Stroke, National Institute of Health, United States
[3] Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, United States
[4] Department of Statistical Science, Duke University, United States

E-mail for correspondence: emsweene@jhsph.edu

Keywords: Supervised Classification; Neuroimaging; Magnetic Resonance Imaging.

Abstract

Magnetic resonance imaging (MRI) can be used to detect lesions in the brains of multiple sclerosis (MS) patients and is essential for evaluating disease-modifying therapies and monitoring disease progression. Many different classification algorithms have been applied to the MS lesion segmentation problem and it is difficult to assess whether improved performance is due to differences in classifiers or in the features used in classification. We evaluate the performance and compare computation times for nine supervised classification algorithms: logistic regression, neural net, support vector machine (svm), quadratic discriminant analysis (QDA), linear discriminant analysis (LDA), gaussian mixture model (GMM), k-nearest neighbors (kNN), Random Forest and Super Learner in the context of lesion classification in structural MRI. We also examine the impact of different feature extractions for this classification: intensity normalization, a candidate voxel selection procedure, spatial smoothing, and local moments. Our findings are that the particular classification algorithm is not important and we focus instead on careful development of the feature space.

1 Introduction

MRI can be used to detect lesions in the brains of MS patients and is essential for evaluating disease-modifying therapies and monitoring disease progression. In practice, lesion load is quantified by manual segmentation of MRI, which is time-consuming, costly, and associated with large inter- and intra- observer variability. Therefore, a sensitive and specific automated method to detect lesions in the brain is essential for the analysis of large MS studies.

Over 80 papers proposing automated lesion segmentation methods have been published in the last 15 years, and yet no solutions has emerged as superior and the problem remains open (García-Lorenzo et al., 2012). This is attributed to a number of factors, including high variability of MS lesion appearance, differences in imaging acquisitions, and the lack of a common framework in which to compare segmentation methods (García-Lorenzo et al., 2012), (Lladó et al., 2011). Many different classification algorithms have been applied to the MS lesion segmentation problem and it is difficult to assess whether improved performance is due to differences in classifiers or in the features used in classification.

We compare supervised classification algorithms for classification of lesion voxels versus healthy tissue in structural MRI. We use the language of pattern recognition. In a supervised learning problem, a response is predicted from a set of features. Feature extraction refers to the process of transforming features by taking functions of the original features. A feature space refers to the space defined by the features and extracted features, in which the supervised classification algorithm is trained and makes predictions.

We examine the effect on lesion segmentation performance in regards to the refinement of the feature space the segmentation is performed in and the use of various supervised classification algorithms. We explore the impact on classification performance for nine supervised classification algorithms as well as refinements of the structural MRI feature space. We fit the model in six different feature spaces, which we refer to as Unnormalized, Normalized, Voxel Selection, Smoothed, Moments, and Smoothed and Moments. The feature extractions used to create these spaces are introduced in the context of MS lesion segmentation in Sweeney et al. (2013). We also introduce moment volumes, a feature space refinement which is to our knowledge is novel in the context of lesion segmentation. The nine classification algorithms we use are logistic regression, neural net, svm, QDA, LDA, GMM, kNN, Random Forest and the Super Learner. Our conclusion is stated in the title: Do not use a cannon to kill a mosquito. Our findings are that the particular classification algorithm is not important and we focus instead on careful development of the feature space. It is our experience that observed differences in algorithms are due to *how* data are used and not to *what* classification approach is used.

2 Materials and Methods

2.1 Structural MRI Data

MRI can be acquired with different pulse sequences to create different imaging contrasts. In this analysis, we focus on MRI studies with three imaging acquisition volumes: the T1-weighted, T2-weighted and fluid-attenuated inversion recovery (FLAIR). Figure 1 shows an example of the three volumes for a single brain MRI study. Each structural MRI volume is an array of 7 million numbers and each of these numbers is referred to as a voxel. The classifiers we are investigating are supervised, so we train and validate on manual lesion delineation made by a technologist. This manual segmentation is also shown in Figure 1.

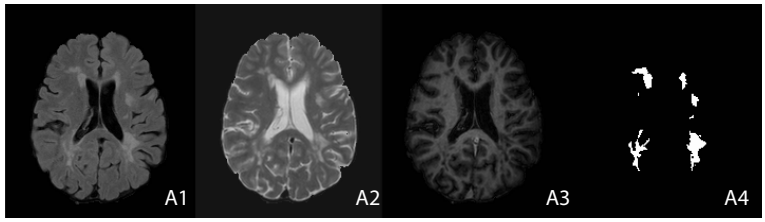


FIGURE 1. A1. FLAIR volume A2. T2-weighted volume A3. T1-weighted volume A4. Manual lesion mask

MRI studies from 98 MS subjects with manual lesion segmentations were used to train and validate the models – 49 studies were randomly assigned to the training set and the remaining 49 studies were used for validation. We fit the models on a set of 500 voxels sampled from each of the training MRI studies (for a total of 24,500 voxels). This was done in order to reduce the amount of time needed to fit each model. We validated on the entire brain volume for each of the 49 studies in the validation set.

2.2 Supervised Classification Algorithms

We performed all statistical modeling in the R environment (version 2.12.0, R Foundation for Statistical Computing, Vienna, Austria). Table 1 show a summary for the supervised classification models fit, including the R package used and the tuning parameters for each model. All tuning parameters were selected using 10-fold cross validation. Each model was fit on a voxel level, assuming independence between voxels.

2.3 Feature Extractions

The feature extractions we use in this analysis are adapted from a proposed lesion segmentation algorithm, OASIS is Automated Statistical Inference

TABLE 1. Summary of the supervised classification models fit in R

Model	R package	Tuning Parameters
Logistic Regression		
LDA	MASS	defaults
QDA	MASS	defaults
GMM	mclust	defaults
SVM , linear kernel	e1071	cost: 1, 10, 100
Random Forest	randomForest	number of trees = 500 mtry = 1 : # of predictors
k -NN	class	k = 1,10, 100
Neural Network	nnet	size = 1, 5, 10 decay= 0, 0.001, 0.1
Super Learner	SuperLearner	all models

for Segmentation (OASIS) (Sweeney et al., 2013). We also introduce local moment volumes, which consist of calculating sample moment over a voxel neighborhood. We use these feature extractions to create six feature spaces: Unnormalized, Normalized, Voxel Selection, Smoothed, Moments, and Smoothed and Moments. A visual representation of the feature spaces we use can be seen in Figure 2.
Here we focus on two of the feature spaces, Unnormalized and Moments. The Unnormalized feature space consists of the raw intensities from the three MRI volumes. The Moments feature space uses intensity normalized MRI volumes, a voxel selection procedure and local moment volumes.

3 Results

Figure 3 shows voxel-level partial Receiver Operating Characteristic (pROC) curves for the validation set with false positive rates of 10% and below for two of the feature spaces: Unnormalized and Moments. The vertical axis of the partial ROC curve shows the true positive rate (sensitivity) for a given threshold of the probability map and the horizontal axis shows the false positive rate (1 - specificity) for this threshold. Figure 3A is a legend for the plots. Figure 3B is the pROC curve for the Unnormalized feature space. Figure 3C is the pROC curve for the Moments feature space.

4 Discussion

In the Unnormalized feature space there is a large difference between the performance of the classifiers. The Super Learner and the GMM perform the best, followed closely by the Neural Net and Random Forest. After the intensity normalization is performed, the voxel selection procedure is

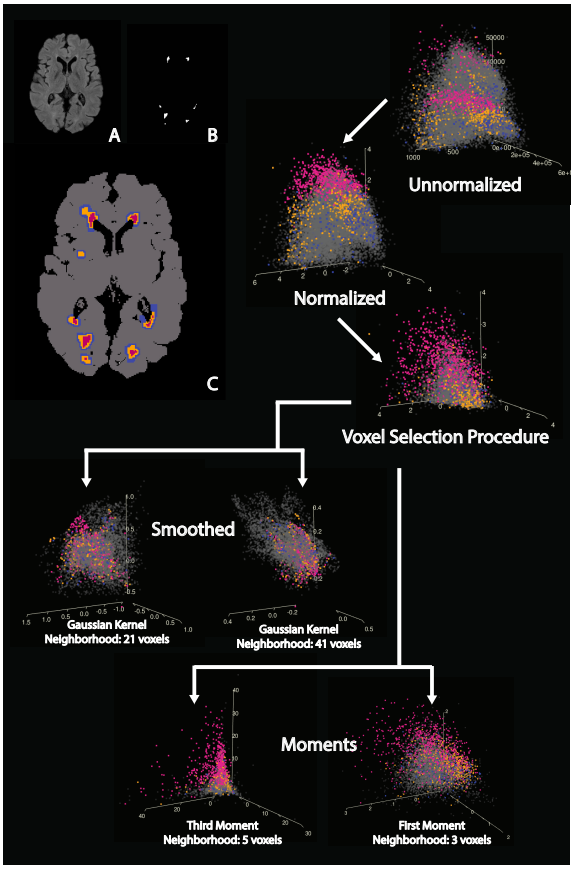


FIGURE 2. Shown are 3-dimensional plots of the FLAIR, T1-weighted, and T2-weighted intensities and functions of these intensities for voxels in 5 randomly sampled subject's MRI studies. For the plots we have randomly sample 10,000 voxels from the 5 randomly sampled MRI studies. Each point in the plot is a voxel from a study. Figure 1A shows a slice of the FLAIR volume for a subject, Figure 1B shows the manual segmentation for this slice, and Figure 1C is a color key for these plots. Lesion voxels are pink, voxels within one voxel (26-connected) of a lesion voxel are orange, and voxels within two voxels (26-connected) of a lesion voxel are blue. All other voxels in the brain are colored grey. The plots are made for the five of the feature spaces that we fit the models on and are labeled as such: Unnormalized, Normalized, Voxel Selection, Smoothed, and Moments.

applied these differences dissappear, and it is important to keep in mind the complexity and time to fit and make predictions for the algorithms, which make simpler methods such as Logistic Regression and LDA more desirable.

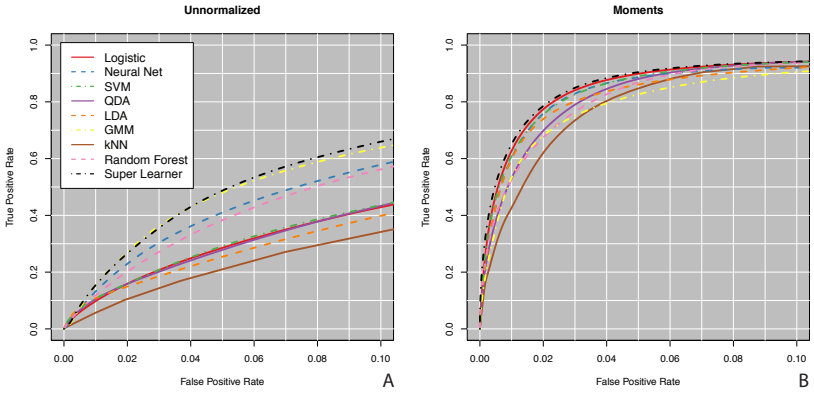


FIGURE 3. The pROC curves for all models on the validation set with false positive rates of 10% and below for two of the feature spaces: A. Unnormalized B. Moments.

5 Conclusion

Our conclusion is stated in the title: Do not use a cannon to kill a mosquito. Our findings are that the particular classification algorithm is not important and we focus instead on careful development of the feature space. It is our experience that observed differences in algorithms is due to *how* data are used and not to *what* classification approach is used.

References

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L. (2012). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Medical Image Analysis, 17, 1-18.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J. C., Quiles, A., Valls, L., Ramió-Torrentá, L., Rovira, A., (2011). Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches, Information Sciences, 186, 164-185.

Sweeney, E.M. , Shinohara, R.T., Shiee, N., Mateen, F.J., Chudgar, A.A., Cuzzocreo, J.L., Calabresi P.A., Pham, D.L., Reich, D.S., Crainiceanu, C.M. (2013) OASIS is Automated Statistical Inference for Segmentation with applications to multiple sclerosis lesion segmentation in MRI. NeuroImage: Clinical, 2, 402-413.

Nonlinear Monotone Regression for High-dimensional Data

Kukatharmini Tharmaratnam[1], Linn Cecilie Bergersen[1], Ingrid K. Glad[1]

[1] Department of Mathematics, University of Oslo, Norway

E-mail for correspondence: kukathat@math.uio.no

Abstract: In recent years, several methods are proposed to model nonlinear relationships in high-dimensional data by using spline basis functions and group penalties. We focus on the special case of nonlinearity as nonlinear *monotone* effects on the response, as is often a natural assumption in medicine and biology. We construct the monotone splines lasso (MS-lasso) method to estimate and select variables using monotone spline basis functions (I-splines). The additive components in the model are represented by the I-spline basis function expansions and the component selection becomes that of selecting the groups of coefficients in the I-spline basis function expansion. We use a recent procedure called cooperative lasso to select sign-coherent groups, that is selecting the groups with either non-negative or non-positive coefficients. This leads to the selection of the important covariates that have nonlinear monotone increasing or monotone decreasing effect on the response in high-dimensional regression problems. Simulated data and real data examples from genomics illustrate the effectiveness of the proposed method. Results indicate that the (adaptive) MS-lasso has excellent properties compared to the other methods both by means of estimation and selection, and can be recommended for high-dimensional monotone regression.

Keywords: Cooperative lasso; I-splines; Lasso; Monotone regression; Nonparametric additive models.

1 Introduction

There has been a major effort in developing methods for monotone regression beyond the strictly linear regression models. These methods are usually concerned with classical situations in which the number of covariates P does not exceed the number of observations n. In the last decade, the massive production of data sets in all areas of science and technology has turned high-dimensional regression problems, where P is much larger than n, into one of the most active research areas within statistics. In certain bio-medical applications it is important to assume that the relationship between an explanatory variable and the outcome is monotonically increasing or decreasing. Very recently, one important contribution has appeared for

monotone regression in high dimensions (Fang and Meinshausen, 2012). In this paper we develop another substantially different tool for this purpose. Given the observations $(y_i, \mathbf{x}_i), i = 1, ..., n$, where y_i is the response and $\mathbf{x}_i = (x_{i1}, ..., x_{iP})^t$ is the vector of covariates for observation i, the additive model is given by

$$y_i = \beta_0 + \sum_{j=1}^{P} g_j(x_{ij}) + \epsilon_i. \tag{1}$$

Here β_0 is the intercept, the g_j's are unknown functions to be estimated and ϵ_i is the unobserved independent random error with mean zero and variance σ^2. We assume $Eg_j(\mathbf{x_j}) = \mathbf{0}$, for $1 \leq j \leq P$, where $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^t$, to ensure unique identification of the g_j's. In Meier et al. (2009); Huang et al. (2010) and Ravikumar et al (2009) each nonparametric component g_j is represented by a linear combination of spline basis functions and the problem can be viewed in terms of a group lasso problem. Combined with the group lasso, the framework becomes a highly flexible alternative to (standard) linear lasso-type methods.
One way of preserving monotonicity is to fit a smooth monotone function via monotone regression splines. Ramsay (1988) introduced integrated splines (I-splines), which essentially are integrated versions of M-splines.

2 Methodology

We propose a new approach to fit nonparametric additive models under the assumption that each component effect $g_j(x_j)$ is monotone. The *monotone splines lasso* (MS-lasso) combines the idea of I-splines with the cooperative lasso (Chiquet et al 2012), and is feasible also in high-dimensional settings where the number of covariates P can exceed the number of observations n. The cooperative lasso is a lasso method where known groups of covariates are treated together, but differs from the standard group lasso in that it assumes that the groups are sign-coherent.
The important advantages of the MS-lasso are indeed that it is not restricted to only monotone *increasing* effects, that is, the estimated monotone functions $\hat{g}_j$ can be either monotone *increasing* or *decreasing* in the same model, and also it is fitting *smooth* monotone functions to each g_j. In this way we are more flexible than the linear model, but more restrictive than the pure nonlinear methods without any shape constraints. Our method is also often biologically more relevant than the adaptive liso, in that we obtain smooth representations of the functions right away. We also suggest a two-step estimator, the adaptive MS-lasso, which leads to less bias and fewer false positives in the final model.

2.1 Monotone Splines Lasso

Suppose we have the additive model in (1) and write each of the individual regression functions as a linear combination of m monotone splines basis functions;

$$g_j(x) = \sum_{k=1}^{m} \beta_{jk} I_k^{(l)}(x), \quad 1 \le j \le P. \tag{2}$$

where $I_k^{(l)}(\cdot)$ is the I-spline basis function and β_{jk} is the k^{th} spline coefficient for the j^{th} covariate. Let $\mathcal{G}_n = \{g : g(x) = \sum_{k=1}^{m} \beta_k I_k^{(l)}(x), \quad \beta_k \ge 0 \text{ or } \beta_k \le 0\}$ be the space of monotone I-spline functions. The constraint for $\{\beta_k\}_1^m$ in $\mathcal{G}_n$ guarantees that each $g \in \mathcal{G}_n$ is monotone. As each component $g_j(x)$ is represented by a spline basis, components that are not selected will have $\beta_{jk} = 0, \forall k$, while selected variables should have spline coefficients that are either all nonnegative or all nonpositive. An I-spline with all $\beta_{jk} \ge 0, \forall k$ will produce a monotone nondecreasing function $g_j(x)$, while an I-spline with $\beta_{jk} \le 0, \forall k$ will produce a nonincreasing function.
We are centering the basis functions to satisfy the identifiability assumption $Eg_j(\mathbf{x_j}) = \mathbf{0}$, for $1 \le j \le P$. Let $z_{ijk} = I_k(x_{ij}) - \bar{I}_{jk}$ be the centered I-spline basis function, where $\bar{I}_{jk} = \frac{1}{n}\sum_{i=1}^{n} I_k(x_{ij})$. Let $\mathbf{Z} = (\mathbf{Z_1}, ..., \mathbf{Z_p})$ be the $n \times (Pm)$ design matrix where all covariates are represented by a centered I-spline basis and $\mathbf{Z_j}$ is the $n \times m$ matrix for the jth covariate. We use the centered response vector $\mathbf{y}$ of length n. Then the MS-lasso estimates of β are defined by minimizing the objective function with respect to β;

$$\hat{\beta}^{MS} = \underset{\beta \in \mathbb{R}^{\mathbf{Pm}}}{\operatorname{argmin}} \left\{ \frac{1}{2} ||\mathbf{y} - \mathbf{Z}\beta||^{\mathbf{2}} + \lambda ||\beta||_{\mathbf{coop}} \right\},$$

where $\lambda \ge 0$ decides the amount of shrinkage and is common to all groups. Here

$$||\beta||_{coop} = ||\beta^+||_{group} + ||\beta^-||_{group} = \sum_{j=1}^{P} w_j \left(||\beta_j^+|| + ||\beta_j^-|| \right)$$

is the cooperative lasso norm with $\beta^+ = (\beta_1^+, ..., \beta_P^+)^t$ and $\beta^- = (\beta_1^-, ..., \beta_P^-)^t$, where $\beta_j^+ = (\beta_{j1}^+, \ldots, \beta_{jm}^+)^t$ and $\beta_j^- = (\beta_{j1}^-, \ldots, \beta_{jm}^-)^t$. Hence β^+ and β^- are the positive and negative parts of β, that is, $\beta_{jk}^+ = \max(0, \beta_{jk})$ and $\beta_{jk}^- = \max(0, -\beta_{jk})$ respectively.
The intuitive idea and motivation of adaptive MS-lasso are the same as for the adaptive lasso and the adaptive group lasso. For example, the adaptive (group) lasso procedures have the property that the solution is at least as sparse as the initial estimator, and can therefore be used to reduce the number of false positives compared to the standard initial procedures. The adaptive procedures also penalize less for components with large initial estimators, implying less biased estimates than for the standard procedures.

3 Results

We generate $w_{i1}, \ldots, w_{ip}$, u_i and v_i independently from $N(0,1)$ truncated to $[0,1]$. The covariates are generated as follows,

$$x_{ij} = \frac{w_{ij} + tu_i}{1+t} \quad \text{for } j \in A, \quad x_{ij} = \frac{w_{ij} + tv_i}{1+t} \quad \text{for } j \notin A,$$

We let $t = 0, 1$ to get independent and dependent covariates. If A is the set of components in the true model, the nonzero and zero components are independent. We take $P = 1000$, the number of replication is 100 and the sample size $n = 50$. The response variable is generated from the following model,

$$\mathbf{y} = g_1(\mathbf{x_1}) + g_2(\mathbf{x_2}) + g_3(\mathbf{x_3}) + g_4(\mathbf{x_4}) + \epsilon.$$

ϵ is generated from normal distribution with mean 0 and variance such that signal to noise ratio $SNR \approx 4$. For the MS-lasso we use a monotone I-splines basis of order two and six evenly distributed knots for all functions g_j. For Huang et al. (2010) method (BS-lasso), we use a quadratic B-spline basis, also with six evenly distributed knots.

TABLE 1. Comparison of the selection performance for these six methods. The proportion of correct selections of each component in the true model, together with the average number of true and false positives.

	Selection					
	g_1	g_2	g_3	g_4	TP	FP
MS-lasso	1.00	0.89	1.00	1.00	3.89	17.72
Ad. MS-lasso	0.98	0.87	1.00	1.00	3.85	2.97
Lasso	0.85	0.72	1.00	1.00	3.57	25.01
Ad. lasso	0.81	0.68	1.00	1.00	3.49	18.40
Ad. liso	0.39	0.98	1.00	1.00	3.37	5.81
BS-lasso	0.00	0.04	0.23	0.93	1.20	1.09

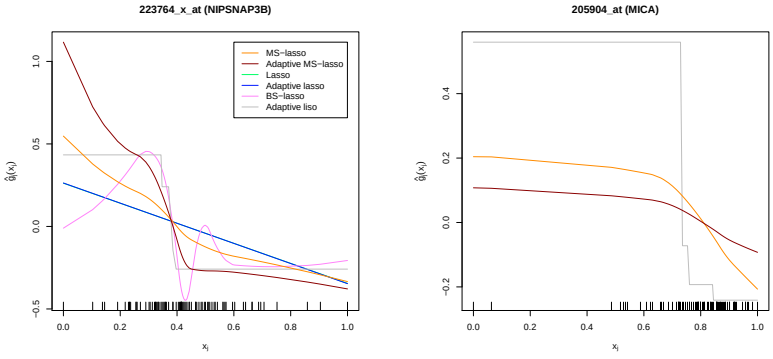
Table 1 shows the MS-lasso is able to select all four components in the true model in almost all of the simulation runs. TP is close to 4 and FP is closed to 18 in MS-lasso. Introducing an adaptive step, reduces the number of false positives.

We also evaluate the estimation error which is given in Table 2. Comparing the MS-lasso and adaptive MS-lasso with their linear competitors, the lasso and the adaptive lasso respectively, we see that the two methods allowing for a nonlinear monotone relationship, estimate the effect of the components with more accuracy. Comparing the MSE for each of the four components individually, we see that the adaptive MS-lasso does much better than all of the other methods.

TABLE 2. Comparison of the estimation performance for these six methods, compute MSE between the fitted function and the true function, averaged over the 100 simulated data sets.

	Estimation			
	g_1	g_2	g_3	g_4
MS-lasso	0.06 (0.03)	0.17 (0.08)	0.15 (0.06)	0.14 (0.06)
Ad. MS-lasso	0.02 (0.03)	0.07 (0.10)	0.03 (0.02)	0.03 (0.04)
Lasso	0.11 (0.04)	0.26 (0.06)	0.35 (0.07)	0.21 (0.07)
Ad. lasso	0.09 (0.05)	0.22 (0.09)	0.28 (0.06)	0.15 (0.07)
Ad. liso	0.12 (0.05)	0.07 (0.06)	0.08 (0.04)	0.05 (0.02)
BS-lasso	0.16 (0.00)	0.32 (0.03)	0.62 (0.22)	0.15 (0.13)

To illustrate our proposed method, we use a bone mineral data set previously studied in Reppe et al. (2010). We consider 84 women who had a trans-iliacal bone biopsy. We take 2000 genes with largest standard deviation. We use the MS-lasso to model the relationship between the bone mineral density and the expression of the 2000 genes and compared the results with already existing methods. In Figure 1(a) we see, the adaptive MS-lasso and the adaptive liso recognize a rapid decrease for midrange values, and it seems that there might be a threshold effect. BS-lasso also seems to be recognizing this rapid decrease, but it is too flexible. We see in Figure 1(b) an example where only the methods assuming monotonicity are selecting the component. Both the monotone splines methods and the adaptive liso estimate a decreasing effect with a breakpoint around 0.7 (in the transformed values).



(a) (b)

FIGURE 1. Estimated functions for two selected genes from the bone data example.

4 Discussion

We have suggested a method for variable selection in high-dimensional regression problems, selecting variables that have a monotonically increasing or decreasing effect on the response. Our method is within the same category as the method of for example Huang et al. (2010), but constructed especially to estimate and select variables showing a (nonlinear) monotone effect. MS-lasso is more flexible than the linear lasso method but more restrictive than the nonlinear methods using B-splines and more flexible than LISO, obtaining a smooth representation of the functions. If the monotonicity assumption is fulfilled, our proposed method MS-lasso gives better results than the other methods.

References

Chiquet, J., Grandvalet, Y., and Charbonnier, C. (2012). Sparsity with sign-ciherent groups of variables via the cooperative -lasso. *Annals of Applied Statistics*, **6**, 795 – 830.

Fang, Z., and Meinshausen, N. (2012). Lasso isotone for high-dimensional additive isotonic regression. *Journal of Computational and Graphical Statistics*, **21**, 72 – 91.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, **38**, 2282 – 2313.

Meier, L., van de Geer, S., and Buhlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics*, **37**, 3779 – 3821.

Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science*, **3**, 425 – 441.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **71**, 1009 – 1030.

Reppe, S., Refvem, H., Gautvik, V. T., et al (2010). Eight genes are highly associated with BMD variation in postmenopausal Caucasian women.*Bone*, **46**, 604 – 612.

A GEE Approach for Correlated Ordinal and Nominal Multinomial Responses

Anestis Touloumis[1]

[1] EMBL-European Bioinformatics Institute, Hinxton, United Kingdom

E-mail for correspondence: anestis@ebi.ac.uk

Abstract: A generalized estimating equations (GEE) approach for correlated multinomial responses was proposed by Touloumis et al. (2013). This GEE approach utilizes marginalized local odds ratios structures to describe the "association" structure and thus, it enables GEE analysis for both ordinal and nominal response categories. To obtain efficient and parsimonious local odds ratios structures, the family of association models (Goodman, 1985) is employed. We discuss the key features of the local odds ratios GEE approach and illustrate its use for marginal modeling of correlated nominal multinomial responses. Finally, we indicate software availability.

Keywords: Association Models; Generalized Estimating Equations; Local Odds Ratios; Multinomial Responses.

1 Introduction

Liang and Zeger (1986) proposed the generalized estimating equation (GEE) method for estimating the regression parameters of a marginal model when the association/correlation structure is of secondary importance. For multinomial responses, Touloumis et al. (2013) recognized that simultaneous modeling of the pairwise association and regression parameters is restricted by the marginal model specification and/or the nature of the response scale and thus, the GEE estimators might not be feasible. These motivated Touloumis et al. (2013) to define $\boldsymbol{\alpha}$, the so-called "association" vector, as a "nuisance" vector that contains the marginalized local odds ratios structure. Instead of relying on the sample local odds ratios, they developed parsimonious and meaningful structures for both ordinal and nominal responses by utilizing the family of association models (Goodman, 1985).

The local odds ratios GEE approach has certain advantages over ordinary GEE approaches. First, it allows marginal modeling of correlated multinomial responses regardless of the response scale. Second, the marginalized local odds ratios are not restricted by the regression parameters and values

of $\boldsymbol{\alpha}$ indicating strong marginalized association patterns are viable independently of the marginal model specification. Third, GEE pitfalls are avoided because $\widehat{\boldsymbol{\alpha}}$ is defined as the unique maximizer of an objective function and not as the solution of an extra set of estimating equations. Finally compared to the independence 'working' model, that is treating all observations as independent, the proposed GEE method seems to increase the efficiency in estimating the marginal regression parameters.
We present the key features of the local odds ratios GEE approach in Section 2, indicate software availability in Section 3 and illustrate its use for correlated nominal multinomial responses in Section 4.

2 The Local Odds Ratios GEE Approach

For notational ease, consider balanced designs but note that the local odds ratios GEE approach can also handle unbalanced designs that satisfy the missing completely at random assumption. Let $Y_{ij} \in \{1, 2, \ldots, C > 2\}$ be the multinomial response variable of observation j $(j = 1, \ldots, J)$ in cluster i $(i = 1, \ldots, N)$ and transform Y_{ij} to the equivalent vector $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ij(C-1)})^T$, where $Y_{ijc} = I(Y_{ij} = c)$. Let $\mathbf{x}_{ij}$ be the covariates matrix associated with observation j in cluster i and let $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \ldots, \mathbf{x}_{iJ}^T)^T$.
Suppose a marginal multinomial generalized linear model holds

$$\mathbf{g}\left[E(\mathbf{Y}_{ij}|\mathbf{x}_i)\right] = \mathbf{g}(\boldsymbol{\pi}_{ij}) = \mathbf{g}(\pi_{ij1}, \ldots, \pi_{ij(C-1)}) = \mathbf{x}_{ij}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is the p-variate parameter vector of interest, $\pi_{ijc} = \Pr(Y_{ij} = c|\mathbf{x}_i)$ and $\mathbf{g}$ is the link vector that respects the nature of the response scale. For example, cumulative link models or adjacent category logit models can be used for ordinal responses and baseline category logit models for nominal. Touloumis et al. (2013) derived the GEE estimator $\widehat{\boldsymbol{\beta}}_G$ by solving

$$\mathbf{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{D}_i \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\pi}_i) = \mathbf{0}$$

where $\mathbf{Y}_i = (\mathbf{Y}_{i1}^T, \ldots, \mathbf{Y}_{iJ}^T)^T$, $\boldsymbol{\pi}_i = E(\mathbf{Y}_i|\mathbf{x}_i)$, $\mathbf{D}_i = \partial\boldsymbol{\pi}_i/\partial\boldsymbol{\beta}$, $\widehat{\boldsymbol{\alpha}}$ is an estimator of $\boldsymbol{\alpha}$, the parameter vector that describes the marginalized local odds ratios structure, and $\mathbf{V}_i = \mathbf{V}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})$ is a $J(C-1) \times J(C-1)$ 'weight' matrix that preserves the form of the true covariance matrix for cluster i. Asymptotically, $\widehat{\boldsymbol{\beta}}_G$ follows the normal distribution provided by Liang and Zeger (1986).
Formally, $\boldsymbol{\alpha}$ includes the marginalized local odds ratios structure $\{\theta_{jcj'c'}\}$, where $\theta_{jcj'c'}$ is the local odds ratio at the cutpoint (c, c') of the contingency table obtained by aggregating the j-th and j'-th observations from each cluster as if no covariates were recoded. Further, they parametrized $\log\theta_{jcj'c'}$ according to one of the following four structures:

1. The uniform structure, $\log \theta_{jcj'c'} = \phi$, where a single parameter measures the association at the full marginalized contingency table.

2. The category exchangeability structure, $\log \theta_{jcj'c'} = \phi_{jj'}$, where one parameter measures the association at each marginalized contingency table. The uniform and the category exchangeability structure are meaningful only for ordinal responses.

3. The time (observation-pair) exchangeability structure, $\log \theta_{jcj'c'} = \phi(\mu_c - \mu_{c+1})(\mu_{c'} - \mu_{c'+1})$, where homogeneous score parameters $\{\mu_c\}$ for the response categories must be estimated.

4. The RC structure, $\log \theta_{jcj'c'} = \phi_{jj'}(\mu_c^{jj'} - \mu_{c+1}^{jj'})(\mu_{c'}^{jj'} - \mu_{c'+1}^{jj'})$ which assumes homogeneous score parameters at each marginalized contingency table. Clearly, the RC structure includes the previous ones as special cases. The time exchangeability and the RC structure is applicable to ordinal and nominal response categories provided that no monotonicity of the score parameters is required.

To address the problem of local odds ratios structure selection, one might inspect the so-called intrinsic parameters $\{\phi_{jj'}\}$ under the *RC* local odds ratios structure. For ordinal (nominal) response categories, it is advisable to use the uniform (time exchangeability) instead of the category exchangeability (RC) structure only when the estimated intrinsic parameters are similar. Simulations in Touloumis et al. (2013) showed that this strategy led to efficiency gains as large as 60% compared to the independence 'working' model.

3 R Package multgee

The local odds ratios GEE approach is implemented at the R package **multgee**. Unlike existing GEE routines, this package offers marginal models for both ordinal and nominal correlated responses. To highlight this, two GEE functions are provided: *ordLORgee* for analyzing ordinal responses and *nomLORgee* for analyzing nominal responses. Additionally, the utility function *intrinsic.pars* can be used in order to choose the local odds ratios structure according to the rule of thumb presented in Section 2.

4 Data Analysis

The San Diego McKinney Homeless Research Project (Hurlburt et al., 1996) was a randomized longitudinal study aimed to evaluate the effectiveness of Section 8 certificates in providing independent housing to severely mentally ill homeless people. People with Section 8 certificates were asked to pay 30% of their income to rent and the local authorities paid the

difference. The repeated response was the housing status, classified on a three-level nominal scale (1=street/shelter living, 2=community housing, 3=independent housing) and measured at baseline and at 6, 12 and 24 months follow-ups (t_j). We restricted our analysis to complete cases and fitted the baseline category logit model

$$\log\left(\frac{\pi_{ijc}}{\pi_{ij3}}\right) = \beta_{0c} + \beta_{1c}t_j + \beta_{2c}t_j^2 + \beta_{3c}s_i + \beta_{4c}(t_j \times s_i) + \beta_{5c}(t_j^2 \times s_i) \quad (1)$$

where π_{ijc} is the probability that subject i at the j-th time point has housing status c and s_i is the Section 8 group indicator, for $i = 1, \ldots, 271$, $j = 1, 2, 3, 4$ and $c = 1, 2$.

We selected the RC structure to model the marginalized local odds ratios because the range of the estimated intrinsic parameters ($0.56 - 3.66$) indicated an underlying time-dependent association pattern. For comparison reasons, Table 1 also displays the GEE estimates of $\boldsymbol{\beta}$ in (1) under the independence 'working' model. Interestingly, the Section 8 group effect in the contrast of community housing versus independent is significant under the RC local odds ratios structure (p-value $= 0.045$) but not under the independence 'working' model (p-value $= 0.166$). This justifies our preference in drawing inference based on the RC local odds ratios structure.

People in the Section 8 group were more likely to be in independent housing than in street or community housing compared to people in the control group. At the end of the study, for example, the estimated odds of independent housing instead of street/shelter living for the Section 8 group

TABLE 1. GEE estimates and standard errors (in parenthesis) for the baseline category logit model. Bold indicates statistical significance at $\alpha = 0.05$.

		Local Odds Ratios Structure	
Contrast	Parameter	Independence	RC
Street	β_{01}	**1.5612** (0.2752)	**1.5586** (0.2721)
vs	β_{11}	**-0.3695** (0.0557)	**-0.3688** (0.0548)
Independent	β_{21}	**0.0103** (0.0020)	**0.0103** (0.0019)
	β_{31}	**-0.6911** (0.3480)	**-0.8438** (0.3346)
	β_{41}	-0.0741 (0.0784)	-0.0457 (0.0794)
	β_{51}	0.0036 (0.0028)	0.0022 (0.0028)
Community	β_{02}	**1.2510** (0.2824)	**1.2479** (0.2811)
vs	β_{12}	-0.0608 (0.0465)	-0.0598 (0.0455)
Independent	β_{22}	0.0003 (0.0016)	0.0003 (0.0015)
	β_{32}	-0.4884 (0.3527)	**-0.6866** (0.3432)
	β_{42}	**-0.2811** (0.0652)	**-0.2054** (0.0616)
	β_{52}	**0.0107** (0.0023)	**0.0081** (0.0021)

TABLE 2. Fitted proportions for the housing status across time.

Group	Status	Baseline	6 months	12 months	24 months
Control	Street	0.515	0.179	0.083	0.115
	Community	0.377	0.584	0.586	0.439
	Independent	0.108	0.237	0.331	0.446
Section 8	Street	0.426	0.152	0.064	0.087
	Community	0.365	0.276	0.183	0.250
	Independent	0.208	0.572	0.753	0.663

were ≈ 1.96 times that for the control group, while the estimated odds of independent housing instead of community housing for the Section 8 group were ≈ 2.60 times that for the control group. The fitted proportions, shown in Table 2, reveal that people in the control group tended to move to community housing right after the baseline and were more likely to obtain independent housing only at the end of study, while people in the Section 8 group obtained independent housing faster. We conclude that the Section 8 certificate was effective in helping people to obtain independent housing.

5 Summary

We discussed a local odds ratios GEE approach which can handle both ordinal and nominal multinomial responses. We illustrated the approach with nominal responses by analyzing the housing dataset, while Touloumis et al. (2013) provided an example with ordinal responses. Touloumis (2013) describes the associated R package **multgee** in a greater detail.

Acknowledgments: This research was based on a collaboration with Prof. Alan Agresti and Prof. Maria Kateri.

References

Goodman, L.A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, **13**, 10 – 69.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13 – 22.

Touloumis, A. (2013). R Package multgee: A Generalized Estimating Equations Solver for Multinomial Responses. *In preparation.*

Touloumis, A., Agresti, A. and Kateri, M. (2013). GEE for multinomial responses using a local odds ratios parameterization. *Biometrics, to appear*.

Through the JASA's Looking-Glass, and What We Found There

Matilde Trevisani[1], Arjuna Tuzzi[2]

[1] University of Trieste, Italy
[2] University of Padua, Italy

E-mail for correspondence: arjuna.tuzzi@unipd.it

Abstract: This study is aimed at identifying sequential patterns for key-words included in chronological corpora and at grouping these word patterns in clusters. A model-based clustering procedure is proposed, with reference to titles of papers published by JASA (1888-2012).

Keywords: chronological corpora, correspondence analysis, curve clustering, functional data analysis, scientific literature

1 Introduction

The *American Statistical Association* is the world's largest community of statisticians and the *Journal of the American Statistical Association* (JASA) has long been considered the premier journal of statistical science. Established in 1888, formerly known as *Publications of the American Statistical Association* (PASA, 1888-1912) and *Quarterly publications of the American Statistical Association* (QASA, 1912-1921), JASA (1922- current) is published quarterly and focuses on statistical applications, theory, and methods in economic, social, physical, engineering, and health sciences.
In this study we explored the opportunities of reading the temporal evolution of concepts, methods and applications, *i.e.*, the history of Statistics, by means of the temporal evolution of key-words included in the papers published by JASA.
In this first attempt we considered the titles of the papers published in the period 1888-2012 in order to retrieve which were in the past and which are today the research fields that JASA covers, from the viewpoint of both methods and application domains.
Main aims of this study are: achieving an overview of the relationship between time and vocabulary to check the existence of a latent temporal pattern (correspondence analysis), identifying key-words showing prototypical temporal patterns, and clustering key-words portraying similar temporal patterns (model-based curve clustering for functional data).

2 Corpus and Data Analysis

Integrating three on-line sources (ASA, ISI, JSTOR) we collected 12,557 items along 125 years, from the first volume No. 1 issue No. 1 (PASA, 1888) to the last volume No. 107 issue No. 500 (JASA, 2012). Some of these items are not articles (*e.g.*, *List of publications*, *News*), some of them do not include content words in the title (*e.g.*, *Comment*, *Rejoinder*) and, since many of them are papers of the past, they often do not include an abstract (abstracts began to appear regularly in the forties). Taking into account only the texts of titles including content words, our corpus is composed of 10,077 titles, 87,060 word-tokens and 7,746 word-types, *i.e.*, this is a small corpus with a relatively limited richness, given that the type-token ratio (V/N) is 8.9% and the number of occurences of word-types (N/V) is 11 on average.

To overcome some of the limitations of an analysis based on simple word-types (forms), we chose analyses based on a recent development of the Porter's stemming algorithm (Porter, 1980); and obtained a new vocabulary including 4834 stem-types (*e.g.*, the word-types: model, models, modeling, and modelling are replaced with the same stem *model*). Moreover, as texts are sequences of words that have different meanings if they are considered in their context of use and alongside the adjacent words, we identified all stem-segments (*i.e.*, sequences of stems, *e.g.*, *model select*, *addit model*, *hierarch model*, *log linear model*, *dynam model*) occurring in the corpus at least twice and composed of minimum 2 and maximum 6 consecutive stems. We sorted the most important stem-segments according to Morrone's IS indexes (Morrone, 1996).

We tagged relevant statistical key-words (stem-types and stem-segments) matching our vocabulary with lists of items included in available on-line Statistics glossaries and, finally, we selected all key-words with frequencies equal to or higher than 10 (*i.e.*, high-frequency stem-types and stem-segments) and discarded function words, *i.e.*, articles, conjunctions, prepositions, pronouns, auxiliary and modal verbs (previous studies show that these are good markers for the writing style whilst content words are suitable to gather topics, *cfr.* Tuzzi, 2010).

In a typical bag-of-words approach, chronological data are organized in words per time-points contingency tables, which show the occurrence of each word at each time-point. In our corpus, the time-point represents a subcorpus of titles published in the same volume, corresponding to the year (or two years for first volumes) of publication. In order to position years and key-words (stem-types and stem-segments) on a map, we conducted on this table a content analysis by means of a classical (lexical) Correspondence Analysis (Murthag, 2005). Figure 1 represents the position of the years on the Cartesian plane generated by the first two factors and shows association among key-words (dots), among years (triangles) and between key-words and years.

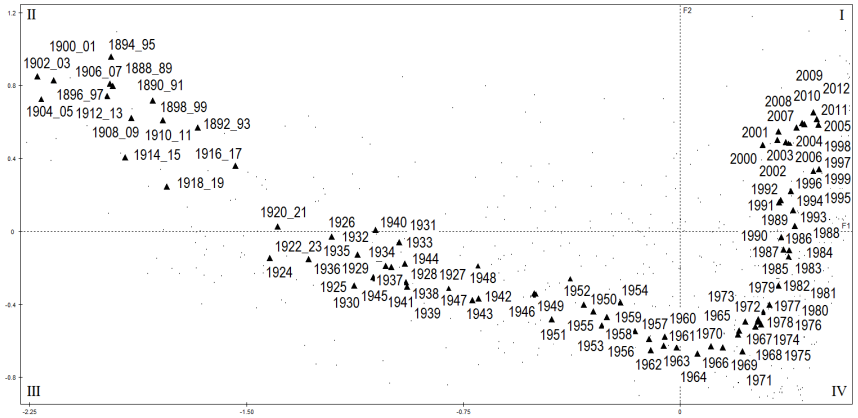


FIGURE 1. First factorial plan of correspondence analysis (30% i.e.).

The temporal evolution of a key-word is expressed by the sequence of its occurrences over years (the rows of the contingency table), and represents an observation of a functional object. In order to reduce the effect of different subcorpus dimension (number of titles available) across years we replaced the frequencies with a politextuality rate (number of titles including the key-word divided by the number of titles). Figure 2 shows the temporal trajectories of some key-words: (top) the most frequent; (bottom-left) identifying main inferential approaches; (right) related to topics of demography and population studies.

In our modeling approach, the temporal evolution of each word i is represented by a curve $y_i(t)$ observed on a set of equally spaced time-points. We assume that the model which generated the data involves, for each curve, a functional effect and a random measurement error term. As we suppose the existence of clusters, each word functional effect should include a cluster-specific functional effect and, in order to handle inter-word variability, an individual functional effect. Thereby we assume a mixed model:

$$y_i(t) = \mu_l(t) + U_i(t) + E_i(t) \tag{1}$$

where $\mu_l(t)$ represents a functional fixed effect that is related to the cluster l to which the word belongs, and $U_i(t)$ is an individual functional random effect modeled as a centered Gaussian process independent from $E_i(t)$, a random measurement error.

Once defined in the functional domain, a classical approach is to convert the original problem into a finite-dimensional one by means of a functional basis representation of the model. Following Giacofci et al. (2013) and Morris and Carroll (2006) we used a wavelet-based representation of the model and the Discrete Wavelet Transform to consider continuous functions on the sole set of sampled points. Several ways of modeling the variance of

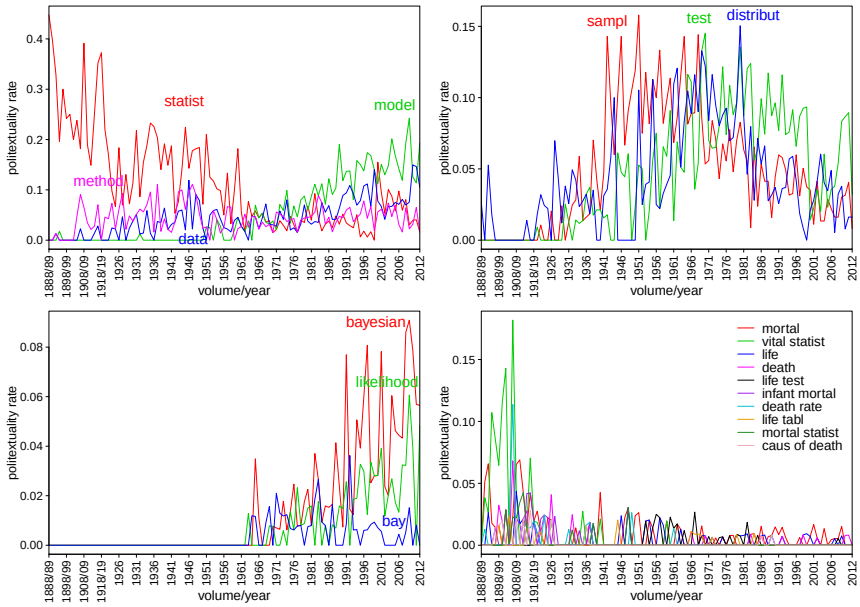


FIGURE 2. Temporal trajectories for some key-words-within-titles.

random effects (constant, group-specific or scale-location varying) are also considered.

3 Results and Conclusions

By means of correspondence analysis the corpus of titles shows a chronological pattern and four different eras in the history of statistics emerge from the first factorial plan: from the origins to the twenties and World War I (II quadrant); from the twenties to World War II and, then, to early sixties (III); from the sixties to early eighties (II); from the late eighties to nowadays (I). Moreover, it shows a progressive reduction in variability along time, *i.e.* the scientific language has become more technical and more specialized in recent decades.

By means of model-based curve clustering we observe the existence of some interesting groups of key-words portraying a similar temporal evolution. We found a first method capable of dealing with irregular curves (peak-like data), the presence of inter-individual (inter-word) variability, and high dimensional curve clustering. At present, a functional clustering mixed model with 10 clusters and constant variance for functional random effects shows the best performance under statistical (according to BIC and ICL criteria) and subject-matter considerations. This means that the selected 900 key-words have a cluster structure and a relevant inter-word variability.

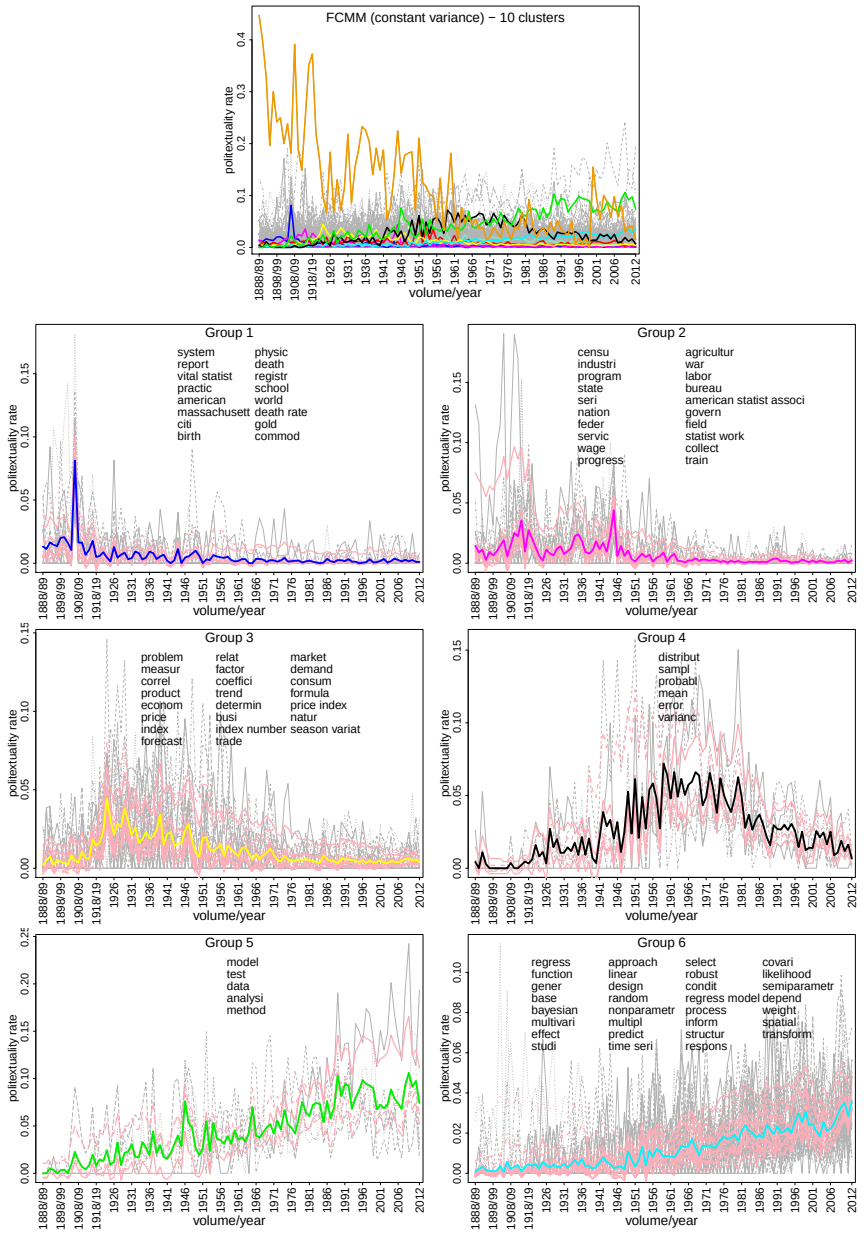


FIGURE 3. The best model with the overall groups (top) and examples of interesting clusters (6 out of 10).

If we exclude clusters with a somewhat flat, undifferentiated trend, and one outlier, other clusters allows us to pass from quantitative analysis to

qualitative reading of results. Some examples are given in Figure 3 where the group-specific functional fixed effect (bold line) and functional random effects (pink lines) are illustrated. The top graph shows the best model with the overall groups. The following six examples of clusters display interesting chronological backgrounds: *Group 1* includes key-words related to demography and population studies (*e.g.*, *birth*, *death*, *vital statistics*, *death rate*, *school*) that refer mainly to the oldest articles at the turn of the century; *Groups 2* and *3* include key-words related to, respectively, public (*e.g.*, *census*, *state*, *federal*, *national*, *bureau*, *government*) and economic (*e.g.*, *product*, *economic*, *price index*, *forecast*, *market*, *trade*, *business*, *consumer*) statistics that were most frequently addressed in the journal during the first decades of the last century and almost disappeared after the sixties; *Group 4* represents the golden age of probability and inference that dominated the second half of the twentieth century; *Groups 5* and *6* show the on going development of, respectively, mainstays (*e.g.*, *model*, *test*, *data*) and methods (*e.g.*, *regression model*, *multivariate*, *time series*, *robust*, *bayesian*, *likelihood*, *nonparametric*, *semiparametric*) of modern statistical sciences.
Although at present our analysis should be considered purely explorative, model-based curve clustering proved promising and one of our research aims is to adopt the same perspective fot analyzing abstracts of recent articles in order to detect which research fields JASA covers nowadays from both the viewpoints of methods and application domains.

References

Giacofci, M., Lambert-Lacroix, S., Marot, G. and Picard, F. (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, **69**, 1, 31 – 40.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *J. Roy. Stat. Soc. B Met.*, **68**, 179 – 199.

Morrone, A. (1996). Temi generali e temi specifici dei programmi di governo attraverso le sequenze di discorso. In: *L'attivit dei governi della repubblica italiana (1947-1994)*, Villone M. and Zuliani A. (Eds.), Bologna: Il Mulino, 351 – 369.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R*. London: Chapman & Hall/CRC.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, **14**, 3, 130 – 137.

Tuzzi, A. (2010). What to put in the bag? Comparing and contrasting procedures for text clustering. *Statistica Applicata - Italian Journal of Applied Statistics*, **22**, 1, 77 – 94.

Modeling the risk of *Borrelia* infection after a tick bite - a Bayesian approach

Jan van de Kassteele[1], Agnetha Hofhuis[2], Peter Teunis[23], Wilfrid van Pelt[2]

[1] Department of statistics, mathematical modeling and data logistics, National Institute for Public Health and the Environment, Bilthoven, the Netherlands
[2] Epidemiology and surveillance unit, Centre for Infectious Disease Control Netherlands, National Institute for Public Health and the Environment, Bilthoven, the Netherlands
[3] Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

E-mail for correspondence: `jan.van.de.kassteele@rivm.nl`

Abstract: Using Bayesian statistical modeling, we investigate risk factors of *Borrelia* infection after a tick bite, such as tick testing for infection with *Borrelia*, and assessment of the duration of the tick's blood meal. We use combined data from two prospective tick bite studies. The data are a mix of real observed and interval censored data, and for some variables, only the maximum is reported. Apart from tick infection with *Borrelia*, tick attachment time and, implicitly, tick engorgement, appear to be an important predictors for *Borrelia* infection.

Keywords: Bayesian modeling; Survival analysis; Tick bite; Risk assessment

1 Introduction

Lyme borreliosis after a tick bite poses a progressive threat to public health [Hubalek, 2009]. To decide upon prophylactic treatment with antibiotics, understanding and quantification of an individual's risk would be of great value.

The individual risk for *Borrelia* infection depends on several factors: the tick infection rate with *Borrelia*, the transmission rate of *Borrelia* from ticks to humans, and tick attachment time. This last one can also be measured as tick engorgement.

To assess the transmission rate of *Borrelia* from ticks to humans, we combine the data from two prospective tick bite studies in the Netherlands. Given the different datatypes of the two studies, modeling the risk of *Borrelia* infection after a tick bite, with regard to tick infection with *Borrelia*, tick attachment time, and tick engorgement, may be problematic using standard statistical techniques. We therefore turn to Bayesian statistical methods.

2 Material and methods

2.1 Data description

Table 1 shows a summary of both datasets and the combined dataset. Information on the number of tick bites and tick attachment duration was collected at three time points: baseline, the six weeks preceding to enrollment, and after a three month follow-up period. Some variables are available as real valued observations in one study, but are observed in categories in the other. Many data are missing.

TABLE 1. Number of records of each variable in Study 1, Study 2 and combined dataset for baseline, historic (six weeks preceding to baseline) and follow-up (three months after baseline). "NA" = missing value. "-" = variable not measured (also NA). "real" = real valued, i.e. not interval censored, observation.

		Study 1			Study 2			Combined		
		baseline	historic	follow-up	baseline	historic	follow-up	baseline	historic	follow-up
Borrelia	0	-	-	260	-	-	186	-	-	446
infection	1	-	-	14	-	-	6	-	-	20
at follow-up	NA	-	-	53	-	-	72	-	-	125
Number	real	327	251	279	264	-	-	591	251	279
of tics	0	0	225	237	0	-	180	0	225	417
	1-3	327	21	35	263	-	11	590	21	46
	4-10	0	5	6	1	-	1	1	5	7
	>10	0	0	1	0	-	0	0	0	1
	NA	0	76	48	0	264	72	0	340	120
Tick	real	295	-	-	-	-	-	295	-	-
attachment	0-12	86	-	-	92	-	-	178	-	-
duration	0-24	-	16	26	-	-	-	-	16	26
(hours)	13-24	126	-	-	82	-	-	165	-	-
	>24	126	5	13	83	-	-	209	5	13
	NA	32	306	288	7	264	264	39	570	552
Tick	empty	87	-	-	73	-	-	160	-	-
engorgement	partially	96	-	-	101	-	-	197	-	-
category	fully	56	-	-	20	-	-	76	-	-
	NA	88	327	327	70	264	264	158	591	591
Tick test	0	183	-	-	202	-	-	385	-	-
Borrelia-	1	77	-	-	44	-	-	121	-	-
positive	NA	67	327	327	18	264	264	85	591	591

If an individual encountered more than one tick at any time point, only the maximum is reported. For example, if a person encountered four ticks at baseline, and a *Borrelia*-positive tick was reported, then it is only known that at least one of them was positive.

2.2 Bayesian model

Given the observed data (real valued, interval censored or the maximum value) and making distributional assumptions on the variables and model parameters, inference can be made about the all unknown parameters as well missing data.

The problem can be formulated as a survival model: during tick attachment, a person is at risk of becoming infected. There is loss to follow-up in case the tick is removed before infection has occurred, so in that case the time

to event is right censored. For infected individuals, it is only known that the event occurred after tick attachment, so the time to event is interval censored.
The survival model can be reformulated as a Binomial regression model with 0/1 outcome [Abbott, 1985], where the probability of infection is a function of the risk factors. Lyme infection status I_i (1 or 0) of individual i ($i = 1, ..., 591$) is modeled by a Bernoulli distribution:

$$I_i \sim Bern(p_{inf_i})$$

The cumulative risk p_{inf_i} that individual i is infected is a function of the baseline risk, historical risk and follow-up time risk:

$$p_{inf_i} = 1 - \left(1 - p_{inf_i}^{base}\right)\left(1 - p_{inf_i}^{hist}\right)\left(1 - p_{inf_i}^{flwp}\right)$$

The cumulative risks are modeled as functions of the number of ticks K, tick attachment duration T, tick engorgement category E, and tick infection with *Borrelia* B. As for any individual only the maximum tick attachment duration T_{max} and maximum tick engorgement category E_{max} were reported, it is assumed that the not reported attachment duration and engorgement category of all other observed ticks are equal to the reported maximum attachment duration and maximum engorgement category.
The efficiency of *Borrelia* transmission is described by the hazard rate λ. As observed in animal experiments, *Borrelia* transmission does not occur at the beginning of the blood uptake [Piesman, 1993], so we set $\lambda = 0$ during an initial period τ of tick attachment. Additionally, we assume that, if a tick tested positive for *Borrelia* ($B = 1$), the hazard rate is constant during tick attachment after τ. Otherwise, if a tick is tested negative ($B = 0$), the hazard rate is assumed zero. As a result, the cumulative risk is given by:

$$p_{inf_i}^{base} = 1 - \exp\Big(- \lambda \max(0, T_{max_i}^{base} - \tau) \sum_{j=0}^{K_i^{base}} B_{ij}^{base} \Big)$$

$$p_{inf_i}^{hist} = 1 - \exp\Big(- \lambda \max(0, T_{max_i}^{hist} - \tau) \sum_{j=0}^{K_i^{hist}} B_{ij}^{hist} \Big)$$

$$p_{inf_i}^{flwp} = 1 - \exp\Big(- \lambda \max(0, T_{max_i}^{flwp} - \tau) \sum_{j=0}^{K_i^{flwp}} B_{ij}^{flwp} \Big)$$

If no tick bites were observed ($K_i = 0$) at one of the three time points, the risk of infection should be zero. This is achieved by setting $T_{max_i} = 0$ and $B_{i0} = 0$ for that time point.
We assume that the maximum tick attachment duration has an Exponential distribution, as this distribution provided the best fit with the observed

maximum tick attachment durations (QQ-plot, figure not shown).

$$T_{max_i}^{base} \sim Exp(\lambda_{T_i}^{base})$$
$$T_{max_i}^{hist} \sim Exp(\lambda_{T_i}^{hist})$$
$$T_{max_i}^{flwp} \sim Exp(\lambda_{T_i}^{flwp})$$

If the maximum tick attachment duration was observed in intervals, the Exponential distribution can be made interval censored for those observations.
The relation between maximum tick engorgement category and the cumulative risk is modeled implicitly by relating the rate parameters of the Exponential distribution for maximum attachment times to the maximum engorgement category E_max:

$$\lambda_{T_i}^{base} = \lambda_{E_1}(E_{max_i}^{base} = 1) + \lambda_{E_2}(E_{max_i}^{base} = 2) + \lambda_{E_3}(E_{max_i}^{base} = 3)$$
$$\lambda_{T_i}^{hist} = \lambda_{E_1}(E_{max_i}^{hist} = 1) + \lambda_{E_2}(E_{max_i}^{hist} = 2) + \lambda_{E_3}(E_{max_i}^{hist} = 3)$$
$$\lambda_{T_i}^{flwp} = \lambda_{E_1}(E_{max_i}^{flwp} = 1) + \lambda_{E_2}(E_{max_i}^{flwp} = 2) + \lambda_{E_3}(E_{max_i}^{flwp} = 3)$$

The maximum tick engorgement category has a Categorical distribution with prevalences p_E:

$$E_{ij}^{base} \sim Cat(p_E)$$

The tick infection with *Borrelia* is modeled for each tick j separately, as this is assumed independent of the individual. Consequently, the tick infection with *Borrelia* has a Bernoulli distribution:

$$B_{ij}^{base} \sim Bern(p_B)$$

under the restriction that only the maximum B_{max} is reported:

$$B_i^{base} = \max_j(B_{ij}^{base})$$

From this the prevalence of a single *Borrelia*-positive tick p_B can be estimated, which can be used to generate data on tick infection with Borrelia of any historical or follow-up tick:

$$B_{ij}^{hist} \sim Bern(p_B)$$
$$B_{ij}^{flwp} \sim Bern(p_B)$$

The same principle, of generating data by sampling from their distribution, conditional its parameter(s), can be applied to any missing values for tick *Borrelia* infection status, maximum tick attachment duration and maximum tick engorgement category.

The number of ticks K is described by a categorical distribution, possibly interval censored if the observed number of ticks is reported as a range.

$$K_i^{base} \sim Cat(p_K^{base})$$
$$K_i^{hist} \sim Cat(p_K^{hist})$$
$$K_i^{flwp} \sim Cat(p_K^{flwp})$$

Finally, (non-informative) priors are assigned to all unknown parameters:

$$\log(\lambda) \sim Norm(0, 0.01) \qquad p_E \sim Dirch(1, 1, 1)$$
$$\tau \sim Gamma(2, 0.05) \qquad p_B \sim Beta(1, 1)$$
$$\lambda_{E_1} \sim Gamma(1, 0.01) \qquad p_K^{base} \sim Dirch(1, 1, ..., 1)$$
$$\lambda_{E_2} \sim Gamma(1, 0.01) \qquad p_K^{hist} \sim Dirch(1, 1, ..., 1)$$
$$\lambda_{E_2} \sim Gamma(1, 0.01) \qquad p_K^{flwp} \sim Dirch(1, 1, ..., 1)$$

The model is implemented in JAGS. All data pre- and post-processing are done in R.

3 Results

We first present parameter estimates together with their 95% credible interval. The hazard rate is 0.0040 h^{-1} (0.0024 - 0.0063). The initial period without blood transmission is 4.3 h (0.7 - 10.2). The mean attachment duration per engorgement category is 21 h (19 - 25), 26 h (22 - 30), and 68 h (54 - 87) for empty, partially and fully engorged ticks, respectively. The probability that one tick is Borrelia positive is 23.8% (20.3 - 27.4).

TABLE 2. Cumulative probability (%) of *Borrelia* infection at different time points and tick engorgement categories.

Engorgement	*Borrelia*-positive	*Borrelia* unknown
Empty	7.1 (0.0 - 28.5)	1.7 (0.0 - 6.8)
Partial	7.7 (0.0 - 30.3)	1.8 (0.0 - 7.1)
Full	20.2 (0.0 - 64.0)	4.8 (0.0 - 15.8)

Given these parameter estimates, Figure 1 is constructed, showing the cumulative risk of Borrelia infection after one tick bite, as a function of tick attachment duration and tick infection status. The influence of engorgement on the risk of *Borrelia* infection is shown in Table 2, where the values in Figure 1 are averaged over the corresponding tick attachment durations.

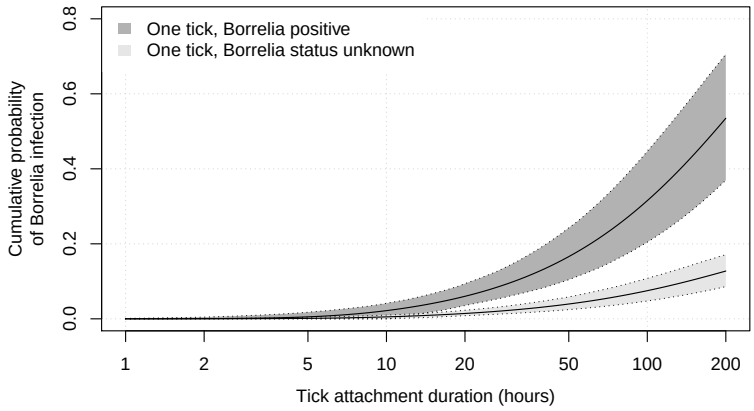


FIGURE 1. Cumulative probability of *Borrelia* infection as a function of tick attachment duration and tick infection status. The dark curve represents the probability in case of one *Borrelia*-positive tick. The light curve represents the probability in case of one tick bite, without information on tick infection with *Borrelia*.

4 Conclusion

Using Bayesian statistical modeling, we can combine information from two prospective studies to estimate the the risk of *Borrelia* infection after a tick bite, with regard to tick infection with *Borrelia*, tick engorgement and tick attachment time, in the presence of missing data, a mix of real observed and interval censored data, and the limitation that only the maximum of some observations is reported. Tick attachment time and, implicitly, tick engorgement, appears to be an important predictor for *Borrelia* infection. If a tick is known to be *Borrelia*-positive, the risk increases with a factor four.

References

Abbott, R.D. (1985). Logistic regression in survival analysis. *Am. J. Epidemiol.*, **121**(3), 465-471.

Hubalek, Z. (2009). Epidemiology of lyme borreliosis. *Curr. Probl. Dermatol.*, **37**, 31-50.

Piesman, J. (1993). Dynamics of Borrelia burgdorferi transmission by nymphal Ixodes dammini ticks. *J. Infect. Dis.*, **167**, 1082-1085.

Joint models for discrete longitudinal outcome and survival

Ardo van den Hout[1], Graciela Muniz[2]

[1] Department of Statistical Science, University College London, UK
[2] Medical Research Council Unit for Lifelong Health and Ageing, London, UK

E-mail for correspondence: ardo.vandenhout@ucl.ac.uk

Abstract: Joint modelling is presented for survival and change in cognitive function in the older population. Because tests of cognitive function often result in discrete outcomes, binomial and beta-binomial mixed-effects regression models are applied to analyse longitudinal measurements. Dropout due to death is accounted for by parametric survival models, where the choice of a Gompertz baseline hazard and the specification of the random-effects structure is of specific interest. Estimation is by marginal likelihood or Bayesian inference. Data from the English Longitudinal Study of Ageing are analysed to illustrate the methods.

Keywords: Beta-binomial distribution; Cognitive function; Gompertz distribution, Marginal likelihood.

1 Introduction

Joint modelling is proposed for longitudinal data on survival and change in cognitive function in the older population. It is assumed that cognitive function is measured repeatedly with a questionnaire test, where performance is quantified by an integer scale for the total sum score. Typically such as test consists of a series of questions with a binary scoring 0/1 for correct/incorrect answers.
Given the population of interest, dropout due to death cannot be ignored when individuals are followed up with respect to a process that is associated with ageing and the proximity of death. Hence the need for joint models for survival and change in cognitive function.
A good overview of past and current research in joint models is presented in Chapters 13-16 of *Longitudinal Data Analysis* (Fitzmaurice et al. 2009). Our joint modelling builds upon the established framework of shared-parameter models as presented and discussed by Rizopoulos (2012). However, model formulation is geared up to the specific features of data for cognitive function in the older population. First, because of the integer test sum score, non-linear mixed-effects models are investigated as alternatives to linear mixed-effects models where the conditional distribution of

the outcome is continuous. In addition to the integer scale, tests of cognitive function often result in skewed distributions due to ceiling effects, and this undermines the assumptions of linear mixed models. Second, because prediction is of specific interest, parametric models will be formulated. The fully parametric approach is an alternative to the semi-parametric modelling of the survival time, which is the default in many joint models. The parametric Gompertz baseline hazard is especially useful in the context of ageing. Third, delayed entry is taken into account in the modelling of survival. Delayed entry is a common feature of longitudinal data for ageing research when age is the time scale and baseline age in the study varies. In that case, individuals are only in the data set if they survived up to or beyond the minimum age defined by the study design.

2 Model for discrete outcome and survival

The response of interest is discrete and takes values in $0, 1, ..., m$. For individual i, $i = 1, ..., N$, with longitudinal response $y_i = (y_{i1}, ..., y_{in_i})$ at times $(t_{i1}, ..., t_{in_i})$ the *measurement model* has linear predictor given by

$$\begin{aligned} \eta_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} + \mathbf{X}_i \gamma \\ \beta_{0i} &= \beta_0 + b_{0i} \\ \beta_{1i} &= \beta_1 + b_{1i}, \end{aligned}$$

for fixed-effects vectors (β_0, β_1) and γ. We assume $b_i = (b_{0i}, b_{1i}) \sim N(0, \mathbf{\Sigma})$. The logit link is $\mu_{ij} = \exp(\eta_{ij})/[1 + \exp(\eta_{ij})]$. The beta-binomial distribution for Y_{ij} with variance parameter θ has $\mathbb{E}[Y_{ij}] = m\mu_{ij}$ and $\mathrm{Var}[Y_i] = m\mu_{ij}(1 - \mu_{ij})[1 + (n_i - 1)\theta/(1 + \theta)]$. Here it is assumed that θ is the same unknown constant for all individuals.

The *hazard model* is a parametric regression model given by

$$h_i(t) = h_0(t) \exp\left[\alpha g(\beta_i) + \mathbf{X}_i^* \gamma^*\right], \tag{1}$$

where time t is age in years, and $g(\beta_i) = g(\beta_i | t)$ is a linear function of random-effects vector $\beta_i = (\beta_{0i}, \beta_{1i})$. The vector γ^* is without an intercept. Examples of parametric forms for the baseline hazard function $h_0(t)$ are exponential: $h_0(t) = \lambda$, Weibull: $h_0(t) = \lambda\tau t^{\tau-1}$ and Gompertz: $h_0(t) = \lambda \exp(\xi t)$, where $\lambda > 0$ and $\tau > 0$.

The *joint model* is a shared-parameter model where the constituent submodels share the random effects. Data for the hazard model are t_{i1}, the time individual i enters the study, and t_i, the time at which either death is observed ($\delta_i = 1$), or death is right-censored ($\delta_i = 0$).

Given a hazard submodel where the baseline hazard depends parametrically on t, a simple form for $\alpha g(\beta_i)$ such as $\alpha\beta_{0i}$ is recommended to prevent identifiability problems with respect to α and the parameters for the baseline hazard.

In joint models where the hazard model is a semi-parametric Cox model, an additional choice for $\alpha g(\beta_i)$ is $\alpha(\beta_{0i} + \beta_{1i}t)$.
Alternatively, joint models can be formulated by linking the function $g(\beta_i)$ to the baseline hazard. For the Gompertz baseline we can define

$$h_i(t) = h_0(t|\beta_i)\exp(\mathbf{X}_i^*\gamma^*) = \lambda\exp(\xi_i t)\exp(\mathbf{X}_i^*\gamma^*), \tag{2}$$

where $\xi_i = \xi_0 + \alpha g(\beta_i)$. The induced heterogeneity is with respect to the effect of time on the hazard.

3 Statistical inference

Assuming independence between the submodels conditional on the random effects, the log-likelihood contribution for individual i is given by

$$\log p(t_i, t_{i1}, \delta_i, y_i|\omega) = \log \int p(t_i, t_{i1}, \delta_i|b_i, \omega)p(y_i|b_i, \omega)p(b_i|\omega)db_i =$$

$$\log \int h(t_i|b_i, \omega)^{\delta_i} P(T \geq t_i|T > t_{i1}, b_i, \omega) \left\{ \prod_j p(y_{ij}|b_i, \omega) \right\} p(b_i|\omega)db_i,$$

where ω is the vector with all the models parameters but the random effects. The log-likelihood allows for left-truncated data by taking into account t_{i1}. Let $\lambda_i^\circ = \lambda\exp[\alpha g(\beta_i) + \mathbf{X}_i^*\gamma^*]$. For the joint model (1) with the Gompertz hazard, we have

$$P(T \geq t_i|T > t_{i0}, b_i, \omega) = \exp\left(-\lambda_i^\circ \xi^{-1}[\exp(\xi t_i) - \exp(\xi t_{i0})]\right).$$

The log-likelihood is computed using Gaussian quadrature for the two-dimensional integral and maximised using a general-purpose optimiser in R.
Consider an individual i, who is not in the data, is alive, and has longitudinal response $\tilde{y}_i$. Let $\tilde{t}_i$ denote age corresponding to the last element of $\tilde{y}_i$, let $\tilde{t}_{i1}$ denote the age corresponding to the first element of $\tilde{y}_i$. Maximum aposteriori estimation (MAP) of random effects b_i for this individual is based upon the conditional density

$$p(b_i|\widehat{\omega}, \tilde{t}_i, \tilde{t}_{i1}, \delta_i = 0, \tilde{y}_i) \propto p(\tilde{t}_i, \tilde{t}_{i1}, \delta_i = 0, \tilde{y}_i|b_i, \widehat{\omega})p(b_i|\widehat{\omega}).$$

Given the estimated random effects, both survival and longitudinal response can be predicted up to an assumed maximum age.
As an alternative to marginal likelihood, Markov Chain Monte Carlo can be used to obtain posterior inference for the model parameters and the random effects. For this, a Gibbs sampler was implemented in R.

4 Application

The English Longitudinal Study of Ageing (`www.ifs.org.uk/ELSA`) contains information on health and quality of life for the English population aged 50 and older. Data are available from waves 1–4 (2002-2009).
Here we focus on the number of words remembered in a recall from a list of ten: "A little while ago, you were read a list of words and you repeated the ones you could remember. Please tell me any of the words that you can remember now." The test score is equal to the number of words remembered. The total sample size of ELSA is 19,834. For the current analysis, we use a random sample of size 1,000.
To protect the identity of the individuals in the ELSA data, ages higher then 90 years are censored. In the current sample, there are 6 individuals with censored age of death, and a further 7 individuals who have age censored during the follow-up. These individuals are removed from the sample. The resulting sample size is 987. The dropout due to death is 9.62% and too substantial to ignore in an analysis of cognitive function.
As a summary of the analysis and to show the benefits of using the beta-binomial and the Gompertz distributions, we compare the Akaike information criterion (AIC) for a series of models. For all the models, the Gaussian quadrature is based upon 11 quadrature points.
The time scale for the extended models is age minus 31. This transformation is used because it results in 1 being the minimal age in the analysis, which simplifies the interpretation of the intercepts and is numerically convenient for fitting the model. Denoting transformed age by t, the basic model is a binomial regression model with an exponential survival model given by

$$\begin{aligned} \eta_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} + \gamma_1 \texttt{sex}_i + \gamma_2 \texttt{educ}_i \\ h_i(t) &= \exp(\gamma_0^* + \alpha\beta_{0i}), \end{aligned}$$

where $\texttt{sex} = 1$ for men, and $\texttt{educ} = 1$ for the higher education level. The λ-parameter for the exponential baseline is estimated by $\exp(\gamma_0^*)$. This joint model has AIC = 12720. Model (1) defined by using a Gompertz baseline hazard specifies

$$h_i(t) = \exp(\gamma_0^* + \xi t + \alpha\beta_{0i}).$$

The AIC is 12634. Choosing the Weibull baseline results in AIC = 12666. As expected, the exponential model is too simple and a time-dependent hazard leads to a better fit. Comparing the Gompertz model with the Weibull model, the AIC favours the former. For this reason, the model with the Gompertz baseline is extended by adding a variance parameter to the measurement model and fitting a beta-binomial regression. This joint model yields AIC = 12631.

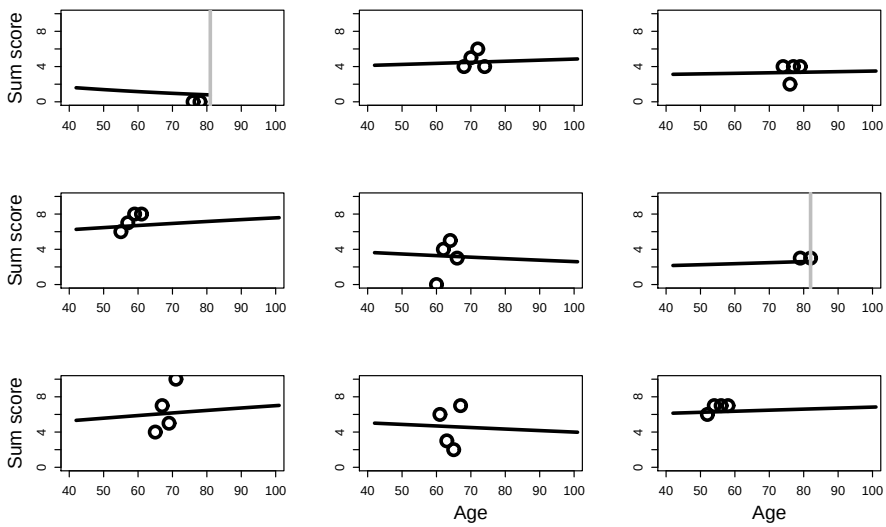


FIGURE 1. Observed number of words remembered and fitted trajectories for 9 individuals (more or less) randomly chosen among those with two or more observations. Grey vertical lines for time of death.

Adding covariates body mass index (`bmi`, a classification into 6 groups: < 20, 20-25, 25-30, 30-35, 35-40, and 40+) and year of birth (`yob`) defines

$$\begin{aligned} \eta_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} + \gamma_1 \texttt{sex}_i + \gamma_2 \texttt{educ}_i + \gamma_3 \texttt{yob}_i \\ h_i(t) &= \exp(\gamma_0^* + \xi t + \alpha \beta_{0i} + \gamma_1^* \texttt{bmi}), \end{aligned}$$

with AIC = 12616.
Using (2) as an alternative random-effects structure, the hazard model in the last model is changed into

$$h_i(t) = \exp(\gamma_0^* + (\xi + \alpha \beta_{0i}) t + \gamma_1^* \texttt{bmi}_i),$$

where the heterogeneity induced by β_{0i} affects the way the hazard is linked to change in t. The assumption in this model is that individuals with the same covariate information have the same hazard at $t = 0$, which corresponds in the present context to 30 years of age. This model has AIC = 12614 and fits better than all previous models.
Parameter α is estimated at a negative value implying that being better at remembering words is associated with better survival. Estimate $\widehat{\xi} > 0$ implies that the hazard for death increases with age. Both these results are as expected.
The estimation of the effect of time and risk factors is robust across the fitted models. There is a clear positive effect for more education ($\widehat{\gamma}_2 >$

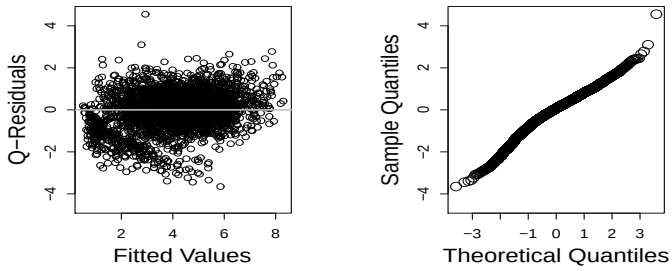


FIGURE 2. Quantile residuals based on MAP estimation of the random effects.

0). Given the effect of education, there is an additional effect when men (`sex = 1`) are compared to women (`sex = 0`): women tend to be better at remembering words than men ($\widehat{\gamma}_1 < 0$). A higher BMI increases the hazard, and being born later is associated with being better in remembering words. Marginal likelihood tends to underestimate the variance components in mixed-effects models. Because of this, Bayesian inference (using a Gibbs sampler and non-informative priors) was compared to the marginal likelihood inference. We did not note any relevant differences.

For the final model, fit to individual data is depicted in Figure 1, where the subset was chosen from a number of randomly generated subsets of individuals with more than one observation. In case of death, the fit is depicted up to time death. The graphs in Figure 1 are based on the MAP estimation of the random effects b_i. Overall the model seems to capture the observed trajectories well.

Given the estimated random effects, further information on goodness of fit of the measurement model can be derived by looking at normalised randomised quantile residuals (Dunn and Smyth 1996; Rigby and Stasinopoulos 2005). Figure 2 shows that the distribution of these residuals is close to the standard normal.

References

Dunn, P.K., Smyth, G.K. (1996). Randomised quantile residuals. *Journal of Computational and Graphical Statistics* 5, 236–244.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Editors) (2009). *Longitudinal Data Analysis*. Chapman & Hall/CRC.

Rigby, R.A, Stasinopoulos, D.M. 2005. Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54, 507–554.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*. Chapman & Hall/CRC.

Bayesian P-spline models for land use raster datasets

Massimo Ventrucci[1], Daniela Cocchi [1], Marian Scott[2]

[1] Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, 40126 Bologna, Italy
[2] School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK

E-mail for correspondence: massimo.ventrucci@unibo.it

Abstract: Land use raster datasets provide useful information for monitoring the development of urban agglomerates. Statistical modelling for raster data is challenged by the high dimensionality of raster maps which requires efficient estimation tools. In this work we propose a Bayesian P-spline modelling framework for binary raster data and some model based tools for identifying boundary changes across space and time.

Keywords: raster; Bayesian P-spline; boundary; urban sprawl.

1 Introduction

Land use datasets are routinely collected by many institutions at a national and international level and provide useful information for many environmental and urban planning purposes, such as the monitoring of urban sprawl (EEA, 2006). The availability of urban land use data collected across several years also allows changes in the evolution of urban agglomerates and their features to be tracked over time.
One objective for urban planners is the identification and monitoring of homogeneous areas characterized by certain types of urban use, for instance residential rather than industrial or commercial use. The boundary region separating an homogeneous urban area from a non urban one is typically characterized by a smooth decay in the *intensity* of urbanization. This is because, at a large scale level, urbanization develops as a continuous process over space, and often the spatial configuration gradually becomes non compact as one moves from the urban centre; this effect is sometimes referred to as urban sprawl (EEA, 2006). This characteristic of the urban process makes it difficult to localize the presence of boundaries or significant changes by simply looking at the observed land use pattern. In this work we identify the position of boundaries separating homogeneous regions,

by searching for significant changes over a fitted surface representing the intensity of the urban phenomena.
We adopt a Bayesian P-spline approach for modelling the surface. Via MCMC we obtain posterior samples of spline coefficients, which is to say the empirical distribution of fitted values is available for computation of the mean surface as well as any α *quantile surface* of interest. By analyzing the posterior empirical distribution of the fitted surface we can investigate significant changes over space, such as identifying the boundary surrounding a hot spot of urbanization, and also investigating possible changes in the boundary over time. This is important to reveal the temporal dynamics of urban development in a given study region.

1.1 Raster data

Land use data are usually collected by satellite imaging or aerial photos. Data come in the form of polygons, each having an associated category of land use. An equivalent raster map can be obtained from the polygon maps by using GIS softwares or the"raster" R package (Hijmans and van Etten, 2012). The new raster dataset consists of a fine grid composed of pixels. Each pixel inherits the category of land use of the polygon in which it falls. In order to preserve the spatial pattern observed in the original polygon map, raster maps must be produced at high resolution resulting in a very large number of pixels.
A Bayesian P-spline logistic model for raster data is described in section 2, where intensity of urbanization is modelled as a smooth surface. In section 3 a case study is presented with the main aim of delineating boundaries for regions characterized by a high urbanization intensity (hot spot) and also to investigate changes in the boundary between two different years. A discussion of further work is given at the end.

2 P-spline Bayesian modelling for binary rasters

As a first step we want to model the large scale spatial trend, or intensity, of urbanization underlying the binary raster map shown in the left panel of Figure 1. Intensity is assumed as a smooth function over space, or surface, which takes values in $(0, 1)$. Raster data come in the form of a matrix Y with say R rows and C columns. Given y_{rc} the binary response in the pixel located at row r and column c, we make the following assumption:

$$y_{rc} \sim Ber(p_{rc}) \tag{1}$$

$$logit(p_{rc}) = (\breve{b}_c \otimes b_r)\boldsymbol{\theta} \tag{2}$$

where p_{rc} is the urban intensity surface evaluated at row r and column c, modelled, in the logit scale, as a weighted sum of B-spline bivariate basis

functions, with $\boldsymbol{\theta}$ the vector of weights or spline coefficients. We follow the P-spline surface representation by Eilers et al. (2006). Operatively, marginal basis matrices $B_{R\times k}$ and $\breve{B}_{C\times q}$ are defined over row $r = 1, ..., R$ and column $c = 1, ..., C$ indices, with k and q depending on the number of equidistant knots chosen along rows and columns respectively. Vector $\breve{b}_c$, of length q, in (2) is the row entry of matrix $\breve{B}$ associated with column index c, while vector b_r, of length q, is the row entry of matrix B associated with row index r. The kronecker product of these two vectors gives a bivariate B-spline evaluated at row r and column c. Traditionally, penalized likelihood is used to estimate $\boldsymbol{\theta}$ conditionally on a fixed smoothing parameter λ; see details in Eilers et al. (2006).

We assume a Bayesian representation of a P-spline surface following the approach presented in Lang and Brezger (2004). A prior model for $\boldsymbol{\theta}$ coefficients gives the stochastic alternative to the traditional penalty approach. We take a 2^{nd} order random walk smoothness prior for $\boldsymbol{\theta}$

$$f(\boldsymbol{\theta}|\tau^2) \propto \exp(-\frac{1}{2\tau^2}\boldsymbol{\theta}^{\mathrm{T}} K \boldsymbol{\theta}) \tag{3}$$

where $K = D^{\mathrm{T}}D$, and D is a 2^{nd} order difference matrix of known coefficients. The variance τ^2 corresponds to the inverse of the smoothing parameter λ, i.e. $\lambda = \frac{1}{\tau^2}$. A fully Bayesian model requires a prior for τ^2, we take a non informative inverse gamma $IG(0.001, 0.001)$.

The posterior for $\boldsymbol{\theta}$ is not tractable and model (2) was estimated via Monte Carlo Markov Chain methods. A Metropolis-Hastings algorithm based on an iterative weighted least squares proposal distribution (Gamerman, 1997) was used for block sampling of the spline coefficients. As discussed in Lang and Brezger (2004) the acceptance rate is usually poor which means many MCMC iterations are needed in order to achieve convergence. In order to speed up the execution of each MCMC iteration we used array algebra methods in Currie et al., (2006) and sparse matrix computation (Furrer and Sain, 2010). The full conditional for λ is a gamma with known parameters, thus a Gibbs sampling step was used to update λ.

2.1 Credible intervals for a surface

The availability of posterior samples for the spline coefficients allows any summary of interest to be computed. Analogously to the use of credible intervals for curves to detect significant change in one dimension, we propose the use of credible intervals for surfaces to detect changes over space.

We define the mean intensity surface at row r and column c as $\hat{\boldsymbol{p}}_{rc,m} = 1/(1+\exp(-(\breve{b}_c \otimes b_r)\hat{\boldsymbol{\theta}}_m)$, with $\hat{\boldsymbol{\theta}}_m$ the posterior sample mean of $\boldsymbol{\theta}$. Analogously, we define the α quantile surface as $\hat{\boldsymbol{p}}_{rc,\alpha} = 1/(1+\exp(-(\breve{b}_c \otimes b_r)\hat{\boldsymbol{\theta}}_\alpha)$, with $\hat{\boldsymbol{\theta}}_\alpha$ the posterior sample quantile corresponding to a probability α. In order to find a hot spot of urbanization it suffices to look for areas where

the $\alpha = 0.05$ quantile surface stays above a fixed probability threshold, say $p = 0.5$. Note, in such a hot spot region 95% of the probability mass of the fitted values distribution is above the threshold p.
Quantile surfaces can also be useful when the objective is to detect regions of significant low intensity, or regions of uncertainty where urbanization is sparse.

3 Results

The data we use come from the administrative database of Emilia Romagna province, Italy. Model (2) has been fitted to data from different years. Left hand panels in Figure 1 display the point patterns of urban residential use in the metropolitan area surrounding Bologna relative to 2008 (top) and 1976 (bottom). Each pixel contains a binary response, 1 indicating the presence of the urban use under study, 0 otherwise. The right hand panels show the associated posterior mean surfaces which give an estimate of the intensity of urban residential use. The black boundary line identifies pixel regions where the $\alpha = 0.05$ quantile surface crosses the threshold $p = 0.5$, meaning that 95% of the probability mass of fitted values are above that threshold. Thus, the contoured area displays a significant hot spot of urbanization. Comparing the boundary lines for the two patterns we see a change occurred between 1976 and 2008. The hot spot area is larger in 1976 than in 2008, as a result of more compactness of the residential pattern in 1976 as compared to 2008. This might be due to a sprawl effect taking place in the study region as regards the residential land use development in later years. A superimposition of the boundary lines in a unique map (here not shown) is useful to visualize the size and the spatial direction of the change.

4 Discussion and future work

Raster land use data has high potential as regards describing patterns of urbanization. P-spline approach can be fruitfully exploited to study spatial and temporal dynamics in these data since it allows efficient smoothing over grids. Bayesian P-spline models give the additional advantage of avoiding automatic selection of λ, as this is assumed as random and estimated via MCMC. As a result, credible intervals which correctly include uncertainty of λ are directly available. This allows a simple method based on quantile surfaces to investigate the presence of spatial features such as boundaries.
Future work will be about two aspects. First, building models coherent with social and economic factors driving urbanization. This requires extension of the model with the introduction of covariates. The second focus will be on diagnostic tools to investigate clustering effects in the residuals. We believe the connection between spatial logistic models and poisson point processes can be exploited to build useful model diagnostics for the model presented.

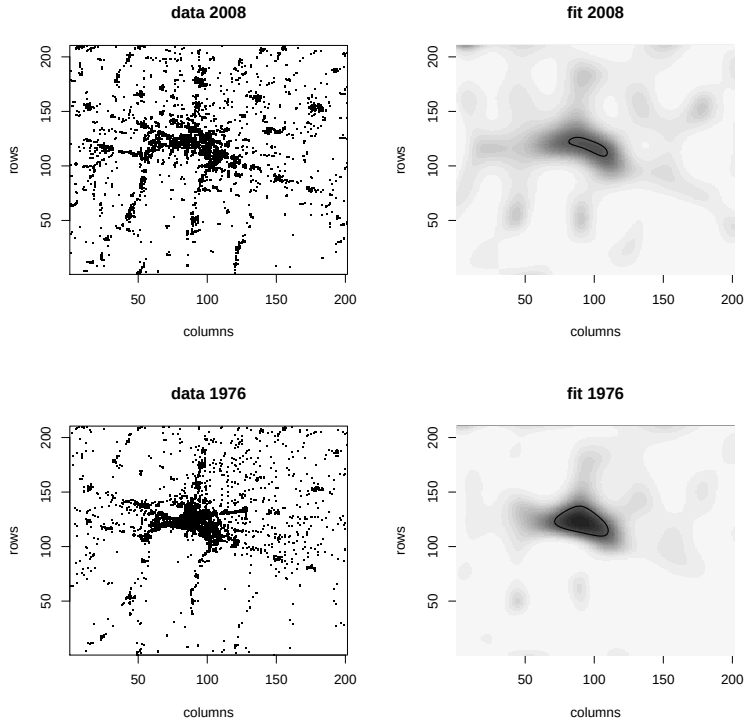


FIGURE 1. Comparison of urban residential patterns in Bologna for 2008 and 1976. On the left the raster binary data: pixel color indicates the presence (black) absence (white) of urban residential use. On the right the associated estimate of the intensity expressed in a continuous grey color scale ranging $(0, 1)$. Boundary black lines localize hot spot regions of urban residential use.

References

Currie, I.D., Durban, M. and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, **68**, 259-280.

EEA (2006). Urban sprawl in Europe. The ignored challenge. *Technical report: Environmental European Agency.*

Eilers, P.H.C., Currie, I.D. and Durban, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, **50**, 61-76.

Furrer, R., Sain, S.R. (2010). spam: A sparse matrix R package with emphasis on MCMC methods for gaussian markov random fields. *Journal of statistical Software*, **36**, 2968-2977.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, **7**, 57-68.

Hijmans R.J., van Etten J. (2012). *raster : Geographic Analysis and Modeling with Raster Data.* URL http://CRAN.R-project.org/package=raster

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183-212.

Bayesian Factorization Model for Analysing Mixed Data

Helga Wagner[1], Regina Tüchler[2]

[1] Department of Applied Statistics and Econometrics, Johannes Kepler University Linz, Austria
[2] Austrian Federal Economic Chamber, Department of Statistics, Vienna, Austria

E-mail for correspondence: helga.wagner@jku.at

Abstract: In many applications multidimensional outcome variables measured on different scales are of interest. In this paper we consider regression modelling of a bivariate response with a normal and a binary component. To model dependence we use a factorization model where we allow for flexible non-linear dependence. We apply this model to analyse material deprivation and household income in Austria.

Keywords: Bayesian Mixed Model; Factorization Model; Data Augmentation; Variable Selection; Living Conditions.

1 Model

Let $\mathbf{y}_i = (y_i^b, y_i^n)$ denote a bivariate response, where y_i^b is a binary and y_i^n a normal component, and $\mathbf{x}_i$ the $1 \times d$ vector of covariates for subject $i = 1, \ldots, N$. We specify a regression model for both outcomes based on the factorization of the joint distribution as

$$p(\mathbf{y}_i|\mathbf{x}_i) = p(y_i^n|\mathbf{x}_i)p(y_i^b|y_i^n, \mathbf{x}_i).$$

The marginal model of the normal component is a standard linear regression model

$$y_i^n = \mu^n \mathbf{x}_i \boldsymbol{\beta}^n + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

To allow for flexible dependence between normal and binary response the predictor for the binary response η_i^b combines a linear function of the covariates with a smooth function f of the error terms of the normal model as

$$\eta_i^b = \mathbf{x}_i \boldsymbol{\beta}^b + f(\frac{y_i^n - \mathbf{x}\boldsymbol{\beta}^n}{\sigma}).$$

We use a logit function to link the predictor to the mean, but other link functions, e.g. probit, robit or complementary log-log are also possible.

To estimate the nonlinear effect of the standardized residual $f(\varepsilon/\sigma)$ we consider a representation of f in terms of a linear combination of B-spline basis functions B_j, as

$$f(\varepsilon/\sigma) = \sum_{j=1}^{J} \gamma_j B_j(\varepsilon/\sigma), \tag{1}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$ denotes the coefficients of this linear combination. Bayesian model formulation is completed by specifying the following prior distributions for the model parameters: standard normal priors for $\boldsymbol{\beta}^n$ and $\boldsymbol{\beta}^b$ and an inverse Gamma prior for the error variance σ^2. For the spline coefficients $\boldsymbol{\gamma}$ we use a smoothness prior specified as a second order random walk,

$$\gamma_j = 2\gamma_{j-1} + \gamma_{j-2} + \nu_j, \quad \nu_j \sim \mathcal{N}(0, \tau^2)$$

and an inverse Gamma prior for the hyper-parameter τ^2.

2 Bayesian inference

Model parameters are estimated by sampling from the posterior distribution with data augmentation and MCMC methods. For the binary logit model we use the representation in terms of a latent utility u_i as

$$u_i = \eta_i^b + \epsilon_i, \qquad y_i^b = I_{(0,\infty)}(u_i),$$

where the error term ϵ_i has a standard logistic distribution. Using the approximation of this error distribution as a finite scale mixture of normal distributions (Frühwirth-Schnatter and Frühwirth, 2010) the regression coefficients $\boldsymbol{\beta}^b$ and the spline coefficients $\boldsymbol{\gamma}$ of the logistic regression model can be estimated from the auxiliary normal model with heteroscedastic errors

$$u_i = \eta_i^b + \tilde{\epsilon}_i, \qquad \tilde{\epsilon}_i \sim \mathcal{N}(0, s_{r_i}^2),$$

where r_i denotes the component indicator for the mixture component. Hence additionally to the model parameters the auxiliary variables $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{r} = (r_1, \ldots, r_n)$ will be sampled. We use the following MCMC scheme:

(I) Sample the regression coefficients $\boldsymbol{\beta}^n$ and the error variance σ^2 of the normal model.

(II) Sample the auxiliary variables $\mathbf{u}, \mathbf{r}$ from $p(\mathbf{u}, \mathbf{r} | \sigma^2, \boldsymbol{\beta}^n, \boldsymbol{\beta}^b, \boldsymbol{\gamma}, \mathbf{y}^n, \mathbf{y}^b)$.

(III) Sample the regression coefficients $\boldsymbol{\beta}^b$ of the logit model from the full conditional $p(\boldsymbol{\beta}^b | \boldsymbol{\beta}^n, \boldsymbol{\gamma}, \sigma^2, \mathbf{y}^n, \mathbf{u}, \mathbf{r})$.

(IV) Sample the spline coefficients $\boldsymbol{\gamma}$ from $p(\boldsymbol{\gamma} | \boldsymbol{\beta}^n, \boldsymbol{\beta}^b, \sigma^2, \mathbf{y}^n, \mathbf{u}, \mathbf{r}, \tau^2)$ and the hyper-parameter τ^2 from $p(\tau^2 | \boldsymbol{\gamma})$.

Note, that due to data augmentation the parameters of the logit model have a normal posterior distribution conditioning on the parameters of the normal model. However, as the likelihood of a binary observation depends on the standardazid residuals and hence on the the parameters of the normal regression model, the posteriors for $\boldsymbol{\beta}^n$ and σ^2 have no closed form and are therefore sampled using a MH-step.

3 Analysis of data on living conditions in Austria

In our application we contribute to research of well-being of societies. This area has become increasingly important as European politics started to focus on indicators that complement GDP. Initiatives that deal with this subject are the "GDP and beyond" initiative, the Stiglitz-Sen-Fitoussi Commission (Stiglitz et al. 2009) and the Sponsorship Group on Measuring Progress, Well-being and Sustainable Development (ESS 2011). In European statistics a scoreboard of indicators is currently developed. Therefore we face an increasing need for models that are able to deal with dependencies between these indicators and that help to analyse driving factors.
We use data from the Survey on Income and Living Conditions (SILC) 2009 to analyse yearly household income and material deprivation in Austria. The household income includes all the money a household has to make its living from, like net personal and capital income, social transfers, alimony, etc., whereas material deprivation is a concept measuring whether a household is capable to meet certain predefined needs, like e.g. TV, phone, holiday away from home.
The factorization model allows to analyse dependence between the continuous outcome variable household income and the binary outcome variable material deprivation. By implementing Bayesian variable selection with spike and slab priors we derive the importance of explanatory variables, like e.g. the age or activity status of the main-income earner, the household type, or migration status. For our analysis we use only households with main income earner not older than 60 years and not retired and household income larger than 1 000 Euro. Thus our final data set includes 3 694 households.

4 Results

The factorization model proposed here is an alternative to the mixed data regression model used in Tüchler and Wagner (2012) and Wagner and Tüchler (2013) on a slightly different version of the data set. The advantage of the factorization model is that it allows for nonlinear dependence. That a more flexible dependence structure is adequate for the data is illustrated in Figure 1 which compares two different specifications of the factorization

model: a model where linear dependence is assumed, i.e.

$$f(\varepsilon/\sigma) = \gamma \frac{\varepsilon}{\sigma}$$

and a model with a smooth f as given in equation (1). The plot shows the posterior mean of $\beta_0^b + f(\varepsilon/\sigma)$, where β_0^b is the intercept in the logit model. The estimated smooth function f has a kink with almost zero slope before and a negative slope after the breakpoint. This implies that the risk of material deprivation changes only little before the breakpoint but decreases quickly afterwards.

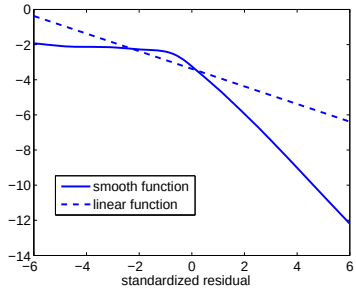


FIGURE 1. Comparing linear and flexible dependence.

Parameter estimates and posterior inclusion probabilities for the flexible model are shown in Table 1. We find that for both response variables the activity status as well as education and migration background of the main income earner are included with a very high probability. All other covariates turn out to have a low inclusion probability for at least some of their categories. As it had to be expected households with a main-income earner working full-time have higher income and are less likely in a situation of material deprivation than households with a main-income earner working part-time or being unemployed or out-of-labour-force. The higher the education level of the main-income earner the higher is the income and the less likely occurs material deprivation. We also find that migration background leads to a higher probability of material deprivation and to a smaller household income.

Figure 2 compares households with all covariates but one equal to the baseline value. The estimated probability of material deprivation is shown for households with/without migration background of the main-income earner in the left panel and for households with lower education/university degree of the main income earner in the right panel. In these plots the non-linear dependence introduced by the function f becomes visible. Low income naturally yields a higher risk of material deprivation. But household income has to rise over a certain value to achieve a considerable decrease in the risk of material deprivation. For households with lower education of the main

TABLE 1. Estimates for the mean and probabilities to be unrestricted.

variable	material deprivation		log(earnings)	
	$\hat{\boldsymbol{\beta}}$	$\Pr(\delta_j = 1)$	$\hat{\boldsymbol{\beta}}$	$\Pr(\delta_j = 1)$
intercept	-3.64	–	9.82	–
gender	0.04	*0.18*	-0.00	*0.07*
(base: male)				
age	-0.00	*0.04*	0.01	*0.01*
(cent. at median 43 y.)				
age^2	0.00	*0.00*	0.00	*0.01*
activity status				
(base: full-time)				
part-time	1.16	*1.00*	-0.25	*1.00*
unemployed	2.45	*1.00*	-0.40	*1.00*
out-of-labour	2.13	*1.00*	-0.56	*1.00*
education				
(base: lower)				
medium	-0.34	*0.68*	0.12	*0.98*
higher	-1.53	*1.00*	0.28	*1.00*
university	-1.71	*1.00*	0.42	*1.00*
migration	1.57	*1.00*	-0.28	*1.00*
(base: no migration)				
type of household				
(base: single)				
2 adults/no children	-0.24	*0.49*	0.18	*1.00*
single-parent	0.20	*0.43*	-0.07	*0.64*
2 adults/1 or 2 Children	-0.03	*0.17*	0.00	*0.01*
2 adults/+3 children	0.02	*0.20*	-0.20	*1.00*
other	-0.03	*0.16*	0.10	*0.97*
type of building				
(base: single-family)				
2 families	0.03	*0.18*	0.00	*0.02*
3 to 9 families	0.66	*0.89*	-0.01	*0.08*
+10 families	0.98	*0.99*	-0.07	*0.86*
other	-0.15	*0.36*	-0.01	*0.05*
population density				
(base: high)				
medium	-0.01	*0.13*	0.00	*0.04*
low	-0.32	*0.65*	-0.01	*0.14*

income earner this breakpoint is about 9 000 Euros with migration background, whereas it is about 12 000 Euros for households with no migration background.

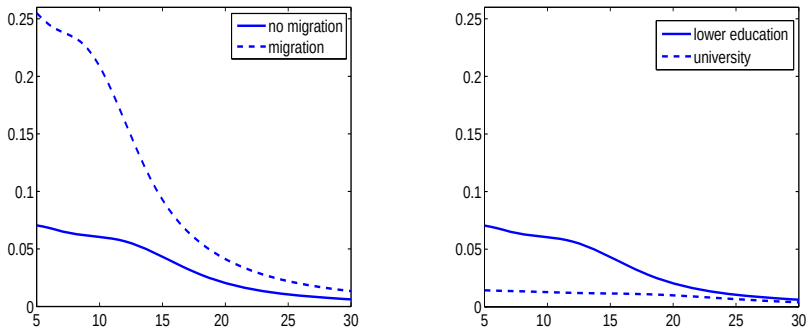


FIGURE 2. Probability of material deprivation conditional on income (in 1000 Euros) for different households

5 Conclusions

We propose a joint regression model for a bivariate response with mixed discrete and continuous type, which allows for flexible, nonlinear dependence between both components. MCMC methods are used for Bayesian inferences, where variable selection and model averaging can be easily incorporated by using spike and slab priors for regression effects.

References

Frühwirth-Schnatter, S. and R. Frühwirth (2010). Data augmentation and MCMC for binary and multinomial logit models. In: T. Kneib and G. Tutz (Eds.), *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, Heidelberg: Physica-Verlag, pp. 111 – 132.

ESS (2011). Sponsorship Group on Measuring Progress, Well-being and Sustainable Development. Final report. *EEA ESSC 2011/11/05/EN*.

Stiglitz, J.E., A. Sen and J.P. Fitoussi (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress. *www.stiglitz-sen-fitoussi.fr*.

Tüchler, R. and H. Wagner (2012). Analysing living conditions in Austria by a Bayesian mixed data model. In: A. Komarek and S. Nagy (Eds.), *Proceedings of the 27th IWSM*, Prague, pp. 321 – 326.

Wagner, H. and R. Tüchler (2013). Sparse Bayesian modeling of mixed econometric data using data augmentation, In: A. R. de Leon and K. C. Chough (Eds.), *Analysis of Mixed Data: Methods and Applications*, Chapman & Hall / CRC, pp. 173 – 188.

Bivariate Bayesian Quantile Regression

Elisabeth Waldmann[1], Anke Stein[2], Thomas Kneib[1]

[1] Chair of Statistics and Econometrics, Georg-August-Universität Göttingen
[2] Chair of Biodiversity, Georg-August-Universität Göttingen

E-mail for correspondence: ewaldma@uni-goettingen.de

Abstract: Quantile regression for conditional random variables has become a widely used tool to analyse relations within data. It provides a detailed description of the conditional distribution, without assuming a distribution type for the conditional distribution. The Bayesian version, which can be implemented by considering the asymmetric Laplace distribution (ALD) as an error distribution is an attractive alternative to other methods, because it returns knowledge on the whole parameter distribution instead of solely point estimations. While for the univariate case there has been a lot of development in the last few years, multivariate responses have only been treated to little extend in the literature, especially in the Bayesian case. By using a multivariate version of the location scale mixture representation for the ALD we are able to apply inference techniques developed for multivariate Gaussian models on multivariate quantile regression and make thus the impact of covariates on the quantiles of more than one dependent variables feasible.

Keywords: Quantile Regression; Bivariate Regression; Bayesian Inference; Biodiversity.

1 Geoadditive Quantile Regression

Quantile regression is a powerful tool to analyse the impact of covariates on a dependent variable. The difference to mean regression is the obvious advantage to gain knowledge about the whole conditional distribution without assuming any restrictive data distribution. Another advantage is the robustness which quantiles possess. A disadvantage – at least in our approach – is the independent estimation of the conditioned quantiles and therefore the possibility of arising quantile crossing. If we want to measure the impact of a set of covariates $\boldsymbol{x}_j, j = 1, \ldots, q$ in some functional form

$$\boldsymbol{\eta_\tau} = \beta_0 + \sum_{j=1}^{p} f(x_j)$$

on the τ-quantile of the conditional distribution of $\boldsymbol{y}|\boldsymbol{X}$ (where $\boldsymbol{X}$ is the matrix of all $\boldsymbol{x}_j$, which are the vectors of the covariates, and $\tau \in (0, 1)$ the

quantile), we have to minimize a criterion different to the least squares. This criterion has to account for the asymmetry of the idea of quantiles. This is implied by the asymmetrically weighted absolute deviations (AWAD):

$$\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{\eta}_\tau) \to \min_{\boldsymbol{\eta}_\tau},$$

where ρ_τ stands for the check function:

$$\rho_\tau(y_i - \boldsymbol{\eta}_\tau) = \begin{cases} \tau|y_i - \boldsymbol{\eta}_\tau| & \text{if } y_i \geq \boldsymbol{\eta}_\tau \\ (1-\tau)|y_i - \boldsymbol{\eta}_\tau| & \text{if } y_i < \boldsymbol{\eta}_\tau. \end{cases} \tag{1}$$

Minimizing the AWAD requires linear programming techniques, which makes inference for more complex predictor functions difficult.

2 Bayesian Quantile Regression

In Bayesian inference, we obviously need an error distribution. We use the ALD, which is defined as follows:

$$f(y|\mu, \delta, \tau) = \tau(1-\tau)\delta \exp(-\rho_\tau(\delta(y-\mu))), \tag{2}$$

where μ denotes the mean and δ the precision. Maximizing the posterior with the ALD as error distribution and imposing noninformative priors on all the parameters leads to the same results as minimizing the check function. The ALD as displayed in formula (2) is not easily accessible, therefore we use the following location scale parametrisation:

$$\boldsymbol{y}|\boldsymbol{W}, \xi, \sigma^2 \sim N\left(\boldsymbol{\eta}_\tau + \xi\boldsymbol{w}, \frac{\sigma^2}{\delta^2}\boldsymbol{W}\right) \Leftrightarrow \boldsymbol{y} \sim ALD(\boldsymbol{\eta}_\tau, \delta, \tau). \tag{3}$$

The weights w_i follow an exponential distribution with rate δ^2 (i.e. the precision of the ALD), $\boldsymbol{W}$ displays the diagonal matrix of the single w_i for $i = 1, \ldots, n$ and the auxillary parameters are such that $\xi = \frac{1-2\tau}{\tau(1-\tau)}$ and $\sigma^2 \frac{2}{\tau(1-\tau)}$. With relation (3) the construction of an MCMC algorithm is straight forward, as it is analogous to the procedure for Gaussian regression.

3 Bivariate Bayesian Quantile Regression

In order to be able to estimate impact of covariates on potentially correlated dependent variables we extended the Gaussian distribution (3) by introducing a coefficient of association ν in the covariance matrix. This leads to a block diagonal matrix $\boldsymbol{\Sigma}$ with matrices

$$\boldsymbol{\Sigma}_{block} = \begin{pmatrix} \frac{1}{\delta_1^2} & \nu \\ \nu & \frac{1}{\delta_2^2} \end{pmatrix}$$

on the diagonal. If we furthermore define two different weight types, which are exponentially distributed with the different δ_k we get a new weight matrix $\boldsymbol{W}$ with vectors $\boldsymbol{w}_i$ which contain two different weights for each individual. Then the marginal distributions of

$$\boldsymbol{y} \sim N(\boldsymbol{\eta} + \xi \frac{\boldsymbol{w}}{\delta^2}, \boldsymbol{W}_{\sigma^2} \boldsymbol{\Sigma} \boldsymbol{W}_{\sigma^2})$$

for the separated dependent variable vectors $\boldsymbol{y}_k$ are ALDs again, so the estimators of the maximized posterior are again the minimizers of the check function. Splitting up the weight matrix in the square roots (i.e. a diagonal matrix $\boldsymbol{W}_{\sigma^2}$ with entries $\sqrt{\sigma^2 w_{ik}}$) leads to the possibility to use an inverse Wishart distribution as prior for $\boldsymbol{\Sigma}_{block}$ and thus also as full conditional. While for the regression parameters themselves the estimation does not change in comparison to the univariate Bayesian quantile regression, the full conditional for the weights w_{ik} is not a closed form distribution anymore, therefore we use a Metropolis-Hastings step.

4 Application on Species Richness Patterns

Explaining large-scale patterns of species richness has been a major goal in macroecology. Although positive relationships between the number of plant and animal species in a region have often been found, it is difficult to disentangle whether these are due to direct producer-consumer interactions or to similar environmental and historical constraints. The empirical relation between the number of plant species on the one hand and animal species (birds, mammals) on the other is displayed on the left side of FIGURE 1. In a study based on structural equation models Jetz et al. (2009) analyzed the influences of environmental covariates like temperature, number of wet days and habitat diversity on this relation. The dataset contains the number of plant and animal species for 639 regions worldwide (for a more detailed data description see Jetz et al (2009) and references therein; we thank Holger Kreft for provision of the data). The environmental influence apparently is not linear and homoscedastic, as can be seen on the right side of FIGURE 1, where the influence of logarithmic temperature on logarithmic number of plants is displayed. Rising temperature has a higher - and almost linear - influence on the higher quantiles, whereas in the lower part of the distribution the effect seems to be shaped differently. Thus we reanalyzed the dataset with the above presented method. The logarithmic temperature was assigned a nonlinear effect, topographic diversity (the maximal elevational range within the region), habitat diversity (the number of different ecosystems in the region) and the number of rainy days

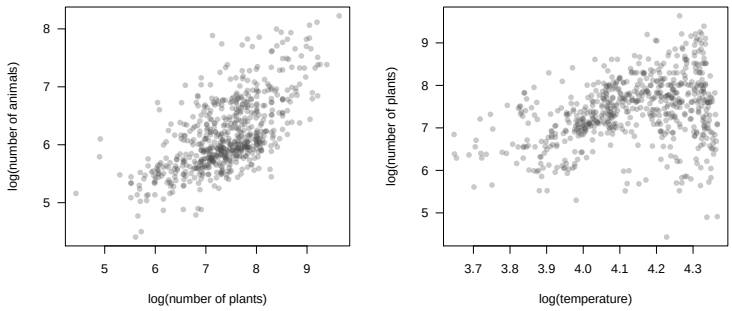


FIGURE 1. Left Plot: Logarithmic number of plants, logarithmic number of animals. Right Plot: Logarithmic temperature, logarithmic number of plants.

were included linearly. Thus we reanalyzed the data set with the above presented method. The logarithmic temperature *temp* was assigned a non-linear effect, topographic diversity (the maximal elevational range within the region), habitatial diversity (the number of different ecosystems in the region) and the number of rainy days where included linearly. The resulting model is:

$$\hat{\boldsymbol{y}}_{k,\tau} = \boldsymbol{X}\boldsymbol{\beta} + f(\boldsymbol{z}_{temp}),$$

where $\boldsymbol{X}$ comprises the above mentioned linear effects and k stands for the two different dependent variables *animals* and *plants*.

The model was estimated for a range of different quantile combinations (τ_1, τ_2), here the case where $\tau_1, \tau_2 \in \{0.5, 0.8, 0.9\}$ is presented. FIGURE 2 displays the boxplots of the samples for the precision matrix for the different combinations. The upper plot displays the precision, τ_1 is the quantile for the first component of the dependend variable, i.e. the logarithmic number of animals, τ_2 respectivly the quantile for which the model was estimated for the logarithmic number of plants. Especially for calculating models with extrem values for both response variables the precisions differ a lot depending on the value of τ for the other component. The lower plot in FIGURE 2 displays the parameter of association ν for the same models. All of these values are positive, but the figure also shows a special behaviour for the value $\tau_1 = \tau_2 = 0.9$. Here the correlation is higher than in all the other combinations. The trend to higher correlation in higher quantiles allready shows in the combination of the two 0.8 quantiles. Thus less of the correlation between the two numbers of species can be explained by the covariates in the higher parts of the distribution. Note that though we cannot really interpret the level of correlation because in our model it is

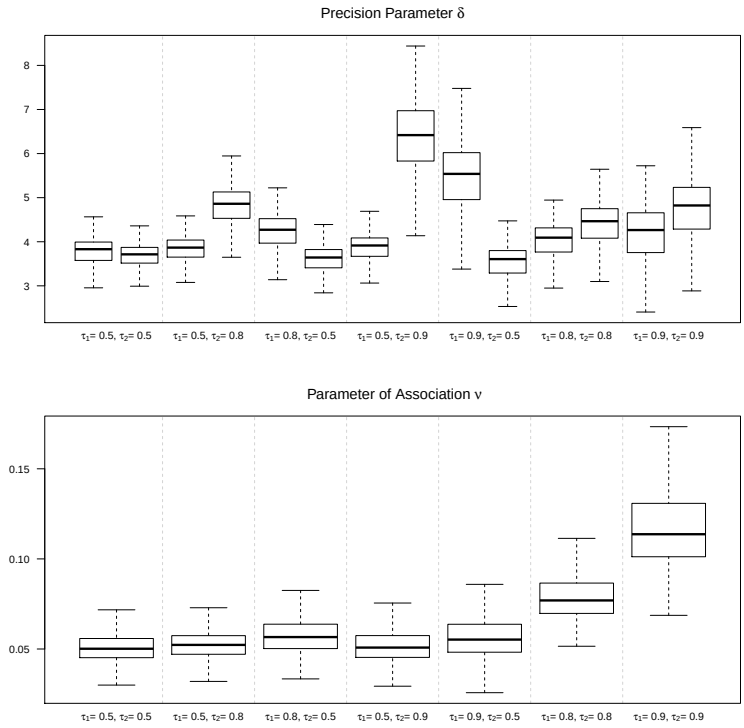


FIGURE 2. Upper Plot: Boxplot for the MCMC samples of the model precisions. Lower Plot: Boxplot of the MCMC samples of the coefficient of association.

estimated based on the weights, we can still compare the values of ν for the different models, as the weights only differ minimally for the individual quantiles over all models.

5 Conclusion

The presented approach unifies Bayesian geoadditive quantile regression with multivariate statistics by using the possibility to extend the location scale mixture of the ALD to a multivariate version. This can be of use in different fields. An extension to a multivariate response should be easily possible.

References

Jetz, W., Kreft, H., Ceballos, G. and Mutke, J. (2009). Global associations between terrestrial producer and vertebrate consumer diversity, *Proc Biol Sci. 276(1655):269-78. doi: 10.1098/rspb.2008.1005.*.

Koenker, R. and Bassett, G. (1978). Regression Quantiles *Econometrica*, **46**, 33 – 50.

Kreft, H. and Jetz, W. (2007). Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 5925 – 5930.

Lang, S., Adebayo, S., Fahrmeir, L. and Steiner, W. 2002. Bayesian Geoadditive Seemingly Unrelated Regression, *Ludwig-Maximilians-Univ., SFB 386*

Waldmann, E., Kneib, T., Yue, Y., Lang, S. and Flexeder, C. (2013). Bayesian Semiparametric Additive Quantile Regression. *accepted at Statistical Modelling.*

GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data

Madawa P. Weerasinghe Jayawardana Rathambalage[1], Georgy Sofronov[1]

[1] Department of Statistics, Macquarie University, Sydney, Australia

E-mail for correspondence: madawa.weerasinghe@mq.edu.au

Abstract: We model DNA read count data obtained through next generation sequencing (NGS) technologies as a multiple change-point process. This means that the data are divided into different segments based on the number of change-points. Each segment of the process is modeled by utilizing the zero-inflated negative binomial (ZINB), as well as the negative binomial (NB) distribution in the Generalized additive models for location, scale and shape (GAMLSS) framework. It is observed that ZINB and NB based models, fit the data better than the competing Poisson model, in which the observed read counts are highly over-dispersed as well as zero-inflated. Moreover, we have considered incorporating auxiliary information to further improve the change-point modelling process by utilizing the GAMLSS framework. The extended Cross-Entropy (CE) method which uses a four-parameter beta distribution is used to estimate the number of change-points as well as their corresponding genome locations. Furthermore, parallel implementation of the procedure results a significant improvement in total running time, in which the procedures are highly computationally intensive. We apply the proposed methodology to find change-points in DNA read count data obtained through Illumina TruSeq exome capture of patients with celiac disease. Our results suggest that the proposed GAMLSS based CE method is an effective methodology to detect change-points in genome-wide data.

Keywords: GAMLSS; Cross-Entropy Method; Change-Point Modelling; Combinatorial Optimization.

1 Introduction

Discovering chromosomal aberrations in the genomic DNA is a widely discussed issue that has been addressed through various scientific techniques based on different perspectives. It is an established fact that the variations in DNA copy number is a source of genetic variation [Campbell et al., 2008] even though the full understanding of the effect of these is still on probe. Recent studies based on microarray technology have identified around 12%

of the human genome and thousands of genes are variable in copy number. It is predicted that the emerging technologies with high sensitivity level of data will further expand this knowledge.

Prior to the advent of next generation sequencing (NGS) technologies, number of methodologies have been developed to detect multiple change-points mainly based on the array Comparative Genomic Hybridization (aCGH) data. Analysis on aCGH data aims to find changes in the mean of the fluorescence color ratios, usually on logarithm scale to detect copy number variations in the human genome. See [Lai et al., 2005] for a review of the aCGH based segmentation methods. However, the introduction of the next generation DNA sequencing technologies and the resulted excess amount of data has increased the complexity level of the process of partitioning the genome in to homogeneous segments to a higher level.

Reviewing the literature on change-point modeling of NGS data, [Xie and Tammi, 2009] proposed a method called CNVseq to identify CNVs on the data generated through shotgun sequencing. Later [Magi et al., 2012] reviewed some of the existing methodologies to detect CNV in read count data. They have normalized the raw read counts and conducted the segmentation based on the techniques mainly developed on aCGH data. They also mentioned that there exist only few statistical procedures that utilize the raw read counts to detect CNVs. In fact, most of the prevailing methods transform the raw read counts by different normalization techniques to a stage, where they can utilize the existing aCGH based segmentation methods. They have not considered utilizing auxiliary information in the generalized linear models (GLM) context to detect multiple change-points in the read count data.

In order to fill this gap in the literature of direct usage of the DNA read counts generated by the NGS platforms, we propose a procedure which utilizes generalized additive models for location, scale and shape (GAMLSS) statistical framework [Rigby and Stasinopoulos, 2005] to incorporate auxiliary information into the modelling process, and extended Cross-Entropy method [Priyadarshana and Sofronov, 2012] to estimate the number of change-points as well as their corresponding genome locations. We observe that the DNA read counts we analyzed are highly over-dispersed as well as zero-inflated. Therefore, the response variable is modelled by utilizing the zero-inflated negative binomial (ZINB) as well as the negative binomial (NB) distribution in the GAMLSS framework.

1.1 Multiple Change-Point Problem, GAMLSS Framework and CE Method

Generalized additive models for location, scale and shape (GAMLSS)

The GAMLSS are a type of semi-parametric regression models use to model univariate response with a set of covariates. It allows for modelling not

only the expected mean but other parameters of the distribution (e.g. location, scale and shape) of the response variable as well. Therefore, it gives more flexibility in modelling process than the generalized additive models (GAMs), and generalized linear models (GLMs).

Multiple Change-Point Problem

Let us formulate the multiple change point problem in mathematical terms. A count data sequence $y = (y_1, y_2, ..., y_L)$ of length L is given. A segmentation of the sequence is specified by the number of change points N and the positions of the change points $c = (c_1, c_2, ..., c_N)$, where $1 = c_0 < c_1 < ... < c_N < c_{N+1} = L + 1$. In this context, a change point is a boundary between two adjacent segments. The value of c_i is the sequence position of the rightmost character of the segment to the left of the i^{th} change-point. Segments are numbered from 0 to N as there will be one or more segments than number of change points. We model each segment of the DNA read count data utilizing ZINB and NB distribution in the GAMLSS regression framework with the use of exon length as the covariate.

Extended Cross-Entropy Method

The Cross-Entropy (CE) method [Rubinstein and Kroese, 2004] is a new generic approach to combinatorial and multi-extremal optimization and rare event simulation. Broadly it can be used to solve estimation and optimization problems. In this study, the process of multiple change point detection is considered as a minimization problem. The CE method is an iterative optimization procedure that starts with a parameterized sampling distribution from which a random sample is generated. Then, each observation or the combinatorial arrangement is scored for its performance, as the solution to a specified optimization problem. A fixed number of best performing combinatorial arrangements are referred to as the elite sample. This elite sample is subsequently used to update the parameters for the sampling distribution. Thus, adaptive parameters are utilized in each iteration. The sampling distribution eventually converges to a degenerate distribution about a locally optimal solution, which ideally will be globally optimal.

In this study we utilize the extended version of the standard CE method, proposed in [Priyadarshana and Sofronov, 2012] with further modifications. We use a stopping criterion (SC) based on Median Absolute Deviation (MAD) as opposed to the variance based SC proposed in the original paper. Furthermore, a multi-core architecture based parallel implementation of the algorithm is implemented in order to carry out calculations more efficiently.

2 Results and Conclusions

In this section, we include results of numerical experiments that illustrate the performance of the proposed method. This example considers a DNA read count data obtained from a study of celiac disease patients. All data were obtained from the Illumina TruSeq exome capture technology. We analyze DNA read count data with respect to chromosome 15 of a patient. We compare the change-point modelling process with and without the effect of auxiliary information utilizing the extended CE method. In the case without any predictor variables, we model the read count data based on both NB and ZINB distributions. In the process of utilizing auxiliary information and the GAMLSS implementation, we consider natural logarithm of the exon length as a predictor variable. In the GAMLSS framework, we carry out the change-point modelling procedure considering the distribution of the response variable as zero-inflated negative binomial (ZINB-GAMLSS) as well as negative binomial (NB-GAMLSS).

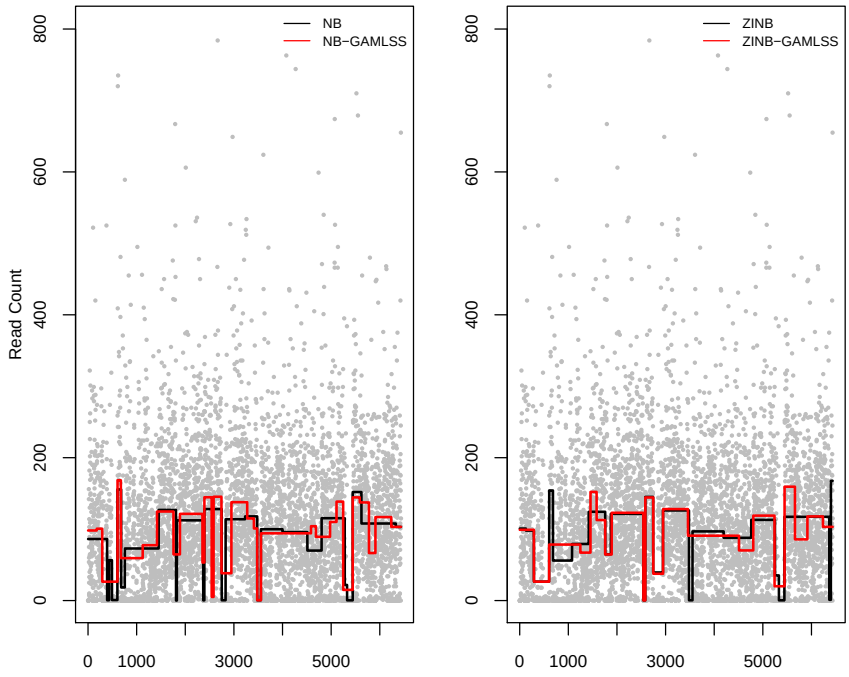


FIGURE 1. Mean profile plots for NB, NB-GAMLSS, ZINB, ZINB-GAMLSS.

Figure 1 shows the mean profile plots of results for the case, with and without any predictor variables. It is visible that in the ZINB set up, the

GAMLSS approach and the ZINB results have a higher level of concordance of estimated change-points, when compared to the NB results. This may be due to the fact that ZINB better models the observed read counts than the NB. It can be further noticed that in general NB based models have estimated more change-points than the ZINB based models for this particular DNA read count data. While the results of this work are encouraging, there are plenty of avenues available for future research work, especially on the implementation of GAMLSS framework and the incorporation of more predictor variables to the modelling process. Furthermore, cluster level implementation of the methodology will certainly improve the processing time, in which all these processes are highly computationally intensive.

Acknowledgments: We thank Dr. Vincent Plagnol (UCL Genetics Institute, University College London, Grower Street, London, UK) for providing celiac disease patients read count data for the analysis.

References

Campbell, P.J., et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*, **6**, 722 – 729.

Lai, W.R., et al. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **19**, 3763 - 3770.

Magi, A., et al. (2012). Read count approach for DNA copy number variants detection. *Bioinformatics*, **4**, 470 - 478.

Priyadarshana, W. J. R. M. and Sofronov, G. (2012). A modified Cross Entropy Method for Detecting Multiple Change Points in DNA Count Data. In: *Proceedings of the IEEE World Congress on Computational Intelligence (IEEE CEC 2012)*, Brisbane, pp. 1020 – 1027.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, **54**, 507 – 554.

Rubinstein, R. and Kroese, D. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York: Springer-Verlag.

Xie, C. and Tammi, M.T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80.

A multivariate Bayesian model for human growth

Sten Willemsen[1], Paul H.C. Eilers[1], Regine Steegers[2], Emmanuel Lesaffre[13]

[1] Department of Biostatistics Erasmus MC Rotterdam, The Netherlands
[2] Department of Obstetrics and Gynaecology Erasmus MC Rotterdam, The Netherlands
[3] L-BioStat, KU Leuven, Belgium

E-mail for correspondence: s.willemsen@erasmusmc.nl

Abstract: We have implemented a Bayesian multivariate version of the SITAR model of Cole et al. This model is applied to the data of the PREDICT study consisting of measurements of embryonic growth taken in the first semester of pregnancy. We demonstrate how the model can be used to find which characteristics influence early growth and how early growth is related to negative birth outcomes.

Keywords: Bayesian statistics; Multivariate statistics; Growth curves

1 Introduction

It is important to have good models of human growth. For example, having a good reference of normal growth during pregnancy can help in deciding whether and when an obstetrical intervention is needed. More generally, abnormal growth patterns can be an indication for diseases later in life. While multivariate growth models exist, in practice growth is usually modelled one dimension at a time. The SITAR model, which is discussed in more detail below, is very suitable to model human growth univariately but because we are also interested in the relationship between the outcomes we needed to extend to incorporate multiple outcomes.
To illustrate our developments we use data from the Predict study in which fetal growth in the first trimester of pregnancy is studied. It has long been thought that growth in the first trimester of a pregnancy is the same for all fetuses. This is why pregnancies are often dated based on measurements done in this period. However, lately this idea is no longer universally accepted. More and more the thought has rooted that there are important differences in fetal growth in the first trimester. In addition, these first trimester differences have been related to birth outcomes. Therefore the study of first trimester growth has become increasingly important.

Contrary to what it usually done we do not want to look at growth only in one dimension (such as the crown rump length or the embryonal volume) or study more dimensions one at a time. We believe that it is important to look at the different aspects of growth in relation to each other. The idea is that normal growth is reflected not only in the separate measurements, but also, and perhaps more importantly, in the relationship of the growth of the different aspects of the fetusses body.

2 The multivariate SITAR model

Our model is based on the 'SuperImposition by Translation And Rotation' model (SITAR) by Cole (2010) which in turn is based on the 'Shape invariant' model of infant growth of Beath (2007). The idea of this model is that there exists a global growth curve that can be modeled by a spline function and that all individual growth profiles can be reduced to this general curve by translating them horizontally and vertically and by stretching them. Mathematically the SITAR model can be expressed as follows:

$$y_{ij} = B\left(\exp(\gamma_{i3})\left[t_{ij} + \gamma_{i1}\right]\right)^{\mathrm{T}}\beta + \gamma_{i2} + \varepsilon_{ij}, \tag{1}$$

where y_{ij} is the outcome of individual i measured at time t_{ij}, the γs are the subject specific effects and express the subject-specific horizontal shift (γ_{i1}), the vertical shift (γ_{i2}) and the stretch (γ_{i3}). The vector $B(x)$ is the basis of a restricted cubic spline evaluated at x. Because of the horizontal shifts, we must be able to evaluate the spline in all points. This is easiest with a restricted spline which is linear beyond the outer knots so we just have to do a linear extrapolation. The SITAR model has been documented to work well in practice. Moreover, it has parameters that are easily interpretable by clinicians.

The univariate SITAR model can easily be extended to model several series of repeatedly measured outcomes. Our multivariate model then is:

$$\begin{aligned} y_{ijk} &\sim N(\gamma_{i2k} + T_{ijk}^{\mathrm{T}}\beta_k, \sigma_k^2), \\ T_{ijk} &= B(\exp(\gamma_{i3k})(t_{ij} + \gamma_{i1k}), \\ \gamma_{i.k} &\sim N(0, \Sigma_\gamma), \end{aligned}$$

where y_{ijk} is the kth response of individual i at time j, the vector T_{ijk} a basis of a natural cubic spline for the kth response of individual i at time j, $B(x)$ is a function that evaluates the basis of a natural cubic spline at time x, β_k the vector of spline coefficients for the kth response, $\gamma_{i.k}$ is the vector with the three subject specific effects for individual i and series k, σ_k is the measurement error variance for series k, and Σ_γ is the variance covariance matrix of the random effects. Because only positive measurements can occur we modelled the log transform of the measurements. An addition advantage of this transform is that it often reduces heteroskedasticity.

We used a Bayesian approach to estimate the model so we must place priors on the parameters. When possible we choose standard noninformative and conjugate priors. The (block) full conditionals for β, σ and Σ_k are of standard form so sampling is straightforward. However, it is not possible to sample directly from the distribution of the subject specific effects so we use a Metropolis algorithm with a multivariate t proposal distribution. This model can be extended to include covariates as follows:

$$\begin{aligned} y_{ijk} &\sim N(\gamma_{i2k} + T_{ijk}\beta_k, \sigma_k^2), \\ T_{ijk} &= B(\exp(\gamma_{i3k})(t_{ij} + \gamma_{i1k}), \\ \gamma_{i..} &\sim N(X_i\alpha, \Sigma_\gamma). \end{aligned} \tag{2}$$

In expression 2 α_k is a parameter that explains how the subject specific effects are related to the covariates X_i. The above model for growth allows to derive reference values for a new subject for a future date given the observed observations from the past. In other words we would like to know the quantiles of the following distribution:

$$p(y_i^F|y_i^P) = \int p(y_i^F|\gamma_i, \beta, \Sigma_\gamma, \sigma)p(\gamma_i|y_i^P, \beta, \Sigma_\gamma, \sigma)p(\beta, \Sigma_\gamma, \sigma)d\gamma_i, \beta, \Sigma_\gamma, \sigma.$$

Here y_i^F refers to the future observation of a new individual while y_i^P for its past observations (ignoring the index for the outcome and the observation times to keep notation simple). In this step we take the distributions of β, Σ_γ and σ as given and do not update them based on y_i^P. This means that for these variables we can use the posterior samples from a model we have already estimated before. Only for γ_i and y_i^F we need to draw new samples from their conditional distributions. As illustrated by Serfling (2002), the concept of quantiles is ambiguous when extended to multiple dimensions in the sense that there are multiple ways in which they could be defined with each of the definitions having some of the attributes that quantiles have univariately but none having all of them. However once we agree on a definition we can use this to obtain multivariate reference values as well. We implemented the sampler for our model in C++ making heavy use of the 'Eigen' and 'Boost' libraries. By writing our own program we have full control over the algorithm. Using C++ results in a considerable speed gain over, for instance, R.

3 Data analysis

To illustrate the multivariate SITAR model we use data from the Predict study, a cohort study that is currently carried out in the Erasmus MC in Rotterdam, the Netherlands. Women are recruited before the sixth week of pregnancy and followed up until delivery. Every week between the sixth and the thirteenth week of pregnancy, they receive a three- dimensional

ultrasound. From these ultrasounds, the Crown-rump length (CRL), Total Arc Length (TAL) and the Embryonal Volume (EV) are determined. The univariate models are run for 300,000 iterations and the multivariate model for 600,000 iterations. For all models we run three chains using different starting values. Convergence was checked visually by examining the trace plots of the Markov chains. We estimate the SITAR model separately for each of the three series (CRL, TAL and EV) and look at the correlation between the subject specific effects obtained in these models. We then estimate the joint model for all the outcomes together. In Figure 3 we see that there is a strong agreement between the horizontal subject specific effects in all three outcomes that is becomes more pronounced when the outcomes are modelled together. This horizontal effect could be interpreted as the measurement error in the Gestational Age. The effect of covariates

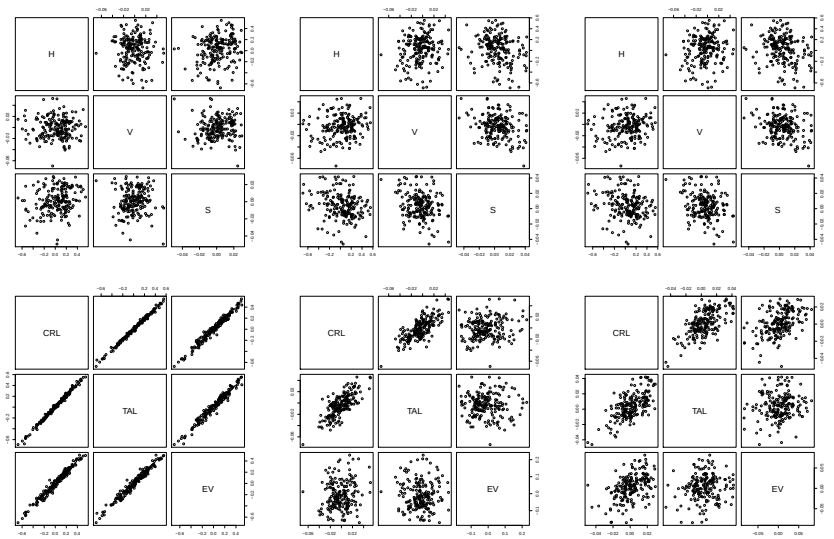


FIGURE 1. Scatterplots of the subject specific effects (On the first row we show the relation between the effects related to the same outcome and on the second row we show the relation between the same kind of effects for the different outcomes)

was investigated. Model validation was performed using posterior predictive checks and by looking at some extensions of the model. Specifically we looked at the skewness and the excess kurtosis of the residual error by replacing its distribution by a skew normal and a t distribution. The usefulness of the multivariate approach is demonstrated by comparing the RMSE and DIC of the univariate models with that of the multivariate model.

4 Discussion

In this study we have used a mixed effects model to study multivariate growth. There are other models for multivariate repeated measurements. We could for example make all coefficients of the spline subject specific. However we think that the interpretation of the subject specific effects is less intuitive in this case. We have shown that the SITAR model can easily be extended to multiple dimensions. This is especially useful to gain insights in the relation between the different aspects of growth. We have tested the multivariate model on data from the PREDICT study. We have also illustrated that our Bayesian estimation procedure allows for altering some of model components easily thereby providing a goodness-of-fit test.

References

Beath, Ken J. (2007). Infant growth modelling using a shape invariant model with random effects *Statistics in Medicine*, **26**, 2547 – 2564.

Cole, Tim J. and Donaldson, Malcolm D C. and Ben-Shlomo, Yoav (2010). SITAR–a useful instrument for growth curve analysis. *Int J Epidemiol*, **39**, 1558 – 1566.

Serfling (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, **56**, 214 – 232.

Inference of non-linear ODE dynamics

Ernst C. Wit [1], Ivan Vujacic [1], Javier Gonzalez [1]

[1] Johann Bernoulli Institute, University of Groningen, Netherlands

E-mail for correspondence: e.c.wit@rug.nl

Abstract: Gene-regulatory systems, signalling pathways and metabolic fluxes are examples in the life-sciences where non-linear dynamics plays an important role. Ignoring single-cell fluctuations, these systems can be described by non-linear systems of differential equations. These models have been very popular in many branches of science due to their flexibility and their ability to describe dynamical systems. Despite the importance of such models in many branches of science they have not been the focus of systematic statistical analysis until recently.
We propose a general approach to estimate the parameters of systems of differential equations measured with noise. Our methodology is based on the maximization of a penalized likelihood where the differential system of equations is used as a penalty. To do so, we use a Reproducing Kernel Hilbert space that allows us to formulate the estimation problem as an unconstrained, easy-to-solve numeric maximization problem. The proposed method is tested in real and simulated examples showing its utility in a wide range of scenarios. We implemented the method as a general purpose package in R.

Keywords: Ordinary differential equations; Reproducing Kernel Hilbert Space; Penalized likelihood; Gene-regulatory systems.

1 Introduction

Ordinary differential equation (ODE) models are the staple modelling tool to represent the dependence on the concentration among mRNA molecules and proteins. Khanin et al. (2007) used a likelihood approach combined with the explicit solution of the ODE to infer the kinetic parameters of the gene regulation model together with the profile of the TF regulator. To estimate the ODE parameters in a gene regulation network, Cao (2008) used a two-step penalization approach originally proposed by Ramsay et al. (2007). Calderhead (2008) proposed a Bayesian method which they applied to the model describing the regulation of genes by the tumour repressor transcription factor protein p53. Auliac et al. (2008) proposed a evolutionary approach for the reverse engineering in gene regulatory networks. Quach et al. (2007) came up with an ODE method to estimate parameters and hidden variables in non-linear state-space models for biological

networks inference. Several differential equation models of transcriptional regulation are studied in Lawrence et al. (2011).
In this article, we propose a new statistical framework to infer the kinetic parameters of gene regulatory network with hidden TFs using time-course expression data. Similar in spirit to the above mentioned likelihood-based approach in Khanin (2007), our procedure is based on the maximization of the data likelihood. The main novelty is that the ODE is interpreted as a constraint. Thus, its solution is not explicitly required to infer the parameters of the system. This simplifies the estimation process and makes the procedure feasible for large scale networks.

2 System and methods

2.1 Modelling transcriptional GRN with ODE models

In gene regulatory networks, the variables of interest are the concentrations of mRNA molecules and the abundance of proteins produced by a set of m genes. For simplicity, in the sequel we will assume that one gene only contains the information to produce one protein. We denote by $\eta(t) = (\eta_1(t), \ldots, \eta_m(t))^T$ the abundance of the proteins (TFs) and by $\mathbf{x}(t) = (x_1(t), \ldots, x_m(t))^T$ the concentrations of the mRNA molecules at time t. We consider that t varies in some time interval $T = [0, \tau]$ in which the GRN is studied. Following mass-action kinetics we assume that the expression of the genes of the network on average satisfy the ODE

$$\dot{x}_k(t) = p(t; \theta_k, \eta) - \delta_k x_k(t), \tag{1}$$

for $k = 1, \ldots, m$, where δ_ks are the degradation rates of mRNAs and $p(t; \theta_k, \eta)$ is a function that describes how the TFs regulate the gene k for some set of parameters θ_k. In general, it is assumed that the TFs satisfy

$$\dot{\eta}_k(t) = \beta_k^{\eta} x_k(t) - \delta_k^{\eta} \eta_k(t),$$

where δ_k^{η} is the protein degradation rate and β_k^{η} is the translational rate for gene k. Let us denote by y_{ki} the measured expression of gene k at a time-point t_i. We assume that the observed gene expression measurements of target genes are conditionally independent given the transcription factor activity and that each target gene k is normally distributed with location parameter $x_k(t)$ and scale parameter $\sigma_k^2(t)$, i.e. $y_{ki} \sim \mathcal{N}(x_k(t_i), \sigma_k^2(t_i))$. Notice that by allowing flexibility in σ^2 with the normal model we can mimic various distributions, *e. g.* when the variance is proportional to the mean, then the normal model can deal with typical log-normal scenarios.

2.2 Discrete formulation

The system of differential equations described by (1) describes the dynamics of the gene regulatory system in the interval T. However, in practical

scenarios, we only have access to a finite number of measurements of the gene expression levels. Following the standard approach in ODE modelling of gene regulatory system, we approximate the rate of gene expression by the first order difference. For $i = 2, \ldots, n-1$ we use the approximation

$$\dot{x}_k(t_i) \approx \frac{x_k(t_{i+1}) - x_k(t_{i-1})}{t_{i+1} - t_{i-1}}. \tag{2}$$

Let us denote by $\mathbf{t} = (t_1 \ldots, t_n)^T$ the measurement time points and $x_k(\mathbf{t}) = (x_k(t_1), \ldots, x_k(t_n))^T$ and $p(\mathbf{t}; \theta_k, \eta) = (p(t_1; \theta_k, \eta), \ldots, p(t_n; \theta_k, \eta))^T$. Then, one can rewrite the dynamics of the gene k in (1) as

$$\mathbf{D}x_k(\mathbf{t}) = p(\mathbf{t}; \theta, \eta) - \delta_k x_k(\mathbf{t}), \tag{3}$$

where the matrix $\mathbf{D}$ is the difference operator

$$\mathbf{D} = \Delta^{-1} \begin{pmatrix} -1 & 1 & & & & \\ -1 & 0 & 1 & & & \\ & & & \ddots & & \\ & & & -1 & 0 & 1 \\ & & & & -1 & 1 \end{pmatrix}, \tag{4}$$

for $\Delta = diag(t_2 - t_1, t_3 - t_1, \ldots, t_n - t_{n-1})$.

2.3 Penalized likelihood

Our aim in this section is to connect (3) with the probabilistic model by means of penalized likelihood. Let us denote by $\mathbf{P}_{\delta_k} = \mathbf{D} + \delta_k \mathbf{I}$, for $\mathbf{I}$ the identity matrix, the difference operator associated to gene k. Rewriting (3) in terms of $\mathbf{P}_{\delta_k}$, we see that the norm

$$\Omega_p(x_k) = \|\mathbf{P}_{\delta_k} x_k(\mathbf{t}) - p(\mathbf{t}; \theta_k, \mu)\|^2 \tag{5}$$

vanishes if and only if $x_k(\mathbf{t})$ satisfies (3). This fact allows us to re-formulate the ODE problem in terms of $\Omega_p(x_k)$. In this sense, following the general idea of penalized likelihood models (Green and Silverman, 1994), one can use (5) to penalize the likelihood,

$$l_{\lambda,k}(S_k; \delta_k, \theta_k, \mathbf{\Sigma}_k, \mu) = l_k(S_k; \delta_k, \theta_k, \mathbf{\Sigma}_k, \mu) + \lambda \Omega_p(x_k) \tag{6}$$

where $\lambda > 0$. By maximizing (6), the fitness of x_k to the data and its smoothness are balanced by means of the parameter λ. Notice that by penalizing the likelihood using $\Omega_p(x_k)$ we explicitly incorporate the information provided by the differential equation to the probabilistic model of the data. However, we do not impose the condition that x_k is a solution of the ODE. This only holds for the extreme case in which we force $\Omega_p(x_k)$ to be zero; this can be done by taking $\lambda \to \infty$. For finite λ, the maximizer of

(6) is generally more stable ill-conditioned problems than those obtained by forcing $\Omega_p(x_k)$ to be zero.
An issue in the approach in (6) is that $\Omega_p(x_k)$ cannot be directly used as a penalty. To overcome this drawback we transform the original problem as follows. Assuming $\mathbf{P}_{\delta_k}$ is invertible, we focus our inferential approach on

$$\tilde{x}_k(\mathbf{t}) = x_k(\mathbf{t}) - \mathbf{P}_{\delta_k}^{-1} p(\mathbf{t}; \theta, \mu), \tag{7}$$

instead of x_k. Note that multiplying by $\mathbf{P}_{\delta_k}$ in both sides of (7) and taking the squared norms we obtain that $\|\mathbf{P}_{\delta_k}\tilde{x}_k(\mathbf{t})\|^2 = \|\mathbf{P}_{\delta_k}x_k(\mathbf{t}) - p(\mathbf{t}; \theta_k, \mu)\|^2$. Equivalently we can write $\Omega_p(x_k) = \Omega_0(\tilde{x}_k) = \|\mathbf{P}_{\delta_k}\tilde{x}_k(\mathbf{t})\|^2$, which is zero when $\tilde{x}_k = 0$. Of course, to focus on $\tilde{x}_k$ requires the transformation of the original observations. Denote by $\mathbf{y}_k = (y_{k1}, \ldots, y_{kn})^T$. Then we define $\tilde{\mathbf{y}}_k = \mathbf{y}_k - \mathbf{P}_{\delta_k}^{-1} p(\mathbf{t}; \theta, \eta)$, for $k = 1, \ldots, m$. It is straightforward to check that $\tilde{y}_{ki} \sim \mathcal{N}(\tilde{x}_k(t_i), \sigma_{ki}^2)$, and therefore the variance of the y_{ki}'s is the same as the variance of the $\tilde{y}_{ki}$'s. Denote by $\tilde{S}_k$ the transformed set of expression measurements associated to the gene k. For a GRN where a single TF η regulates all genes in the network the penalized-log-likelihood can be written as

$$l_\lambda(\Delta, \Theta, \mathbf{\Sigma}, \mu | \tilde{S}) \quad = \quad \sum_{k=1}^{m} l_k(\tilde{S}_k; \delta_k, \theta_k, \mathbf{\Sigma}_k, \mu) - \lambda \sum_{k=1}^{m} \Omega_0(\tilde{x}_k) \tag{8}$$

where $\tilde{S} = \{\tilde{S}_1, \ldots, \tilde{S}_m\}$ represents the whole sample available for the network; Θ represent all sets of kinetic parameters θ_k, $\Delta = \{\delta_1, \ldots, \delta_m\}$, $\mathbf{\Sigma}$ stands for all scale parameters of the normal distribution and μ is the set of weights corresponding to the representation of the TF activity in a spline basis.

2.4 Reproducing kernel Hilbert space framework

It is worthwhile mentioning some of the properties of expression (8). It has been well studied that a penalized likelihood model with a penalty involving derivatives can be understood as a regularization problem in a Reproducing Kernel Hilbert Space (RKHS) (e.g. Berlinet and Thomas-Agnan, 2005).
Consider the penalty $\Omega_0(\tilde{x}_k)$ corresponding to each gene of the network. Then we can write that

$$\Omega_0(\tilde{x}_k) = \|\mathbf{P}_{\delta_k}\tilde{x}_k(\mathbf{t})\|^2 = \langle \mathbf{P}_{\delta_k}\tilde{x}_k(\mathbf{t}), \mathbf{P}_{\delta_k}\tilde{x}_k(\mathbf{t}) \rangle = \tilde{x}_k(\mathbf{t})^T \mathbf{P}_{\delta_k}^T \mathbf{P}_{\delta_k} \tilde{x}_k(\mathbf{t}).$$

Denote by $\mathbf{K}_{\delta_k} = (\mathbf{P}_{\delta_k}^T \mathbf{P}_{\delta_k})^{-1}$. By construction and assuming that $\mathbf{P}_{\delta_k}$ is non-singular, it is guaranteed that the matrix $\mathbf{K}_{\delta_k}$ always exist, it is symmetric and positive definite. That is $\mathbf{K}_{\delta_k}$ is a covariance operator or kernel that defines uniquely a reproducing kernel Hilbert space $\mathcal{H}_K$. Functions in $\mathcal{H}_K$ are characterized by vectors in $\mathbb{R}^n$. Therefore

$$\Omega_0(\tilde{x}_k) = \|\tilde{x}_k(\mathbf{t})\|_K^2 = \alpha_k^T \mathbf{K}_{\delta_k} \alpha_k,$$

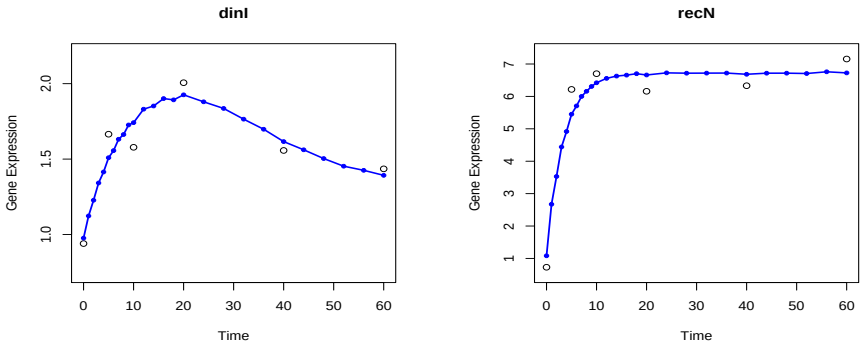


FIGURE 1. Both dinI and recN are upregulated after UV radiation. The levels of the former decline after minute 20, whereas the other levels out.

for α_k the vector in $\mathbb{R}^n$ characterizing $\tilde{x}_k$. See Steinke and Scholkopf (2008) for details. Also, in the RKHS framework, the transformed expression level of each gene k is given by

$$\tilde{x}_k(\mathbf{t}) = \mathbf{K}_{\delta_k}\alpha_k = \mathbf{S}_{\lambda,k}\mathbf{y}_k, \tag{9}$$

where $\mathbf{S}_{\lambda,k} = \mathbf{K}_{\delta_k}(\mathbf{K}_{\delta_k} + 2\lambda\boldsymbol{\Sigma}_k)^{-1}$. Estimates of the gene expression profiles $x_1, \ldots, x_m$ can be recovered using (7).

3 Conclusion and results

The data set used for this experiment is made up of 14 expression genes (dinF, dinI, lexA, recA, recN, ruvA, ruvB, sbmC, sulA, umuC, umuD, uvrB, yegG and ijW) of the Escherichia coli SOS system. The 14 genes are targets of the master repressor LexA and their expression is studied under UV exposure (40 J/m^2) in both wild-type cells and lexA1 mutants. The abundance of the mRNA molecules associated to the genes was measured at six time points, precisely in 0, 5, 10, 20, 40 and 60 minutes. The reconstructed gene profiles show a good fit with the data in the 14 cases. In Figure 1 we show the data and the estimated profiles for the genes dinI and recN. These two genes were selected because they exhibit a different types of profiles.

In this work, we have presented a new ODE-based approach for inferring GRN with one hidden TF from time-course expression measurements. The proposed approach does not require that the transcription factor activity has a predefined shape and a general spline representation allows it to capture the dynamics of the TF. The proposed method has been successfully applied in the reconstruction of the SOS repair system in Escherichia Coli.

In this example, the reconstructed TF exhibits a similar behavior to (independent) experimentally measured profiles. In addition the gene expression data are fitted properly and the results are coherent with those obtained in previous references.

References

Auliac, C., Frouin, V., Gidrol, X. and dAlche Buc, F. (2008) Evolutionary approaches for the reverse-engineering of gene regulatory networks: A study on a biologically realistic dataset. *BMC Bioinformatics*, 9.

Berlinet, A. and Thomas-Agnan, C. (2005) *Reproducing kernel hilbert spaces in probability and statistics*. New York: Springer.

Calderhead, B., Girolami, M. and Lawrence, N. (2008) Accelerating bayesian inference over nonlinear differential equations with gaussian processes. *Neural Information Processing Systems*, **22**.

Cao, J. and Zhao, H. (2008) Estimating dynamic models for gene regulation networks. *Bioinformatics*, **24**, 1619-1624.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.

Khanin, R., Vinciotti, V., Mersinias, V., Smith, C., and Wit, E.C. (2007). Statistical reconstruction of transcription factor activity using Michaelis-Menten kinetics. *Biometrics*, **63**, 816-823.

Lawrence, N.D., Rattray, M., Honkela, A. and Titsias, M. (2011) Gaussian process inference for differential equation models of transcriptional regulation. In: M. P. H. Stumpf, D. J. Balding and M. Girolami (eds) *Handbook of Statistical Systems Biology*, **22**, 376-394.

Quach, M., Brunel, N. and dAlche Buc, F. (2007) Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference. *Bioinformatics*, **23**, 3209-3216.

Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007) Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B*, **69**, 741-796.

Steinke, F. and Scholkopf, B. (2008) Kernels, regularization and differential equations. *Pattern Recognition*, **41**, 3271-3286.

Part III - Contributed Papers (volume 2)

Preface volume 2

This second proceedings volume contains the papers presented as posters at the 28th International Workshop on Statistical Modelling held in Palermo.
It is a sign of the successful evolution of the IWSM that the number of contributions has increased over the years with such a progression that the proceedings book, a small and handy volume in the early editions of the workshop, has grown so thick that carrying it to the conference room and browsing it during the talks has become impractical. As a consequence, we have decided to continue the editorial choice made for the first time at the 27th IWSM in Prague, and have split the proceedings book into two parts: volume 1 containing papers accompanying the invited and orally presented contributed talks, and volume 2, which is composed of papers being presented as posters.

Due to the high number of submissions, two afternoons are entirely devoted to poster sessions to allow the authors to present and discuss their work. This confirms the importance attributed by the workshop to the poster presentations.

We thank all authors of papers included in this volume for participating in the workshop, and for carefully preparing their manuscripts and posters.

Vito Muggeo
Vincenza Capursi
Giovanni Boscaino
Gianfranco Lovison

Palermo, May 2013

Extending frailty models applied to infectious disease epidemiology

Steven Abrams[1], Niel Hens[1,2]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[2] Centre for Health Economics Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

E-mail for correspondence: steven.abrams@uhasselt.be

Abstract: It has been shown that individual heterogeneity in the acquisition of infectious diseases has a large impact on the estimation of important epidemiological parameters such as the (basic) reproduction number. Therefore frailty modelling has become increasingly popular in infectious disease epidemiology. However, so far, using frailty models, it was assumed infections confer lifelong immunity after recovery, an assumption which is untenable for non-immunizing infections. Our work concentrates on refining the existing frailty models to encompass infection processes with reinfections and waning immunity. Shared gamma frailty models, which are frequently used in practice, and correlated gamma frailty models that have proven to be a valuable alternative are considered. We show that naively assuming lifelong immunity in frailty models introduces substantial bias in the estimation of the basic and effective reproduction number. We illustrate our work using Belgian cross-sectional serological data on parvovirus B19 (PVB19) and varicella zoster virus (VZV). Whereas it is typically assumed that lifelong immunity holds for VZV, more recently, empirical evidence for PVB19 indicates waning of immunity after infection, leading to potential reinfections with the virus.

Keywords: shared and correlated gamma frailty models; social contact rates; SIRS transmission model; mass action principle; serological data.

1 Introduction

In recent years, frailty modelling has become increasingly popular in survival analysis to model multivariate event times. Even more so, as individuals differ greatly in their risk of acquiring infections, frailty models found their way into the field of infectious disease epidemiology. Farrington et al. (2001) considered the shared gamma frailty model in the context of bivariate current status data. However, due to its severe limitations, the more flexible correlated frailty model was used by Hens et al. (2009), at the cost of assuming a parametric baseline hazard. From an epidemiological point

of view, frailty models rely on the assumption of lifelong immunity after recovery which becomes untenable for non-immunizing infections. Furthermore, as individual heterogeneity inflates estimates for the basic reproduction number, a correct assessment of heterogeneity, and therefore a correct specification of the infection process, is of utmost importance to obtain reliable estimates for this quantity. In our work, we focus on shared and correlated frailty models for non-immunizing infections. The methodology is illustrated using Belgian current status data on parvovirus B19 (PVB19) and varicella zoster virus (VZV) collected between 2001 and 2003. In addition, a parametric baseline hazard of infection is derived from the mass action principle in which transmission of the pathogen is related to social contact data obtained from the Belgian POLYMOD survey.

2 Materials and methods

Consider bivariate current status data (y_1, y_2, a) with y_i the observed immunological status with respect to infection $i = 1,2$ and a the age of the subject at the cross-sectional sampling time. The binary random variables Y_i, given age a, follow a binomial distribution with probability of being seropositive equal to $\pi_i(a) = 1 - S_i(a)$, and $S_i(a)$ is the proportion susceptible of age a. The age-dependent seroprevalence for both infections can be modelled using frailty models, thereby estimating model parameters $\boldsymbol{\theta}$ while maximizing the multinomial loglikelihood with contribution

$$\begin{aligned} ll(y_1, y_2, a|\boldsymbol{\theta}) &= y_1 y_2 \log\left(1 - S_1(a|\boldsymbol{\theta}) - S_2(a|\boldsymbol{\theta}) + S_{12}(a|\boldsymbol{\theta})\right) + \\ & \quad y_1(1-y_2)\log\left(S_2(a|\boldsymbol{\theta}) - S_{12}(a|\boldsymbol{\theta})\right) + \\ & \quad (1-y_1)y_2\log\left(S_1(a|\boldsymbol{\theta}) - S_{12}(a|\boldsymbol{\theta})\right) + \\ & \quad (1-y_1)(1-y_2)\log\left(S_{12}(a|\boldsymbol{\theta})\right), \end{aligned}$$

From this point onwards, dependence on the model parameters $\boldsymbol{\theta}$ is suppressed from notation. Let Z_i represent a frailty with unit mean and variance σ_{if}^2. For infections in endemic equilibrium and without loss of natural immunity, the susceptible proportion of age a with frailty Z_i is given by

$$S_i(a|Z_i) = \exp\left(-\int_0^a Z_i \lambda_{i0}(u)du\right) = \exp\left(-Z_i M_{i0}(a)\right), \quad i = 1, 2$$

under the proportional hazards assumption (PHA). The unconditional survival functions equal $S_i(a) = \mathbf{L}_i(M_{i0}(a))$, expressed in terms of the Laplace transform $\mathbf{L}_i$ of Z_i and the integrated baseline hazard function $M_{i0}(a)$. Solving the system of ordinary differential equations associated with the mathematical SIRS compartmental model yields:

$$\begin{aligned} S_i(a) &= \exp\left(-\int_0^a \sigma_i(u)du\right)\mathbf{L}_i(M_{i0}(a)) + \\ & \quad \int_0^a \sigma_i(u)\exp\left(-\int_u^a \sigma_i(v)dv\right)\mathbf{L}_i(M_{i0}(a) - M_{i0}(u))du. \end{aligned}$$

when individuals are allowed to flow back from the recovered to the susceptible state at a replenishment rate $\sigma_i(a)$. The bivariate unconditional survival function $S_{12}(a)$ is derived assuming conditional independence of the infection times, given the frailty terms Z_i. In the shared gamma frailty setting, $Z_1 = Z_2 \equiv Z$, where $Z \sim \Gamma(1/\sigma_f^2, 1/\sigma_f^2)$. In the correlated frailty model, we have $Z_1 = \sigma_{1f}^2 (Y_0^* + Y_1^*)$, $Z_2 = \sigma_{2f}^2 (Y_0^* + Y_2^*)$, where $Y_l^* \sim \Gamma(k_l, 1)$ $(l = 0,1,2)$ are independent random variables. The gamma frailty distribution is preferred due to its mathematical convenience and closed-form expression for the Laplace transform.

The time homogeneous mass action principle, which is briefly described here, links the available information on social contact behaviour to the baseline hazard $\lambda_{i0}(a)$. In the presence of individual frailty terms the mass action principle can be rendered as follows (Farrington et al., 2001):

$$\lambda_i(a, Z_i) = \frac{ND_i}{L} \int_0^\infty \int_0^\infty \beta_i(a, Z_i; a', Z_i') \lambda_i(a', Z_i') S_i(a'|Z_i') \phi(a') f_i(Z_i') da' dZ_i'$$

where f_i is the density function of Z_i', $\beta_i(a, Z_i; a', Z_i')$ equals the per capita rate at which an infectious individual of age a' and frailty Z_i' makes an effective contact with a susceptible individual of age a and frailty Z_i, and $\phi(a')$ represents the probability of being alive at age a'. In addition, N, D_i and L are the population size, the mean duration of infectiousness for infection i and the life expectancy, respectively. Under the PHA, $\beta_i(a, Z_i; a', Z_i') = Z_i Z_i' \beta_{i0}(a, a')$ and $\lambda_i(a, Z_i) = Z_i \lambda_{i0}(a)$. Moreover, $\beta_{i0}(a, a')$ is decomposed into a proportionality factor $q_i(a, a'|c)$, representing transmission potential upon a contact, and $c(a, a')$, the annual per capita rate at which individuals of age a' contact individuals of age a. An iterative procedure is used to solve the mass action principle and to derive the baseline hazard of infection thereof. The basic reproduction number R_{i0}, $i = 1,2$, is defined as $(1+\sigma_{if}^2)$ times the dominant eigenvalue of the next generation matrix (Diekmann et al., 1990).

3 Data application

Three shared gamma frailty models are fitted to the serology from PVB19 and VZV. Despite potential reinfections with PVB19, VZV infections are assumed to confer lifelong immunity since accounting for more complex infection dynamics did not improve model fit. The model relying on the assumption of lifelong immunity for both infections is denoted by M1. In addition, model M2 allows for replenishment of the susceptible compartment at a constant rate σ_1 solely for PVB19. Finally, model M3 simply extends model M2 by introducing an age-dependent dichotomous replenishment for PVB19 based on a cut-off value of 35 years.

The results in Table 1 indicate that the models with SIRS dynamics for PVB19 (M2 and M3) outperform the traditional SIR model (M1) based

on AIC-values. Furthermore, the frailty variance is seriously overestimated in model M1 which is reflected as well in the estimated basic reproduction numbers $\hat{R}_{i0}$. Misspecification of the underlying infection process for one infection also influences the estimated reproduction number for the other one.

TABLE 1. ML estimates with regard to PVB19 ($i = 1$) and VZV ($i = 2$) with 95% bootstrap-based CI and corresponding AIC-values.

Model				$\hat{R}_{i0}$		AIC
M1	q_{10}	0.073	[0.069, 0.077]	3.59	[3.27, 3.90]	4537.28
	q_{20}	0.209	[0.189, 0.232]	12.07	[10.46, 13.74]	
	σ_f^2	0.158	[0.102, 0.210]			
M2	q_{10}	0.072	[0.068, 0.075]	3.17	[2.94, 3.43]	4477.98
	σ	0.011	[0.007, 0.014]			
	q_{20}	0.177	[0.162, 0.196]	9.15	[8.07, 10.53]	
	σ_f^2	0.036	[5.4e-7, 0.086]			
M3	q_{10}	0.072	[0.069, 0.075]	3.13	[2.95, 3.38]	4474.39
	σ_1	0.016	[0.010, 0.022]			
	σ_2	0.008	[0.005, 0.012]			
	q_{20}	0.173	[0.161, 0.191]	8.82	[8.01, 10.13]	
	σ_f^2	0.021	[3.6e-7, 0.071]			

4 Discussion

We showed that the use of traditional frailty models results in biased estimates of important epidemiological parameters when incorrectly relying on the assumption of lifelong immunity. Henceforth, frailty models comprising more general infection processes should be considered instead when evidence against natural immunity exists.

References

Diekmann, O. et al. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, **28**, 365 – 382.

Farrington, C. P. et al. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Applied Statistics*, **50**, 251 – 292.

Hens, N. et al. (2009). The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. *Statistics in Medicine*, **27(14)**, 2785 – 2800.

The student *talent* in a random effects Quantile Regression Model for university performance

Giada Adelfio[1], Giovanni Boscaino[1]

[1] Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli studi di Palermo, Italia

E-mail for correspondence: giada.adelfio@unipa.it

Abstract: This paper is a development of a previous work on performance of Italian university students. A Quantile Regression was carried out on a proposed performance indicator, based on a transformation of the median of the marks weighted by credits. Results suggested to investigate the role of students peculiar features on their performance, measured by the transformation of the weighted marks. Therefore, a random intercept Quantile Regression model is fitted on real data concerning graduates over the legal university duration set in Italy, enrolled in 2002 in two Degree Courses of the University of Palermo. Results show that the student performance seems to be influenced by the student's aptitudes, motivation, interest or more in general the students *talent*, rather than socio-demographic characteristics.

Keywords: student performance; random intercept; Regression Quantile.

1 Introduction

In Adelfio et al. (2013) we investigated the determinants of students performance at the end of their educational path. In particular, we introduced a new performance indicator based on a rescaled version of the median of marks weighted by credits. It was robust to outliers and to the asymmetry of the distribution of the marks. That indicator was A quantile regression (QR) approach (Koenker, 2005) was proposed to study its determinants. The results suggested how the proposed approach could be really an useful tool for the policy makers to improve the performance of the university students, above all of those students who can not get the graduation in legal time - a crucial problem for the Italian University System.
In this paper we want to explore the use of QR for the analysis of higher education performance data accounting for the subject-specific features that are different among students. In other words, we want to investigate the effect of the individual student specific source of variability that can affect in a different way each quantiles of the distribution of a scaled version of

weighted marks (i.e. a proper performance indicator), by quantile regression with random effects (Koenker (2004), Geraci and Bottai (2007)).
In a general dependency random effect model, the subject variability, due to the correlation between subject-specific observations, is therefore captured by a random parameter. Usually the simple random intercept is added to the model and its variance informs about the individual specific source of variability, or unobserved heterogeneity, that is not adequately controlled for by other covariates in the model.
With respect to the students performance, we think that this kind of approach could be a useful tool for a deeper analysis, because we account also for all those unobservable features (like aptitudes and motivation).
The paper is organized as follow. Section 2 the used statistical model is described. In Section 3 data, analysis and conclusions are presented.

2 Quantile Regression for repeated measurements

Let $(\mathbf{x}_{ij}^{\mathrm{T}}, y_{ij})$, for $j = 1, ..., n_i$ and $i = 1, ..., N$, be repeated measurements data, where $\mathbf{x}_{ij}^{\mathrm{T}}$ are row p-vectors of a known design matrix and y_{ij} is the jth measurement of a continuous random variable on the ith subject.
According to the considered approach the linear mixed quantile functions of the response y_{ij} is:

$$G_{y_{ij}|u_i}(\tau|\mathbf{x}_{ij}, u_i) = \mathbf{x}_{ij}^{\mathrm{T}}\beta + u_i,\ j = 1, \ldots, n_i,\ i = 1, \ldots, N \tag{1}$$

where $G_{y_{ij}|u_i}(\cdot) \equiv F^{-1}_{y_{ij}|u_i}(\cdot)$ is the inverse of the cumulative distribution function of the response conditional on a location-shift random effect u_i. For this model the location-shift effects are assumed random and identically and independently distributed according to some density f_u, usually $u_i \sim N(0, \alpha)$, characterized by a τ-dependent parameter $(\alpha(\tau))$. Moving away from the penalized approach provided by Koenker (2004), Geraci and Bottai (2007) assume that y_{ij}, conditionally on u_i are independently distributed according as an Asymmetric Laplace Distribution (ALD):

$$f(y_{ij}|\beta, u_i, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right)\right\}$$

where $\mu_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\beta + u_i$ is the linear predictor of the τth quantile, fixed and known, and σ is the usual scale parameter. The random effects, that induce a correlation structure among observations on the same subject, are assumed to be
That is a likelihood-based approach to the estimation of the regression quantiles based on the ALD and it is better then the penalized fixed effects based approach in terms of mean squared error of the QR estimators. Alternative models with non-normally distributed residuals were developed (Seltzer and Choi, 2002).

We have chosen to use the approach based on ALD, which provides an automatic choice of the optimal level of penalization, also because it represents a suitable error law for the least absolutes estimator and therefore a natural choice in a quantile regression approach.

3 Data Analysis

We apply a linear QR model with a subject-specific random intercept that accounts for within-group correlation with respect to a cohort of students enrolled in 2002 and graduated from 5 up to 7 years after (the out-of-legal-duration graduates), in two Degree Courses: the most numerous Degree Course of the Faculty of Economics - that is Economics and Finance (E) with 131 students - and the most numerous Degree Course of the Faculty of Sciences - that is Life Sciences (L) with 98 students - both of the University of Palermo, Italy.
In Adelfio et al. (2013) the students performance was summarized by the median of a transformation of their own wighted marks. Results suggested to consider the transformed weighted marks distribution of each student. In our opinion each student differs from others not only for socio-economical aspects but also for aptitudes, abilities, interests, motivations, etc. All these aspects can be caught by the variability of marks rather than by a synthetic intensity indicator as the median. In fact, different marks distributions can have the same median and with it we can not able to distinguish different performances. Different performances, measured by different marks distributions, could be explanatory of different students *talent*. The transformed marks weighted by credits, as reported in Adelfio et al. (2013), for $i-th$ student and $j-th$ course, is:

$$m'_{ij} = \frac{12}{max_j(m^w_{ij}) - min_j(m^w_{ij})} \times (m^w_{ij} - min_j(m^w_{ij})) + 18 \qquad (2)$$

where $m^w_{ij} = \frac{m_{ij}C_j}{\sum_{j=1}^{J} C_j}$, with m_{ij} is the mark and C_j is the credit for the course j. We model m'_{ij} as a function of the following covariates in a with random intercept QR model: High school diploma type (Lyceum vs Not Lyceum), Residence (Living in Palermo vs Not living in Palermo), Gender (Female vs Male), Degree Course (Life Sciences vs Economics), and High school diploma mark (centered at the mean).
In figure 1 results of analysis are reported, with respect to the fixed coefficients estimates. To assess the suitability of (1), results are also commented in the light of those reported in Adelfio et al. (2013), not reported here for the sake of brevity. For each of the estimated coefficients we plot the QR estimates of the fixed parameters of (1), conditional to each quantile τ ($\tau = 0.05, 0.25, 0.50, 0.75, 0.95$), by the dashed curve with filled dots. These points may be interpreted as the impact of a unit-change of each covariate on the response variable, fixed the others. The grey area represents the

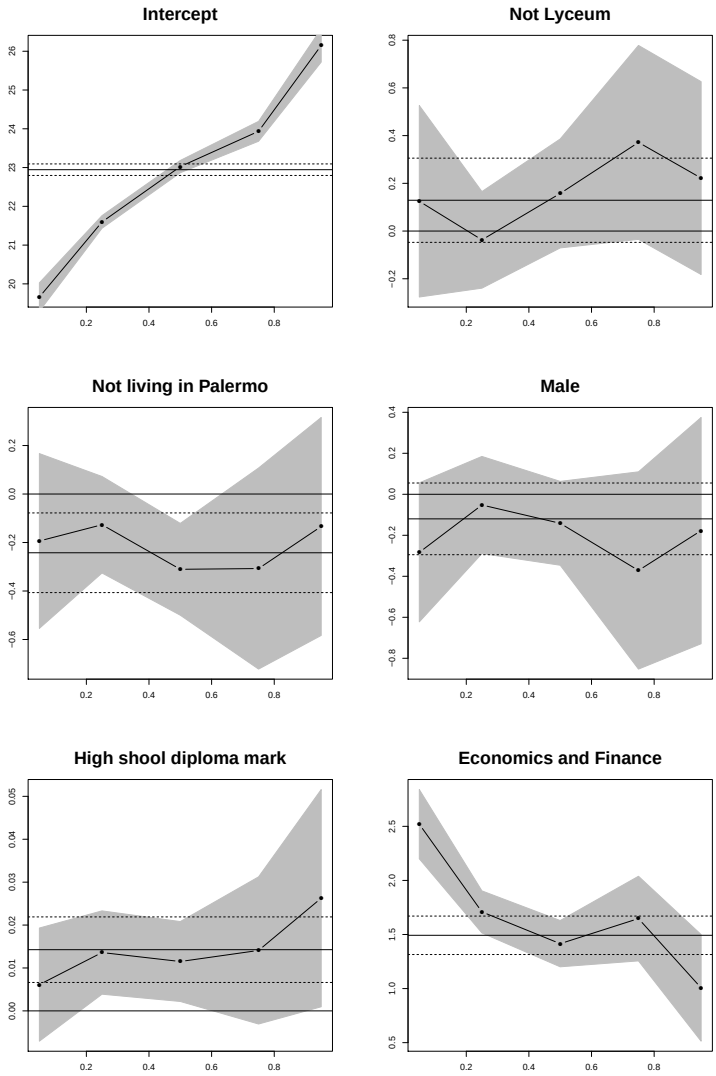


FIGURE 1. Fixed coefficients estimates of random intercept QR model.

95% pointwise confidence band. The solid horizontal line, together with its 95% confidence intervals (horizontal dashed lines), refers to the estimate for the Linear Mixed Model. The Intercept panel refers to the expected m'_{ij} conditional to each quantiles for the female, living in Palermo, with a Lyceum diploma, enrolled in Life Sciences Degree Course, with a mean High school diploma mark equals to 89.85 student. It is steeper than the

TABLE 1. τ dependent estimated variance of u_i random intercept

τ	0.05	0.25	0.50	0.75	0.95
$\hat{\alpha}(\tau)$	0.151	0.104	0.132	0.454	0.378

QR model without random effect: considering the m'_{ij}'s distribution rather than their median allows to appreciate the variation of the expected performance (vertical axis) when we move between two consecutive quantiles (horizontal axis). With the exception of the last one, other panels show no significative effect – in most of the quantiles – of the covariates. This it is a partial confirmation of that the student performance is mainly due to the student own *talent* rather than socio-demographic characteristics. In fact, the significative effect of Degree Course covariate reflects the different performances among students. As reported in Adelfio et al (2013), we suppose that students that enrolled in Life Sciences might be more motivated than those of Economics and Finance. In fact, E and L offer quite different educational study plans: the first one offers a study plan with subjects as mathematics, law, economic history, etc., while the second one offers a study plan with more specific subjects such as physics, chemistry, botany, etc.. In our opinion, who chooses L might have a real interest and passion for the topic, while who chooses E might be also motivated by job market opportunities and the general-interest educational study plan. These students peculiar characteristics are not directly measured, therefore the random intercept aims to catch their effects on performance. In fact, the τ dependent estimated variances of u_i random intercept of (1) are in $[0.104, 0.454]$ and they reflect the heterogeneity among students due to their own unobservable *talent* (tab.1).

References

Adelfio, G., Boscaino, G., Capursi, V. (2013) Quantile Regression on a new indicator for higher education performance. Legal Deposit n.2545TR2013

Geraci M. and Bottai M. (2007) Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, **8**, 140 – 154.

Koenker, R. (2004) Quantile Regression for Longitudinal Data *Journal of Multivariate Analysis*, **91**, 74 – 89.

Koenker, R. (2005) Quantile Regression, *Cambridge University Press.*

Seltzer, M. and Choi, K. (2002) Model checking and sensitivity analysis for multilevel models. In: *N. Duan and S. Reise (Eds.), Multilevel modeling: Methodological advances, issues, and applications*, Hillsdale, NJ: Lawrence Erlbaum.

Item selection via Bayesian graded response model

Serena Arima[1]

[1] Dipartimento di Metodi e Modelli per l'economia, il territorio e la finanza, Sapienza Universitá di Roma, Roma, Italy

E-mail for correspondence: serena.arima@uniroma1.it

Abstract: The number of items included in a questionnaire is usually large, leading to a time consuming and expensive administration, and possibly inaccurate response. The main goal of this paper is to define a model-based procedure for reducing the number of items in a questionnaire so that its reduced version has the same characteristics in terms of latent trait evaluation of the complete one. We propose a mixed cumulative logit models, known in the psychometrics literature as graded response model: the responses to the different items are modelled as function of the individual latent trait and as function of items characteristics, such as their difficulty and their discriminant power. We model the discriminant and the difficulty parameters jointly using a mixture of k Normal distributions. Mixture components correspond to disjoint groups of items and items belonging to the same component can be considered equivalent in terms of difficulty and discriminant power. According to decision criteria, we select a subset of items such that the reduced questionnaire is able to provide the same information of the complete one. The model is estimated using a fully Bayesian approach and the choice of the number of mixture components is justified according to information criteria. The proposed method is applied to a questionnaire for the quality of life in dysarthric speakers and compared with competing models proposed in literature.

Keywords: Graded response model; mixture distribution; MCMC.

1 Introduction

In recent years evaluating people using tests is very common and there has been a considerable interest among psychologists and statisticians in developing a theory that allows to improve educational and psychological tests. One of the main difficulty in using tests is that the number of items included in a questionnaire is usually large, leading to a time consuming and expensive administration. Moreover, a large amount of questions make the respondents tired and may lead to possibly inaccurate responses. For this reason, it is of interest to develop a methodology that allows us to select a subset of items such that the reduced questionnaire has the same

characteristics of the full one. Item response theory (IRT) is the area of psychometry that deals with the problem of tests construction, item calibration and with the evaluation of latent ability the test aims at measuring. Item response models are latent trait models in which the probability of correct responses are modelled as function of examinees' ability and as function of items characteristics, such as their difficulty levels and their discriminant powers. Motivated by our data, coming from a questionnaire for the quality of life in dysarthric speakers, we will focus on the graded response model (Samejima (1969)). We propose an extension of the graded response model in which the discriminant and difficulty parameters are modelled jointly as a $k-$component mixture distribution. The mixture allows us to cluster homogeneous items in terms of discriminant power and difficulty: a decision rule is proposed in order to select the subset of items to be used in the reduced version of the questionnaire.

2 The proposed model

Let Y_{ij} represent the response variable of the subject $i-$th for the item $j-$th, with $i = 1, ..., n$ and $j = 1, ..., r$. The variable Y_{ij} is a categorical response with H possible ordered categories. The n subject are assumed independent. Let

$$p_{ijh} = P(Y_{ij} = h|\Theta = \theta) \quad h = 1, ..., H$$

denote the probability that a subject i with latent trait (or ability) level θ responds by category h to item j. Obviously we assume that $\sum_{h=1}^{H} p_{ijh} = 1$. We will focus on a Bayesian version of the so-called cumulative logit model defined as the following multi-stage model (Wollack et al. (2012)):

Stage.1 $logit[P(Y_{ij} \geq h|\theta, \gamma, \beta)] = \gamma_j(\theta_i - \beta_{jh})$

Stage.2 $log(\gamma_j)|\theta, \beta \sim N(m_\gamma, s^2_\gamma)$

$\theta_i|\gamma, \beta \sim N(0, 1)$

$\beta_{jh}|\gamma, \theta \sim N(\beta_j, s^2_j)$ with the constraint $\beta_{j1} < .. < \beta_{jH}$

Stage.3 $\beta_j \sim N(\mu_\beta, \sigma^2_\beta)$

In the first stage we define the likelihood of the model. The second stage is devoted to the specification of the prior distributions: item discriminant power is modelled with a log-normal distribution and the latent trait as a standard normal distribution, for model identifiability. With respect to the difficulty parameter, constrained prior distributions must be specified in order to have $\beta_{j1} < .. < \beta_{jH}$. Hyperparameters $m_\gamma, s^2_\gamma, \mu_\beta, \sigma^2_\beta$ are fixed in order to have non-informative prior distributions.

In Bartolucci et al. (2012), items are selected according to their discriminant power: the reduced questionnaire consists of the subset of items whose discriminant power is larger than a fixed cut-off value. However, items are characterized by both discriminant and difficulty parameters: hence, in order to select subsets of equivalent items, in terms of both difficulty and discriminant power, we propose to model jointly these parameters according to a $K-$components mixture distribution

$$\begin{pmatrix} log(\gamma_j) \\ \beta \end{pmatrix} \sim \sum_{k=1}^{k} \pi_k N(\mu_k, \Sigma)$$

The mixture weights π_k are given a Dirichlet distribution with K parameters equal to 1; non-informative hyperprior distributions have been used for μ and Σ. In this context, items belonging to the same mixture component can be considered as equivalent in terms of both discriminant and difficulty power and may be selected in a reduced version of the questionnaire. The posterior distributions of the parameters of interest cannot be obtained analytically and ad-hoc MCMC algorithms are used in order to simulate samples from them and identifiability constraints adopted for the mixture components (Celeux et al. (2000)).

2.1 Item selection criteria

Once the model has been fitted, we have to select the subset of items in the reduced questionnaire. This choice is strongly related with the final goal of the analysis:

- we aim at reproducing a reduced equivalent version of the questionnaire, with m items ($m < n$);
- we aim at obtaining a reduced version of the questionnaire containing items with high level of difficulty and high discriminant power.

Both solutions may be of interest for the practitioners. In the first situation, we believe that our original questionnaire measures the latent trait appropriately and we would like to have an equivalent but reduced version of it, as in the case of the evaluation of the quality of life in dysarthric speakers. Indeed, this solution defines a questionnaire with a variety of items, including simple or less discriminant items that could have been included in the original questionnaire as distractors or to encourage the respondent. Hence, let $w_k = \frac{n_k}{n}$ where n_k is the number of items belonging to $k-$th cluster: once the items have been ordered according to their discriminant power, we will select $m \cdot w_k$ items from each group. In this way, the reduced questionnaire should reproduce the same characteristics of the original one. The second solution aims at obtaining an optimized version of the questionnaire, selecting the first more discriminant and difficult m items. In

practice, we will select m items from the first mixture components. A similar strategy has been adopted in Bartolucci et al. (2012), when the items have been selected only according to the discriminant power.
Another important point is the selection of the number of mixture components: we will fit several models with increasing number of mixture components and select the optimal number of mixture components according to information criteria such as DIC.

3 Data description and Results

Quality of life (QOL) is an important concept in healthcare and deals with how people feel about their health. In this work we will focus on a questionnaire aimed at evaluating the quality of life of patients affected by dysarthria, a motor speech disorder resulting from neurological injury of the motor component of the motor-speech system. The questionnaire consists of 40 items and it has been administrated to 105 patients: each question has 4 possible ordered answers about the quality of life (bad, moderate, sufficient, good, high). Practitioners aim at having a reduced version of the questionnaire with $m = 15$ items., The proposed model has been fitted using $k = 1, 5, 10, 15, 20$ groups and we obtain the lowest DIC for $k = 5$. Both criteria presented in Subsection 2.1 have been applied and compared in terms of predicted latent trait: both reduced questionnaires estimate the latent trait satisfactory and the estimates are in agreement with those obtained with the complete questionnaire.
We compare the results with the method in Bartolucci et al. (2012) highlighting pro and cons of both methods.

References

Bartolucci,F., Montanari,G.E., and Pandolfi,S. (2012) Domensionality of the latent structure and item selection via latent class multidimensional IRT models. *Psychometrika* **77(4)**, 782 – 802.

Celeux, G., Hurn, M., Robert, C. P. (2000) Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957 – 970.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.

Wollack, J.A., Bolt, D.A., Cohen, A.S., Lee, Y.S. (2012). Recovery of Graded Response Model Parameters: A Comparison of Marginal Maximum Likelihood and Markov Chain Monte Carlo Estimation, *Applied Psychological Measurement* , **36**, 399 – 419.

Two valid and interpretable metrics to summarize raw accelerometry data

Jiawei Bai[1], Bing He[1], Thomas A. Glass[2] and Ciprian M. Crainiceanu[1]

[1] Department of Biostatistics, Johns Hopkins University, USA
[2] Department of Epidemiology, Johns Hopkins University, USA

E-mail for correspondence: jbai@jhsph.edu

Abstract: We introduced a pair of explicit metrics for human activity based on high density acceleration recordings from a hip worn tri-axial accelerometer: 1) Time Active, a measure of the length of time when activity is distinguishable from rest and 2) Activity Intensity, a measure of relative amplitude of activity relative to rest. They both measure the level of activity of human, but characterize it in different aspect. They are normalized (having the same interpretation across subjects and days), easy to explain and implement, and reproducible across platforms and software implementations. Metrics were validated by visual inspection of results, quantitative in-lab replication studies.

Keywords: Tri-axial Accelerometer; Physical Activity; Signal Processing; Accelerometry; Validation.

1 Introduction

A commonly used outcome measure in aging research is the capacity to engage in activities of daily living (ADLs), or sentinel behaviors required to live independently. Conventional methods for measuring ADL include self-reported questionnaires or clinician ratings based on observed behavior, which have several limitations, including recall bias and the lost of minute by minute information. A better approach is to use "accelerometers", which allows collection of real-time, densely sampled information on movement. However, translating information from high volume and complex data from wearable sensors into acceptable measurements can be done only by careful standardization and transformation to guarantee the validity and reproducibility of the measurements. Current measurements produced by software that accompanies these devices are usually expressed in "activity counts", which though sharing the same name, often have different definitions for different manufactures (Ancoli-Israel et al., 2003). Moreover, other devices produces activity counts that, while formally defined,

do not have a clear interpretation, and may not capture sufficient variability in older subjects. Here we propose data normalization and a pair of novel, explicit, and interpretable metrics that can be used in medical and epidemiological studies.
The manuscript is laid out as follows. Section 2 describes the structure of our data and introduces necessary notations. Section 3 covers in detail the metrics that we proposed. We also discussed the validity of these metrics using the data of the in-lab study.

2 Data

Our data were collected from elder men and women from an ongoing cohort study, the Baltimore Memory Study. Subjects were asked to wear the accelerometer during waking hours for 4-5 consecutive days, removing the device during showering and swimming. They are the data that we mainly focus on. There was also an in-lab session for some of the subjects, in which they were asked to perform a series of activities including walking and chair stands. Later we would do the validation using the data from the in-lab sessions. The accelerometer generated tri-axial voltage in three orthogonal axes, as proxies of acceleration in three direction, at a sample rate of 10Hz. Denote the data by $\mathbf{X}_i(t) = \{X_{i1}(t), X_{i2}(t), X_{i3}(t)\}$, $t = 1, 2, ..., T_i$, where T_i is the length of the accelerometer time series for Subject i. In this paper we used field data from 34 subjects and each subject was observed from 4 to 5 days. So $i = 1, 2, \ldots, 34$ and T_i is very large. We define the activity label time series $L_i(t) \in \{0, 1\}$ as $L_i(t) = 1$ when Subject i is active at time point t and $L_i(t) = 0$ otherwise. $L_i(t)$ was estimated by a tailored method originally proposed in Bai 2011. Once $L_i(t)$'s are estimated, we divide the entire recorded time period of Subject i into two sets of time points, T_i^A and T_i^I, corresponding to active and inactive time periods. Specifically, $\forall t \in T_i^A$, $L_i(t) = 1$ and $\forall t \in T_i^I$, $L_i(t) = 0$. We also denote by J_i the number of days when Subject i is observed, while T_i is the total number of time points where the subject is observed. The number of days, J_i varies between 3 and 5 days. Let t_{ij}^0 be the time index for start of day j, which has a total of T_{ij} data points. The number of data points per day, T_{ij}, can depend on the particular day, j, because it only describes the time when the subject is not in bed, which can vary every day.

3 Methods

Based on the raw data and the label $L_i(t)$, 2 basic metrics were proposed: Time Active and Activity Intensity. These metrics are defined so that they reflect different aspects of the physical activity condition of human.
Time Active (TA) ($\mathrm{TA}_i(k) = \frac{\sum_{s=1}^{W} L_i\{(k-1)W+s\}}{W}$) is the proportion of time that were declared active in the fixed length window $[(k-1)W+1, kW]$

for Subject i. It measures the overall active level in any given time interval. Length of the interval provides a lot of flexibility of summarizing the data. Panel 1 in Figure 1 shows the TA bars of Subject 3092 in every 15-minute interval. Each TA bar ranges from 0 to 1, and is rendered in light blue (TA $\leq$ 0.3), red (0.3 $<$ TA $<$ 0.7) or purple (TA $\geq$ 0.7). Substantial difference of TA between day and night can be observed, while TA during the day has values in a wide range but that during the night is simply near 0.

Activity Intensity (AI) ($\mathrm{AI}_i(t) = \max\left\{\frac{1}{3}\sum_{m=1}^{3}\frac{\sigma_{im}(t)-\overline{\sigma}_{im}}{\overline{\sigma}_{im}}, 0\right\}$) is similar to TA, which measure the overall active level, but focus on the actual intensity of movement (amplitude of signal) instead of the binary measurement of active/inactive. It is computed by considering the variation of the tri-axial signal σ_{im}, after removing the systematic noise of the device $\overline{\sigma}_{im}$. Panel 2 of Figure 1 visualizes the AI of the same subject. Each AI bar is colored light blue, red or purple according to their TA values as in Panel 1. Apparently, two panels in Figure 1 share a lot of similarity regarding to the active level of Subject 3092, but Panel 2 helps distinguish between long-lasting low intensity activities and short-lasting high intensity activities (i.e. long-time walking versus short-time running).

The probability density functions of AI during walking and chair stands of the in-lab session are shown in Figure 2. AI is calculated for every second and displayed as black bars under the corresponding raw-data plots. The densities of AI during walking are quite consistent within- and between-subjects. Similar results were found for all 10 subjects with in-lab data. The density curve of AI for chair-stands is different, though it displays

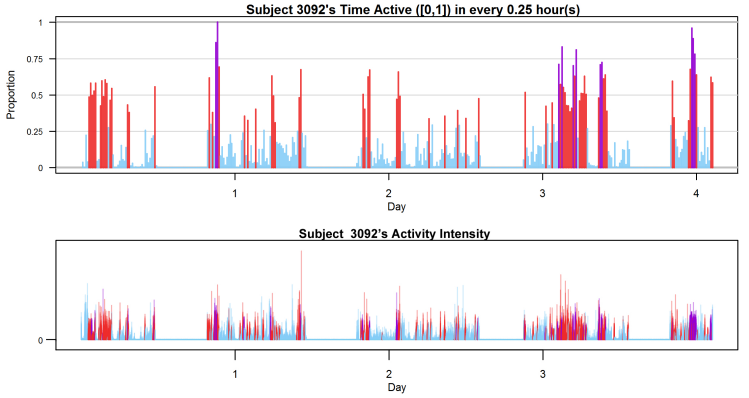


FIGURE 1. Panel 1: Time Active (TA) bars for Subject 3092 (panel 1). Each TA bar has a value between 0 to 1, and is colored light blue (TA $\leq$ 0.3), red (0.3 $<$ TA $<$ 0.7) or purple (TA $\geq$ 0.7). Panel 2: 3-day Activity Intensity (AI) for Subjects 3092. AI bars are colored light blue, red or purple according to their TA values as in Panel 1.

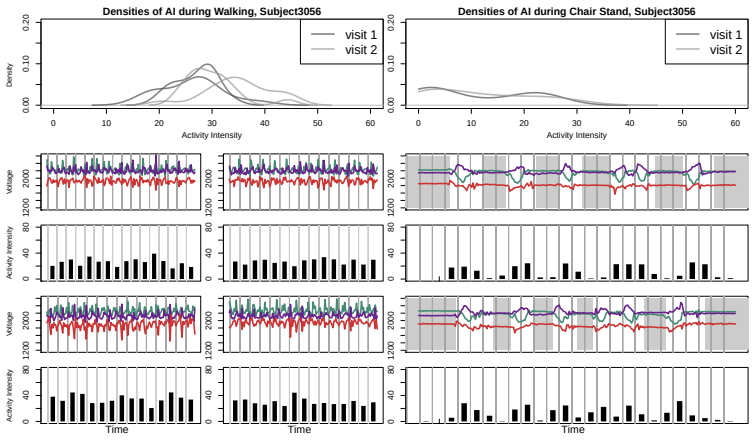


FIGURE 2. The plots for the metrics validation for Subject 3056 during two different visits. In each visit, there are two replicates of walking and three replicates chair stands. The figure shows the density curves of the AI in the top panel and the raw accelerometry signal as well as the AI during the activities.

a lot of similarity within subjects with more variability across subjects. The difference in histogram shapes between walking and chair standing is probably due to the fact that chair standing consists of three different sub-activities: resting, standing-up and sitting-down. The AI for chair-stands is low during inactive periods and high during active periods. AI during these sub-activities were quite similar within subjects across visits.

4 Conclusion

The need of accurately measuring human activity (especially daily human activity) urged researchers to deploy accelerometers into the studies. However, the information extraction from the raw data was a tough task, especially while the consistency and reproducibility are required to be guaranteed. We proposed Time Active and Activity Intensity, which are valid, transparent and reproducible, as summaries of the raw data. They reflected the characteristics of human activities in different aspects.

References

Bai, J. (2011). Accelerometer-based prediction of activity for epidemiological research. *Master's Thesis*. Johns Hopkins University.

Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W. and Pollak, C. (2003). The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, **26**(3), 342 – 392.

Recursive Residuals Application in Linear Mixed Models

Ahmed S. Bani-Mustafa [1], Kenan M. Matawie[2]

[1] ALFAISAL University-PSCJ, Jeddah, KSA
[2] University of Western Sydney, Sydney, Australia

E-mail for correspondence: a.ajjour@pscj.edu.sa

Abstract: This paper presents recursive residuals definition, formulae and application to Linear Mixed Models (LMM). The approach of estimation is based on the estimation method of fitting-of-constants . Model fit is also assessed through a graphical display of the recursive residuals and their Cumulative Sums (CUSUM).

Keywords: Fitting-of-Constant; LMM; Recursive Estimation; Recursive Residuals.

1 Introduction

Recursive residuals are useful and powerful analysis tools for a wide variety of fixed effect models, particularly in providing diagnostic tests for detecting serial correlation, heteroscedasticity, functional misspecifications and structural change in regression models. Together with estimation of the model parameters they have the best statistical properties (including independency) and provide intuitive graphical tools for investigating changes of model parameters overtime using the CUSUM test. The approach we consider here is to present the recursive residuals and their estimates for LMM based on well-known LMM estimation, Henderson's fitting-of-constants method.

2 Recursive Residuals for LMM

Recursive estimation is a technique for updating parameter estimates where the resulting change in the estimates is proportional to the recursive residuals. The recursive residual corresponding to an observation Y_t at time t, is the scaled difference between Y_t and its best predictor using observations recorded prior to time t. Thus, current and successive predictors of Y_t are computed recursively based on parameter estimates from observations prior to t.

Let Y_t be the continuous observation on the dependent variable at time t corresponds to vectors of regression variables x_t and s_t. The LMM we consider at time t can be expressed as

$$\mathbf{Y}_t = \mathbf{x}_t\boldsymbol{\beta} + \mathbf{s}_t\mathbf{u} + E_t \ , \quad t = 1, 2, ..., n. \tag{1}$$

where at time t there are t observations (the first t observations) in the following matrices and vectors and the remaining observations $n - t$ are considered zeros (i.e., $t + 1$,$t + 2$,...,n), y is the response vector $n \times 1$, X is the observed design matrix for the fixed effect $n \times p$ matrix, $\boldsymbol{\beta}$ is the unobserved parameter vector of fixed effects $p \times 1$, $\mathbf{S}$ is the observed design matrix for the random effect $n \times r$, $\mathbf{u}$ is the vector of unobserved random effect $r \times 1$, with $E(\mathbf{u}) = \mathbf{0}$ and $Var(u) = \sigma_u^2\mathbf{I} = \mathbf{G}$, $\boldsymbol{\varepsilon}$ is the error term vector $n \times 1$, assumed to be independent and normally distributed with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2\mathbf{I} = \mathbf{R}$. Assume that all levels of upertain to the same source of variation, such that $Var(u) = \sigma_u^2\mathbf{I} = \mathbf{G}$ and $Cov(\boldsymbol{u;\varepsilon}) = \mathbf{0}$. the recursive estimates $\widehat{\boldsymbol{\beta}}_t$ and $\widehat{\mathbf{u}}_\mathbf{t}$ for $\boldsymbol{\beta}$ and $\mathbf{u}$ at time t are the solutions Henderson's equations for the model in (1) recursively, which may be obtained as follows:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_t \\ \widehat{\mathbf{u}}_t \end{bmatrix} = \mathbf{H}_\mathbf{t}^{-\mathbf{1}} \begin{bmatrix} \mathbf{X}_t^T\mathbf{Y}_t \\ \mathbf{S}_t^T\mathbf{Y}_t \end{bmatrix} \tag{2}$$

where

$$\begin{aligned} \mathbf{H_t} &= \begin{bmatrix} \mathbf{X}_t^T\mathbf{X}_t & \mathbf{X}_t^T\mathbf{S}_t \\ \mathbf{S}_t^T\mathbf{X}_t & \mathbf{S}_t^T\mathbf{S}_t \end{bmatrix} \\ &= \mathbf{H}_{t-1} + \mathbf{z_t}\mathbf{z_t^T} \ , \end{aligned}$$

and

$$\mathbf{z_t} = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{s}_t \end{bmatrix}.$$

using the Sherman-Morrison formula (Bartlett 1951) and the recursive notation of McGilchrist et al. (1983), $\mathbf{H}_t^{-1}$ can be written as

$$\mathbf{H}_{t-1}^{-1} - c_t^{*-1}\mathbf{g}_t^*\mathbf{g}_t^{*T} \tag{3}$$

where

$$c_t^* = 1 + \mathbf{z}_t^T\mathbf{H}_{t-1}^{-1}\mathbf{z}_t \ \text{ and } \ \mathbf{g}_t^* = \mathbf{H}_{t-1}^{-1}\mathbf{z}_t.$$

The method of calculating recursive residuals, always starts with initial estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$ as a vector of zeros and correspondingly, for $t = 0$, $\mathbf{H}_t = \mathbf{0}$, its inverse $\mathbf{H}_t^{-1}$ and their product $\mathbf{H}_t^{-1}\mathbf{H}_t$ are taken to $(p + r) \times (p + r)$ matrices of zeros. As observations are added, $\mathbf{X}_{t-1}, \mathbf{S}_{t-1}, \mathbf{y}_{t-1}$ are replaced by $\mathbf{X}_t, \mathbf{S}_t, \mathbf{y}_t$ and the rank of $\mathbf{H}_t$ remains the same or can

increase by one. Thus, the parameters and recursive residuals are estimated progressively using the following recursive residuals formula

$$W_t = \begin{cases} c_t^{*-1/2} \left[Y_t - \mathbf{z}_t^T \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{t-1} \\ \widehat{\mathbf{u}}_{t-1} \end{pmatrix} \right] & \text{rank stay the same} \\ 0 & \text{rank increases} \end{cases}$$

3 Example

To illustrate the application and computation of these developed formula for recursive residuals and estimates for LMM we used Nobre & Singer (2007) data. The data relates to a comparison of the capacity to remove bacterial plaque with continuous daily use with a low cost monoblock toothbrush against a conventional toothbrush. Indices of plaque in 32 children (aged 4 - 6 years) were measured before and after tooth brushing at four evaluation sessions. The data is an example of repeated measurements, taken on the same experimental units over four evaluation sessions, adjusting for pretreatment bacterial plaque indices (Nobre & Singer, 2007).
Our reframing of Nobre & Singer (2007) LMM in a matrix form at time t as:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{S}_t\mathbf{u} + \boldsymbol{\varepsilon}_t \ , \quad t = 1, 2, ..., 128 \tag{4}$$

y is a 128×1 response vector of post-treatment bacterial plaque indices, $\mathbf{X}$ is a 128×3 fixed effects design matrix (intercept, two types of toothbrush and log(pre-treatment)) and $\mathbf{S}$ is a 128×32 random effects design matrix (subject effect), $\boldsymbol{\varepsilon}$ is a 128 errors vector normally distributed with zero mean and $\sigma_\varepsilon^2\mathbf{I}$ variance. The vector $\boldsymbol{\beta}$ is a 3×1 vector of fixed effects that are unknown and u is a 32×1 vector of random effects normally distributed with zero mean and $\sigma_u^2\mathbf{I}$ variance and $\mathbf{u} \perp \boldsymbol{\varepsilon}$.
The LMM obtained was significant, with t-values of (-10.027,16.087,2.98) for the intercept, pretreatment and treatment effects respectively. The LMM appeared to be satisfactory. A plot of the standardized residuals $\widehat{\boldsymbol{\varepsilon}}$ versus the fitted values, and the normal quantile plot for the residuals were satisfactory except for an indication of a possible two outliers. These results are similar to those given by Nobre & Singer (2007) who investigated three techniques of residuals analysis for LMM using the same data.
Recursive residuals normal quartile plots and their CUSUM and normal quartile plots appear in Figures (1a, 1b). The CUSUM at observation i is $\sum_{t=1}^{i} W_t$ where W_t is the recursive residual at observation t.
The normal probability plot Figure (1a) shows an approximately straight line with the data points at the two ends of the line indicating possible low values and/or outliers. However, the CUSUM plot Figure (1b) shows an initial upward trend followed by a downward trend indicating some sort of

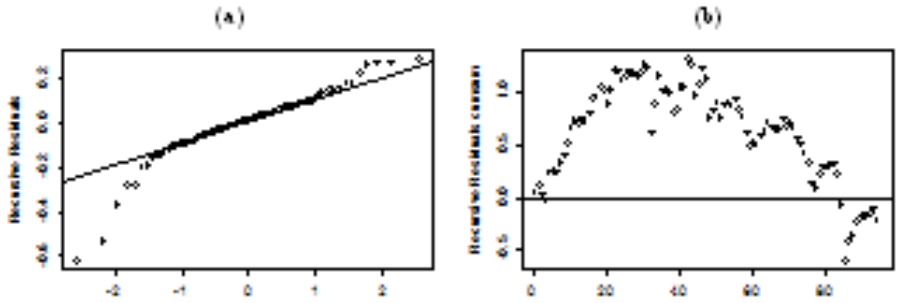


FIGURE 1. (a) Normal Quantile-Quantile plot and (b) Recursive Residuals CUSUM graph

model misfit with a negative CUSUM ($\sum W_t$ = -0.231). Model misfit such as this may be due to many reasons includes outliers, an omitted variable, incorrect model specification or incorrect model underlying distributional assumptions. For more details see Hawkins (1991) and Kianifard & Swallow (1996). Dealing with model misfit may improve the above model, but this is not the aim of this example.
It should pointed out that in their initial data paper, Nobre & Singer (2007), also identified the same two outliers, which were identified by standardized residuals and recursive residuals plots shown above. Removing these two outliers did not improve the fit of the model or CUSUM plot.

References

Bartlett M. (1951). An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*, 107 – 111.

Hawkins D. (1991). Diagnostics for use with regression recursive residuals. *Technometrics* **33(2)**, 221 – 234.

Henderson C. (1953). Estimation of variance and covariance components. *Biometrics*, 226 – 252.

Kianifard F., Swallow W. (1996). A review of the development and application of recursive residuals in linear models. *Journal of the American Statistical Association*, **91**, 391 – 400.

McGilchrist C., Sandland R., Hennessy J. (1983). Generalized inverses used in recursive residuals estimation of the general linear model. *Australian & New Zealand Journal of Statistics*, **25(2)**, 321 – 328.

Nobre, J., Singer, J. (2007). Residual analysis for linear mixed models. *Biometrical journal*, **49**, 863 – 75.

Modelling the growth of Abdominal Aortic Aneurysms using mixed effects regression with autocorrelated residuals

Paul D. Baxter[1], Marc A. Bailey[2], D. Julian A. Scott[2]

[1] Division of Biostatistics, University of Leeds, UK
[2] Leeds Vascular Institute, Leeds Teaching Hosptials Trust, UK

E-mail for correspondence: p.d.baxter@leeds.ac.uk

Abstract: An Abdominal Aortic Aneurysm (AAA) occurs when the walls of the abdominal aorta weaken. Once an AAA develops, its size increases, although growth varies considerably between individuals for reasons that are not yet well understood. If an AAA ruptures then severe internal bleeding occurs, which is often fatal. However, corrective surgery caries risk (particularly amongst the elderly, co-morbid patients that typically develop AAA) and would not usually be considered until risk of rupture (which relates directly to aneurysm size) exceeds surgical risk. Understanding the growth process to identify how it might be monitored and modified is therefore of great interest. We conclude mixed effects regression models with autocorrelated residuals provide a parsimonious and stable approach to growth modelling. The approach is illustrated using data from the Leeds Aneurysm Development Study (LEADS).

Keywords: Autocorrelation; Mixed Effects Regression.

1 Background

An AAA ultrasound screening programme is currently being introduced across the National Health Service in England, UK. The screening programme will invite all men for screening during the year they turn 65. Patients with AAA greater than 5.5cm in diameter are (subject to co-morbidities) offered either open or minimally invasive corrective surgery. For AAAs between 3.0cm and 5.5cm, the risk of death caused by surgery is higher than the risk of rupture. Patients therefore undergo regular repeat screening. Longitudinal data from ultrasound screening of 297 AAA patients in LEADS is available for secondary analysis. The screening protocol used at Leeds Teaching Hospital Trust is shown in Table 1. Adherence to the protocol is shown as a boxplot in Figure 1(a). Descriptive statistics for the 297 patients is shown in Table 2 and example growth trajectories from 49 randomly selected patients in Figure 1(b).
Using the LEADS data described above, we aim to investigate how to model growth whilst respecting the data structure of longitudinal measurements nested within patients and allowing for the possibility of non-linear growth.

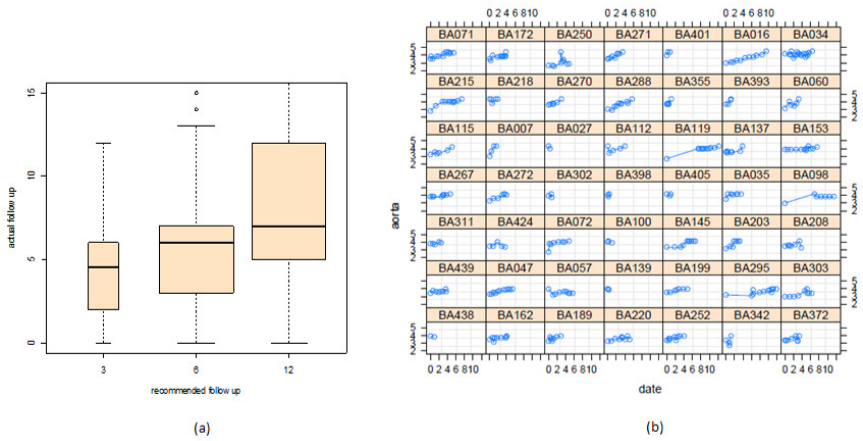


FIGURE 1. (a) boxplot of recommended follow up time (months) versus actual follow up time (months); (b) plot of example growth trajectories, date from initial diagnosis (years) versus AAA diameter (cm).

2 Methods

Simple approaches to growth modelling such as fitting a single regression line with fixed slope and intercept pooled across patients do not allow for departures from linearity and violate the assumption of independent observations. However, they are widely used (see Bailey et al., 2011). Non-linear models of quadratic and logistic growth are also of interest, but

TABLE 1. AAA screening protocol used at Leeds Teaching Hospital Trust.

AAA diameter (cm)	Recommended scan protocol (months)
< 3.0	no follow up
$\geq 3.0, < 4.0$	annual
$\geq 4.0, < 5.0$	6 months
$\geq 5.0, < 5.5$	3 months

TABLE 2. Descriptive statistics for 297 patients in the LEADS data.

Measurement	Median	IQR
Initial AAA diameter (cm)	3.7	1.0
Final AAA diameter (cm)	5.0	1.2
Observation period (years)	3.2	3.3
Number of AAA scans	6	5

again suffer from violation of independence assumptions when fitted as fixed effects models pooled across patients.

Linear mixed effects models (see, for example, Chapter 2 of Pinheiro and Bates, 2000) can be fitted, for example using the `nlme` library in R (R Core Team, 2012). Linear mixed models with random slopes and intercepts can be considered a reasonable basic model that respects the structure of growth measurements nested within patients (see, for example, Sweeting et al., 2010). However, departures from linear growth are ignored in this approach. Extension to non-linear mixed effects models is possible (see Chapter 6 of Pinheiro and Bates, 2000), though (in our experience) can suffer convergence issues.

As an alternative modelling approach to allow non-linearity, we consider a linear mixed effects model with random slopes and intercepts where the model residuals are autocorrelated (see Section 5.3 of Pinheiro and Bates, 2000). Autocorrelation can be addressed by Box-Jenkins autoregressive moving average (ARMA) structures, although these do now explicitly allow for the unequal spacing of observations through time. Alternatively, isotropic variogram models of spatial autocorrelation (e.g. exponential, Gaussian, spherical) can be considered that explicitly allow for unequal spacing.

3 Results

Table 3 shows degrees of freedom, AIC, BIC and log-likelihood for:

1. A linear mixed effects model with random and fixed slope and intercept.
2. A non-linear mixed effects model with random and fixed slope, quadratic time effect and intercept.
3. A linear mixed effect model with random and fixed slope and intercept & AR(2) residuals.
4. A linear mixed effect model with random and fixed slope and intercept & residuals modelled by a Gaussian isotropic variogram model.

TABLE 3. Model fit statistics for linear and non-linear mixed effects models.

Model	df	AIC	BIC	LogLik
1	6	1937	1971	-962
2	10	1832	1888	-906
3	8	1834	1879	-909
4	8	1848	1894	-916

A non-linear mixed effects logistic model (see Equation 6.7 of Pinheiro and Bates, 2000) did not converge - we speculate due to the unequal spacing of observations in the LEADS data (a design feature of the screening protocol). Exponential and spherical isotropic variogram models gave the same AIC and BIC as the Gaussian isotropic variogram model shown in Table 3. Depending on whether AIC or BIC is preferred, either the quadratic model (2) or the AR(2) residuals model (3) is parsimonious.

4 Conclusions

Mixed effects models allow a simple approach to modelling AAA growth that respects the hierarchical structure of the data - growth measurements nested within patients. Allowing for non-linearity in the growth process is challenging due to the unequal spacing of observations through time. A simple approach to allow for non-linearity by modelling autocorrelation in residuals shows promise, although further work is required to identify the most appropriate autocorrelation structure.

Acknowledgments: We are grateful to the Investigators of the Leeds Aneurysm Development Study (LEADS) for providing access to the data.

References

Bailey, M.A., et al. (2011). A systematic review of the methodology employed to calculate abdominal aortic aneurysm growth rate. *Ultrasound*, **19**, 197 – 202.

Pinheiro, J.C., and Bates, D.M. (2000). *Mixed Effects Models in S and S-PLUS*, New York: Springer.

R Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Sweeting, M.J., Thomson, S.G., Brown, L.C., and Greenhalgh, R.M. (2010). Use of angiotensin converting enzyme inhibitors is associated with increased growth rate of abdominal aortic aneurysms, *Journal of Vascular Surgery*, **52**(1), 1 – 4.

Finding profiles in time-course gene expressions

Marco Bazzi[1], Paola Tellaroli[1]

[1] Department of Statistics, University of Padova

E-mail for correspondence: tellaroli@stat.unipd.it

Abstract: Recent developments in the high-throughput screening techniques lead to collect an enormous amount of data in biological experiments. Sophisticated statistical tools are needed in order to investigate these data. Biologists are often interested in measuring the gene expression evolution over time. Identifying genes that have a similar pattern over time, it is a crucial step to determine the genes that play a role in a specific cellular function. Due to this temporal data structure, the usual clustering techniques are not applicable. The main contribution of this work is to propose a new clustering technique that lead to determine the suitable number of groups in which the genes follow a similar temporal path. In order to assess the efficiency of our method, we compare it with the one most used in literature, called microarray Significant Profiles (maSigPro). We present the results of our analysis conducted on real data.

Keywords: B-splines; Clustering algorithms; Distance measure; DNA microarray; Gene expression data.

1 Introduction

1.1 The problem

Genomic experiments generate large and complex multivariate data sets, so the scientific community is making great efforts in developing new statistical methods *ad hoc* to handle this kind of data. Furthermore, thanks to advanced technologies we are now able to observe over time the evolution of the gene expression.

In time-course microarray experiments, the expression of a certain cell is measured in some time points during a particular biological process. By knowing groups of genes that are expressed in a similar fashion through a biological process, biologists are able to infer gene function and gene regulation mechanisms (Quackenbush, 2001; Slonim, 2002). Another peculiar feature of this experiment is the small number of time points available, both because multiple arrays are very expensive, both because, even if prices go down, short time series experiments would remain prevalent since in many studies it is prohibitive to obtain large quantities of biological material.

1.2 Clustering methods

The direct application of the standard clustering methods to the analysis of temporal profiles is difficult because they typically assume that the observation for each gene in different experiments is independent. This assumption holds when expression measures are taken from independent biological samples, such as different subjects or different experimental conditions. However, it is not valid when the observations may be related to the ones at the previous time points.
We follow the proposal of Abraham (2003) where the temporal evolution of the phenomenon is considered as a curve, fitted by particular spline. After estimating the coefficients of these non parametric functions, they identify groups of genes applying clustering techniques directly to coefficients.
Several authors use a model-based clustering, for example James (2003) proposed a mixed effects spline model for analyzing yeast cell cycle data or Ray (2006) suggested a non-parametric Bayesian method where a mixture of Dirichlet processes is used to determine the number of groups and the elements in each of them. However, the probabilistic specification of these models is a problematic issue and the computational time could be long, especially with high dimensional data sets.
Conesa et al. (2006) propose a general regression-based approach for the analysis of single or multiple microarray time series. This methodology, named maSig-Pro (microarray Significant Profiles) is a two-step regression strategy where model parameters have to be adjusted according to the data under study and the specific interests of the researcher.

1.3 Cross-clustering tecnique

As main contribution of our work, we developed a simple technique called Cross-clustering which combines two well-established hierarchical clustering algorithms (complete linkage and Ward algorithm). Our approach permits to identify a suitable number of clusters according to an intuitive optimization criterion and it isolates outlier genes, not forcing them into a group.

1.4 Paper organization

Our paper is organized as follows: in Section 2 the general methodological ideas is presented, introducing data transformation based on B-splines coefficients and our Cross-clustering algorithm; in Section 3 we show the application on real data; while Section 4 is devoted to the discussion and final remarks.

2 Methods

2.1 Data normalization

The underlying idea of clustering analysis is that genes acting together belong to similar, or at least related, functional categories. If two gene profiles show the same trend of up- or down-regulating, then it is reasonable to consider these two co-regulated, even if their levels of expression are quite different. As underlined by Huber (2008), microarray measurements in most cases carry no meaningful physical units and this happens because no universal units are associated with the feature intensity values measured on a microarray. For these reasons, it is a common choice to consider the log-ratio transformation of data. For the $i-$th gene, separately for each experimental condition l, we consider

$$y_{til} = \log_2 \frac{z_{til}}{z_{1il}} \quad t = 1, \cdots, T,$$

where z_{til} is the gene expression observed at time t. The use of the logarithm in base 2 is a popular convention in the field.

2.2 B-splines coefficients

Abraham (2003) proposed to consider the evolution of gene expression in a certain time interval as a functional data. Thus, it becomes advantageous to exploit this functional structure for partitioning the observations into groups. This method consists in two steps: first to fit the curves by non-linear non parametric models and then to cluster the estimated model coefficients. Although linear models could be used for this purpose, they are often too restrictive to capture the underlying phenomenon. On the other hand, the flexibility of tools such as splines permits to reach satisfactory results with a limited number of coefficients.
Formally, we assume that the measurements of the curves $G_i(t)$ are done with independent random errors ε_{it} which can be thought as added to the underlying smooth evolution of gene expression. The aim of the procedure is to remove this noisy part, focusing on the smooth interesting part. After summarizing each curve by a few coefficients, which capture the smooth part of the process with enough flexibility, we have to partition these coefficients. Abraham (2003) suggest to use B-splines to fit the curves and this choice seems to be good also in this context. In many applications, a linear combination of third-degree B-splines is enough flexible to investigate the non-linear evolution on time of several phenomena.

2.3 Cross-clustering algorithm

To identify a proper number of clusters and to assign genes to them, we combine in an efficient way two of the most known and well-established

hierarchical clustering algorithms: Ward and Complete linkage. The first seems to be able to identify a suitable number of clusters. The latter does not select the number of clusters in an appropriate way but it has the great fashion to isolate very clearly outliers. Comparing clusters obtained with these two procedures, we can find a proper number of clusters and, at the same time, remove from these groups isolated genes.

3 Application

In a real experiment mice were divided into two groups: the first group was subjected to a pharmacological procedure to block the nervous system that controls the muscle activity, called *denervation*, while the second one is the control group. The genetic material was extracted from two muscles: a frequently active (slow) and a sporadically active (fast). In this way, four experimental conditions are obtained. The expression of genes was recorded every 4 hours for 24 hours (0h, 4h, 8h, 12h, 16h, 20h). Biologists are interested in genes that change over time, called *oscillating genes*, according to two conditions: at least 1 of the 6 values must be above 100 and the ratio between maximum and minimum expression value must be greater than 1.5 in all conditions. We have a data set containing 1164 *oscillating* genes organized in 24 samples (4 experimental conditions $\times$ 6 time points).

4 Conclusion

In this work, we propose a new algorithm which bypass the limitations arising when we apply classical clustering techniques to time dependent data. With a simple iterative procedure, we group those genes with a similar temporal path, avoiding to classify genes with a strange trend, considered outliers. We compared slow and fast non denervated (1vs3), and slow versus fast denervated (2vs4). Cross-clustering finds less clusters with respect to maSigPro procedure, but it classifies more genes. We also calculated the Dunn Index, which is an internal validation measure which consists in the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. It takes values between zero and ∞, and should be maximized. In our case, the Dunn Index is extremely higher for the Cross-clustering method.

Acknowledgments: Special thanks to Silvio Bicciato (Department of Biomedical Sciences, University of Modena and Reggio Emilia), Alessandra Brazzale (Department of Statistics, University of Padua) and Kenneth Dyar (Venetian Institute of Molecular Medicine) for their useful support which led to significant improvement of this work.

References

Abraham, C. and Cornillon, P.A. and Matzner-Lber, E. and Molinari, N. (2003). Unsupervised Curve Clustering using B-Splines *Scandinavian Journal of Statistics*, **30**, 3, 581 – 595.

Conesa, A., Nueda, M. J., Ferrer, F. and Talon, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments *Bionformatics*, **22**,9, 1096 – 1102.

Curry, H.B. and Schoenberg, I.J. (1966). On Plya frequency functions IV: the fundamental spline functions and their limits *Journal d'analyse mathmatique*, **17**,1 71 – 107.

Hardle, W.K. and Simar, L. (2012). *Applied multivariate statistical analysis.* Springer

Huber, W. (2008). Fold-Changes, Log-Ratios, Background Correction, Shrinkage Estimation, and Variance Stabilization In: *Bioconductor case studies*, Springer, pp. 63 – 82,

James, G.M. and Sugar, C.A. (2003). Clustering for sparsely sampled functional data *Journal of the American Statistical Association*, **98**, 462, 397 – 408.

Quackenbush, J. (2001). Computational analysis of microarray data *Nature Reviews Genetics*, **2**,6 418 – 427.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods *Journal of the Royal Statistical Society: Series B*, **68**,2 305 – 332.

Slonim, D.K. (2002). From patterns to pathways: gene expression data analysis comes of age *Nature genetics*, **32**, 502 – 508.

Wasserman, L. (2005). *All of Nonparametric statistics.* Springer

A note on improved random effects prediction in GLMMs

Ruggero Bellio[1], Paolo Vidoni[1]

[1] Department of Economics and Statistics, University of Udine, via Tomadini 30/a, I-33100 Udine (Italy)

E-mail for correspondence: paolo.vidoni@uniud.it

Abstract: This paper concerns prediction of random effects, and in particular of expected responses, in generalized linear mixed models, with emphasis on the construction of prediction intervals having conditional coverage probability closed to the target nominal value. Some theoretical results are briefly presented, and some easy-to use formulas are applied to obtain improved random effects prediction intervals in the logistic-normal model.

Keywords: Coverage probability; Generalized linear mixed models; Prediction interval.

1 Introduction and preliminaries

The prediction of the value of a future random variable, based on an observed sample, is usually expressed in terms of prediction intervals. This paper presents some results on prediction of random effects, and in particular of expected responses, in generalized linear mixed models, abbreviated as GLMMs (see e.g. Booth and Hobert, 1998, Skrondal and Rabe-Hesketh, 2009, and references therein). The general aim here is to define a simple procedure for defining prediction intervals which have a conditional coverage probability closed to the target nominal value, and improve on the simple plug-in solution.

Let us assume that response data are arranged in $k \geq 1$ groups and that the random variable Y_{ij} describes the response of unit j in the i-th group, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$. The m-dimensional continuous random vector $U_i = (U_{i1}, \ldots, U_{im})^T$, $i = 1, \ldots, k$, is the unobservable random effect associated with the i-th group; $U_1, \ldots, U_k$ are independent, identically distributed, and the pairs (Y_i, U_i), $i = 1, \ldots, k$, with $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$, are independent. Moreover, conditional on $U_i = u_i$, the responses in the i-th group are independent, having conditional density, with respect to a suitable dominating measure,

$$p_{ij}(y_{ij}|u_i; \beta, \lambda) = c(\lambda, y_{ij}) \exp[\lambda\{y_{ij}\theta_{ij} - K(\theta_{ij})\}], \quad y_{ij} \in \mathcal{Y} \subseteq \mathbf{R},$$

where $\theta_{ij} = x_{ij}^T\beta + z_{ij}^T u_i$ is the linear predictor, with $x_{ij} = (x_{ij1}, \ldots, x_{ijq})^T$ and $z_{ij} = (z_{ij1}, \ldots, z_{ijm})^T$ known covariate values and $\beta = (\beta_1, \ldots, \beta_q)^T$ a q-dimensional parameter. Indeed, the mean is $\mu_{ij} = \mu(\theta_{ij})$, whereas $\lambda \in \Lambda \subseteq \mathbf{R}^+$ is the index parameter and $\sigma^2 = 1/\lambda$ is the dispersion parameter. The class of GLMMs is obtained by considering a monotonic differentiable link function $g(\cdot)$ such that $g(\mu_{ij}) = x_{ij}^T\beta + z_{ij}^T u_i$. Here, $\mu_{ij} = g^{-1}(x_{ij}^T\beta + z_{ij}^T u_i)$ and $\theta_{ij} = \theta(g^{-1}(x_{ij}^T\beta + z_{ij}^T u_i))$, with $g^{-1}(\cdot)$ and $\theta(\cdot)$ the inverse of $g(\cdot)$ and $\mu(\cdot)$, respectively. If the canonical link function $g(\cdot) = \theta(\cdot)$ is considered, we obtain $\theta_{ij} = x_{ij}^T\beta + z_{ij}^T u_i$. With regard to the random effects, we assume that $U_i = (U_{i1}, \ldots, U_{im})^T$, $i = 1, \ldots, k$, follows a m-dimensional Gaussian distribution with null mean vector and $\Sigma = \mathrm{diag}(\gamma)$, $\gamma = (\sigma_1^2, \ldots, \sigma_m^2)^T$, as variance matrix.

The interest here is on prediction of one-dimensional transformations $R = R(U_i, \omega)$ of the random effects U_i and, in particular, on linear combinations of the form $x_{ij}^T\beta + z_{ij}^T U_i$ or the corresponding mean response $R = g^{-1}(x_{ij}^T\beta + z_{ij}^T U_i)$. In particular, prediction should be based on the conditional density of R given $Y_i = y_i$

$$f(r|y_i, \omega) = \frac{f(r;\omega)\prod_{j=1}^{n_i} c(\lambda, y_{ij}) \exp[\lambda\{y_{ij}g(r) - K(g(r)\}]}{L_i(\omega; y_i)}, \qquad (1)$$

with $\omega = (\beta^T, \sigma^2, \gamma)$. Function $f(r;\omega)$ is the marginal density of R, which may be computed explicitly since in this case the linear combination $x_{ij}^T\beta + z_{ij}^T U_i$ follows a Gaussian distribution with mean $x_{ij}^T\beta$ and variance equal to $\sum_{s=1}^m z_{ijs}^2\sigma_s^2$. If the normalizing function $L_i(\omega; y_i)$, which corresponds to the i-th likelihood component, is not known explicitly, an estimate based on numerical or approximation-based techniques is required.

2 Improved prediction limits

The aim is to provide a relatively simple procedure for predicting random effects U_i, $i = 1, \ldots, k$, and associated one-dimensional transformations R, by means of prediction limits with good coverage properties. More precisely, prediction fit evaluation is done conditionally on the observed value of the response Y_i, as proposed by Booth and Hobert (1998) for point predictors of expected responses in GLMMs.

A simple solution involves the estimative or plug-in prediction limit $\hat{r}_\alpha = r_\alpha(\hat{\omega}, y_i)$ obtained by substituting the unknown parameter ω with an asymptotically efficient estimator $\hat{\omega} = \hat{\omega}(Y)$ (usually the maximum likelihood estimator) in the α-quantile of the conditional distribution of R given $Y_i = y_i$. That is, $r_\alpha(\omega, y_i) = F^{-1}(\alpha|y_i;\omega)$, with $F^{-1}(\cdot|y_i;\omega)$ the inverse of the distribution function $F(\cdot|y_i;\omega)$ associated to (1). It is well-known that estimative prediction limits are usually imprecise, since the associated conditional coverage error can be substantial. It is possible to prove that the

conditional coverage probability of $\hat{r}_\alpha$ is

$$\begin{aligned}\hat{\alpha}(\omega, y_i) &= P_{Y,R|Y_i}\{R \leq \hat{r}_\alpha | Y_i = y_i\} = E_{Y|Y_i}\left[F\{\hat{r}_\alpha | Y_i; \omega\} | Y_i = y_i\right] \\ &= \alpha + O(\max\{n_i^{-1}, k^{-1}\}),\end{aligned}$$

where the expectation is with respect to the conditional distribution of Y given $Y_i = y_i$. The component of order $O(n_i^{-1})$ of the coverage error term is related to the Laplace approximation for the i-th likelihood component $L_i(\omega; y_i)$, and it can be nil when accurate computation is employed.
Moreover, in order to reduce the asymptotic order of the coverage error term, the Ueki and Fueda's (2007) procedure has been applied, giving the following simple modified estimative prediction limit

$$\tilde{r}_\alpha(\hat{\omega}, y_i) = 2r_\alpha(\hat{\omega}, y_i) - r_{\hat{\alpha}(\omega, y_i)}(\hat{\omega}, y_i),$$

where the estimative coverage probability $\hat{\alpha}(\omega, y_i)$, since usually unknown, may be estimated by means of a suitable bootstrap parametric technique conditional on $Y_i = y_i$. It is possible to prove that the conditional coverage error of this modified prediction limit is of order $o(\max\{n_i^{-1}, k^{-1}\})$, thus reduced with respect to that of the estimative solution.

3 The logistic-normal model

As an illustrative example, we consider the logistic-normal example already analysed by Booth and Hobert (1998). In particular, they analysed data from a multicenter clinical trial on a two-treatment comparison, fitting a logistic model with random clinic effects. Namely, the linear predictor is given by

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij} + u_i\,,$$

where x_{ij} is the binary treatment indicator, and $U_i \sim N(0, \sigma_1^2)$ the random intercept, $i = 1, \ldots, 8$, $j = 1, \ldots, n_i$ $(13 \leq n_i \leq 73)$, and a logistic regression model is adopted for the binary response given the random effect.
The method of the previous section has been applied to obtain modified predictive limits for the clinic random effects u_i with improved coverage properties. In particular, the integrals required for the likelihood function and the predictive distribution $f(u_i|y_i; \omega)$ have been approximated by Gaussian quadrature, and the improved prediction limits computed with 2,000 bootstrap replications. The results are illustrated in Figure 1, for all the clinics.
The modification is noticeable, for the primary reason that the number of groups is just $k = 8$. A small scale simulation study has been run, obtaining an estimated coverage for the estimative procedure below the target level, being actually between 66% and 85% for the various groups. This is actually in good agreement with the simulation studies reported

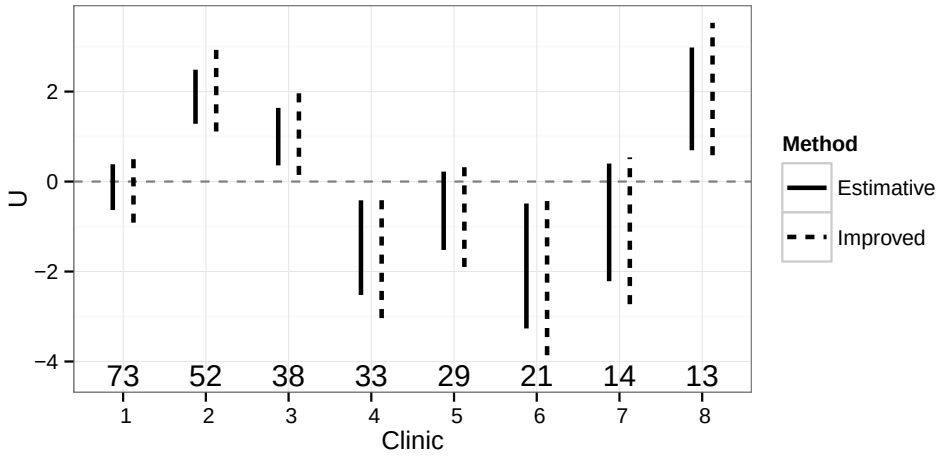


FIGURE 1. 95% prediction limits for clinic random effects, with annotated clinic sample size.

in Ten Have and Localio (1999), that illustrate the need for an improved prediction procedure.
Research on speeding-up the computation of the modified limits is currently ongoing, as efficient implementations is required for estimating the coverage properties of the improved method. Indeed, future research will consider computational aspects as well as applications to some models of interest, for various formulation of $R(U, \omega)$.

References

Booth, J.G. and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, **93**, 262 – 272.

Ten Have, T.R. and Localio, A.R. (1999). Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics*, **55**, 1022 – 1029.

Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society, Series A*, **172**, 659 – 687.

Ueki, M. and Fueda, K. (2007). Adjusting estimative prediction limits. *Bio-metrika*, **94**, 509 – 511.

Calibration model with scale mixtures of skew-normal distributions

Betsabé G. Blas Achic[1], Marcos Antonio A. Pereira[1]

[1] Federal University of Pernambuco, Recife-Brazil

E-mail for correspondence: betsabe@de.ufpe.br

Abstract: This work presents a new statistical linear calibration model with replicate measurements by assuming that the error model follows the family of scale mixtures of skew-normal distributions, which is a class of asymmetric thick-tailed distributions that includes the skew-normal distribution. The parameter estimates are found using an EM algorithm. The new approach is applied to a real dataset from chemical analysis.

Keywords: Calibration Curve; Mixture Scale; EM algorithm.

1 Introduction

The calibration model is applied in different areas, and it is composed of two stages. In chemical analysis, the purpose is establish a quantitative relationship over the two stages, for the first stage it is between known concentrations of an analyte and their observed response variables, and on the second stage it is between an unknown concentration and the related observed response variable. The main interest is estimate this unknown concentration (Blas *et al.*,2007).

This paper discusses a new calibration model with replicate measurements by assuming that the error model follows a family of scale mixtures of skew-normal (SMSN) distributions, as introduced by Branco and Dey (2001). We write $Y \sim SMSN(\mu, \sigma^2, \lambda)$, which means, the random variable Y follows a SMSN distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ. The probability density function of Y is $f(y) = 2\int_0^\infty \phi(y|\mu, \sigma^2\kappa(u))\Phi\left(\lambda\frac{y-\mu}{\sigma}\right) dH(u;\tau)$, where $y \in \mathbb{R}$, $\phi(.)$ denotes the density of univariate normal distribution with mean μ and variance $\sigma^2 > 0$ and $\Phi(.)$ is the distribution function of the standard univariate normal distribution. U is a random variable with distribution function $H(.,\tau)$ and density $h(.,\tau)$ and τ is a scalar or vector parameter indexing the distribution of U. In this work we consider $\kappa(u) = 1/u$, which leads to good mathematical properties. The SMSN family is a flexible class of distributions for robust estimation. One particular case is the skew-normal (SN) distribution which is arrived

when H is degenerated, with $u = 1$. The SMSN class also includes distributions such as the skew-t (ST), skew-slash (SSL) and the skew-potenial exponential(SPE) distribution.

2 Linear Calibration Model with SMSN distributions error

The SMSN linear calibration model is given by

$$y_{ij} = \alpha + \beta x_i + \epsilon_{ij},\ i = 1, 2, \ldots, n \text{ e } j = 1, 2, \ldots, r_i, \quad (1)$$
$$y_{0i} = \alpha + \beta x_0 + \epsilon_{0i},\ i = n+1, n+2, \ldots, n+m, \quad (2)$$

where y_{ij} and y_{0i} are observed responses for the fixed value x_i and the unknow quantity x_0, respectively. α, β and x_0 are unknown parameters. ϵ_{ij} and ϵ_{0i} are independent and identically distributed (iid) SMSN with 0 location parameter, scale parameter σ^2 and skewness parameter λ. The EM algorithm for the proposed model parameters are presented on the following.
The model (1-1) can be written hierarchically as

$$\begin{aligned}
Y_{ij}|T_{ij} = t_{ij}, U_{ij} = u_{ij}, &\overset{\text{iid}}{\sim} N\left(\alpha + \beta x_i + t_{ij}\frac{\sigma\lambda\kappa(u_{ij})}{\sqrt{s}}, \frac{\sigma^2\kappa(u_{ij})}{s}\right)\\
U_{ij} \overset{\text{iid}}{\sim} H(u_{ij};\tau) \quad &\text{and} \quad T_{ij} \overset{\text{iid}}{\sim} NH(0,1),\ i = 1, \ldots, n \text{ e } j = 1, \ldots, r_i,\\
Y_{0i}|T_{0i} = t_{0i}, U_{0i} = u_{0i}, &\overset{\text{iid}}{\sim} N\left(\alpha + \beta x_0 + t_{0i}\frac{\sigma\lambda\kappa(u_{0i})}{\sqrt{s_0}}\frac{\sigma^2\kappa(u_{0i})}{s_0}\right)\\
U_{0i} \overset{\text{iid}}{\sim} H(u_{0i};\tau) \quad &\text{and} \quad T_{0i} \overset{\text{iid}}{\sim} HN(0,1),\ i = n+1, \ldots, n+m,
\end{aligned}$$

where $HN(0,1)$ denotes the half-$N(0,1)$ distribution, $s = 1 + \lambda^2\kappa(u_{ij})$ and $s_0 = 1 + \lambda^2\kappa(u_{0i})$. The parameter τ from the mixing variable is fixed previously, as recomended by Lange K. L. *et al.* (1989). Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_R^\top)$, $\boldsymbol{u} = (u_1, \ldots, u_R)^\top$, $\boldsymbol{t} = (t_1, \ldots, t_R)^\top$, $\boldsymbol{y_0} = (y_{01}, \ldots, y_{0m})^\top$, $\boldsymbol{u_0} = (u_{01}, \ldots, u_{0m})^\top$, $\boldsymbol{t_0} = (t_{01}, \ldots, t_{0m})^\top$ and $R = \sum_{i=1}^{n} r_i$. Then, under the hierarchical model (1-1), it follows the complete log-likelihood function $\ell_c(\boldsymbol{\theta}|\boldsymbol{y_c})$ associated with $\boldsymbol{y_c} = (\boldsymbol{y}^\top, \boldsymbol{y_0}^\top, \boldsymbol{u}^\top, \boldsymbol{u_0}^\top, \boldsymbol{t}^\top, \boldsymbol{t_0}^\top)^\top$.
Let $\boldsymbol{\theta}^{(p)} = (\alpha^{(p)}, \beta^{(p)}, \sigma^{2(p)}, \lambda^{(p)}, x_0^{(p)})^\top$ be the estimates of $\boldsymbol{\theta}$ at the pth iteration. Then we have the conditional expectation of the complete log-likelihood function $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = \mathbb{E}\left[\ell_c(\boldsymbol{\theta}|\boldsymbol{y_c})|\boldsymbol{y}, \boldsymbol{y_0}, \widehat{\boldsymbol{\theta}}^{(p)}\right]$. Thus, we have the following EM algorithm steps:

E-step: Given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(p)}$, compute $\widehat{t_{ij}}^{(p)}, \widehat{t_{ij}^2}^{(p)}, \widehat{t_{0i}}^{(p)}$ and $\widehat{t_{0i}^2}^{(p)}$, where $\widehat{t_{ij}} = \mathbb{E}(T_{ij}|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{ij})$, $\widehat{t_{ij}^2} = \mathbb{E}(T_{ij}^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{ij})$, $\widehat{t_{0i}} = \mathbb{E}(T_{0i}|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{0i})$ and $\widehat{t_{0i}^2} = \mathbb{E}(T_{0i}^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, y_{0i})$.

M-step: Update $\hat{\boldsymbol{\theta}}^{(p+1)}$ by maximizing $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ over $\boldsymbol{\theta}$, which leads to the following closed form expressions:

$$
\begin{aligned}
\widehat{\alpha}^{(p+1)} &= \left[\widehat{\boldsymbol{\kappa}}^{(p)\top}\mathbf{1}_{\boldsymbol{R}} + \widehat{\boldsymbol{\kappa_0}}^{(p)\top}\mathbf{1}_{\boldsymbol{m}} + (R+m)\lambda^{(p)2}\right]^{-1} \\
&\quad \left[\left(\boldsymbol{y}^\top - \beta^{(p)}\boldsymbol{x}^\top\right)\widehat{\boldsymbol{\kappa}}^{(p)} + \lambda^{(p)}\left(\lambda^{(p)}\boldsymbol{y}^\top - \widehat{\boldsymbol{t}}^{(p)\top} - \lambda^{(p)}\beta^{(p)}\boldsymbol{x}^\top\right)\mathbf{1}_{\boldsymbol{R}} + \boldsymbol{y_0}^\top\widehat{\boldsymbol{\kappa_0}}^{(p)}\right. \\
&\quad \left. + \left(\lambda^{(p)2}\boldsymbol{y_0}^\top - \lambda^{(p)}\widehat{\boldsymbol{t_0}}^{(p)\top} - \beta^{(p)}x_0^{(p)}\widehat{\boldsymbol{\kappa_0}}^{(p)\top}\right)\mathbf{1}_{\boldsymbol{m}} - m\beta^{(p)}x_0^{(p)}\lambda^{(p)2}\right] \\
\widehat{\beta}^{(p+1)} &= \left[\boldsymbol{x}^\top\left(\boldsymbol{D}(\widehat{\boldsymbol{\kappa}}^{(p)}) + \lambda^{(p)2}\mathbf{I}_{\boldsymbol{R}}\right)\boldsymbol{x} + x_0^{(p)2}\left(\widehat{\boldsymbol{\kappa_0}}^{(p)\top}\mathbf{1}_{\boldsymbol{m}} + \lambda^{(p)2}m\right)\right]^{-1} \\
&\quad \left\{\boldsymbol{x}^\top\left[\boldsymbol{D}(\widehat{\boldsymbol{\kappa}}^{(p)})\boldsymbol{y} - \alpha^{(p)}\widehat{\boldsymbol{\kappa}}^{(p)} + \lambda^{(p)2}\left(\boldsymbol{y} - \alpha^{(p)}\mathbf{1}_{\boldsymbol{R}}\right) - \lambda^{(p)}\widehat{\boldsymbol{t}}^{(p)}\right] + x_0^{(p)} \times\right. \\
&\quad \left.\left[\boldsymbol{y}_0^\top\widehat{\boldsymbol{\kappa_0}}^{(p)} + \left(\lambda^{(p)2}\boldsymbol{y_0}^\top - \alpha^{(p)}\widehat{\boldsymbol{\kappa_0}}^{(p)\top} - \lambda^{(p)}\widehat{\boldsymbol{t_0}}^{(p)\top}\right)\mathbf{1}_{\boldsymbol{m}} - m\lambda^{(p)2}\alpha^{(p)}\right]\right\} \\
\widehat{\sigma^2}^{(p+1)} &= [2(R+m)]^{-1}\left[\left(\boldsymbol{\eta}^{(p)\top}\boldsymbol{D}(\widehat{\boldsymbol{\kappa}}^{(p)}) + \lambda^{(p)2}\boldsymbol{\eta}^{(p)\top} - 2\lambda^{(p)}\widehat{\boldsymbol{t}}^{(p)\top}\right)\boldsymbol{\eta}^{(p)} + \widehat{\boldsymbol{t}^2}^{(p)\top}\mathbf{1}_{\boldsymbol{R}}\right. \\
&\quad \left. + \left(\boldsymbol{\eta_0}^{(p)\top}\boldsymbol{D}(\widehat{\kappa}_0^{(p)}) + \lambda^{(p)2}\boldsymbol{\eta_0}^{(p)\top} - 2\lambda^{(p)}\widehat{\boldsymbol{t}}_{\boldsymbol{0}}^{(p)\top}\right)\boldsymbol{\eta_0}^{(p)} + \widehat{\boldsymbol{t}_0^2}^{(p)\top}\mathbf{1}_{\boldsymbol{m}}\right] \\
\widehat{\lambda}^{(p+1)} &= \left[\boldsymbol{\eta}^{(p)\top}\boldsymbol{\eta}^{(p)} + \boldsymbol{\eta_0}^{(p)\top}\boldsymbol{\eta_0}^{(p)}\right]^{-1}\left[\widehat{\boldsymbol{t}}^{(p)\top}\boldsymbol{\eta}^{(p)} + \widehat{\boldsymbol{t}}_{\boldsymbol{0}}^{(p)\top}\boldsymbol{\eta_0}^{(p)}\right] \\
\widehat{x}_0^{(p+1)} &= \left[\beta^{(p)}\left(\widehat{\boldsymbol{\kappa_0}}^{(p)\top}\mathbf{1}_{\boldsymbol{m}} + m\lambda^{(p)2}\right)\right]^{-1} \\
&\quad \left[\boldsymbol{y_0}^\top\widehat{\boldsymbol{\kappa_0}}^{(p)} - \left(\alpha^{(p)}\widehat{\boldsymbol{\kappa_0}}^{(p)\top} + \lambda^{(p)}\widehat{\boldsymbol{t}}_{\boldsymbol{0}}^{(p)\top}\right)\mathbf{1}_{\boldsymbol{m}} + \lambda^{(p)2}\left(\boldsymbol{y_0}^\top\mathbf{1}_{\boldsymbol{m}} - m\alpha^{(p)}\right)\right].
\end{aligned}
$$

where $\boldsymbol{\eta}^{(p)} = \boldsymbol{y} - \alpha^{(p)}\mathbf{1}_{\boldsymbol{R}} - \beta^{(p)}\boldsymbol{x}$, $\boldsymbol{\eta_0}^{(p)} = \boldsymbol{y_0} - \left(\alpha^{(p)} + \beta^{(p)}x_0^{(p)}\right)\mathbf{1}_{\boldsymbol{m}}$, $\boldsymbol{D}(\boldsymbol{A}) = \text{Diag}(a_1, a_2, \dots)$, $\mathbf{1}_{\boldsymbol{k}}$ denotes an k-dimensional column vector of ones and $\mathbf{I}_{\boldsymbol{R}}$ is an identity matrix of order R.
The Fisher-information matrix is used to calculate the covariance matrices associated to the maximum-likelihood estimates.

3 Aplication

We fit the SMNS calibration model to the real data set discussed by Neto *et al.* (2007) for the SN, ST, SSL and SPE distributions. These datase are given in Table 1. We use the triplicate absorbance readings $\boldsymbol{y_0} = (14.804, 14.861, 14.731)$ for the second stage data of the calibration model and the rest of the dataset will be used for the first stage. Table 2 it is presented the parameter estimates, the asymptotic standard errors and the expanded uncertainty $U(\widehat{x_0})$, which is the confidence interval amplitude divided by 2, for the four distributions, and also it is given the AIC and BIC criteria. In this table we observe that among the four distributions the SPE distribution has the smallest error and the ST distribution has the smallest standard error for the estimator of x_0. According to both information criteria the ST distribution is more adequate than the those other distributions.

TABLE 1. Zinc concentration (mg/l) and triplicate absorbance readings.

Concentration	Absorbance		
x_i	y_{i1}	y_{i2}	y_{i3}
0.0	0.696	0.696	0.706
0.5	7.632	7.688	7.603
1.0	**14.804**	**14.861**	**14.731**
2.0	28.895	29.156	29.322
3.0	43.993	43.574	44.699

TABLE 2. Parameter estimates for the *ST, SN, SSL* and *SPE* distributions.

Distribution	Parameters				Criteria	
	α	β	x_0	$U(\widehat{x_0})$	AIC	BIC
ST	0.497	14.195	1.002	0.003	-80.35	-76.80
($\tau = 2$)	(0.019)	(0.011)	(0.002)			
SN	0.300	14.290	1.004	0.045	-7.99	-4.45
	(0.224)	(0.112)	(0.023)			
SSL	0.491	14.198	1.002	0.004	-70.47	-66.93
($\tau = 1.5$)	(0.002)	(0.007)	(0.002)			
SPE	0.312	14.285	1.001	0.043	-9.03	-5.59
($\tau = 0.8$)	(0.251)	(0.125)	(0.022)			

4 Conclusions

The proposed SMSN calibration model adjusts skewness and heavy-tailedness simultaneously. In the application it is shown that the alternative distributions of the SN are more adequate.

Acknowledgments: The first author thanks **FACEPE** partial financial support. The second author thanks the scholarship support from **CAPES**.

References

Blas, B., Sandoval, M. C. an Yoshida, O. S. (2007). Homoscedastic controlled calibration model. *Journal of Chemometrics*, **21**, 145 - 155.

Branco,M.D. and Dey, D.K. (2001). A class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, **79**, 99 – 113.

Lange, K.L., Little, J.A. and Taylor, M.G. (1989). Robust Statistical modeling using the t distribution. *Journal of the American Statistical Association*, **84**, 881 – 896.

Neto, B. B., Scarminio, I. S. and Bruns, R. E. (2007). *Como fazer experimentos: pesquisa e desenvolvimento na ciência e na indústria*. São Paulo: Editora da Unicamp.

Distributed fusion filtering using correlated missing observations from multiple sensors

Raquel Caballero-Águila[1], Irene García-Garrido[2], Josefa Linares-Pérez[2]

[1] Dpto. Estadística e I.O., Universidad de Jaén, Spain
[2] Dpto. Estadística e I.O., Universidad de Granada, Spain

E-mail for correspondence: irenegarciag@ugr.es

Abstract: A distributed fusion filtering algorithm for discrete-time linear stochastic systems with missing observations coming from multiple sensors is proposed. At each sensor, the Bernoulli random variables describing the phenomenon of missing observations are assumed to be correlated at instants that differ m units of time. For each sensor subsystem, local least-squares linear filtering estimators are given and the estimation error cross-covariance matrices between any two sensors are derived, then the proposed distributed fusion filter is obtained based on the distributed fusion criterion weighted by matrices in the linear minimum variance sense.

Keywords: Least-squares estimation; Distributed fusion estimation; Missing observations; Multiple sensors.

1 Introduction

Estimation problems in multi-sensor systems are recently motivating a significant amount of research due to their increasing applicability in many engineering fields (see e.g. Hespanha et al. (2007) and references therein). A basic matter in systems with multiple sensors is how to fuse the measurement data from the different sensors to address the estimation problem. For this purpose, several authors have proposed distributed fusion methods, in which each sensor estimates the state based on its own measurement data, and sends such estimate to the fusion center for fusion according to a certain information criterion. For example, under the assumption of normal distribution, a distributed fusion estimator is proposed in Kim (1994) based in maximum likelihood criterion, and the distributed fusion criterion weighted by matrices in the linear minimum variance sense is established in Sun and Deng (2004), which is equivalent to the maximum likelihood fusion criterion under normality assumption. Recently, attention is being focused on distributed fusion estimation problems in networked systems with unreliable network transmission for multi-sensor systems with random delays

(see Feng and Zeng (2011)) and packet dropouts (see Zhang et al. (2012)). Nevertheless, to the best of the authors knowledge, the literature concerning distributed fusion estimation in multi-sensor systems with missing measurements is relatively scarcer. In this paper, the least-squares (LS) linear distributed fusion estimation problem in multi-sensor systems with missing measurements is addressed assuming that the Bernoulli random variables modeling the missing measurements at each sensor are correlated at instants that differ m units of time. This special form of correlation allows us to consider certain class of systems in which the state cannot be missing in $m+1$ consecutive observations; specifically, sensor networks where sensor failures may happen and a failed sensor is substituted m sampling times after having failed.

2 Model description

Consider the following discrete-time linear multi-sensor stochastic systems with missing measurements:

$$x_k = F_{k-1}x_{k-1} + w_{k-1}, \quad k \geq 1,$$
$$y_k^i = \theta_k^i H_k^i x_k + v_k^i, \quad k \geq 1, \quad i = 1, \ldots, r,$$

where $x_k \in \mathbb{R}^n$ is the state and $y_k^i \in \mathbb{R}$ is the measurement provided by sensor i at sampling time k. $\{w_k;\ k \geq 0\}$ and $\{v_k^i;\ k \geq 1\}$, $i = 1, \ldots, r$, are zero-mean white sequences with covariances $Cov[w_k] = Q_k$ and $Cov[v_k^i] = R_k^i$, respectively, and $\{\theta_k^i;\ k \geq 1\}$, $i = 1, \ldots, r$, are Bernoulli random variables with $P[\theta_k^i = 1] = \overline{\theta}_k^i$. For $i = 1, \ldots, r$, the variables θ_k^i and θ_s^i are independent for $|k-s| \neq 0, m$, and $Cov[\theta_k^i, \theta_s^i] = K_{k,s}^{\theta^i}$ are known for $|k-s| = 0, m$. The initial state x_0 is a random vector with $E[x_0] = \overline{x}_0$ and $Cov[x_0] = P_0$. Moreover, x_0 and the noise processes $\{w_k;\ k \geq 0\}$, $\{v_k^i;\ k \geq 1\}$ and $\{\theta_k^i;\ k \geq 1\}$, for $i = 1, \ldots, r$, are mutually independent.

From these properties, $D_k = E[x_k x_k^{\mathrm{T}}]$ is recursively calculated by

$$D_k = F_{k-1}D_{k-1}F_{k-1}^{\mathrm{T}} + Q_{k-1}, \quad k \geq 1; \qquad D_0 = P_0 + \overline{x}_0\overline{x}_0^{\mathrm{T}}.$$

3 Distributed fusion estimation

Our aim is to solve the LS estimation problem of the state x_k based on the measurements $\{y_1^i, y_2^i, \ldots, y_k^i\}$, $i = 1, 2, \ldots, r$, by using the distributed fusion method.

For this purpose, firstly the following recursive algorithm to obtain local LS linear filters, $\widehat{x}_{k/k}^i$, for $i = 1, 2, \ldots, r$, along with their estimation error covariance matrices, is established. Also, recursive formulas for the error cross-covariance matrices between any two local estimates are presented.

The local LS linear filter, $\widehat{x}_{k/k}^i$, is obtained by

$$\widehat{x}^i_{k/k} = F_{k-1}\widehat{x}^i_{k-1/k-1} + S^i_{k,k}(\Pi^{ii}_{k,k})^{-1}\nu^i_k, \quad k \geq 1; \quad \widehat{x}^i_{0/0} = \overline{x}_0,$$

The innovation, ν^i_k, is given by

$$\begin{aligned}
\nu^i_k &= y^i_k - \overline{\theta}^i_k H^i_k F_{k-1}\widehat{x}^i_{k-1/k-1}, \quad k \leq m,\\
\nu^i_k &= y^i_k - \overline{\theta}^i_k H^i_k F_{k-1}\widehat{x}^i_{k-1/k-1}\\
&\quad - \Psi^i_{k,k-m}\left(\nu^i_{k-m} - \sum_{l=1}^{m-1} T^{i\mathrm{T}}_{k-l,k-m}(\Pi^{ii}_{k-l,k-l})^{-1}\nu^i_{k-l}\right), \ k > m,
\end{aligned}$$

where $\Psi^i_{k,k-m} = K^{\theta^i}_{k,k-m}H^i_k\mathbb{F}_{k,k-m}D_{k-m}H^{i\mathrm{T}}_{k-m}(\Pi^{ii}_{k-m,k-m})^{-1}$, with $\mathbb{F}_{k,i} = F_{k-1}\cdots F_i$.
The matrices $T^i_{k,k-l}$ are determined by

$$T^i_{k,k-l} = \overline{\theta}^i_k H^i_k\mathbb{F}_{k,k-l}S^i_{k-l,k-l}, \quad 2 \leq k \leq m, \quad 1 \leq l \leq k-1,$$

$$T^i_{k,k-l} = \overline{\theta}^i_k H^i_k\mathbb{F}_{k,k-l}S^i_{k-l,k-l} - \Psi^i_{k,k-m}T^{i\mathrm{T}}_{k-l,k-m}, \quad k > m, \quad 1 \leq l \leq m-1.$$

The innovation covariance matrix, $\Pi^{ii}_{k,k}$, satisfies

$$\begin{aligned}
\Pi^{ii}_{k,k} &= \overline{\theta}^i_k\left(1-\overline{\theta}^i_k\right)H^i_k D_k H^{i\mathrm{T}}_k + R^i_k + \overline{\theta}^i_k H^i_k S^i_{k,k}, \quad k \leq m,\\
\Pi^{ii}_{k,k} &= \overline{\theta}^i_k\left(1-\overline{\theta}^i_k\right)H^i_k D_k H^{i\mathrm{T}}_k + R^i_k + \overline{\theta}^i_k H^i_k S^i_{k,k} + \overline{\theta}^i_k S^{i\mathrm{T}}_{k,k}H^{i\mathrm{T}}_k\\
&\quad - (\overline{\theta}^i_k)^2 H^i_k P^{ii}_{k/k-1}H^{i\mathrm{T}}_k - \Psi^i_{k,k-m}\Bigg(\Pi^{ii}_{k-m,k-m}\\
&\quad + \sum_{l=1}^{m-1} T^{i\mathrm{T}}_{k-l,k-m}\left(\Pi^{ii}_{k-l,k-l}\right)^{-1}T^i_{k-l,k-m}\Bigg)\Psi^{i\mathrm{T}}_{k,k-m}, \ k > m.
\end{aligned}$$

The matrix $S^i_{k,k}$ is derived by the following expression

$$\begin{aligned}
S^i_{k,k} &= \overline{\theta}^i_k P^{ii}_{k/k-1}H^{i\mathrm{T}}_k, \quad k \leq m,\\
S^i_{k,k} &= \overline{\theta}^i_k P^{ii}_{k/k-1}H^{i\mathrm{T}}_k - \Bigg(\mathbb{F}_{k,k-m}S^i_{k-m,k-m}\\
&\quad - \sum_{l=1}^{m-1}\mathbb{F}_{k,k-l}S^i_{k-l,k-l}\left(\Pi^{ii}_{k-l,k-l}\right)^{-1}T^i_{k-l,k-m}\Bigg)\Psi^{i\mathrm{T}}_{k,k-m}, \quad k > m,
\end{aligned}$$

where $P^{ii}_{k/k-1}$, the prediction error covariance matrix, is obtained by

$$P^{ii}_{k/k-1} = F_{k-1}P^{ii}_{k-1/k-1}F^{\mathrm{T}}_{k-1} + Q_{k-1}, \quad k \geq 1,$$

with $P^{ii}_{k/k}$, the filtering error covariance matrix, satisfying

$$P^{ii}_{k/k} = P^{ii}_{k/k-1} - S^i_{k,k}\left(\Pi^{ii}_{k,k}\right)^{-1}S^{i\mathrm{T}}_{k,k}, \quad k \geq 1; \quad P^{ii}_{0/0} = P_0.$$

The filtering error cross-covariance matrix, $P^{ij}_{k/k}$, is specified by

$$P_{k/k}^{ij} = P_{k/k-1}^{ij} + S_{k,k}^{i}(\Pi_{k,k}^{ii})^{-1}\Pi_{k,k}^{ij}(\Pi_{k,k}^{jj})^{-1}S_{k,k}^{j\mathrm{T}}$$
$$- \left(S_{k,k}^{j} - L_{k,k}^{ij}\right)(\Pi_{k,k}^{jj})^{-1}S_{k,k}^{j\mathrm{T}} - S_{k,k}^{i}(\Pi_{k,k}^{ii})^{-1}\left(S_{k,k}^{i} - L_{k,k}^{ji}\right)^{\mathrm{T}},\ k \geq 1,$$
$$P_{k/k-1}^{ij} = F_{k-1}P_{k-1/k-1}^{ij}F_{k-1}^{\mathrm{T}} + Q_{k-1}, \quad k \geq 1; \quad P_{0/0}^{ij} = P_0.$$

Secondly, the matrix-weighted distributed fusion estimator $\widehat{x}_{k/k}^{0}$, is presented below by applying the optimal information fusion criterion weighted by matrices in the linear minimum variance sense (Sun and Deng (2004)).

The LS distributed fusion filter, $\widehat{x}_{k/k}^{0}$, is given by

$$\widehat{x}_{k/k}^{0} = A_{k,k}^{1}\widehat{x}_{k/k}^{1} + \cdots + A_{k,k}^{r}\widehat{x}_{k/k}^{r},$$

where $\widehat{x}_{k/k}^{i}$ $(i = 1, 2 \ldots, r)$, the local LS linear filters, are calculated by the previous recursive algorithm.

The optimal matrix weights $A_{k,k}^{i}$ $(i = 1, 2, \ldots, r)$ are computed by

$$A_{k,k} = \Sigma_{k/k}^{-1}e\left(e^{\mathrm{T}}\Sigma_{k/k}^{-1}e\right)^{-1},$$

where the matrices $A_{k,k} = \left[A_{k,k}^{1}, \ldots, A_{k,k}^{r}\right]^{\mathrm{T}}$ and $e = [I, \ldots, I]^{\mathrm{T}}$ are both $nr \times n$ matrices, and $\Sigma_{k/k} = (P_{k/k}^{ij})_{i,j=1,\ldots,r}$ is a symmetric positive definite matrix of dimension $nr \times nr$.

The error covariance matrix of the distributed weighted fusion estimator is computed by $P_{k/k}^{0} = \left(e^{\mathrm{T}}\Sigma_{k/k}^{-1}e\right)^{-1}$.

Acknowledgments: This research is supported by Ministerio de Ciencia e Innovación (Programa FPU and grant No. MTM2011-24718) and Junta de Andalucía (grant No. P07-FQM-02701).

References

Feng, J. and Zeng, M. (2011). Descriptor recursive estimation for multiple sensors with different delay rates. *International Journal of Control*, **84 (3)**, 584 – 596.

Hespanha, J., Naghshtabrizi, P. and Xu, Y. (2007). A survey of recent results in networked control systems. *Proceedings of the IEEE*, **95 (1)**, 138 – 162.

Kim, K.H. (1994). Development of track to track fusion algorithms. In: *Proceedings of the American Control Conference*, Maryland, pp. 1037 – 1041.

Sun, S.L. and Deng, Z.L. (2004). Multi-sensor optimal information fusion Kalman filter. *Automatica*, **40**, 1017 – 1023.

Zhang, W.-A., Feng, G. and Yu, L. (2012). Multi-rate distributed fusion estimation for sensor networks with packet losses. *Automatica*, **48**, 2016 – 2028.

Optimal least-squares linear centralized filter for systems with autocorrelated and cross-correlated noises

Raquel Caballero-Águila[1], Irene García-Garrido[2], Josefa Linares-Pérez[2]

[1] Dpto. Estadística e I.O., Universidad de Jaén, Spain
[2] Dpto. Estadística e I.O., Universidad de Granada, Spain

E-mail for correspondence: raguila@ujaen.es

Abstract: The optimal least-squares linear centralized estimation problem is addressed for a class of discrete-time multi-sensor linear systems with autocorrelated and cross-correlated noises. The process noise and all the sensor noises are assumed to be one-step autocorrelated, different sensor noises are one-step cross-correlated, and, for each sensor, the process noise and the measurement noise are two-step cross-correlated. By using an innovation approach, a recursive algorithm for the optimal least-squares linear centralized filter is derived.

Keywords: Least-squares estimation; Centralized fusion estimation; Multi-sensor systems.

1 Introduction

During the past decades, there has been an increasing interest in the filtering problem in multi-sensor systems with correlated noises. For example, the optimal Kalman filtering fusion problem in systems with cross-correlated sensor noises is addressed in Song et al. (2007), while Feng and Zeng (2012) study the same problem in systems with cross-correlated process noises and measurement noises; in these papers correlated noises at the same sampling time are considered. In general, the assumption of correlation and cross-correlation of the noise process and measurement noises in different sampling times makes difficult the identification of optimal estimators; this limitation has encouraged research into suboptimal Kalman-type estimation problems. In Song et al. (2008), a Kalman-type recursive filter is presented for systems with finite-step correlated process noises, and the filtering problem with multi-step correlated process and measurement noises is investigated in Fu et al. (2008). The problem of distributed weighted robust Kalman filter fusion is studied in Feng et al. (2013), for a class of uncertain systems with autocorrelated and cross-correlated noises.

Our aim is to address the optimal least-squares (LS) linear centralized fusion estimation problem in multi-sensor systems with autocorrelated and cross-correlated noises. Unlike most previous results with correlated noises, in which suboptimal Kalman-type estimators are proposed, in this paper a recursive algorithm for the optimal LS linear filter is obtained by using an innovation approach which provides a simple derivation of the estimation algorithms due to the fact that the innovations constitute a white process.

2 System model

Consider the following multi-sensor system:

$$x_k = F_{k-1}x_{k-1} + w_{k-1}, \quad k \geq 1,$$
$$y_k^i = H_k^i x_k + v_k^i, \quad k \geq 1, \quad i = 1, \ldots, r,$$

where $x_k \in \mathbb{R}^n$ is the state, $y_k^i \in \mathbb{R}$, $i = 1, \ldots, r$, is the measurement collected by sensor i at sampling time k, $\{w_k;\, k \geq 0\}$ and $\{v_k^i;\, k \geq 1\}$, $i = 1, \ldots, r$, are noise sequences. Next, the statistical properties assumed about the initial state and noise processes are specified:

(i) The initial state x_0 is a random vector with $E[x_0] = \overline{x}_0$ and $Cov[x_0] = P_0$ and it is independent of the additive noises.

(ii) The process additive noise $\{w_k;\, k \geq 0\}$ and the measurement noises $\{v_k^i;\, k \geq 1\}$, $i = 1, \ldots, r$, are zero-mean sequences with covariances and cross-covariances:

$$Cov[w_k, w_s] = Q_{k,k}\delta_{k-s} + Q_{k,s}\delta_{k-s+1} + Q_{k,s}\delta_{k-s-1},$$
$$Cov[v_k^i, v_s^j] = R_{k,k}^{ij}\delta_{k-s} + R_{k,s}^{ij}\delta_{k-s+1} + R_{k,s}^{ij}\delta_{k-s-1},$$
$$Cov[w_k, v_s^i] = S_{k,k}^i\delta_{k-s} + S_{k,s}^i\delta_{k-s+1} + S_{k,s}^i\delta_{k-s+2}.$$

Our aim is to solve the optimal LS linear estimation problem of the state x_k based on the measurements $\left\{y_1^i, y_2^i, \ldots, y_k^i\right\}$, for $i = 1, \ldots, r$, by using centralized fusion method to process the measured sensor data; for this purpose, the observation equation is rewritten in a stacked form as follows:

$$y_k = H_k x_k + v_k, \quad k \geq 1,$$

where $y_k = (y_k^1, \ldots, y_k^r)^{\mathrm{T}}$, $v_k = (v_k^1, \ldots, v_k^r)^{\mathrm{T}}$ and $H_k = (H_k^{1\mathrm{T}}, \ldots, H_k^{r\mathrm{T}})^{\mathrm{T}}$.

The following properties are easily inferred from the model assumptions:

(I) The noise $\{v_k;\, k \geq 1\}$ is a zero-mean process, independent of x_0, and satisfies:

$$Cov[v_k, v_s] = R_{k,k}\delta_{k-s} + R_{k,s}\delta_{k-s+1} + R_{k,s}\delta_{k-s-1},$$
$$Cov[w_k, v_s] = S_{k,k}\delta_{k-s} + S_{k,s}\delta_{k-s+1} + S_{k,s}\delta_{k-s+2},$$

where $R_{k,s} = \left(R_{k,s}^{ij}\right)_{i,j=1,\ldots,r}$ and $S_{k,s} = (S_{k,s}^1, \ldots, S_{k,s}^r)$.

(II) The state vector x_k and the measurement noise vector v_k are correlated with $E_k = E[x_k v_k^{\mathrm{T}}]$ satisfying:

$$E_k = F_{k-1}S_{k-2,k} + S_{k-1,k}, \quad k \geq 2; \qquad E_1 = S_{0,1}.$$

3 Optimal LS linear centralized fusion estimation

As known, the optimal LS linear filter $\widehat{x}_{k/k}$ is the orthogonal projection of the state x_k over the linear space spanned by $\{y_1, y_2, \ldots, y_k\}$. These observations are generally non-orthogonal vectors, but the Gram-Schmidt orthogonalization procedure allows us to substitute them by a set of orthogonal vectors, called *innovations*, defined as the difference between each observation and its one-stage predictor. Due to the orthogonality property of the innovations and since the innovation process is uniquely determined by the observations, the LS linear filter, $\widehat{x}_{k/k}$, can be calculated as linear combination of the innovations; namely,

$$\widehat{x}_{k/k} = \sum_{s=1}^{k} \mathcal{X}_{k,s}\Pi_{s,s}^{-1}\mu_s, \quad k \geq 1,$$

where $\mu_s = y_s - \widehat{y}_{s/s-1}$ are the innovations, with $\widehat{y}_{s/s-1}$ the observation predictor, $\Pi_{s,s} = E[\mu_s\mu_s^{\mathrm{T}}]$, and $\mathcal{X}_{k,s} = E[x_k\mu_s^{\mathrm{T}}]$. This expression provides the starting point to derive the following recursive filtering algorithm for the centralized fusion estimation problem.

The optimal LS linear centralized filter $\widehat{x}_{k/k}$ is obtained as

$$\widehat{x}_{k/k} = \widehat{x}_{k/k-1} + \mathcal{X}_{k,k}\Pi_{k,k}^{-1}\mu_k, \quad k \geq 1; \quad \widehat{x}_{0/0} = \overline{x}_0,$$

where the state predictor, $\widehat{x}_{k/k-1}$, satisfies

$$\widehat{x}_{k/k-1} = F_{k-1}\widehat{x}_{k-1/k-1} + \mathcal{W}_{k-1,k-1}\Pi_{k-1,k-1}^{-1}\mu_{k-1}, \ k \geq 2; \quad \widehat{x}_{1/0} = F_0\widehat{x}_{0/0},$$

with $\mathcal{W}_{k,k} = Q_{k,k-1}H_k^{\mathrm{T}} + S_{k,k}$, $k \geq 1$.

The innovation, μ_k, is calculated by

$$\mu_k = y_k - H_k\widehat{x}_{k/k-1} - \mathcal{V}_{k,k-1}\Pi_{k-1,k-1}^{-1}\mu_{k-1}, \ k \geq 2; \quad \mu_1 = y_1 - H_1\widehat{x}_{1/0},$$

where $\mathcal{V}_{k,k-1} = S_{k-2,k}^{\mathrm{T}}H_{k-1}^{\mathrm{T}} + R_{k,k-1}$, $k \geq 2$.

The matrix $\mathcal{X}_{k,k}$ is given by

$$\mathcal{X}_{k,k} = P_{k/k-1}H_k^{\mathrm{T}} + E_k - \mathcal{X}_{k,k-1}\Pi_{k-1,k-1}^{-1}\mathcal{V}_{k,k-1}^{\mathrm{T}}, \ k \geq 2; \quad \mathcal{X}_{1,1} = P_{1/0}H_1^{\mathrm{T}} + E_1,$$

where $\mathcal{X}_{k,k-1} = F_{k-1}\mathcal{X}_{k-1,k-1} + \mathcal{W}_{k-1,k-1}$, $k \geq 2$.

The prediction error covariance matrix, $P_{k/k-1}$, is obtained by

$$\begin{aligned} P_{k/k-1} &= F_{k-1}P_{k-1/k-1}F_{k-1}^{\mathrm{T}} + Q_{k-1,k-1} + F_{k-1}\mathcal{J}_{k-1} + \mathcal{J}_{k-1}^{\mathrm{T}}F_{k-1}^{\mathrm{T}} \\ &\quad - \mathcal{W}_{k-1,k-1}\Pi_{k-1,k-1}^{-1}\mathcal{W}_{k-1,k-1}^{\mathrm{T}}, \quad k \geq 2; \\ P_{1/0} &= F_0 P_{0/0} F_0^{\mathrm{T}} + Q_{0,0}, \end{aligned}$$

where $\mathcal{J}_k = Q_{k-1,k} - \mathcal{X}_{k,k}\Pi_{k,k}^{-1}\mathcal{W}_{k,k}^{\mathrm{T}}, \quad k \geq 1.$

The filtering error covariance matrix, $P_{k/k}$, is given by

$$P_{k/k} = P_{k/k-1} - \mathcal{X}_{k,k}\Pi_{k,k}^{-1}\mathcal{X}_{k,k}^{\mathrm{T}}, \quad k \geq 1; \qquad P_{0/0} = P_0.$$

The innovation covariance matrix, $\Pi_{k,k}$, satisfies

$$\Pi_{k,k} = R_{k,k} + H_k\mathcal{X}_{k,k} + \mathcal{X}_{k,k}^{\mathrm{T}}H_k^{\mathrm{T}} - H_k P_{k/k-1}H_k^{\mathrm{T}} - \mathcal{V}_{k,k-1}\Pi_{k-1,k-1}^{-1}\mathcal{V}_{k,k-1}^{\mathrm{T}}, \ k \geq 2;$$
$$\Pi_{1,1} = R_{1,1} + H_1\mathcal{X}_{1,1} + \mathcal{X}_{1,1}^{\mathrm{T}}H_1^{\mathrm{T}} - H_1 P_{1/0}H_1^{\mathrm{T}}.$$

Acknowledgments: This research is supported by Ministerio de Ciencia e Innovación (Programa FPU and grant No. MTM2011-24718) and Junta de Andalucía (grant No. P07-FQM-02701).

References

Song, E., Zhu, Y. Zhou, J. and You, Z. (2007). Optimal Kalman filtering fusion with cross-correlated sensor noises. *Automatica*, **43**, 1450 – 1456.

Feng, J. and Zeng, M. (2012). Optimal distributed Kalman filtering fusion for a linear dynamic system with cross-correlated noises. *International Journal of Systems Science*, **43**, 385 – 398.

Song, E., Zhu, Y. and You, Z. (2008). The Kalman type recursive state estimator with a finite-step correlated process noises. In: *Proceedings of the IEEE International Conference on Automation and Logistics*, pp. 196 – 200.

Fu, A., Zhu, Y. and Song, E. (2008). The optimal Kalman type state estimator with multi-step correlated process and measurement noises. In: *The 2008 International Conference on Embedded Software and Systems*, pp. 215 – 220.

Feng, J., Wang, Z. and Zeng, M. (2013). Distributed weighted robust Kalman filter fusion for uncertain systems with autocorrelated and cross-correlated noises. *Information Fusion*, **14**, 78 – 86.

Nonparametric method for treatment comparison in the winemaking process

Marcos H. Cascone[1], Larissa A. Matos[1], José A. G. Campos[1]

[1] Departament of Statistics, Campinas State University, Brazil

E-mail for correspondence: marcos.cascone@gmail.com

Abstract: Recent studies are showing that winemaking process depends of other elements that might help the quality be better than the usual way of production. One of these elements is the quantity of sugar present in the grape. Many techniques of production has been implemented, and the ground cover with different type of glasses has been the one that produce great results. The aim of this paper is to find how type of glass makes with the grape has more quantity of sugar, using the GAM models in which ensure good flexibility and no assumptions about the structure of variables.

Keywords: Generalized Additive Models; Winemaking process; Model selection.

1 Introduction

The fermentation process is a catalyst function which converts the grape juice to alcoholic beverage. During this process, the yeast interacts with the sugar present in the juice to generate a mixture of ethyl alcohol and carbon dioxide. Champagnol (1984) shows that the temperature and the speed of fermentation are also important factors in the winemaking, as well as oxygen levels in the juice early in the fermentation. The control of the levels of sugar during the fermentation is extremely important, since the amount of sugar is not large enough, wine quality may be affected and, according with Jackson (2008), as more the quantity of sugar the fermentation process generate, better will be the quality of the wine.
Lincoln Univeristy's Center for Viticulture and Oenology, New Zealand, have been developing a project to control sugar levels in the grapes, stating that the amount of sugar is directly related to the level of solar radiation that they receive. The proposed technique is to cover the ground with glass slides subdivided into 3 categories: clear glass, brown glass and mixed colored glass (in which basically consists of the colors green and brown).
The Generalized Additive Models, introduced by Hastie and Tibshirani (1986), provide a good alternative to deal with this kind of problems, once the flexibility of this model allow us work with no assumption about the

structure of variables and covariates. Furthermore, the GAM models can be used as a tool to select models and variables.

2 Methodology

Although Generalized Linear Models, introduced by Nelder and Wedderburn (1972), are very flexible when the response Y is non Gaussian, this is possible only when the linear predictor between the response and covariate is linear. The Generalized Additive Models are an extension of the GLMs in which the linear predictor is not obliged to be linear in the covariates, but a sum of smoothing functions applied to the covariates. These class of models ensure good flexibility and may be used as a tool to model selection. The inference of these models is likelihood-based, where the estimators are obtained through likelihood function maximization. We may notice that GAM are a kind of GLM extension, in which the hypothesis are extended to a class where the relation between response and explanatory variables is not particularly linear. The GAM has the expression

$$\eta_i = \alpha + \sum_{j=1}^{k} f_j(X_{ij}),$$

where η_i is the link function and $E(f_j(X_{ij})) = 0$. Note that $X_i^\top \beta = \sum_{j=1}^{k} X_{ij}\beta_j$ was replaced by $f_j(X_{ij})$ in the GLM, where $f_j(X_{ij})$ is a non parametric function with unknown structure and the estimate is given by smoothing splines.

Therefore, neither the assumption of linearity between $g(\mu_i)$ and explanatory variables, nor the structure of this relation needs to be known, and may be estimated from a data set. The estimated function $\hat{f}_j(X_{ij})$, also known by smoothing spline, generally is a mean of Y_i in the neighborhood of X_i and the smoothing spline allow us to describe such structure. Besides that, it is possible to reveal non linearities in that relation. A simple estimate of $f(X_i)$ is the mean of Y near of the corresponding values of X_i, and this estimate is usually called by moving average estimate.

Formally, $\hat{f}(X_i) = \text{mean}_{j \in V^S(X_i)}$, where $V^S(X_i)$ is a symmetric neighborhood of X_i, that is, $V^S(X_i) = \{max(i-k,1), \ldots, i-1, i, i+1, \ldots, min(i+k, n)\}$. Hastie and Tibshirani (1986) shows a large variety of smoothers, as the class of "cubic smoothing splines", in which it's more sophisticated than the moving average smoothers. The smoother's choice involves the type of mean of Y to be evaluated and the size of the neighborhood k. The last one, also called smoothing parameter, determine the relation between the bias and variance of the smoother.

3 Application

The data represent the amount of light received by the glass in wavelength and the amount of light that is reflected by them. The control group, denoted by *Treatment A*, is characterized by the ground with no glasses on the ground, and the treatments with glasses are: ground covered with transparent glass, denoted by *Treatment B*, ground covered with brown glass, denoted by *Treatment C* and ground covered with mixture of green and brown glasses, denoted by *Treatment D*.
The goal of the adjustment was to verify if there exists difference after applying the glass mat, and if so, which of the three types of glasses is better to generate more quality to the grape. Since the relation between the wavelength solar arriving to the ground and the amount of radiation that is transmitted to the grape is unknown, the GAM was chose to ensure greater flexibility in the analysis.
The model used to the fit the data is given by

$$Radiation = \eta + Treatment + \sum_{j=1}^{k} f_j(\text{Wavelength}),$$

where η is the intercept. The variable $Treatment$ represents the factor corresponding to the treatments B, C and D, while the treatment A is the reference treatment, and the variable $Wavelength$ represents the wavelength arriving to the grape. The first step of the analysis was to verify if the treatments were significantly different compared with the ground with no treatment. If they are, we have to choose which is more efficient. The base number chose for the cubic smoothing splines was $k = 40$. More details about this choose, see Wood (2004) and Wood (2012).
The result obtained is given by Table 1 (other results are ommited by lack of space). Note that the $p - value$ of the treatments are all significant, suggesting that all of them are significantly different of the treatment with no glasses. Besides that, the chose of GAM was appropriate, once the estimate smoothing parameter's degrees of freedom is 30.9 and is statistically significant (p-value $< 2e - 16$). Figure 1 shows that treatment B was more effective in the reflection than the other treatments.

TABLE 1. Summary of the adjust for the treatments.

	Estimate	S.E.	p-value
Intercept	-2.4320	0.0147	<2e-16
Treatment B	0.6389	0.0203	<2e-16
Treatment C	-0.4238	0.0203	<2e-16
Treatment D	-0.2514	0.0203	<2e-16

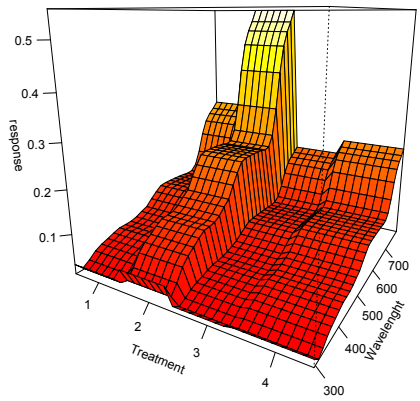


FIGURE 1. Fit of GAM for the four treatments.

4 Conclusion

This paper has shown an application of GAM in a situation where was desirable a flexible model but with enough complexity to this application. Other results were ommited, but we have seen (Figure 1) that Treatment B was better to produce sugar in the grapes and, thus, better wine quality.

References

Champagnol, F. (1984). *Elements of the physiology of the vine and of general viticulture.* Franois Champagnol.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 297 – 310.

Jackson, R.S. (2008). *Wine science: principles and applications.* Academic Press.

Nelder, J.A. and Wedderburn, R.W. (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 370 – 384.

Wood, S.N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* ,**99**, 673 – 686.

Wood, S.N. (2012). Package "mgcv": *Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation*, from http://cran.r-project.org/web/packages/mgcv/mgcv.pdf.

Marginal modeling of dependent paired comparison data

Manuela Cattelan[1], Cristiano Varin[2]

[1] University of Padova, Italy
[2] University Ca' Foscari Venezia, Italy

E-mail for correspondence: manuela.cattelan@unipd.it

Abstract: In the analysis of sport data it is often of interest to investigate whether some features of the players or teams influence the results of the contests. This analysis can be performed through models developed for paired comparison data. However, the assumption of independence of observations typical of traditional models for regression analysis of paired comparison data appears unrealistic. Here, a marginal model that accounts for dependence among observations with common wrestlers is proposed. The model is fitted by means of the hybrid pairwise likelihood method.

Keywords: Hybrid pairwise likelihood, Optimal estimating equations, Paired comparisons, Sumo.

1 A model for dependent paired comparison data

In the analysis of sport data it is often of interest to assess which features of teams or players influence the results of the contests. Here, we consider the results of sumo matches played by the top professional division in Japan in 2010 with the aim of determining whether some physical features of the wrestlers have an impact on the results of the matches. Often, this type of data are analyzed by means of models developed for paired comparison data (Cattelan, 2012). Traditional models for paired comparison data assume that all observations are independent, but it is unrealistic to assume that matches with a common wrestler are independent. We propose a marginal model for paired comparison data that accounts for dependence between observations with common sumo wrestlers.
Let Y_{ij} denote the result of the match between wrestlers i and j and, for notational convenience, assume that all paired comparisons (i, j), $i = 1, \ldots, n-1$ and $j = i+1, \ldots, n$ are observed. $Y_{ij} = 1$ if i wins against j and is 0 otherwise. Traditional models for paired comparison data define the probability that i wins against j as

$$p_{ij} = F(\lambda_i - \lambda_j),$$

where λ_i denotes the "ability" of wrestler i, $i = 1, \ldots, n$ and $F(\cdot)$ denotes the cumulative distribution function of a zero-symmetric random variable. The Thurstone (1927) model assumes that F is the normal cumulative distribution while the Bradley-Terry (1952) model assumes that F is the logistic cumulative distribution function. We further assume that a vector $\boldsymbol{\beta}_i = (\beta_1, \ldots, \beta_d)$ of d wrestler-specific explanatory variables is available. Accordingly, the ability of the ith wrestler is specified as the linear combination

$$\lambda_i = x_{1i}\beta_1 + \cdots + x_{di}\beta_d.$$

Traditional models are estimated assuming independence among all observations, so the estimates are obtained by solving the likelihood equations

$$\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{p}) = \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{y} = (y_{12}, y_{13}, \ldots, y_{n-1\,n})$, $\boldsymbol{p} = (p_{12}, \ldots, p_{n-1\,n})$, $\boldsymbol{D}$ is the Jacobian of $\boldsymbol{p}$ with respect to the components of $\boldsymbol{\beta}$ and $\boldsymbol{V}$ is the variance of $\boldsymbol{Y}$. When independence among all observations is assumed, $\boldsymbol{V}$ is a diagonal matrix with entries $p_{ij}(1-p_{ij})$. Under the independence assumption, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{\mathrm{ind}}$ has asymptotic normal distribution with mean $\boldsymbol{\beta}$ and variance $(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}$.

A traditional measure of dependence in binary data is the cross-ratio defined as

$$\text{CROSS-RATIO}(Y_{ij}, Y_{ik}) = \frac{\mathrm{pr}(Y_{ij}=1, Y_{ik}=1)\mathrm{pr}(Y_{ij}=0, Y_{ik}=0)}{\mathrm{pr}(Y_{ij}=0, Y_{ik}=1)\mathrm{pr}(Y_{ij}=1, Y_{ik}=0)}.$$

It is sensible to assume that only matches sharing a wrestler are dependent, while matches involving all different wrestlers are independent. Moreover, we assume a common cross ratio φ for all matches with at least a common wrestler.

Models for paired comparison data should satisfy a symmetry property since $\mathrm{pr}(Y_{ij} = 1)$ should be equal to $\mathrm{pr}(Y_{ji} = 0)$. Furthermore, $\mathrm{pr}(Y_{ij} = 1, Y_{ik} = 1)$ should be equal to $\mathrm{pr}(Y_{ji} = 0, Y_{ik} = 1)$. This property implies that $\text{CROSS-RATIO}(Y_{ij}, Y_{ik}) = 1/\text{CROSS-RATIO}(Y_{ji}, Y_{ik})$. As a consequence, the cross-ratio is specified as follows

$$\text{CROSS-RATIO}(Y_{ij}, Y_{kl}) = \begin{cases} \varphi, & \text{if } i = k \text{ or } j = l, \\ 1/\varphi, & \text{if } i = l \text{ or } j = k, \\ 1, & \text{if } i \neq j \neq k \neq l. \end{cases}$$

Using results from (Dale, 1986), we can derive the bivariate probability of observing a win for wrestler i against both j and k as

$$\mathrm{pr}(Y_{ij}=1, Y_{ik}=1) = \begin{cases} p_{ij}p_{ik}, & \text{if } \varphi = 1, \\ \dfrac{1 + (p_{ij} + p_{ik})(\varphi - 1) - H(p_{ij}, p_{ik}, \varphi)}{2(\varphi - 1)}, & \text{if } \varphi \neq 1, \end{cases} \tag{2}$$

where $H(q_1, q_2, \varphi) = \sqrt{\{1 + (q_1 + q_2)(\varphi - 1)\}^2 + 4\varphi(1 - \varphi)q_1 q_2}$. The probabilities of the other three possible combinations of results for the two fights are computed from the marginal univariate probabilities and formula (2). The specification of a complete multivariate model requires particular care because of the relation between the cross-ratios deriving from the symmetry properties that must be fulfilled. Moreover, a full multivariate model may be difficult to estimate. Accordingly, in the next section we discuss a fitting method that relies only on marginal univariate and bivariate probabilities.

2 Hybrid pairwise likelihood estimation

Hybrid pairwise likelihood is an estimating method proposed by Kuk (2007) that iterates between optimal estimating equations for the regression parameters and pairwise likelihood for the dependence parameter. The regressors are estimated employing optimal estimating equations similar to (1) but with a covariance matrix $\boldsymbol{V}_2$ that includes dependence between observations. The non-diagonal elements of the matrix $\boldsymbol{V}_2$ are computed as $\text{cov}(Y_{ij}, Y_{ik}) = \text{pr}(Y_{ij} = 1, Y_{ik} = 1) - \text{pr}(Y_{ij} = 1)\text{pr}(Y_{ik} = 1)$, so only marginal bivariate probabilities, which are computed as shown in formula (2), are necessary. The hybrid pairwise likelihood method requires to solve formula (1) with $\boldsymbol{V}$ replaced by $\boldsymbol{V}_2$ for a fixed dependence parameter φ, and then estimate φ from the pairwise likelihood with $\boldsymbol{\beta}$ fixed. Pairwise likelihood is a type of composite likelihood (Varin *et al.*, 2011) constructed as the product of marginal bivariate probabilities

$$L_{\text{pair}}(\varphi; \boldsymbol{Y}) = \prod_{(i,j)} \prod_{(k,l)} \text{pr}(Y_{ij} = y_{ij}, Y_{kl} = y_{kl}; \varphi).$$

The hybrid pairwise likelihood estimators of the regression coefficients are asymptotically normally distributed with mean $\boldsymbol{\beta}$ and variance $(\boldsymbol{D}^{\top}\boldsymbol{V}_2^{-1}\boldsymbol{D})^{-1}$ (Kuk, 2007).

3 Application and conclusions

The described methodology is applied to the results of the 2010 tournaments of the sumo Makuuchi division. The data include 1806 matches played by 61 different wrestlers. The available covariates are the height and weight of the wrestlers, their age and the year in which the wrestlers started to fight.

The traditional Bradley-Terry model assuming independence among observations is fitted to the data. The first three columns in Table 1 show the results of the best model. The significant covariates are the height of the wrestlers, the square of the height, the age of the wrestler, and the number of years the wrestler has been fighting.

TABLE 1. Estimates, standard errors and absolute z values of the sumo data employing the Bradley-Terry model (`Independent`) and the proposed model (`Dependent`).

	Independent			Dependent		
	Est.	S.e.	$\|z\|$	Est.	S.e.	$\|z\|$
Height	0.60	0.20	2.99	0.70	0.35	1.99
Height$^2 \times 10^{-2}$	-0.20	0.05	3.08	-0.20	0.10	2.06
Age	0.08	0.01	5.40	0.10	0.03	3.62
Experience	-0.04	0.01	3.08	-0.04	0.03	1.73

However, the Bradley-Terry model does not account for dependence between observations. Columns 4-6 in Table 1 show the estimates, standard errors and absolute z values of the proposed marginal model that accounts for dependence. The estimate of the dependence parameter is $\hat{\varphi} = 1.157$. The inclusion of dependence in the model leads to a lower absolute z value of all the covariates. In particular, the regression parameter of the years of wrestling experience is not significant when dependence is accounted for. It is important to include dependence in paired comparison data because the significance of the parameters may change. The model proposed and the estimating method employed require only the specification of marginal bivariate probabilities. This method can be applied in many other fields in which paired comparison data arise.

References

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, **39**, 324 – 345.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, **27**, 412 – 433.

Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909 – 917.

Kuk, A.Y.C. (2007). A hybrid pairwise likelihood method. *Biometrika*, **94**, 939 – 952.

Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review*, **34**, 368 – 389.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5 – 42.

Functional Linear Mixed Models for Sparsely and Irregularly Sampled Data

Jona Cederbaum[1], Sonja Greven[1]

[1] Institute of Statistics, Ludwig-Maximilians University Munich, Germany

E-mail for correspondence: Jona.Cederbaum@stat.uni-muenchen.de

Abstract: We propose a nonparametric estimation procedure to analyze correlated functional data which are observed irregularly or even sparsely. The model we propose is a functional linear mixed model which can be seen as a functional analogue to the linear mixed model. Estimation is based on dimension reduction via functional principal component analysis. Our procedure allows the decomposition of the variability of the data as well as the estimation of main effects of interest and borrows strength across curves. The method is motivated by data from speech production research.

Keywords: Functional Principle Component Analysis; Functional Data; Sparse Data; Mixed Models.

1 Introduction

Conventional regressions approaches for functional data (fd) often assume that the functional observations are independent and observed on a fine, regular grid. This may be very restrictive in practice, where functional observations often are correlated and frequently evaluated on – possibly only few – irregularly spaced points. Sources of correlation may be, as in the multivariate case, repeated measurements (longitudinal functional data) or grouping in the data.
We propose an estimation procedure that allows analyzing irregularly and sparsely sampled fd with an additional correlation structure. We build on two existing approaches. First, we extend the functional linear mixed model for longitudinal fd of Greven et al. (2010) to more general correlated fd which are not sampled on a fine, regular grid. Second, we generalize the work of Yao et al. (2005), who consider sparse independent fd, to correlated functions.

2 The General Functional Linear Mixed Model

Scalar correlated data are frequently analyzed using linear mixed models. A functional analogue to the standard linear mixed model is given by

$$Y_i(t) = \mu(x_i, t) + z_i^T B(t) + E_i(t) + \varepsilon_{it}, \quad i = 1, \ldots, n, \tag{1}$$

where $Y_i(t)$ is a vector of random functions observed at arguments $t \in \mathcal{T}$, a closed interval in $\mathbb{R}$. $\mu(x_i, t)$ is a curve specific smooth mean function dependent on a vector of covariates x_i. The effects of the covariates may be linear or smooth. The random effects in the linear mixed model are replaced by a vector-valued zero mean square integrable random process $B(t)$. z_i is a covariate vector. $E_i(t)$ is a curve specific deviation in form of a smooth residual curve, and ε_{it} is white noise measurement error with variance σ^2. We assume that $B(t)$, $E_i(t)$, and ε_{it} are independent for all curves $i = 1, \ldots, n$.

3 Motivating Data Application

In this application, the aim is to understand under which conditions the sounds "sch" and "s" in the German language overlap when they subsequently appear in a word. We analyze data from an experiment in which nine subjects speak out loud different imaginary compound words containing "sch" and "s" (such as "Calla*sSch*immel" or "Gula*schS*impel"). The subjects repeat each word up to five times while their tongue movement is measured and summarized in a one-dimensional index over time. The index development for one subject is depicted in figure 1. In the left, we show the development for compound words with "sch" following "s" and in the right, the one for words with "s" following "sch". Index values of 1 indicate that "s" is spoken where as index values of "-1" stand for the pronunciation of "sch". Due to differing reading lengths, the time scale is standardized to a [0,1] interval resulting in irregular spacing of the measurements. The data consist of curves which are correlated for each subject as well as for each compound word which is why we propose a functional linear mixed model with crossed random effects of the form

$$Y_{ijh}(t) = \mu(x_{ij}, t) + B_i(t) + C_j(t) + E_{ijh}(t) + \varepsilon_{ijht}, \tag{2}$$

with $Y_{ijh}(t)$ denoting the summarizing index for subject i, compound word j and repetition h at time $t \in \mathcal{T} = [0, 1]$. $\mu(x_{ij}, t)$ is a curve-specific smooth mean function, x_{ij} are known covariates such as the order of the sounds "s" and "sch" or which syllables are stressed. $B_i(t)$ and $C_j(t)$ are random functional intercepts for subjects and words, respectively. Speaker-, word-, and repetition-specific deviations are modeled by the smooth curve-specific residual term $E_{ijh}(t)$. ε_{ijht} is white measurement error with variance σ^2.

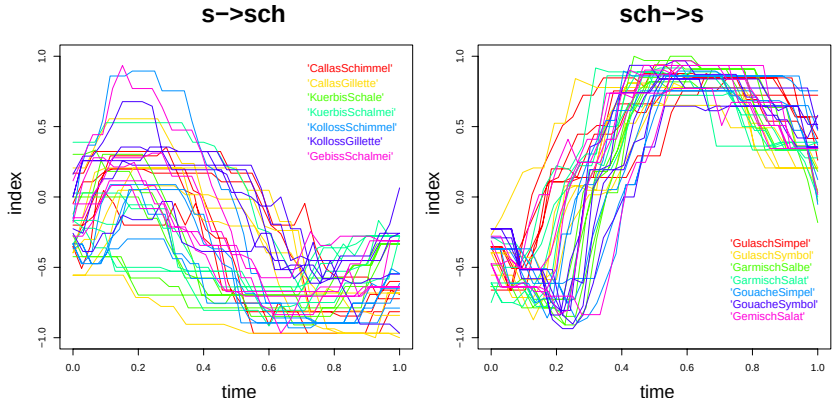


FIGURE 1. Index development over time for *one* subject. Left: compound words with "sch" following "s". Right: compound words with "s" following "sch". The curves belonging to one word are the same color.

4 Estimation

So far, the GFLMM does not differ between densely and irregularly or sparsely sampled data. The nature of the grid comes into play in parameter estimation and implementation is more challenging in our case. For reasons of simplicity we focus the presentation of the algorithm on model (2).

Step 1 The fixed main effects function is estimated using penalized splines under a working independence assumption. $Y_{ijh}(t) = \mu(x_{ij}, t) + \varepsilon_{ijht}$. Given an estimator for $\mu(x_{ij}, t)$ the data can be centered $\tilde{Y}_{ijh}(t) = Y_{ijh}(t) - \hat{\mu}(x_{ij}, t)$ for all i, j, h, and t.

Step 2 The crucial step is the estimation of the auto-covariance functions of processes $B_i(t), C_j(t)$, and $E_{ijh}(t)$. We use the following variance decomposition

$$Cov\{\tilde{Y}_{ijh}(s), \tilde{Y}_{i'j'h'}(t)\} = K_B(s,t)\delta_{i,i'} + K_C(s,t)\delta_{j,j'} + \left[K_E(s,t) + \sigma^2\delta_{s,t}\right]\delta_{i,i'}\delta_{j,j'}, \quad (3)$$

with $K_B(s,t) = Cov\{B_i(s), B_i(t)\}$, $K_C(s,t) = Cov\{C_j(s), C_j(t)\}$,and $K_E(s,t) = Cov\{E_{ijh}(s), E_{ijh}(t)\}$. $\delta_{i,i'}$ is the Kronecker delta.

We estimate $K_B(s,t), K_C(s,t)$, and $K_E(s,t)$ by bivariate smoothing of the respective cross-products in s and t (for $s \neq t$). Strength is borrowed across curves which is of particular importance when curves are sampled sparsely. We evaluate the estimators also for $s = t$. σ^2 is estimated using the mean difference of $\tilde{Y}_{ijh}(t)^2$ and the diagonal of the smoothed auto-covariance of $\tilde{Y}_{ijh}(t)$. The error

variance can easily be extended to be time-varying. Note that for random intercept models, the estimation of the auto-covariances is similar to that in Di and Crainiceanu (2010) but the proposed method can be generalized more straightforwardly to a crossed design.

Step 3 We use the auto-covariances evaluated on a regular grid to obtain estimates of the eigenvalues and of the eigenfunctions using the eigendecomposition of the auto-covariance matrices. The Karhunen-Loève expansion is then used to obtain parsimonious expansions of the random processes of the form

$$B_i(t) = \sum_{k=1}^{N_B} \xi_{ik}^B \phi_k^B(t), C_j(t) = \sum_{k=1}^{N_C} \xi_{jk}^C \phi_k^C(t), E_{ijh}(t) = \sum_{k=1}^{N_E} \xi_{ijhk}^E \phi_k^E(t),$$

where ξ_{ik}^B, ξ_{jk}^C, and ξ_{ijhk}^E are uncorrelated random variables with zero mean and variances corresponding to the ordered eigenvalues and $\phi_k^B(t), \phi_k^C(t)$, and $\phi_k^E(t)$ are the corresponding eigenfunctions. The numbers of eigenfunctions can be chosen by the proportion of variance explained (Greven et al., 2010). For fixed N_B, N_C, and N_E, model (2) is a linear mixed model.

Step 4 The subject-, word-, and visit-specific scores ξ_{ik}^B, ξ_{jk}^C, and ξ_{ijhk}^E can then be obtained directly as estimated BLUPS. Note that the traditional way via numerical integration does not work for sparse fd.

5 Discussion and Future Work

We have presented an estimation procedure for correlated fd which are sampled irregularly or sparsely. We want to point out that the presented work is work in progress and that further investigations are necessary. We plan to perform more general simulations and to compare our method to others. We consider to implement an iterative estimation procedure to improve accuracy of mean and covariance estimates for sparse data.

References

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, **4**, 1022 – 1054.

Di, C.-Z. and Crainiceanu, C. (2010). Multilevel Sparse Functional Principle Component Analysis. Johns Hopkins University, Dept. of Biostatistics. Available at: http://works.bepress.com/di/1

Yao, F., Müller, H., and Wang, J. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, **100(470)**, 577 – 590.

Testing Linearity for Nonlinear Count Time Series Models

Vasiliki Christou[1], Konstantinos Fokianos[1]

[1] University of Cyprus, Nicosia, Cyprus

E-mail for correspondence: christou.vasiliki@ucy.ac.cy

Abstract: We consider testing linearity against two special classes of nonlinear models. In particular, we are interested in Poisson or negative binomial processes for count time series. The score test is our preferable method, since it requires estimation only under the null hypothesis. It can be shown that if all the parameters of the non linear model are identified under the null, then the score statistic is asymptotically $\mathcal{X}^2$ distributed. But when the model has non identifiable parameters under the null, then the classical asymptotic theory does not hold and we employ a supremum type of test.

Keywords: Nonidentifiability, Quasi likelihood, Score test.

1 Introduction

We consider testing linearity against two special classes of nonlinear alternatives for count time series data. The first class contains models which do not face the problem of nonidentifiability, that is all the parameters of the model are identified under the null hypothesis. For this class of models and under the null hypothesis of linearity, the score test statistic possesses an asymptotic $\mathcal{X}^2$ distribution. The second class of nonlinear models consists of models in which a nonnegative nuisance parameter exists under the alternative hypothesis but not when linearity holds. In this particular case the testing problem is nonstandard and the classical asymptotic theory for the score test does not apply.

We focus on count time series autoregressive models based on either the Poisson or the negative binomial distribution. After parameterizing suitably the negative binomial distribution so that it has the same mean as the Poisson, we employ quasi likelihood inference to get the consistent estimators. Once the estimators are obtained, we calculate the score test statistic and we investigate the size and the power of the test by a simulation study, based on a parametric bootstrap procedure.

2 Autoregressive Modeling and Inference

Assume that $\{Y_t, t \geq 1\}$ is a count time series and let $\{\lambda_t, t \geq 1\}$ be an unobserved sequence of mean processes. Denote by $\mathcal{F}_t^{Y,\lambda}$ the history of the response process up to and including time t. We suppose that the conditional distribution of $Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}$ is either the Poisson or the negative binomial distribution. The negative binomial distribution is suitably reparameterized to have the same mean as the Poisson distribution.
We consider a linear model for the mean process λ_t given by

$$\lambda_t = d + a\lambda_{t-1} + bY_{t-1}, \tag{1}$$

and two nonlinear models defined by

$$\lambda_t = \frac{d}{(1+Y_{t-1})^{\gamma}} + a\lambda_{t-1} + bY_{t-1}, \tag{2}$$

and

$$\lambda_t = d + a\lambda_{t-1} + (b + c\exp(-\gamma Y_{t-1}^2))Y_{t-1}, \tag{3}$$

where all the parameters d, a, b, γ, c are assumed to be positive.
We suggest to use Poisson based score estimating function for estimating the unknown parameters of the model. This methodology avoids complicated likelihood function and it ensures that the regression parameters are estimated consistently. In addition, it can be shown that the estimators are asymptotically normally distributed. For the case of the negative binomial distribution, the additional parameter ν is estimated consistently by a method of moments estimator. For more details see Christou and Fokianos (2012).

3 Testing Linearity

We carry out testing using the score (or Lagrange Multiplier) test. The main advantage of this test is that it requires estimation only for the constrained model. In other words, it requires estimation for the simple linear model. Denote by $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$ the unknown parameter and let $\boldsymbol{S}_n = (\boldsymbol{S}_n^{(1)}, \boldsymbol{S}_n^{(2)})$ be the corresponding partition of the score function.
The test of interest is

$$H_0 : \boldsymbol{\theta}^{(2)} = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\theta}^{(2)} > \mathbf{0}, \qquad \text{componentwise.}$$

We use the score test statistic proposed by Amendola and Francq (2009) and Breslow (1990), which is given by

$$LM_n = \boldsymbol{S}_n^{(2)'}(\tilde{\boldsymbol{\theta}}_n)\widetilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\boldsymbol{\theta}}_n)\boldsymbol{S}_n^{(2)}(\tilde{\boldsymbol{\theta}}_n),$$

where $\tilde{\boldsymbol{\theta}}_n = (\tilde{\boldsymbol{\theta}}_n^{(1)}, \mathbf{0})$ is the maximum quasi-likelihood estimator of $\boldsymbol{\theta}$ for the linear model and $\widetilde{\boldsymbol{\Sigma}}$ is an appropriate estimator for the covariance matrix $\boldsymbol{\Sigma} = \mathrm{Var}(\frac{1}{\sqrt{n}} \boldsymbol{S}_n^{(2)}(\tilde{\boldsymbol{\theta}}_n))$.

If all the parameters are identified under the null hypothesis, the standard asymptotical theory holds and the score statistic follows asymptotically a $\mathcal{X}^2_{m_2}$ distribution under the null, where m_2 is the length of the subvector $\boldsymbol{\theta}^{(2)}$ (see Francq and Zakoïan (2010)). Note that model (2) belongs to this class.

But in many cases we have to test the linearity assumption for non linear models that contain nuisance parameters that are not identified under the null (see model (3)). The lack of identification affects also the score test and the classical asymptotic theory does not apply. Davies (1987) proposed a supremum test to solve this problem. Consider that Γ is a grid of values for the nuisance parameter, denoted by γ. Then the sup-score test statistic is given by

$$LM_n = \sup_{\gamma \in \Gamma} LM_n(\gamma).$$

TABLE 1. Empirical size for sample sizes $n = 500$ and $n = 1000$. Data are generated from the linear model (1) with true values $(d, a, b) = (1.5, 0.05, 0.6)$. Results are based on B = 499 bootstrap replicates and 200 simulations.

	Bootstrap test for $n = 500$		Bootstrap test for $n = 1000$	
Nominal significance level	Poisson	NegBin ($\nu = 4$)	Poisson	NegBin ($\nu = 4$)
$\alpha = 1\%$	0.005	0.000	0.000	0.015
$\alpha = 5\%$	0.050	0.025	0.055	0.060
$\alpha = 10\%$	0.130	0.080	0.100	0.095

4 Simulations and Case Study

Based on parametric bootstrap procedure, we investigate the size and the power of the test. We simulate data either from the Poisson or the negative binomial distribution. Tables 1–3 summarize the result of a simulation study. All results demonstrate the validity of our approach.

Acknowledgments: Work supported by Cyprus Research Promotion Foundation TEXNOLOGIA/THEPIS/0609(BE)/02.

References

Amendola, A. and Francq, C. (2009). Concepts and Tools for Nonlinear Time-Series Modelling. *Handbook of Computational Econometrics.*

TABLE 2. Empirical power for sample sizes $n = 500$ and $n = 1000$. Data are generated from the non linear model (2) with true values $(d, a, b) = (1.5, 0.05, 0.6)$ and $\gamma \in \{0.3, 0.5, 1\}$. Results are based on B = 499 bootstrap replicates and 200 simulations. The nominal significance level is $\alpha = 5\%$.

	Bootstrap test for $n = 500$		Bootstrap test for $n = 1000$	
Nonlinear (2) γ	Poisson	NegBin ($\nu = 4$)	Poisson	NegBin ($\nu = 4$)
$\gamma = 0.3$	0.207	0.157	0.271	0.212
$\gamma = 0.5$	0.450	0.424	0.740	0.688
$\gamma = 1$	0.924	0.837	1.000	0.995

TABLE 3. Empirical power for sample sizes $n = 500$ and $n = 1000$. Data are generated from the non linear model (3) with true values $(d, a, b) = (0.5, 0.3, 0.2)$, $c_1 \in \{0.2, 0.4\}$ and $\gamma \in \{0.05, 0.5\}$. Results are based on B = 499 bootstrap replicates and 200 simulations. The nominal significance level is $\alpha = 5\%$.

		Bootstrap test for $n = 500$		Bootstrap test for $n = 1000$	
Nonlinear (3) c	γ	Poisson	NegBin ($\nu = 4$)	Poisson	NegBin ($\nu = 4$)
$c = 0.2$	$\gamma = 0.05$	0.140	0.220	0.315	0.355
$c = 0.2$	$\gamma = 0.5$	0.111	0.122	0.312	0.265
$c = 0.4$	$\gamma = 0.05$	0.755	0.739	0.985	0.985
$c = 0.4$	$\gamma = 0.5$	0.420	0.469	0.855	0.775

John Wiley & Sons, Ltd. pp. 377 – 427.

Breslow, N. (1990). Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-Likelihood Models. *Journal of the American Statistical Association*, **85**(410), 565 – 571.

Christou, V. and Fokianos, K. (2012). Quasi–Likelihood Inference for Negative Binomial Time Series Models. *submitted for publication.*

Davies, R. B. (1987). Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative. *Biometrika*, **74**(1), 33 – 43.

Francq, C. and Zakoïan, J.-M. (2010). *GARCH models: Stracture, Statistical Inference and Financial Applications.* UK: Wiley.

Bayesian inference for a family based on the Weibull and the power series distributions

Juliana Cobre[1], Mário de Castro[1]

[1] Universidade de São Paulo, São Carlos-SP, Brazil

E-mail for correspondence: jucobre@icmc.usp.br

Abstract: In this work we deal with Bayesian inference for the parameters of some distributions in the Weibull power series family. For statistical modeling purposes, this class of three parameter distributions allows great flexibility. The density function can be bimodal. Furthermore, the hazard rate function accommodates increasing, decreasing and upside down bathtub shapes. We base our inferences on the Markov chain Monte Carlo (MCMC) simulation methods. Results from a simulation study aimed to assess some frequentist properties of the estimators are reported.

Keywords: Lifetime distributions, Weibull distribution, Power series distribution, Bayesian inference, MCMC methods.

1 Introduction

Distributions for modeling the life length of individuals and materials are extensively studied in the statistical literature under the headings of reliability and survival analysis. The general decreasing failure rate (DFR) family proposed by Chahkandi and Ganjali (2009) includes several especial cases existing in literature.

By compounding the Weibull and the power series distributions, Morais and Barreto-Souza (2011) proposed the Weibull power series (WPS) class of distributions. Their construction runs as follows. Let $Y_1, \ldots, Y_Z$ be a random sample from the Weibull distribution with density function $f(y|\alpha,\beta) = \alpha\beta y^{\alpha-1}\exp(-\beta y^{\alpha})$, for $\alpha > 0, \beta > 0$ and $y > 0$, where Z follows the truncated at 0 power series distribution with probability function $p(z|\theta) = \frac{a_z\theta^z}{C(\theta)}$, $z = 1, 2, \ldots$, where $\theta \in \Theta$, $a_z > 0$ does not depend on θ and $C(\theta) = \sum_{k=1}^{\infty} a_k\theta^k$ (see Table 1). Let $X = \min\{Y_1, \ldots, Y_Z\}$, so that the distribution of $X|Z = z$ is Weibull with parameters α and βz. The marginal distribution of X is termed the WPS distribution, whose density function is given by

$$f(x|\alpha,\beta,\theta) = \frac{\alpha\theta\beta x^{\alpha-1}e^{-\beta x^{\alpha}}C'(\theta e^{-\beta x^{\alpha}})}{C(\theta)}, \; x \geq 0. \tag{1}$$

Besides having as particular cases the exponential power series family of distributions, the WPS class of distributions has increasing, decreasing and

TABLE 1. Some truncated at 0 distributions in the power series family.

Distribution	a_z	$C(\theta)$	$C'(\theta)$	Θ	$\pi(\theta\|c_1, d_1)$
Geometric	1	$\frac{\theta}{1-\theta}$	$\frac{1}{(1-\theta)^2}$	$(0,1)$	$\text{Be}(\theta;\, c_1, d_1)$
Poisson	$1/z!$	$e^{\theta}-1$	e^{θ}	$(0,\infty)$	$\text{Ga}(\theta;\, c_1, d_1)$
Logarithmic	$1/z$	$-\log(1-\theta)$	$\frac{1}{(1-\theta)}$	$(0,1)$	$\text{Be}(\theta;\, c_1, d_1)$
Binomial	$\binom{m}{z}$	$(1+\theta)^m-1$	$m(\theta+1)^{m-1}$	$(0,\infty)$	$\text{Ga}(\theta;\, c_1, d_1)$

upside down bathtub shaped failure rate function. Therefore, the WPS family is a more attractive class of distributions than the DFR family introduced by Chahkandi and Ganjali (2009).

Here we develop inferential tools under a Bayesian viewpoint for the parameters of the WPS class of distributions proposed by Morais and Barreto-Souza (2011). Our paper unfolds as follows. In Section 2 we present the required steps to draw samples from the posterior distribution. Preliminary results of a simulation study are reported in Section 3. We end up with some remarks in Section 4.

2 Bayesian inference

We assume that α, β and θ are *a priori* independent, that is,

$$\pi(\alpha, \beta, \theta|\boldsymbol{h}) = \pi(\alpha|c_0, d_0)\pi(\beta|c_1, d_1)\pi(\theta|c_2, d_2), \tag{2}$$

where $\boldsymbol{h}$ denotes the vector of hyperparameters. We postulate $\pi(\alpha|c_0, d_0) = \text{Ga}(\alpha; c_0, d_0)$ and $\pi(\beta|c_1, d_1) = \text{Ga}(\beta;\, c_1, d_1)$, where $\text{Ga}(\cdot; c, d)$ denotes the density function of the gamma distribution with mean equal to c/d. The prior specification for θ is specified in Table 1, where $\text{Be}(\cdot; c, d)$ denotes the density function of the beta distribution. The vector of hyperparameters $\boldsymbol{h}$ is chosen to ensure vague prior knowledge.

From (1) and (2), the posterior distribution of (α, β, θ) is given by $\pi(\alpha, \beta, \theta|\boldsymbol{x}, \boldsymbol{h}) \propto \pi(\alpha, \beta, \theta|\boldsymbol{h}) \prod_{i=1}^{n} f(x_i|\alpha, \beta, \theta)$. However, to ease the computations, we resort to data augmentation. The observed data $\boldsymbol{X} = (X_1, \ldots, X_n)$ is augmented by $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$. Then, the density function of the complete data $(\boldsymbol{X}, \boldsymbol{Z})$, observable and unobservable variables, respectively, is given by $f(x, z|\alpha, \beta, \theta) = \frac{a_z \theta^z}{C(\theta)} \alpha\beta z x^{\alpha-1} \exp(-\beta z x^{\alpha}),\ x \geq 0, z = 1, 2, \ldots$. The likelihood function corresponding to the complete data $(\boldsymbol{X}, \boldsymbol{Z})$, with $\{(X_i, Z_i)\}_{i=1}^{n}$ conditionally independent given (α, β, θ), has expression

$$L(\alpha, \beta, \theta; \boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{n} \frac{a_{z_i} \theta^{z_i}}{C(\theta)} \alpha\beta z_i x_i^{\alpha-1} \exp(-\beta z_i x_i^{\alpha}). \tag{3}$$

After combining the likelihood function in (3) with the prior distribution, the joint posterior distribution of (α, β, θ) results to be $\pi(\alpha, \beta, \theta|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{h}) \propto \pi(\alpha|c_0, d_0)\pi(\beta|c_1, d_1)\pi(\theta|c_2, d_2)L(\alpha, \beta, \theta; \boldsymbol{x}, \boldsymbol{z})$.

Taking into account the prior distribution in (2) and the likelihood function in (3), the full conditional distributions turn out to be

$$\pi(\alpha|\beta,\boldsymbol{x},\boldsymbol{z},c_0,d_0) \propto \alpha^n\Big(\prod_{i=1}^{n} x_i\Big)^{\alpha-1} \exp\Big(-\beta\sum_{i=1}^{n} x_i^{\alpha} z_i\Big)\pi(\alpha|c_0,d_0), \quad (4)$$

$$\pi(\beta|\alpha,\boldsymbol{x},\boldsymbol{z},c_1,d_1) = \mathrm{Ga}(\beta; c_1+n, d_1+\sum_{i=1}^{n} x_i^{\alpha} z_i), \quad (5)$$

$$\pi(\theta|\boldsymbol{z},c_2,d_2) \propto \frac{\theta^{\sum_{i=1}^{n} z_i}}{\{C(\theta)\}^n}\pi(\theta|c_2,d_2) \quad (6)$$

$$\text{and} \quad \pi(z_i|x_i,\alpha,\beta,\theta) = \frac{z_i \exp\{-\beta x_i^{\alpha}(z_i-1)\}a_{z_i}\theta^{z_i-1}}{C'(\theta e^{-\beta x_i^{\alpha}})}, \; z_i = 1,2,\ldots. \quad (7)$$

Samples from z in (7) are drawn by applying the rejection method Devroye (1986). In (4), the distribution is log-concave and the sampling is straightforward with the adaptive rejection method (Gilks and Wild, 1992). We stress that the distribution of β in (5) is the same, whichever the distribution in Table 1. For the geometric distribution, θ in (6) is sampled from a $\mathrm{Be}(c_2+\sum_{i=1}^{n} z_i - n, d_2+n)$ distribution, whereas the remaining distributions in Table 1 require Metropolis steps. MCMC computations in the simulations (Section 3) were implemented using the FORTRAN language.

3 Results

In this section we present the results of a simulation study. Our study comprises the Weibull geometric (WG) distribution, the Weibull Poisson (WP) distribution and the Weibull logarithmic (WL) distribution. Some frequentist properties of the Bayesian estimators are assessed. The hyperparameters in (2) were set at $c_0 = d_0 = 0.01$, $c_1 = d_1 = 0.01$ and $c_2 = d_2 = 1$. In Table 1 summaries from 500 replications. For each replication of the WG and the EG distributions, after discarding the first 2,000 iterations of the Gibbs sampler, we used 15,000 iterations with thinning equal to 5, thus obtaining 3,000 samples for each parameter. The simulations comprise the WL distribution with similar assumptions and illustrate the study written in the BUGS language. In general, the average of the posterior standard deviations and the root mean squared error of the posterior means are close. For the WG and WL distributions, even when $n = 500$, some bias still remains in the estimator of θ and the coverage probability of the 95% highest posterior density interval (HPD) is not so close to the nominal value.

4 Conclusion

Frequentist properties of the Bayesian estimators are close to what was presented by Tahmasbi and Rezaei (2008) and Chahkandi and Ganjali (2009),

TABLE 2. Posterior estimates from 500 replications and different samples sizes n=500 (True: true value of the parameter, Est.: average of the posterior means, SD: average of the posterior standard deviations, RMSE: root mean squared error of the posterior means and CP: coverage probability of the 95% HPD interval).

	Parameter	True	Est.	SD	RMSE	CP
EG	β	1.00	1.02	0.12	0.13	0.954
	θ	0.60	0.58	0.07	0.07	0.946
WG	α	3.00	3.02	0.20	0.19	0.944
	β	1.50	1.51	0.25	0.23	0.946
	θ	0.45	0.42	0.17	0.15	0.928
WL	α	3.00	2.99	0.21	0.27	0.934
	β	1.50	1.54	0.19	0.18	0.931
	θ	0.45	0.44	0.22	0.14	0.966

except for the coverage probabilities not presented in these studies. In an extended version of the paper a more detailed simulation study will be carried out. Furthermore, different distributions in the family will be fitted to real data sets.

Acknowledgments: Special thanks to CNPq, Brazil.

References

Chahkandi, M. and Ganjali, M. (2009). On some lifetime distributions with decreasing failure rate. *Comput. Statist. Data Anal.*, **53**, 4433 – 4440.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* New York: Springer.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, **41**, 337 – 348.

Morais, A. and Barreto-Souza, W. (2011). A compound class of Weibull and power series distributions. *Comput. Statist. Data Anal.*, **55**, 1410 – 1425.

Tahmasbi, R. and Rezaei, S. (2008). A two-parameter lifetime distribution with decreasing failure rate. *Comput. Statist. Data Anal.*, **52**, 3889 – 3901.

Corrected Profile Likelihood in Heteroscedastic Symmetric Nonlinear Model

Audrey H.M.A. Cysneiros[1], Mariana C. Araújo[1]

[1] Departamento de Estatística, Universidade Federal de Pernambuco, Brazil

E-mail for correspondence: audrey@de.ufpe.br

Abstract: In this paper, our goal is to develop a Bartlett correction to the likelihood ratio and modified profile likelihood ratio statistics, respectively, in the class of heteroscedastic symmetric nonlinear model (HSNLM) proposed by Cysneiros et al. (2010). We present numerical evidence on the finite-sample behavior of the different associated likelihood ratio tests. The results favor the tests we propose.

Keywords: Bartlett correction; Heteroscedastic symmetric nonlinear model; Likelihood ratio test; Profile likelihood.

1 Heteroscedastic Symmetric Nonlinear Model

Cysneiros *et al.* (2010) proposed a class of heteroscedastic symmetric nonlinear model (HSNLM) assuming that the random variables $Y_1, \ldots, Y_n$ are independent and each Y_i has a symmetric distribution with mean parameter $\mu_i \in IR$ and dispersion parameter $\phi_i > 0$ and density function

$$\pi(y_i; \mu_i, \phi_i) = \frac{1}{\sqrt{\phi_i}} g(u_i),\ y \in IR, \tag{1}$$

where $u_i = (y_i - \mu_i)^2/\phi_i,\ i = 1, \ldots, n$, and the generating density function $g : IR \to [0, \infty)$ is such that $\int_0^\infty g(u)du < \infty$. We will denote $Y_i \sim S(\mu_i, \phi_i, g)$.
The heteroscedastic symmetric nonlinear model is given by

$$Y_i = \mu_i + \sqrt{\phi_i}\epsilon_i,\ i = 1, \ldots, n, \tag{2}$$

where ϵ_i are i.i.d. with $\epsilon_i \sim S(0, 1, g)$ and both mean and dispersion parameters, respectively μ_i and ϕ_i, vary across the observations through nonlinear regression structures as follows: the mean response is $\mu = (\mu_1, \ldots, \mu_n)^\top$ with $\mu_i = f(x_i; \beta)$, where $\mathbf{x}_i$ is an $m \times 1$ vector of known explanatory variables associated with the ith response, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ $(p < n)$ vector of unknown regression parameters and $f(\cdot; \cdot)$ is twice continuously differentiable function in β. Moreover, the $n \times p$ matrix of derivatives of μ with respect to β is denotated by $\tilde{X} = \partial\mu/\partial\beta$ and assumed to be

full rank. A systematic component for the dispersion parameter vector $\phi = (\phi_1, \ldots, \phi_n)^\top$ given by $\phi_i = h(\tau_i)$ is assumed, where $h(\cdot)$, commonly named a dispersion link function, is a known one-to-one continuously diferentiable function of the dispersion linear predictor defined by $\tau_i = \mathbf{z}_i^\top \gamma$, where $\mathbf{z}_i$ is a $q \times 1$ vector of explanatory variables that can have components in common with $\mathbf{x}_i$ and $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ is a $q \times 1$ vector of unknown parameters. The dispersion link function $h(\cdot)$ should be a positive-value function and one possible choice for $h(\cdot)$ is $h(\tau) = \exp(\tau)$.

2 Bartlett Correction

To test hypothesis of interest in regression models is frequently used the likelihood ratio test which is based on first order asymptotics, since it relies on a large sample approximation. Under the null hypothesis, the likelihood ratio statistic (LR) is distributed as χ^2_q up to an error of order n^{-1} where q is the number of imposed restriction under the null hypothesis. However, it is well known that for small sample size the approximation of LR to χ^2_q distribution can be not satisfactory. An alternative to improve this approximation is to incorporate a correction factor proposed by Bartlett (1937) to the LR statistic. Therefore, the expected value of the corrected statistic $LR^* = LR/(1+c)$ is closer to the one from χ^2_q distribution than the expected value of LR, an error of order n^{-2}. Thus, in matrix notation we obtain the Bartlett correction factor c to the likelihood ratio statistic of the test $H_0 : \gamma = \gamma^{(\mathbf{0})}$ versus $H_1 : \gamma \neq \gamma^{(\mathbf{0})}$ considering multiplicative heteroscedasticity, that is, $\phi_i = \exp\{\mathbf{z_i}^\top \gamma\}$, is given by $c = \epsilon_q + \epsilon_{p,q}$, with

$$
\begin{aligned}
\epsilon_q &= tr\left((Q_3 + Q_4 + Q_5 + Q_6) Z^{(2)}_{\gamma d}\right) + (N_1 + N_3 + N_9)\mathbf{1}^\top \Lambda_2 Z^{(3)}_\gamma \Lambda_2 \mathbf{1} \\
&+ (N_2 - 2N_9)\mathbf{1}^\top \Lambda_1 Z^{(3)}_\gamma \Lambda_2 \mathbf{1} + (N_4 + N_9)\mathbf{1}^\top \Lambda_1 Z^{(3)}_\gamma \Lambda_1 \mathbf{1} \\
&+ (N_5 + N_9)\mathbf{1}^\top \Lambda_2 Z^{(2)}_{\gamma d} Z_\gamma \Lambda_2 \mathbf{1} + N_7 \mathbf{1}^\top \Lambda_2 Z^{(2)}_{\gamma d} Z_\gamma \Lambda_1 \mathbf{1} \\
&+ (N_6 - 2N_9)\mathbf{1}^\top \Lambda_1 Z^{(2)}_{\gamma d} Z_\gamma \Lambda_2 \mathbf{1} + (N_8 + N_9)\mathbf{1}^\top \Lambda_1 Z^{(2)}_{\gamma d} Z_\gamma \Lambda_1 \mathbf{1},
\end{aligned}
$$

$$
\begin{aligned}
\epsilon_{p,q} &= -\frac{1}{\delta_{(0,1,0,0,0)}} tr\left((Q_1 + Q_2) Z_{\beta d} Z_{\gamma d}\right) - N_{18} \mathbf{1}^\top \Lambda_3 Z_{\beta d} Z_\gamma Z_{\gamma d} \Lambda_2 \mathbf{1} \\
&- N_{19} \mathbf{1}^\top \Lambda_3 Z_{\beta d} Z_\gamma Z_{\gamma d} \Lambda_1 \mathbf{1} - N_{20} \mathbf{1}^\top \Lambda_2 Z_{\gamma d} Z_\gamma Z_{\beta d} \Lambda_3 \mathbf{1} \\
&- N_{21} \mathbf{1}^\top \Lambda_1 Z_{\gamma d} Z_\gamma Z_{\beta d} \Lambda_3 \mathbf{1} + (2N_{22} + N_{24}) \mathbf{1}^\top \Lambda_3 Z_\gamma Z^{(2)}_\beta \Lambda_3 \mathbf{1} \\
&+ N_{23} \mathbf{1}^\top \Lambda_3 Z_{\beta d} Z_\gamma Z_{\beta d} \Lambda_3 \mathbf{1},
\end{aligned}
$$

where $Z_\beta = \tilde{X}(\tilde{X}^\top \Lambda \tilde{X})^{-1} \tilde{X}^\top$, $Z_\gamma = \tilde{P}(\tilde{P}^\top V \tilde{P})^{-1} \tilde{P}^\top$, $\tilde{P} = \partial\tau/\partial\gamma$, $Z_{\beta d} = \mathrm{diag}\{z_{\beta_{11}}, \ldots, z_{\beta_{nn}}\}$, and $Z_{\gamma d} = \mathrm{diag}\{z_{\gamma_{11}}, \ldots, z_{\gamma_{nn}}\}$ are matrices of dimension $n \times n$. We denote $Z^{(3)}_\gamma = Z^{(2)}_\gamma \odot Z_\gamma$, $Z^{(2)}_\gamma = Z_\gamma \odot Z_\gamma$, where $\odot$ denotes Hadamard product. The matrices Q_1–Q_6, N_1–N_9, N_{18}–N_{24} are of dimension $n \times n$ and will be ommited from the paper to economize space.

3 Corrected Profile Likelihood Ratio Statistic

For models with nuisance parameters is common practice to make inferences based on a profile likelihood, which is a function of the likelihood genuine involving only the parameters of interest. In this direction, we are interested in making inferences on the interest parameters in the HSNLM, where we shall use the results in Cox and Reid (1987) to obtain modified profile likelihood ratio statistic (LR_m) distributed as χ^2_q, under the null hypothesis, up to an error of order n^{-1}. The modified profile likelihood ratio statistics for the test of $H_0 : \gamma = \gamma^{(\mathbf{0})}$ is given by

$$LR_m = -2\{l_{CR}(\gamma^{(\mathbf{0})}) - l_{CR}(\hat{\gamma})\},$$

where

$$l_{CR}(\gamma) = l(\gamma, \hat{\beta}_\gamma) - \frac{1}{2}\log|j_{\beta\beta}(\gamma, \hat{\beta}_\gamma)|, \tag{3}$$

being $j_{\beta\beta}(\gamma, \hat{\beta}_\gamma)$ the observed information matrix corresponding to β when γ is fixed, given by

$$j_{\beta\beta} = -\frac{\partial^2 l(\theta)}{\partial\beta_j\partial\beta_l} = -\sum_{i=1}^{n} t^{(2)}_{(z_i)}\frac{1}{\phi_i}(j,l)_i + \sum_{i=1}^{n} t^{(1)}_{(z_i)}\frac{1}{\sqrt{\phi_i}}(jl)_i,$$

where $t(z_i) = \log g(z_i^2)$, with $z_i = \frac{(y_i-\mu_i)}{\sqrt{\phi_i}}$ and $t^{(k)}_{(z_i)} = \partial^k t(z_i)/\partial z^k$, with $k = 1, 2$. A Bartlett correction factor, proposed by DiCiccio and Stern (1994), also can be incoporated to LR_m in order to improve that approximation. The Bartlett correction leads to an modified statistic, $LR^*_m = LR_m/(1 + c_m)$, distributed as χ^2_q, under the null hypothesis, up to an error of order n^{-2}. Thus, in matrix notation we obtain the Bartlett correction fator c_m to the LR^*_m statistic for the test $H_0 : \gamma = \gamma^{(\mathbf{0})}$ versus $H_1 : \gamma \neq \gamma^{(\mathbf{0})}$ considering multiplicative heteroscedasticity, given by

$$c_m = \frac{1}{4}tr(M_1 H_d^{(2)}) + \frac{1}{4}\mathbf{1}^\top H_d M_4 H M_4 H_d \mathbf{1} + \frac{1}{6}\mathbf{1}^\top M_4 H^{(3)}\mathbf{1}, \tag{4}$$

where $H = \{h_{ij}\} = -\mathbf{Z}[\mathbf{Z}^\top V\mathbf{Z}]^{-1}\mathbf{Z}^\top$, with $\mathbf{Z} = (z_1, \ldots, z_n)^\top$, $H_d = \text{diag}\{h_{11}, \ldots, h_{nn}\}$, $H_d^{(2)} = \text{diag}\{h_{11}^2, \ldots, h_{nn}^2\}$ and $H^{(3)} = (h_{ij})^3$.

4 Numerical Evidence

In Table 1, we present the powers of the tests of the null hypothesis $H_0 : \gamma = \gamma_0$. The values of γ used ranged from 0.1 to 0.7 for the power exponential model (with $k = 0.3$) and 0.1 to 1.0 for Student–t model (with $\nu = 4$). The tests were performed using size-corrected critical values (obtained from the size simulations) in order to force all tests to have the same size. The simulations were carried out using n = 35, $\alpha = 0.10\%$, $q = 3$ and $p = 5$. (All

entries are in percentages.) We note that LR and LR^* tests are less powerful than LR_m and LR_m^*. The numerical results showed that the modified profile likelihood ratio test was better performed that the likelihood ratio test, also the best performing test is the Bartlett-corrected modified profile likelihood ratio test.

TABLE 1. Nonnull rejection rates, inference on γ.

	Student-t ($\nu = 4$)				Power exponential (k=0.3)			
γ	LR	LR^*	LR_m	LR_m^*	LR	LR^*	LR_m	LR_m^*
0.1	8.3	8.5	13.3	13.3	10.9	10.6	10.5	10.6
0.2	10.9	10.5	24.5	24, 6	16.4	16.0	23.2	23.2
0.3	17.9	18.8	43.6	43.6	28.5	28.5	42.2	42.2
0.4	31.0	31.0	63.4	63.4	59.7	59.9	75.1	75.1
0.5	49.2	49.4	79.7	79.7	76.8	76.7	86.8	86.9
0.6	68.5	68.5	91.3	91.3	88.6	89.0	93.8	93.8
0.7	80.7	80.7	95.8	95.8	97.8	97.8	98.9	98.9
0.8	92.6	92.9	99.1	99.1				
0.9	97.1	97.3	99.7	99.7				
1.0	99.4	99.4	99.9	99.9				

Acknowledgments: Special thanks to CNPq and FACEPE, for the financial support.

References

Bartlett, M. S. (1937). Properties of suficiency and statistical tests. *Proceedings of the Royal Society of London A,* **160**, 268 – 282.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society B*, **49**, 1 – 39.

Cysneiros, F. J. A., Cordeiro, G. M. and Cysneiros, A. H. M. A. (2010). Corrected maximum likelihood estimators in heteroscedastic symmetric nonlinear models. *Journal of Statistical Computacional and Simulation*, **80**, 451 – 461.

DiCiccio, T. J. and Stern, S. E. (1994). Frequentist and Bayesian Bartlett correction of test statistics based on adjusted profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological),* **56**, 397 – 408.

Mortality forecasting for related populations using Lee-Carter type models

Ivan Luciano Danesi[1], Steven Haberman[2], Pietro Millossovich[2]

[1] University of Padova, Padova, Italy
[2] Cass Business School, London, England

E-mail for correspondence: danesi@stat.unipd.it

Abstract: Forecasting mortality is not a straightforward issue as the drop in mortality data is not uniform over either age or time. The Lee-Carter model and some of its extensions are applied to the mortality data of Nordic countries. Some of the models assume independence of the populations, while others try to model some common factors in the mortality dynamics. Unlike most approaches in the literature, we model mortality improvement rates rather than the rates themselves. The comparison between the models is performed with respect to predictive capacity.

Keywords: improvement rates; Lee-Carter model; mortality rates.

1 Introduction

The level of mortality rates determines several aspects of our society. As a matter of fact, the private and the public retirement systems, as well as other components of the social security system, are planned and modified according to the values assumed by mortality rates.
The mortality rate $m_{x,t}$ referred to age x and year t can be obtained dividing the number of deaths $D_{x,t}$ by the exposure to risk $\mathrm{ETR}_{x,t}$. The rates $m_{x,t}$ are computed for every age x and year t and are organized into a matrix which has ages on rows and years on columns. A great number of models were proposed for evaluating mortality tables in future years. One of the most influential model is that introduced in Lee and Carter (1992), which has since received great deal of attention and has been extended in several directions.
The aim of this paper is to forecast the mortality rates of mainland Scandinavian countries (Denmark, Norway and Sweden) and Finland, which represent the most of the population of the Nordic countries. These four countries share common traits in their respective societies. For this reason the inhabitants of Nordic countries can be considered as related populations. In Li and Lee (2005) the importance of forecasting mortality in

a coherent way for related population was highlighted and, subsequently, other approaches were proposed.
In Section 2 some Lee-Carter type models are presented, in Section 3 the models are applied to the data and are evaluated with respect to the predictive capacity. In Section 4 there are some concluding remarks.

2 The models

The original formulation of the model presented in Lee and Carter (1992) is

$$\ln m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t}, \quad \sum_x b_x = 1, \ \sum_t k_t = 0.$$

The logarithm of $m_{x,t}$ is specified as a function of a_x, which is the general shape across age of the mortality, a bilinear term $b_x k_t$ plus an error term $\varepsilon_{x,t} \sim N(0, \sigma^2)$. The bilinear term is composed by k_t, an index of the level of mortality across years, and b_x, which describes the level of deviations from the general shape shape a_x in response to variations of k_t. In this formulation the random errors are homoskedastic, which is commonly a strong and often unrealistic hypotesis. To solve this problem, Brouhns et al. (2002) proposed an application of the Lee-Carter model using a Poisson random variable for the number of deaths. In this case the target is the force of mortality $\mu_{x,t}$ (we remind that under some assumptions commonly adopted $\mu_{x,t} = m_{x,t}$). The number of deaths $D_{x,t}$ is described by $D_{x,t} \sim \text{Poisson}\,(\text{ETR}_{x,t}\,\mu_{x,t})$ where

$$\mu_{x,t} = e^{\alpha_x + \beta_x k_t}, \quad \sum_x b_x = 1, \ \sum_t k_t = 0$$

which has the form of the Lee-Carter model, apart from the error term.
Sometimes, in the literature, it is the improvement in mortality rates, rather than the rate itself, which is modelled. An example is Haberman and Renshaw (2012), who consider

$$z_{x,t} = 2\frac{1 - m_{x,t}/m_{x,t-1}}{1 + m_{x,t}/m_{x,t-1}}.$$

The values of $z_{x,t}$ are modelled as realizations of independent Gaussian random variables $Z_{x,t}$ assuming constant dispersion, hence $Z_{x,t} \sim N(\eta_{x,t}, \sigma^2)$. We consider the following first moment predictor structure:

$$\eta_{x,t} = \beta_x k_t, \quad \sum_x \beta_x = 1.$$

The fitting is done by minimising the model deviance, defined as $Dev = \sum_{x,t}(z_{x,t} - \eta_{x,t})^2$. This model can be estimated even assuming variable dispersion, introducing weights ϕ_x. In this case the variance of $Z_{x,t}$ is $\phi_x \sigma^2$

and the squared residuals $r_{x,t} = (z_{x,t} - \eta_{x,t})^2$ are modelled as independent gamma responses introducing a two stage iterative estimating procedure (for further details about the computation procedure of the weights ϕ_x see Haberman and Renshaw (2012)).

3 Application and results

From now on an index $i = 1, \ldots, 4$ is introduced to indicate the following countries: Denmark, Finland, Norway and Sweden. The models listed below are applied to female mortality data for the ages 20-89 and for the years 1965-1994.

LCP. $D^i_{x,t} \sim \text{Poisson}(\text{ETR}^i_{x,t}\,\mu^i_{x,t})$, $\log \mu^i_{x,t} = \alpha^i_x + \beta^i_x k^i_t$, $\sum_x b^i_x = 1$, $\sum_t k^i_t = 0$, forecast of the time varying coefficients k^i_t: independent random walks with drift.

MIR. $Z^i_{x,t} \sim N(\eta^i_{x,t}, \sigma^2_i)$, $\eta^i_{x,t} = \beta^i_x k^i_t$, $\sum_x \beta^i_x = 1$, forecast of the time varying coefficients: four independent AR(1).

MIRM. $Z^i_{x,t} \sim N(\eta^i_{x,t}, \sigma^2_i)$, $\eta^i_{x,t} = \beta^i_x k^i_t$, $\sum_x \beta^i_x = 1$, forecast of the time varying coefficients: VAR(1).

MIR1. $Z^i_{x,t} \sim N(\eta^i_{x,t}, \sigma^2_i)$, $\eta^i_{x,t} = \beta^i_x k_t$, $\sum_{x,i} \beta^i_x = 1$, forecast of the time varying coefficient: AR(1).

MIRϕ. $Z^i_{x,t} \sim N(\eta^i_{x,t}, \phi^i_x \sigma^2_i)$, $\eta^i_{x,t} = \beta^i_x k^i_t$, $\sum_x \beta^i_x = 1$, forecast of the time varying coefficients: four independent AR(1).

MIRMϕ. $Z^i_{x,t} \sim N(\eta^i_{x,t}, \phi^i_x \sigma^2_i)$, $\eta^i_{x,t} = \beta^i_x k^i_t$, $\sum_x \beta^i_x = 1$, forecast of the time varying coefficients: VAR(1).

MIR1ϕ. $Z^i_{x,t} \sim N(\eta^i_{x,t}, \phi^i_x \sigma^2_i)$, $\eta^i_{x,t} = \beta^i_x k_t$, $\sum_{x,i} \beta^i_x = 1$, forecast of the time varying coefficient: AR(1).

Regarding the models where mortality improvement rates are used, the forecast values of the $m^i_{x,t}$, denoted $\hat{m}^i_{x,t}$, are obtained applying iteratively the formula $\hat{m}^i_{x,t} = \hat{m}^i_{x,t-1}(2 - z^i_{x,t})/(2 + z^i_{x,t})$ for $t = 1995, \ldots, 2009$, starting from the values of $m^i_{x,t}$ in $t = 1994$.
The results of the analyses are compared with the actual mortality rates of years 1995-2009 applying the mean absolute percentage error

$$\text{MAPE}_i = \frac{1}{15 \cdot 70} \sum_{x,t} \left| \frac{m^i_{x,t} - \hat{m}^i_{x,t}}{m^i_{x,t}} \right|.$$

The results are summarized in Table 1.

TABLE 1. Values of MAPE of the forecast for Denmark (DK), Finland (FI), Norway (N) and Sweden (SE).

	LCP	MIR	MIRM	MIR1	MIRϕ	MIRMϕ	MIR1ϕ
DK	0.3025	0.2952	0.2951	0.2934	0.2907	0.2915	0.3071
FI	0.5081	0.2637	0.2635	0.2634	0.2511	0.2487	0.2490
N	0.1538	0.2327	0.2325	0.2330	0.2382	0.2378	0.2362
SE	0.1362	0.2373	0.2374	0.2352	0.2151	0.2149	0.2136

4 Discussion

Based on the results in Table 1, some comments follow below, although it should be borne in mind that the considered populations have relative small sizes and hence there is a great variability in the mortality path.
No model performs better than the others, but the model LCP presents the widest range in the quality of the results (from 13.62% to 50.81%). The models work quite generally better for Sweden, which has a population amount that is almost double in size than the other three populations considered. If there are n related populations, it is likely that less than n different paths of general mortality level (k_t^i) should be considered. This can be seen by noting that the quality of the results is not worsened when considering one time varying coefficient (MIR1 with respect to MIR) and could actually improve (MIR1ϕ with respect to MIRϕ).

Acknowledgments: Special thanks, for the important advices, to Prof. Nicola Torelli and Prof. Ermanno Pitacco.

References

Brouhns, N., Denuit, M., and Vermunt, J.K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31(3)**, 373 – 393.

Haberman, S. and Renshaw, A. (2012). Parametric mortality improvement rate modelling and projecting. *Insurance: Mathematics and Economics*, **50(3)**, 309 – 333.

Lee, R.D. and Carter L.R. (1992). Modeling and Forecasting U. S. Mortality. *Journal of the American Statistical Association*, **87(14)**, 659 – 671.

Li, N. and Lee, R. (2005). Coherent Mortality Forecasts for a Group of Populations: An Extension of the Lee-Carter Method. *Demography*, **42(3)**, 575 – 594.

Multilevel factor models: Identification of Three-level Model Parameters for the Study of Regional Development in Argentina

Maria del Pilar Díaz[1], José M. Vargas[23], Margarita Díaz[3]

[1] Biostatistics Unit, Medical Sciences Department, UNC, Córdoba, Argentina
[2] ICBA, UNVM, Villa María, Argentina
[3] Institute of Mathematics and Statistics, UNC, Córdoba, Argentina

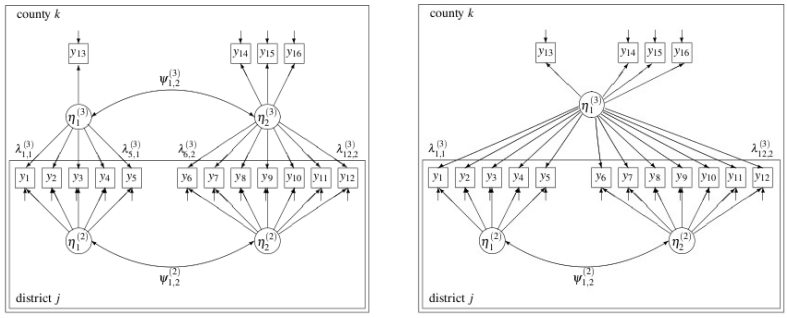
E-mail for correspondence: pdiaz@fcm.unc.edu.ar

Abstract: Based on a theoretical-social model which states that Communication and Information Technologies (CITs) infuence the human development, due to the impact they have on economic growth, this work explores the relationship between the social, economic and technological dimensions or constructs, in a province, Córdoba, of Argentina. Confirmatory common framework (GLLAMMs, Skrondal & Rabe-Hesketh, 2004) was performed considering three level modeling. This approach allowed us to identify some indicators that measure constructs and help characterize the level of development of our region using hierarchical information. Since the estimation method was based on full information maximum likelihood, we devoted special attention to the identifiability of the parameters. Two constructs to describe the socioeconomic and technological (SET) development at the district level were obtained. Due to correlation between the two latent variables at department level was near one, a new common factor model containing only one dimension was appropriate.

Keywords: GLLAMM;latent variable; CFA; socio-technological dimensions; identifiability

1 Motivation and Modeling

The measurement of access and use of ICTs, as well as its dynamics, is indispensable to understand the development of today's information societies and to support adequate design of policies. The present work contributes to this measurement, since it identifies the factors associated with SET development and it studies its regional distribution in Córdoba, Argentina. It examines the dimensions of SET constructs in the nested political divisions of Córdoba, and analyzes the relationship between them. We used confirmatory factor analysis (CFA) to explore the dimensionality of constructs. In the multidimensional case, an important example of a restricted model is that of a complexity one model or independent clusters model, where Λ



(a) First model. (b) Second model.

FIGURE 1. Models considered. Short arrows pointing to boxes y_i correspond to residuals ϵ_i; long arrows from circles $\eta_s^{(l)}$ pointing to boxes y_i correspond to loadings $\lambda_{i,s}^{(l)}$, where l stands for levels 2 or 3.

has enough elements set to zero so that each indicator measures one and only one factor. Such a configuration makes sense if one set of indicators is designed to measure one factor and another one is designed to measure another factor. A single-level factor model would not be appropriate in the present study because our information was hierarchical (municipal districts arranged into departments). Thus, we used a three-level factor model, considering districts as level 2 and departments as level 3, with two latent variables in each one. Twelve indicators measured at the district level were used to define two latent variables at each level. A graphical representation of the model is given in figure 1(a). Latent variables are represented by circles, observed variables by rectangles and arrows connecting circles; rectangles also represent regressions residuals being the short arrows pointing at circles or rectangles. Curved double-headed arrows connecting two variables indicate that they are correlated. It is typically assumed that common and unique factors have multivariate normal distributions. Four additional indicators were included at the department level in order to generate the department constructs, giving a total number of indicators of $I = 16$. The three-level model chosen for our study has the following matrix formulation:

$$\mathbf{y} = \beta + \Lambda^{(2)}\eta^{(2)} + \Lambda^{(3)}\eta^{(3)} + \epsilon,$$

where β is the expectation of the observed variables $\mathbf{y}$ (indicators), $\Lambda^{(2)}$, $\Lambda^{(3)}$ denote the 16×2 factor loadings matrices at second and third level respectively, $\eta^{(2)}$, $\eta^{(3)}$ denote the 2×1 latent factors at second and third level respectively, and ϵ is the error term that is assumed independent of latent factors. Additionally we introduce the following matrices of parameters

$$\Psi^{(l)} = \mathbb{V}[\eta^{(l)}] = \mathbb{E}[\eta^{(l)}\eta^{(l)\prime}],\ l = 2, 3,$$
$$\Theta = \mathbb{V}[\epsilon] = \mathbb{E}[\epsilon\epsilon'],$$

with the usual convention that $\mathbb{E}[\eta^{(l)}] = 0$ and $\mathbb{E}[\epsilon] = 0$. Then the covariance matrix of the observed variables is

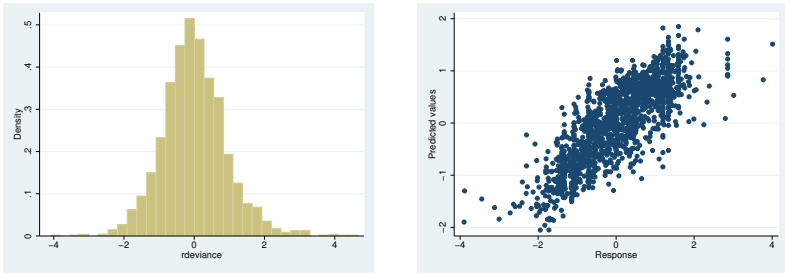
$$\Sigma = \mathbb{V}[\mathbf{y}] = \Lambda^{(2)}\Psi^{(2)}\Lambda^{(2)'} + \Lambda^{(3)}\Psi^{(3)}\Lambda^{(3)'} + \Theta. \tag{1}$$

2 Identification of Three-level Model Parameters

The identification of parameters becomes an issue given the relative large number of them and we understand it as described in Skrondal(2004), p. 135-138, and O'Brien(1994). Bollen(1989) summarizes and extends several rules that establish the identifiability of models; O'Brien(1994) extends those rules even further. These rules apply only to models with a factor complexity of one; that is, models in which each indicator loads only on a single latent variable. Our model being of complexity two, does not satisfies these conditions requiring a new proof specially tailored for the case at hand. Nevertheless, the identification process of parameters is essentially a hierarchical one and this will prove to be enough. We show that parameters of our model can be determined uniquely from information of measured variables. Through algebraic manipulation, we show that if $\Lambda^{(l)}$, $\Psi^{(l)}$ and Θ exist such that the relationship above for Σ holds, then $\Lambda^{(l)}$, $\Psi^{(l)}$ and Θ must be unique. First of all, we notice for further convenience that levels two, three, and latent factors, induce a block partition on the above equation for Σ, $\Lambda^{(2)}\Psi^{(2)}\Lambda^{(2)'}$, $\Lambda^{(3)}\Psi^{(3)}\Lambda^{(3)}$ and Θ have a block partition of shape 3 by 3. We label those blocks $A, D', E', D, B, F', E, F, C'$, from left to right and top to bottom respectively, where A is 5×5, B is 7×7, C is 4×4, D is 7×5, E is 4×5 and F is 4×7. We will refer to those blocks in any of the matrices involved to locate entries that will be of interest to us. We proceed in steps: first we prove that third level parameters can be identified by blocks E, F and C of Σ, that is, all of $\Lambda^{(3)}\Psi^{(3)}\Lambda^{(3)'}$ and block C of Θ can be identified, then we show that all remaining parameters in $\Lambda^{(2)}\Psi^{(2)}\Lambda^{(2)'}$ and blocks A and B of Θ can be identified. Proceeding in this way, we show that the identification of parameters in the first model proposed with four latent factors, $\eta_1^{(2)}$, $\eta_2^{(2)}$, $\eta_1^{(3)}$ and $\eta_2^{(3)}$ is guaranteed; the second model proposed with one factor in the third level, can be shown to be identified by similar algebraic manipulations.

3 Results

The performance of model was suitable. Figure 2 illustrates the behavior of deviance residuals and predicted values versus indicators. For the first one we also observed that percentiles 5, 50 and 95 were respectively -1.4113, -0.0492 and 1.5953, which confirm appropriate representation of our model. Our results indicated that two constructs were significant to



(a) Deviance residuals. (b) Predicted values vs indicators.

FIGURE 2. Behavior of deviance residuals and predicted values versus indicators

describe the socio-economic and technological development at the district level ($\hat{\rho}^{(2)} = 0.55$. All the indicators included in the common factor model were significant ($p < 0.05$) at this lower level. However, at the department level, the correlation between the two latent variables was near one ($\hat{\rho}^{(3)} = 0.88$), suggesting a new common factor model containing only one dimension at this level (figure 1(b)). The percentage of households with telephone was the most important socio-economic indicator at the district level, while the percentage of households with health insurance coverage the least important. Also at this level, the most important technological indicator was the percentage of the population with secondary-level education or beyond. When the new common factor model, with only a latent variable at the department level, was fitted similar factor loading estimates were obtained at both levels. We have chosen this model even though a negligible change was obtained for AIC statistic. All the socio-economic items were significant, except for percentage of homeowner at a higher level, whereas percentage of municipalities with web-site was the only non significant technological item. In addition, a correlation equal to 0.58 was observed at the district level, confirming two constructs for it.

References

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

O'Brien, R. M. (1994). *Identification of Simple Measurement Models with Multiple Latent Variables and Correlated Errors*. In: *Sociological Methodology*, Vol. 24.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: multilevel, longitudinal, and structural equation models.*. Chapman & Hall/CRC.

Assessing the impact of non-additive noise on modelling transcriptional regulation with Gaussian processes.

Vinny Davies[1], Dirk Husmeier[1]

[1] School of Mathematics and Statistics, University of Glasgow, Scotland

E-mail for correspondence: v.davies.1@research.gla.ac.uk

Abstract: In transcriptional regulation, transcription factors (TFs) are often unobservable at mRNA level or may be controlled outside of the system being modelled. Gaussian processes are a promising approach for dealing with these difficulties as a prior distribution can be defined over the latent TF activity profiles and the posterior distribution inferred from the observed expression levels of potential target genes. However previous approaches have been based on the assumption of additive Gaussian noise to maintain analytical tractability. We investigate the influence of a more realistic form of noise on a biologically accurate system based on Michaelis-Menten kinetics.

Keywords: Transcriptional regulation, Gaussian processes, additive and multiplicative noise, Michaelis-Menten kinetics

1 Introduction

A particular challenge in the quantitative modelling of transcriptional regulation is that transcription factors (TFs), the regulatory proteins at the heart of the process, are frequently subject to post-translational modification, which may affect their DNA binding capability. Consequently, gene expression levels of TFs contain only limited information about their actual activities. A promising approach to deal with these difficulties was proposed in Gao et al. (2008), inspired by the work of Barenco et al. (2006). The authors advocate the use of Gaussian processes to define prior distributions over the latent TF activity profiles. Inference is soundly based on the principles of non-parametric Bayesian statistics, consistently inferring the posterior distribution of the unknown TF activities from the observed expression levels of potential target genes, and inferring regulatory network structures after marginalizing over the unknown TF activity profiles.
The choice of a non-parametric prior distribution from the Gaussian process family is not a restrictive modelling assumption. Somewhat more restrictive is the assumption of additive Gaussian noise, which can be found in all

previous applications (Gao et al. (2008), Honkela et al. (2010), etc.). Previous work by Rocke and Durbin (2001) showed that mRNA concentrations obtained from microarray experiments are of a more complex form and the purpose of this work is to investigate what effect this deviation from additive Gaussianity has on the inference in transcriptional regulation.

2 Method

A linear model of gene expression was proposed by Barenco et al. (2006)

$$\frac{dx_i(t)}{dt} = B_i + S_i f(t) - D_i x_i(t) \tag{1}$$

where $i \in \{1, \ldots, G\}$ is a set of genes regulated by the same TF, $x_i(t)$ are the (unknown) true gene expression levels at time point t, $f(t)$ is the (unknown) TF activity, B_i is the basal transcription rate of gene i, S_i is the sensitivity to binding of TF, and D_i is a decay rate. We assume that (noisy) measurements of $x_i(t)$ can be obtained, however TF activity is unknown and therefore $f(t)$ is assumed to be unobservable.
Eq. (1) has the analytical solution:

$$x_i(t) = \frac{B_i}{D_i} + S_i \int_0^t \exp(-D_i(t-u)) f(u) du. \tag{2}$$

Gao et al. (2008) proposed a non-parametric Bayesian approach to inference in this model by placing a Gaussian process prior with a squared exponential covariance matrix on the unknown TF activities $\mathbf{f} = (f(t_1), \ldots, f(t_T))$ at timepoints $\mathbf{t} = (t_1, \ldots, t_T)$. The linear form of the model implies that the joint prior distribution of the expression profiles of all regulated genes, $\mathbf{x}_i$, is described by a Gaussian process prior with a covariance matrix, $\mathbf{K}$, that depends on the hyperparameters of the prior, θ_h, as well as the parameters that characterise the transcriptional regulation processes via eq. (2):

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}') &= \mathcal{N}(\mathbf{B}./\mathbf{D}, \mathbf{K}); \quad \mathbf{K} = \mathbf{K}(\boldsymbol{\theta}') \\ \boldsymbol{\theta}' &= (\theta_h, B_1, \ldots, B_G, S_1, \ldots, B_G, D_1, \ldots, D_G) \end{aligned} \tag{3}$$

where $B./D$ is a point-wise vector division. See Davies and Husmeier (2013) for details.
To relate the unknown true gene expression profiles $\mathbf{x}_i = (x_i(t_1), \ldots, x_i(t_T))$ to noisy measurements $\mathbf{y}_i = (y_i(t_1), \ldots, y_i(t_T))$, Gao et al. (2008) assumed additive Gaussian noise of constant variance σ^2. The marginalisation over $\mathbf{y}$ is analytically tractable and gives:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I}) \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{K}(\boldsymbol{\theta}')) d\mathbf{x} = \mathcal{N}(\mathbf{y}|\mathbf{B}./\mathbf{D}, \mathbf{K}(\boldsymbol{\theta}') + \sigma^2\mathbf{I}) \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}', \sigma^2)$. Inference of the parameters $\boldsymbol{\theta}$ can then be achieved in a maximum likelihood or Bayesian framework; see Bishop (2006).

However Rocke and Durbin (2001) showed that the noise in transcriptional profiling with microarrays has the following more general form:

$$y_i(t) = c + x_i(t)\exp(\epsilon_\mu) + \epsilon_t \quad \text{where} \quad \epsilon_j \sim \mathcal{N}(0, \sigma_j^2) \tag{5}$$

where c is mean background noise, and σ_μ^2 and σ_t^2 are unknown variance parameters. Replacing $\mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I})$ in eq. (4) by the noise in eq. (5) does not give a closed-form solution, and this has therefore been ignored in previous work. The objective of the present study is to quantify the effect the deviation from additive Gaussianity has on the inference of the transcriptional regulation.

3 Data

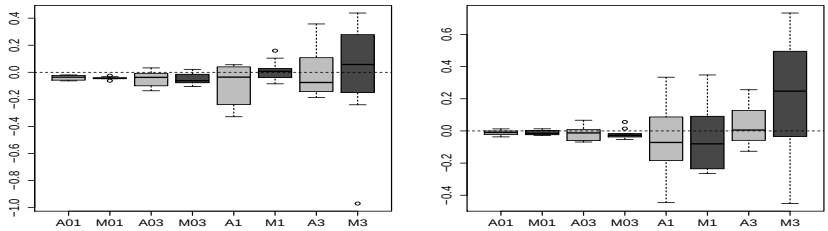
We combined a simple regulatory network for three genes with a protein signalling pathway from Vyshemirsky and Girolami (2008); see Davies and Husmeier (2013) for details. The active form of the TF is unobservable due to post-translational modification, and the processes leading to the formation of active TF is controlled outside of the subsystem being modelled. The transcriptional profiles of the downstream genes were generated by solving eq. (2) with the different kinetic parameters. 18 values from these expression profiles were then subjected to either additive Gaussian noise, or the more complex noise of eq. (5). For the non-Gaussian noise the standard deviations were chosen on a roughly log scale such that $\sigma_\mu, \sigma_t = (0.01, 0.03, 0.1, 0.3)$, with equivalent values chosen for the additive noise model to allow for a fair comparison. This was repeated 10 times for each standard deviation size and noise model.

4 Results

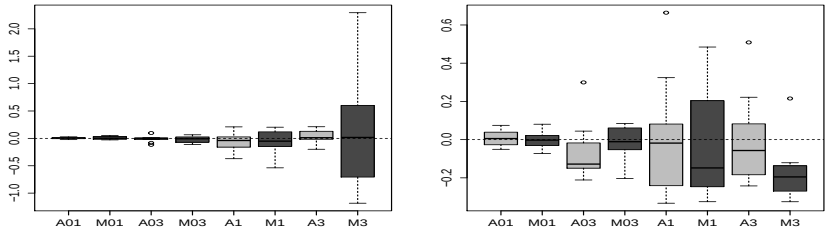
For a relatively small data set, our results, given in Figure 1, have shown that the deviation from additive Gaussian noise has little negative effect when $\sigma_\mu, \sigma_t = (0.01, 0.03)$. For larger standard deviations the results show a consistent deterioration in the case of non-Gaussian noise, although this cannot be easily quantified until $\sigma_\mu, \sigma_t = (0.3)$. For this level of variance, Figure 1, as well as similar results for the kinetic parameter estimates, show a roughly four fold increase in the median error.

5 Conclusion

Our work has considered the implications of having non-Gaussian noise when using Gaussian processes for modelling transcriptional regulation. This noise model violates some of the modelling assumptions and causes a deterioration in the ability of the model to perform parameter inference. We have shown that the effect of this noise is not as significant as first assumed and the negative effect only becomes apparent for larger variances.



(a) Gene Profile, x_1 (b) Gene Profile, x_2



(a) Gene Profile, x_3 (b) Gene Profile, f

FIGURE 1. Box plots of the error of gene and TF profile predictions. Box plots for the additive Gaussian, 'A', and non-Gaussian, 'M', noise are given in light and dark grey respectively. The standard deviations used for σ_μ and σ_t are given under each box plot and represent the values (0.01,0.03,0.1,0.3)

References

Barenco, M. et al. (2006). Ranked prediction of p53 targets using hidden variable dynamic modelling. *Genome Biology*, **7(3)**, R25.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Davies, V. and Husmeier, D. (2013). Modelling transcriptional regulation with Gaussian processes. Technical Report. University of Glasgow *www.maths.gla.ac.uk/∼dhusmeier/MyPapers/bookChapterVinny.pdf*

Gao, P. et al. (2008). Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, **24**, 70 – 75.

Honkela, A. et al. (2010). Model-based method for transcription factor target identification with limited data. *PNAS*, **107(17)**, 7793 – 7798.

Rocke, D.M. and Durbin, B. (2001). A model for measurement error of gene expression analysis. *J. Comput. Biol.*, **8**, 557 – 569

Vyshemirsky, V. and Girolami, M.A. (2008). Bayesian ranking of biochemical system models. *Bioinformatics*, **24(6)**, 833 – 839.

Model Comparisons for RNA-Seq data

Margaret R. Donald[1], Ashwin Unnikrishnan[1], John E. Pimanda[1], Susan R. Wilson[12]

[1] University of New South Wales, Australia
[2] Australian National University, Australia

E-mail for correspondence: Margaret.Donald@unsw.edu.au

Abstract: Various strategies for finding differentially expressed genes in high-throughput genomic studies have been proposed. Using a set of paired patient data, we have explored some of the many possible models. Results differ dependent on whether the data are normalised or not, and on the method of normalisation selected. Models that allow for over-dispersion fitted the data better. There was little consistency in the declaration of differentially expressed genes between the various approaches.

Keywords: paired data; negative binomial; false discovery rate (FDR); differentially expressed genes (DE genes).

1 Introduction

RNA-Seq technology (Mortazavi et al, 2008) is a recent development to measure gene expression that is considered more accurate than microarray measurements, and essentially is in the form of counts for the genes. Here we consider 15 patients, eight suffering from Myelodysplastic Syndrome (MDS) and the other seven from Chronic Myelomonocytic Leukaemia (CMML), who were all treated with the DNA hypomethylating drug, Vidaza (AZA). Patients who showed a favourable clinical response to the treatment were classified as responders, while the rest were classified as non-responders. RNA-Seq data were obtained both before and after 6 cycles of AZA treatment. In all, 35868 genes or gene variants were obtained that reduced to 16862 after filtering to exclude very low counts. Here the focus is on finding differentially expressed genes before and after treatment dependent on whether the patient is a responder or not.
RNA-seq data count data are generally modelled as either from a Poisson distribution or from a negative binomial distribution, which can be derived as a hierarchical Poisson-gamma distribution. The negative binomial allows modelling of 'overdispersion' and hence is a favoured option for modelling these data. The variance of the negative binomial is expressed as $\sigma^2 = \mu + \phi\mu^2$, where μ is the mean, and ϕ is the dispersion parameter.

Many packages are available for finding differentially expressed genes, and we compare results from the packages EdgeR (Robinson et al, 2010), and DESeq (Anders and Huber, 2010). Further, using the total deviance of fitted SAS models, we compare results from Poisson and negative binomial models fitted using maximum likelihood.
For assessing the significance of terms modelling the mean in the SAS models, likelihood ratio tests were used. Bonferroni adjustments and locfdr (Efron et al, 2011) were used to determine differentially expressed genes with an FDR of 5% for the SAS models. Both DESeq & edgeR use Benjamini-Hochberg (1995) adjustments. Shared DE genes between models were computed using both FDR adjustments and α of 0.05. These adjustments do not change the ordering of the gene by gene p-values. Hence, we also found the first 100 most differentially expressed genes for each method and compared the number of shared DE genes per method.
Normalised were compared with unnormalised results. The SAS models used normalisations of the median count, the total count, the upper quartile, the PoissonSeq normalisation (Li 2012) and the TMM of edgeR.
Dispersion estimates in SAS are the MLE estimates from the model. We have used the default normalisation for edgeR, TMM, and for DESeq, the median of the ratios of observed counts (Anders & Huber, 2010).
The data are paired data, and the pairing is accounted for in the analysis to allow more efficient testing by removing some of the heterogeneity associated with each patient. This is done by fitting nuisance parameters, the patient specific effects, p_i ($i = 1...15$). The model to describe the mean count for each gene consisted of a patient specific intercept, a time effect (t), and a responder by time interaction effect (tr). Thus, if the count for patient i, at time t (t=0,1), associated with patient i, a responder r (r=0,1) is $y_{i(r)t}$, the model for the mean is given as $\mu = \exp[M_{it}(p_i + \alpha t + \beta tr)]$, where M_{it} is a normalisation factor or 'library size' for the RNA sequencing for patient i at time t. (When total counts differ markedly, an adjustment needs to be made using an 'offset' or normalisation variable.)

2 Results & Conclusions

Total deviances across the gene by gene analyses for the various SAS analyses showed that for these data, negative binomial models were required, and that, not surprisingly, when most genes are 'null', total deviances were not useful for choosing between normalisations.
Tables (not shown) of shared genes across models indicated reasonable proportions of shared genes across the SAS models, which typically showed a five to ten fold increase in the number of 'DE' genes in comparison with edgeR and DESeq. The non-paired SAS model, however, had fewer shared DE genes with both the other SAS models and DESeq and edgeR. This shows that despite the apparently non-significant deviance difference, pairing significantly affects gene probability orderings. Most of the edgeR DE

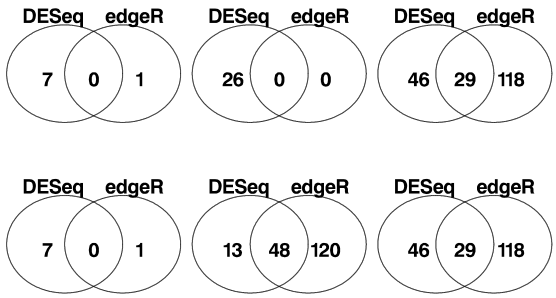


FIGURE 1. Shared DE genes between DESeq & edgeR: upper - from the correlated parameter ('crossed') estimates and the joint testing; lower - from the uncorrelated parameter ('within') estimates.

genes were shared by the SAS models. The DE genes picked by the Bonferroni adjustments compared with those picked from the Benjamini-Hochberg adjustments showed the same pattern. Given SAS' ten-fold difference in the number of DE declared genes, we would have expected all the DE genes of DESeq to be found in the SAS models, but just two-thirds of the DESeq DE genes were found by the SAS MLE models. This seems to indicate a problem with DESeq. Tables showing the sharing across the first 100 genes showed similar patterns.

Parameter estimates for t & tr were highly correlated across the 16862 gene model fits ($\hat{\rho} = -0.78$, SD=.012). Reparameterising the model so that each time term was expressed as time within responder gave estimates which were uncorrelated. The correlated parameterisation masked many DE genes found by the uncorrelated parameterisation. Joint testing (available in both DESeq & edgeR) resolves parameterisation differences (Figure 1). However, there were few shared DE genes between these two packages.

With 30 observations per gene and 13 residual degrees of freedom in the model, MLE should be a reasonable method for finding DE genes. Our results showed that pairing/non-pairing & choice of normalisation can make considerable differences in the gene by gene ordering of p-values in MLE estimation. As in all statistical modelling, the choice of model is fundamental to the conclusions. Where several effects are of interest, joint testing should be undertaken. Simulations under the null for these more complex models need to be undertaken.

Acknowledgments: This research is supported in part by the National Health & Medical Research Council Grant 525453, the NHMRC and the Leukaemia Foundation. AZA was provided by Celgene under a compassionate access scheme.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data, *Genome Biology*, **11** R106.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289 – 300.

Efron, B., Turnbull, B.B. and Narasimhan, B.B. (2011). locfdr: Computes local false discovery rates, (R package version 1.1-7).

Li, J. (2012). PoissonSeq: Significance analysis of sequencing data based on a Poisson log linear model, (R package version 1.1.2).

Mortazavi,A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature Methods*, **5**(7), 621 – 628.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010). EdgeR: a bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, **26**(1), 139 – 140.

SAS Institute (2010), SAS Version 9.2.2, SAS Institute Inc., Cary, NC., USA.

Marginal models from exponential family mixed models with nonormal random effect distribution

Lizandra C. Fabio[1,2], Francisco José A. Cysneiros[2], Gilberto A. Paula[3]

[1] Departamento de Estatística, Universidade Federal da Bahia, Brazil,
[2] Departamento de Estatística, Universidade Federal de Pernambuco, Brazil
[3] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

E-mail for correspondence: cysneiros@de.ufpe.br

Abstract: In this paper we present a class of exponential family (EF) mixed models in which the random intercept follows a generalized log-gamma (GLG) distribution. For specific hierarchical models and particular parameter settings for the random intercept distribution, marginal models are derived in closed-form. An application with real data is given for illustration.

Keywords: Inverted Dirichlet distribution; Multivariate negative binomial distribution; Binomial-GLG distribution.

1 Introduction

Fabio et al. (2012) proposed the random intercept Poisson generalized log-gamma (Poisson-GLG) model for accommodating overdispersion and capturing skew forms for the random intercept distribution. For a particular parameter setting of the GLG distribution the multivariate negative binomial distribution is derived as a marginal model. In this paper we propose the random intercept exponential family generalized log-gamma (EF-GLG) models for which we derive in closed-form two marginal models by specifying the hierarchical model and the parameter setting of the random intercept distribution.

1.1 Generalized log-gamma distribution

Let y be a random variable following a generalized log-gamma distribution. The probability density function (pdf) of y is given by

$$f(y;\mu,\sigma,\lambda)=\begin{cases}\frac{c(\lambda)}{\sigma}\exp\left[\frac{(y-\mu)}{\lambda\sigma}-\frac{1}{\lambda^2}\exp\left\{\frac{\lambda(y-\mu)}{\sigma}\right\}\right], & \text{if} \quad \lambda\neq 0,\\ \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, & \text{if} \quad \lambda=0,\end{cases} \tag{1}$$

where $y \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$ and $\lambda \in \mathbb{R}$ are, respectively, the location, scale and shape parameters and $c(\lambda) = \frac{|\lambda|}{\Gamma(\lambda^{-2})}(\lambda^{-2})^{\lambda^{-2}}$ with $\Gamma(\cdot)$ being the gamma function. We will denote $y \sim \text{GLG}(\mu, \sigma, \lambda)$. The extreme value distribution is a particular case of (1) when $\lambda = 1$. For $\lambda < 0$ the pdf of y is skew to the right and for $\lambda > 0$ it is skew to the left.

1.2 The random intercept EF-GLG model

Let y_{ij} denote the jth outcome measured for the ith cluster (subject), $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. We will assume the following random intercept EF-GLG model:

(i) $y_{ij}|b_i \overset{\text{ind}}{\sim} \text{EF}(u_{ij}, \phi)$,

(ii) $g(u_{ij}) = \eta_{ij} + b_i$ and

(iii) $b_i \overset{\text{iid}}{\sim} \text{GLG}(0, \lambda, \lambda), \quad \lambda > 0,$

where $g(.)$ is the link function, $\eta_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}$ is the linear predictor with $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^{\top}$ contains values of explanatory variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\top}$ is the parameter vector of the systematic component and ϕ^{-1} is the dispersion parameter. Let $f_{Y|b}(y_{ij}|b_i; \boldsymbol{\beta})$ and $f_b(b_i; \lambda, \lambda)$ be the probability mass (or density) function (pmf or pdf) of $y_{ij}|b_i$ and pdf of b_i, respectively. Then, the marginal pmf (or pdf) of $\boldsymbol{y} = (\boldsymbol{y}_1^{\top}, \ldots, \boldsymbol{y}_n^{\top})^{\top}$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})^{\top}$, is given by

$$f_Y(\boldsymbol{y}_i; \boldsymbol{\beta}, \lambda) = \int_{-\infty}^{+\infty} \prod_{j=1}^{m_i} f_{Y|b}(y_{ij}|b_i; \boldsymbol{\beta}) f_b(b_i; \lambda, \lambda) db_i. \tag{2}$$

In the next section we will derive marginal models from (1) by specifying the hierarchical model. We have used the NLMIXED procedure available in the SAS and R software to perform the maximization of the approximate log-likelihood function.

1.3 Marginal models from the random intercept EF-GLG model

(a) Multivariate negative binomial (MNB) model

The MNB distribution was derived from the random intercept Poisson-GLG model (see Fabio et al. (2012)). The marginal pmf of $\boldsymbol{y}_i \sim \text{MNB}(\mu_{\text{ij}}, \phi)$ is given by

$$f_Y(\boldsymbol{y}_i; \boldsymbol{\beta}, \phi) = \frac{\Gamma(\phi + y_{i+})\phi^{\phi}}{(\prod_{j=1}^{m_i} y_{ij}!)\Gamma(\phi)} \frac{\exp(\sum_{j=1}^{m_i} y_{ij} \log \mu_{ij})}{(\phi + \mu_{i+})^{\phi + y_{i+}}},$$

in that $\phi = \lambda^{-2}$, $y_{i+} = \sum_{j=1}^{m_i} y_{ij}$ and $\mu_{i+} = \sum_{j=1}^{m_i} \mu_{ij}$ for $y_{ij} = 0, 1, \ldots$, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Others marginal models derived from the random intercept EF-GLG class are presented in the sequel.

(b) Inverted Dirichlet distribution

We obtain this multivariate distribution from the random intercept Gamma-GLG model. In this case, (i) $y_{ij}|b_i \overset{\text{ind}}{\sim} \text{Gamma}(u_{ij}, \phi)$ and (ii) $\log(u_{ij}^{-1}) = \eta_{ij} + b_i$ in the hierarchical model. After some algebraic manipulation the marginal pdf of $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})^\top$ takes the form

$$f(\boldsymbol{y}_i; \beta, \phi) = \frac{\Gamma[\phi(m_i+1)]}{\Gamma(\phi)^{(m_i+1)}} \frac{\prod_{j=1}^{m_i} (\mu_{ij} y_{ij})^{\phi-1} \mu_{ij}}{\left(1 + \sum_{j=1}^{m_i} \mu_{ij} y_{ij}\right)^{\phi(m_i+1)}}, \tag{3}$$

where $\phi = \lambda^{-2}$, $\mu_{ij} = \exp(\eta_{ij})$ and $y_{ij} > 0$. Making the transformation $z_{ij} = \mu_{ij} y_{ij}$, the marginal pdf in (3) reduces to the standard inverted Dirichlet distribution (see, for example, Kotz et al. (2000)).

(c) Binomial-GLG distribution

This multivariate distribution is derived from the random intercept Binomial-GLG model. In this case, (i) $y_{ij}|b_i \overset{\text{ind}}{\sim} \text{Binomial}(m_i, u_{ij})$ and (ii) $\log\{-\log(1-u_{ij})\} = \eta_{ij} + b_i$. Algebraic manipulation (see, for example, Fog (2008)) leads to the following marginal pmf for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})^\top$:

$$\begin{aligned} f(\boldsymbol{y}_i; \beta, \phi) \quad = \quad & (-1)^k \left\{ \prod_{j=1}^{m_i} \binom{m_i}{m_i y_{ij}} \right\} \sum_{k=0}^{m_i y_{ij}} \cdots \sum_{k=0}^{m_i y_{ij}} \left\{ \prod_{j=1}^{m_i} \binom{m_i y_{ij}}{k} \right\} \\ & \times \left(\phi^{-1} k \mu_{i+} + \phi^{-1} m_i \mu_{i+} + \phi^{-1} m_i \sum_{k=0}^{m_i} y_{ij} \mu_{ij} + 1 \right)^{-\phi}, \end{aligned} \tag{4}$$

where $\phi = \lambda^{-2}$, for $y_{ij} = 0, 1, \ldots$, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Closed-form expressions (omitted here) for the marginal distributions from the random intercept Normal-GLG and Inverse-Gaussian-GLG models were also obtained.

2 Application

We will present a comparative study among diabetic groups discussed by Cysneiros and Paula (2004). The group1 (control), group2 (diabetic without complications) and group3 (diabetic with hypertension) were considered. For each patient the response was a physical task measured at the times 1,..., 6, 8 and 10 min. Let y_{ijk} be the observed physical task for the *kth* patient of the *ith* group at the time j. Figure 1(d) shows us evidence that the variability of the random intercept in relation to its average is skew to the left ($\lambda > 0$). We will assume the following random

intercept EF-GLG model: (i) $y_{ijk}|b_i \overset{\text{ind}}{\sim} \text{Gamma}(u_{ij}, \phi)$ and (ii) $\log(u_{ij}^{-1}) = \mu+\alpha_i+b_i$, for $i = 1, 2, 3$ and $\alpha_1 = 0$, which reduces for $\lambda > 0$ to the inverted Dirichlet distribution (3). The parameter estimates (approximate s.e.) are: $\hat{\alpha}_2 = -0.2998(0.1648)$, $\hat{\alpha}_3 = 0.6773(0.1641)$ and $\hat{\phi} = 12.4504(1.3387)$. From the fitted estimates there are indications that the group2 has a physical task mean larger than the group1, whereas the diabetic group3 has a smaller physical task mean. These results are in accordance with the profiles presented in Figures 1(a)-(c). The Figure 1(e) give us evidence that the model is well fitted.

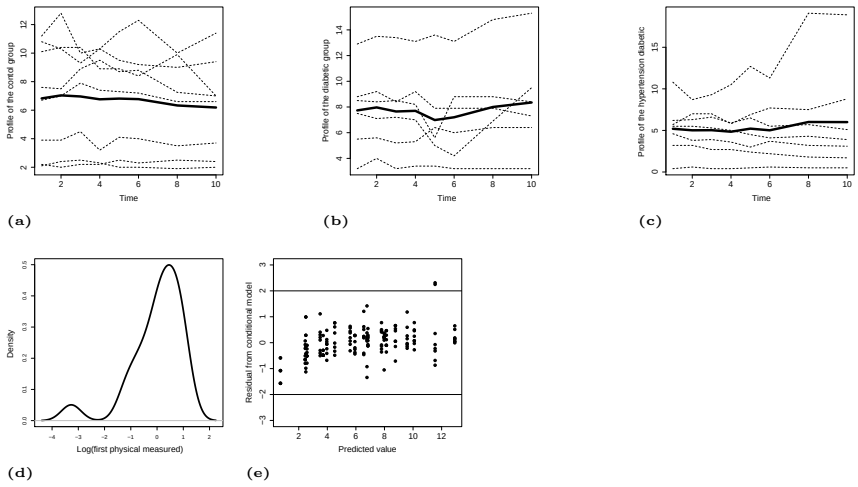


FIGURE 1. Profile of the three groups ((a): group1, (b): group2 and (c): group3), the density of the first physical task measure(d) and conditional residual plot (e).

Acknowledgments: The authors are grateful to CNPq and FACEPE, Brazil, for the financial support.

References

Fabio, L.C., Paula, G.A and de Castro, M. (2012). A Poisson mixed model with nonormal random effect distribution. *Computational Statistics Data Analysis*, **56**, 1499 – 1510.

Fog, A. (2008). Calculation methods for Wallenius noncentral hypergeometric distribution. *Communications in Statistics, Simulation and Computation*, **37**, 258 – 273.

Cysneiros, F.J.A. and Paula, G. (2004). One-sided tests in linear models

with multivariate t-distribution. *Communications in Statistics, Simulation and Computation*, **33**, 747 – 771.

Kotz, S. and Balakrishnan, N. and Johnson, N.L. (2000). *Continuous Multivariate Distributions*. Canada: John Wiley & Sons.

Sparse Penalised Methods in Phenology

Zhou Fang[1]

[1] Biomathematics and Statistics Scotland, Edinburgh, UK

E-mail for correspondence: zfang@bioss.ac.uk

Abstract: We are often interested in relating daily temperature records to the timing of key events, such as flowering times and appearance of certain insect species. When biologically motivated models are unavailable or difficult to estimate, regression techniques may provide an useful alternative. We propose a methodology based on the Lasso, using L1 penalisation of a hockey stick basis, as well as an extension using a two-stage procedure to enhance results. We conduct an empirical simulation illustrating the effectiveness of this proposal relative to an existing method.

Keywords: Phenology; Functional Analysis; Lasso.

1 Introduction

Many biological or ecological events have a connection to environmental variables, such as the temperature record, and we might often wish to estimate this connection. Cases where the underlying biological processes are well understood lead to ideas like the spring warming and sequential models for flowering times, and fitting might proceed by the estimation of parameters in these models.

However, it can be helpful to instead consider regression based techniques, where we assume a functional relationship between the timing of the event, and the measured temperature over a period. Model fitting proceeds by the estimation of this relationship. To simplify computation and avoid overfitting, we might constrain the relationship of the event time and the temperature at each time point to be linear, deriving a functional linear model:

$$Y_i = \sum_{t=A_i}^{B_i} X_t f(t - A_i) + \varepsilon_i.$$

Here, Y_i is the time of the i-th event, X_t corresponds to the temperature at time t, A_i and B_i denote start and end times for the temperature record considered for each event, while ε is a random error. For simplicity, we usually assume that ε_i is i.i.d. Normal. A_i and B_i are usually provided a-priori, though if $f(t) = 0$, those portions of the temperature record would have no relevance to the prediction.

The advantages of regression methods are that they can be easy to use, are flexible, and in comparison to biologically based models, involving many parameters with non-linear effects, may be applied effectively to smaller datasets. They are also useful when we lack prior knowledge of the mechanisms involved, which is helpful for less studied events such as insect arrival. However, the estimation of f is challenging. One avenue, here, that has not been often considered is the application of sparse regression techniques (such as the Lasso). In this work, we develop some novel methods and compare them to existing methods in simulated datasets.

2 Methods

Previously suggested methods for regression in phenology include Stepwise Regression and Penalised Signal Regression (PSR), as well as Fusion. In Stepwise Regression, we aggregate the X_t to weekly or monthly totals, and use these as potential terms in a multiple linear regression, selecting the terms using standard ANOVA techniques. In PSR, we conduct a regression procedure penalising squared differences in slope between neighbouring time points. Using the P-spline signal regression variant of Marx et al (1999), we end up with a smoothly varying estimate of f, which can often perform well. Roberts (2012) also suggested a strategy using Fusion. This is done by conducting a penalised regression penalising the sum of the absolute value of the differences of coefficients between adjacent time points. The result of this methodology is that it produces an estimate of f that consists of several constant steps, with discontinuities between them. Our approach is related, but has several distinct differences.

We propose to begin by transforming the problem, and conducting a L1 penalised maximum likelihood procedure:

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \sum_{t=A_i}^{B_i} X_t Z(t - A_i)\beta \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

Z is a basis expansion matrix for f. In our case, the columns are centred versions of hockey sticks —thus, for some grid of knot points, t_j, $j = 1, \ldots, p$, which in our experiments we have chosen to mark one week intervals, we have $Z^{(j)}(s) = (s - t_j)_+ - C$.

The $\widehat{\beta}$ thus computed provides an estimate of f, as $\widehat{f}(t) = Z(t)\widehat{\beta}$.

However, a multi-stage procedure can be advantageous. Here, we compute weights w_j by aggregating $|\widehat{\beta}|$ over an appropriately chosen epoch their corresponding t_j belongs to, (such as a season, or a year), so that if $F(t)$ returns, say, the season index of t,

$$w_j = \sum_{k=1}^{p} |\widehat{\beta}_k| I\left(F(t_j) = F(t_k)\right).$$

Then we may compute a second stage analysis as

$$\widetilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \sum_{t=A_i}^{B_i} X_t Z(t - A_i)\beta \right)^2 + \mu \sum_{j=1}^{p} |\beta_j| / w_j,$$

forming $\widetilde{f}(t) = Z(t)\widetilde{\beta}$.

Both of these optimisations are Lasso calculations and can be easily and quickly carried out by existing methods, such as the coordinate descent algorithm of Friedman et al (2007). Tuning parameters μ and λ are chosen by separate cross validations at each stage.

The penalty creates sparsity amongst the coefficients β, which means that the calculation attempts to find appropriate knots for a piecewise linear estimate of f. This provides an advantage over PSR if there is not many years of data, and relatively long periods of potential interest, as defined by $|B - A|$. Producing piecewise linear estimates, instead of the step function estimates of the Fusion procedure, enforces continuity.

The use of the weights in the two stage procedure allows for a potential additional improvement, by helping better filter the temperature record for time points that do not appear to have an effect on Y. This can be significant if $|B - A|$ is excessively long. More detail is given in the poster.

3 Simulations

We illustrate the method by simulating datasets of flowering times. To generate data for this, we opted for a formulation based on that in Roberts (2012). For a 'spring warming' model, this is specified by a start day d_i, a threshold temperature T, and target number of degree days F. Under this model, Y_i is the minimal y such that

$$\sum_{t=d_i}^{y} \max(X_t - T, 0) > F,$$

with an added Normal random noise component. The temperature data was generated by the sum of a seasonal component and AR2 noise.

We use a training set of 30 simulated years, with $[A_i, B_i]$ for each year comprising the first 180 days of the year of flowering plus two preceding years. To be effective, the algorithm must be able to pick out periods of the temperature record that have no effect on the flowering time.

To evaluate performance, we generate an additional test set of 1000 years (this time without noise), and measure the performance on predicting flowering days on this new additional set of records through mean squared error. We also calculate an 'ideal' regression curve, by fitting a PSR estimate to a much larger set of generated observations. We can then compare the estimates we obtain to this curve.

In the poster, we also consider a more complex 'sequential model', which incorporates a requirement that a cool Autumn or Winter precedes the Spring warming, as well discuss applicability to Aphid arrival times.

4 Results

In example fits, which we present in the poster, we see that the Lasso methods choose a small number of knot points, enabling a simple piecewise linear function to be fitted. The reweighting can be effective in reducing the erroneous fits made in areas that have no relationship to the event time. Over 100 replications, we find that the Lasso procedures generally perform well, attaining some improvements in terms of MSE (Table 1). The reweighting technique can be effective, so long as the relevant sections of the temperature records take place over a short period of time relative to the whole of the temperature records considered.

TABLE 1. Prediction MSE over 100 replications (With std deviation in brackets)

	Ideal	PSR	Lasso	2-Stage Lasso
Average	2.04 (0.05)	4.06 (1.12)	3.58 (0.84)	3.10 (0.67)
% PSR	52.9 (10.3)	100.0 (0.0)	91.0 (21.7)	78.7 (15.1)

5 Conclusions

Sparse regression techniques can provide useful results in phenology, especially when used together with appropriately chosen basis expansions. The extended 2-stage methodology can be effective. Further work is in progress.

Acknowledgments: This work was supported by PhD funding via the EPSRC, and Scottish Government Rural and Environment Science and Analytical Services Division research funding.

References

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise Coordinate Optimization. *Annals of Applied Statistics*, **1**, 302 – 332.

Marx, B.D. and Eilers, P.H.C. (1999). Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach. *Technometrics*, **41**, 1 – 13.

Roberts, A.M.I. (2012). Comparison of regression methods for phenology. *International Journal of Biometeorology*, **56**, 707 – 717.

Modelling Occupational Stress and Burnout in Portuguese University Teachers by using Structural Equation Models

Susana Faria[1], A. Manuela Gonçalves[1], Rui Gomes[2]

[1] Department of Mathematics and Applications, CMAT-Centre of Mathematics, University of Minho, Portugal;
[2] Department of Applied Psychology, University of Minho, Portugal

E-mail for correspondence: mneves@math.uminho.pt

Abstract: The aim of the present study is to test a hypothetical model to analyze the mediating role of cognitive appraisal on the relation between occupational stress and burnout. A sample of 333 teachers was collected from a public university in the north of Portugal. The participants answered a protocol with measures of stress on academic staff, cognitive appraisal of work activity, and a burnout inventory for educators. Structural equation modelling (SEM) was used to evaluate the mediating effects. Results confirmed that cognitive appraisals partially explained the relationship between occupational stress and burnout at work, turning these variables a promising underlying mechanism for explaining adaptation at work.

Keywords: Occupational stress; Burnout; Structural Equation Models; Mediation model; University Teachers.

1 Introduction

Using SEM, the main goal of this study was analyzed by the hypothesis that tested the mediating effect of cognitive appraisal on the relation between stress and burnout. More specifically, it was assumed that primary cognitive appraisal mediate the relationship between occupational stress and burnout. The relation stressor-strain is assumed by some important theoretical frameworks, being also supported by empirical findings that demonstrate the relation between stress and burnout. Hypothesis one postulated that stress was positively related to burnout.

2 Methods

2.1 Participants and Procedure

The total sample consisted of 333 teachers working in a public university in the north of Portugal, being 129 males (39.95%) and 194 females (60.1%).

Participant's ages varied between 23 and 65 years old ($M = 42.67$; $SD = 6.87$), being 4.2% lecturers, 10.5% assistants, 62.6% assistants professor, 18.9% associate professor, and 3.8% full professors. Most of the teachers had full time contract at the university (90.6%) and had tenured contract without term (58.7%). All the participants that wanted information on their results filled in their name and address for further contact. Altogether, 893 questionnaires were distributed, and 333 were collected and considered valid, which showed a return rate of 37%.

2.2 Measures

Demographic Questionnaire. This questionnaire assessed personal (e.g., age, sex) and professional (e.g, years of work, category of employee, employment status) characteristics of teachers.
Stress Questionnaire for Academic Staff (SQAS; Gomes, 2010). This instrument evaluates the sources of stress that teachers face in their activity, including 32 items distributed for eight stress dimensions.
Cognitive Appraisal Scale (CAS; Gomes, 2008). This instrument evaluates primary and secondary cognitive appraisal. Primary cognitive appraisal was assessed with three dimensions.
Maslach Burnout Inventory - Educators Survey (MBI-ES) (Maslach et al., 1996). The questionnaire includes 22 items divided into three subscales.

2.3 Statistical Analysis

SEM was used to test the hypotheses. All analyses were conducted in AMOS 20. To assess model fit of the structural models, we used the χ^2 goodness-of-fit statistic, the root mean square error of approximation (RMSEA), the Tucker-Lewis index (TLI) and the comparative fit index (CFI). Finally, the bootstrap procedure of AMOS was also used to obtain 95% confidence intervals around parameter estimates. Bootstrapping is considered a powerful resampling method to obtain parameter estimates and confidence intervals being not assumed that the variables are normally distributed.

3 Results

3.1 Mediation Models

Regarding the model that tested the relation between stress, primary cognitive appraisal, and burnout, the fit of the 1-factor model with all items from the thirteen study variables loading on a single latent variable was compared with that of a 4-factor model that included stress, threat perception, challenge perception, and burnout. The 4-factor model fitted well to the data, $\chi^2(616df) = 999.7, p < 0.01$; RMSEA = 0.044 (pclose = 0.972);

CFI = 0.94; NFI = 0.90; TLI = 0.93, and its fit was superior to that of the 1-factor model $\chi^2(80df) = 3752.7; p < 0.001$. All standardized factor loadings were significant, ranging from 0.23 to 0.85. These results confirmed the validity of the 4-factor specified measurement model.

3.2 Structural Models

For the test of the structural models, it was compared the fit of a mediated model to the fit of a direct model. In the mediated model it was established a relation between stress, cognitive appraisal, and burnout. In the direct model it was established a relation from stress and cognitive appraisal to burnout. Also, it was analyzed which type of mediation (e.g., partial or full) could best explain the data. In the partial mediation model, we added direct paths from stress to cognitive appraisal. Finally, in the full mediation model, we removed the direct paths from stress to burnout.

The direct effects model did not fit the data successfully. The RMSEA (0.057) deviated significantly from 0.50(pclose < 0.01). The full mediation model showed acceptable fit indices (RMSEA = 0.056, CFI = 0.90, TLI = 0.89), but the partial mediation model, in which all direct and indirect effects were included, appeared to have the best fit indices (RMSEA = 0.054 (pclose = 0.053); CFI = 0.91; TLI = 0.90). The difference in chi-square between the fully and partially mediated model was significant $\chi^2(1) = 42.55; p < 0.001$), indicating that the direct effects cannot be ignored.

Table 1 presents the standardized effects for the partial mediation of the, namely the parameter estimates of the structural paths' coefficients and the squared multiple correlation coefficients. The estimates of the direct and indirect effects were based on 1000 bootstrap samples, being presented in parenthesis the corresponding 95% confidence intervals of these bootstrap estimates.

TABLE 1. Standardized effects (95% confidence intervals) in partial mediation model

	Primary cognitive apparaisal		Burnout	
	Threat perception	Challenge perception	Indirect effect	Direct effect
Stress	0.382^{**} (0.267; 0.490)	-0.350^{**} (-0.456; -0.239)	0.212^{**} (0.126; 0.330)	0.524^{**} (0.339; 0.696)
Threat perception				0.339^{**} (0.153; 0.497)
Challenge perception				-0.237^{**} (−0.401; −0.056)
R^2	0.15^{**} (0.071; 0.240)	0.12^{**} (0.057; 0.208)		0.69^{**} (0.455; 0.990)

4 Conclusions

It can be observed that the partial mediation model explained 15% of the variance associated with threat perception and 12% of the variance associated with challenge perception. Also, this model explained 69% of the variance in the burnout experience. Stress increased burnout both directly and indirectly, being confirmed the partial mediation effect of primary cognitive appraisal on the relation between stress and burnout. As predicted, occupational stress was related to primary cognitive appraisal, both positively (threat perception) and negatively (challenge perception). Also, primary cognitive appraisal was positively (threat perception) and negatively (challenge perception) related to burnout.

Acknowledgments: A. Manuela Gonçalves and Susana Faria were partially financed by FEDER Funds through "Programa Operacional Factores de Competividade - COMPETE" and by Portuguese Funds through FCT- "Fundação para a Ciência e a Tecnologia", within the Project Est-C/MAT/UI0013/2011.

References

Byrne, B.M. (2001). *Structural equation modeling with AMOS: Basic concepts. Applications, and programming.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cooper, C.L., Dewe, P.J. and O'Driscoll, M.P. (2001). *Organizational stress: A review and critique of theory, research, and applications.* Thousand Oaks, CA: Sage.

Gomes, A.R. (2008). *Escala de Avaliação Cognitiva.* Relatório técnico não publicado. [Cognitive Appraisal Scale]. Unpublished technical report. Braga: Instituto de Educação e Psicologia, Universidade do Minho.

Gomes, A. R. (2010). *Questionário de Stress nos Professores do Ensino Superior (QSPES).* [Stress Questionnaire for Academic Staff]. Unpublished technical report. Braga: Escola de Psicologia, Universidade do Minho.

Maslach, C., Jackson, S.E., and Schwab, R.L. (1996). *Maslach Burnout Inventory - Educators Survey (MBI-ES).* In C. Maslach, S.E. Jackson, and M.P. Leiter (Eds.), MBI Manual (3rd ed., pp. 27 – 32). Mountain View, CA: CPP, Inc.

Measuring the component overlapping in mixtures of linear regressions

Susana Faria[1], Gilda Soromenho[2]

[1] Department of Mathematics and Applications, CMAT-Centre of Mathematics, University of Minho, Portugal;
[2] Institute of Education, University of Minho, Portugal

E-mail for correspondence: sfaria@math.uminho.pt

Abstract: Entropy-type measures for the heterogeneity of data have been used for a long time. In a mixture model context, entropy criterions can be used to measure the overlapping of the mixture components. In this paper we study an entropy-based criterion in mixtures of linear regressions to measure the closeness between the mixture components.
We show how an entropy criterion can be derived based on the Kullback-Leiber distance, which is a measure of distance between probability distributions. To investigate the effectiveness of the proposed criterion, a simulation study was performed.

Keywords: Mixtures of linear regressions; entropy criterion; Kullback-Leiber information; simulation study

1 Introduction

Finite mixture models are a well-known method for modelling data that arise from a heterogeneous population (e.g., see McLachlan and Peel, 2000; Fruhwirth-Schnatter, 2006 for a review). The study of these models is a well-established and active area of statistical research and mixtures of regressions have also been studied fairly extensively. Mixtures of linear regressions have also been studied extensively, especially when no information about membership of the points assigned to each line was available.
The mixture of linear regression model is given as follows:

$$y_i = \begin{cases} \mathrm{x}_i^T\beta_1 + \epsilon_{i1} & \text{with probability } \pi_1, \\ \mathrm{x}_i^T\beta_2 + \epsilon_{i2} & \text{with probability } \pi_2, \\ \vdots & \\ \mathrm{x}_i^T\beta_J + \epsilon_{iJ} & \text{with probability } \pi_J \end{cases} \tag{1}$$

where y_i is the value of the response variable in the ith observation; x_i^T $(i = 1, \ldots, n)$ denotes the transpose of the (p+1)-dimensional vector

of independent variables for the ith observation, β_j $(j = 1, \ldots, J)$ denotes the (p+1)-dimensional vector of regressor variables for the jth component, π_j are the mixing probabilities ($0 < \pi_j < 1$, for all $j = 1, \ldots, J$ and $\sum_j \pi_j = 1$). Finally, ϵ_{ij} are the random errors; under the assumption of normality, we have $\epsilon_{ij} \sim N(0, \sigma_j^2)$, $(i = 1, \ldots, n; j = 1, \ldots, J)$.

2 Statistical Entropy

The Kullback-Leibler(KL) information (Kullback, 1959), also known as Kullback's directed divergence, is the measure of information discrepancy between $f(x)$ and $g(x)$, which is defined as

$$KL(f : g) = \int f(x) log \frac{f(x)}{g(x)} dx$$

where $f(x)$ is referred to as the reference distribution. The Kullback's directed divergence can be considered as a kind of a distance between the two probability densities, though it is not a real distance measure because it is not symmetric. An alternative directed divergence is the Kullback's symmetric divergence defined as the sum of two directed divergences (Frühwirth-Schnatter, 2006),

$$J(f : g) = KL(f : g) + KL(g : f) = \int f(x) log \frac{f(x)}{g(x)} dx + \int g(x) log \frac{g(x)}{f(x)} dx$$

(Leisch,2004) uses the Kullback-Leibler(KL) information to diagnose which components overlap in a mixture model.
Although there are many possible distance measures between two densities available in literature, the Kullback's symmetric divergence is attractive because of its simplicity and analytical tractability for mixtures models.

2.1 An entropy-based criterion

Consider a two-component mixture of linear regressions,

$$f(y|\mathbf{x}) = \pi_1 f_1(y; \mathbf{x}^T \beta_1, \sigma_1^2) + \pi_2 f_2(y; \mathbf{x^T} \beta_1, \sigma_2^2).$$

Based on Kullback's symmetric divergence, we define a criterion (EC) to study the overlapping of the mixture normal components in a two-component mixture of linear regressions,

$$\begin{aligned} EC(\pi_1 f_1 : \pi_2 f_2) &= KL(\pi_1 f_1 : \pi_2 f_2) + KL(\pi_2 f_2 : \pi_1 f_1) = \\ &= 2\pi_1 ln(\tfrac{\pi_1}{\pi_2}) + ln(\tfrac{\pi_2}{\pi_1}) + \pi_1 KL(f_1 : f_2) + \pi_2 KL(f_2 : f_1) \end{aligned}$$

with

$$KL(f_i : f_j) = \frac{n}{2}(ln \frac{\sigma_j^2}{\sigma_i^2} + \frac{\sigma_i^2}{\sigma_j^2}) + \frac{1}{2\sigma_j^2}(\mathbf{x^T} \beta_\mathbf{i} - \mathbf{x^T} \beta_\mathbf{j})^\mathbf{2} - \frac{\mathbf{n}}{\mathbf{2}}$$

3 Simulation study

To investigate the effectiveness of the proposed criterion, a simulation study was performed. We used the freeware R to develop the simulation program. Consider the mixtures of linear regressions,
mixture model 1: $f(y|\mathbf{x}) = \pi_1 f_1(y; \mathbf{x}^{\mathbf{T}}\beta_{f1}, \sigma^2_{f1}) + \pi_2 f_2(y; \mathbf{x}^{\mathbf{T}}\beta_{f2}, \sigma^2_{f2})$
mixture model 2: $g(y|\mathbf{x}) = \pi_1 g_1(y; \mathbf{x}^{\mathbf{T}}\beta_{g1}, \sigma^2_{g1}) + \pi_2 g_2(y; \mathbf{x}^{\mathbf{T}}\beta_{g2}, \sigma^2_{g2})$.
Samples of size $n = 100$ were generated for each set of true parameter values shown on Table 1 and the mixing proportion $\pi_1 = \{0.2, 0.4, 0.8\}$. We considered two typical configurations of the true regression lines: parallel and concurrent. For each type of simulated data set, 200 samples of size n were simulated.

TABLE 1. True parameter values for the essays

Configuration	Mixture 1						Mixture 2				
	β_{01}	β_{11}	β_{02}	β_{12}	σ^2_{f1}	σ^2_{f2}	β_{01}	β_{11}	β_{02}	β_{12}	σ^2_{f2}
concurrent/concurrent	2	1	4	-1	0.35	0.5	0	1	2	-1	0.5
paralell/paralell	1	1	3	1	0.35	0.5	3	1	5	1	0.5
paralell/concurrent	2	1	4	-1	0.35	0.5	3	1	5	1	0.5
concurrent/concurrent	0	1	2	-1	0.15	0.35	4	1	6	-1	0.5
paralell/paralell	1	1	4	1	0.3	0.35	3	1	6	1	0.5
paralell/concurrent	1	1	4	-1	0.3	0.35	2	1	4	-1	0.5

The simulation process consists of the following steps:

- Create a data set of size $n = 100$ of mixture model 1. Fit a mixture of linear regression models to the data using the EM algorithm. Save the estimated parameters $\hat{\Psi} = \{(\hat{\pi}_1, \hat{\pi}_2, \hat{\beta}_{f1}, \hat{\beta}_{f2}, \hat{\sigma}^2_{f1}, \hat{\sigma}^2_{f2})\}$ and calculate the estimated $\widehat{EC}$.
- Determine σ^2_{g1} so that the two mixture models (mixture model 1 and mixture model 2) have the same estimated EC value.
- Create a data set of size $n = 100$ of mixture model 2. Fit a mixture of linear regression models to the data using the EM algorithm. Save the estimated parameters $\hat{\Psi} = \{(\hat{\pi}_1, \hat{\pi}_2, \hat{\beta}_{g1}, \hat{\beta}_{g2}, \hat{\sigma}^2_{g1}, \hat{\sigma}^2_{g2})\}$ and calculate the estimated $\widehat{EC}$.
- Calculate the Mahalanobis distance between estimated and true parameters values in mixture model 1 and in mixture model 2.
- The Mann-Whitney test is used for testing the equality of the two Mahalanobis distances.

4 Conclusion

In this article, we define an entropy-based criterion in two component mixtures of linear regressions to measure the overlapping of the mixture components.
We note the following general findings:

- When the true regression lines are parallel in two mixtures models and the degree of overlap between two components is the same, there is no differences between the two estimates of the parameters of mixtures of linear regressions ;
- When the true regression lines are concurrent in two mixtures models and the degree of overlap between two components is the same, there is no differences between the two estimates of the parameters of mixtures of linear regressions ;
- When the true regression lines are parallel in one mixture model and concurrent in another mixture model and the degree of overlap between two components is the same, there is no differences between the two estimates of the parameters of mixtures of linear regressions.

We may conclude that the configurations of the true regression lines does not affect the performance of the EM algorithm, but only its degree of overlapping.

Acknowledgments: This research was financed by FEDER Funds through "Programa Operacional Factores de Competitividade COMPETE" and by Portuguese Funds through "FCT - Fundação para a Ciência e Tecnologia" , within the Project Est-C/MAT/UI0013/2011.

References

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, Heidelberg.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Leisch, F. (2004). *Exploring the Structure of Mixture Model Components*. In J. Antoch (ed.),Compstat 2004 Proceedings in Computational Statistics, pp. 1405-1412. Physika Verlag, Heidelberg, Germany.

Mclachlan, G. J., and Peel, P. (2000). *Finite Mixture Models*. Wiley, New York.

Traffic policies and air quality in Italian cities

Alessandro Fassò [1]

[1] University of Bergamo, Italy

E-mail for correspondence: alessandro.fasso@unibg.it

Keywords: Spatio-temporal intervention analysis; Congestion charge; Particulate matters; Nitrogen oxides.

1 Introduction

Inter alias, traffic policies are often motivated by the need to improve air quality in terms of pollutants concentration reduction. Hence a fundamental step is to assess if a given policy actually obtains first a relevant reduction of emissions of pollutants in the atmosphere and therefore a relevant reduction of pollutants concentrations. Emissions are sometimes assessed using computational models for car traffic intensity, typology and ageing.
In this paper, we apply a general statistical methodology for spatiotemporal impact assessment based on concentrations observed trhough air quality monitoring networks. In this context, we consider changes of concentrations observed in traffic stations and pertaining to primary pollutants as proxies of emission changes. Moreover, changes in secondary pollutants and/or ground stations do prompt for more fundamental changes of population exposure.
The general approach proposed is illustrated by to important case studies. The first one is related to the introduction of a socalled "congestion charge" in the city of Milan. According to it, from January 16, 2012, drivers are rquired to pay a fee of five Euros for entering the central area, known as "*Area C*" and reported in Figure 1. As discussed in Fasssò (2013), in the first two months, car traffic decreased by 36% in Area C and, at the overall city level, Municipality reported a traffic reduction by 6% which started at the beginning of January, before the congestion charge. This preemptive reduction may be partly due to the preparatory campaign played out by the administration and partly to an overall decrease in gas consumption at the national level caused by the economic crisis. January and February were cold and heavily polluted months with a large number of days exceeding the thresholds fixed by the European regulations (see Arduino and Fassò, 2012). Although the congestion charge is intended as a measure of traffic control, the question which arises is whether there has been an impact on air quality or not. We will focus here on two important compounds entering

most air quality indexes, namely particulate matters and nitrogen oxides which are important for their toxicity. Particulate matters are predominantly secondary pollutants. Hence they have a background level which is large in percentage and difficult to reduce through local traffic policies. Moreover their health effects are known to depend not only on their concentration but also on particle size, composition and black carbon content, therefore, having extensive data on particle numbers (see e.g. Hong-di and Wei-Zehn, 2012) or on black carbon content (see e.g. Janssen et al. 2011) would be very useful to understand air quality from a health protection point of view. To a large extent, total nitrogen oxides are compounds of primary pollutants. Hence they are closely linked to local traffic and have a background level which is smaller in percentage than particulate matters.
The second case study is related to the city of Turin and the various restrictions relized since 2010. Some restrictoins are related to the central ZTL area, other are global measures. We than have to face a multi-intervention analysis problem.
The rest of the paper is organized as follows. In Section 2 we introduce a general spatiotemporal model, which is capable of various levels of complexity according to the information content of the underlying monitoring network. The model allows us to estimate the impact on air quality and the reduction of human exposure following the considered environmental policy. Maximum likelihood estimation is obtained by a version of the EM algorithm and impact testing is proposed as a likelihood ratio test. In section 3, the above approach is applied to the introduction of the congestion charge in Milan city. In particular the result of Fassò (2013) are here generalized to the multivariae case. In section 4, the data of Turin traffic restrictions are considered.

2 Spatio temporal surveillance model

We consider here a general multivariate spatio-temporal model given by

$$y(s,t) = \beta(s,t)' x(s,t) + \varepsilon(s,t) \tag{1}$$

where $y(s,t)$ is a $q-$dim vector of measurements of interest at location $s \in R^2$ and time $t = 1,2,...,T$ (e.g. days or hours); $x(s,t)$ is a $k-$dim covariate vector (including unity); $\beta(s,t)$ is a $q \times k$ stochastic coefficient matrix whose elements are given by

$$\beta_{ij}(s,t) = \beta_{ij} + \gamma_{ij} Z_{ij}(t) + \delta_{ij} \omega_{ij}(s,t)$$

where β_{ij}, γ_{ij} and δ_{ij} are coefficients to be estimated; moreover $Z(t) = vec(Z_{ij}(t))$ is the common temporal component given by a vector Gaussian Markovian process

$$Z(t) = GZ(t-1) + \eta(t)$$

with standardized innovations $\eta\ NID(0, I)$ and diagonal persistence matrix G; $\omega(\cdot, t)$ is the spatial component given by iid replicates of a standardized Gaussian linear coregionalization model (LCM) with exponential correlation parameter θ; finally $\varepsilon(s, t)$ is a Gaussian white noise in space and time.
This model is quite flexible and has been used for computing the distribution of airborne pollutants human exposure in Finazzi et al (2013) . Moreover it has been shown useful for spatial clustering of geolocated time series as in Finazzi and Scott (2013) . Maximum likelihood estimation of this model is based on the EM algorithm as in Fassò and Finazzi (2011) and Finazzi and Fassò (2013).
In order to cover for spatio-temporal analysis, we modify equation (1) by adding the impact α as follows:

$$y(s, t) = -\alpha(s, t) + \beta(s, t)' x(s, t) + \varepsilon(s, t) \tag{2}$$

Denoting the intervention time by t^*, the spatial impact of the traffic reduction is defined as

$$\alpha(s, t) = \begin{cases} 0 & t < t^* \\ \alpha(s) & t \geq t^* \end{cases}$$

where $\alpha(s) \geq 0$ defines a step impact, allowing spatial variations. In general, α may be a random impact or a deterministic one. For example Fassò (2013) use $\alpha(s) = \alpha_1$ for s in city center and $\alpha = \alpha_2$ elsewhere.

3 Milan case study

Using model (2) , Fassò (2013) found a reduction of 8% with $se = 0.035$ for particulate matters (PM_{10}) and a reduction of 19% with $se = 0.032$ for Nitrogen oxides (NO_x). Here, using a multivariate extension we will exploit the maximum information of Milan data in order to reduce uncertainty.

4 Turin case study

The city of Turin, in recent years, experienced a number of different measures to control and reduce car traffic and veichular emissions. In particular, a number of green sundays have been established, which bar car use daytime; moreover different selective restrictions to aged cars have been carried out, both at city level and limited to ZTL, which is reported in Figure 2.
Following this scheme, we will analyse the problem by means of a multi-intervention spatio-temporal analysis. This will help clarifying if these policies have an impact on local emission changes only or may be aimed at reducing population exposure at large.

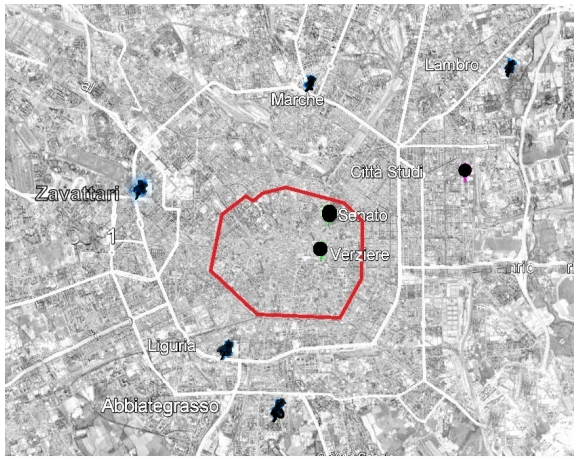


FIGURE 1. City of Milan; Area C is delimited by the red ring.

FIGURE 2. City fof Turin; ZTL is the area delimited by blue lines.

References

Fassò A (2013) Statistical assessment of air quality interventions. *Stochastic Environmental Research and Risk Assessment*. In printing.

Fassò A, Finazzi F, (2011) Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics*. **22**:6, 735-748.

Finazzi F, Fassò A (2012) DSTEM - A statistical software for multivariate space-time environmental data modeling. In Goncalves A.M. et al. (Ed's 2012), *Proceedings of the International Workshop on Spatio-Temporal Modelling (METMA VI)*. Guimaraes, 12-14 September 2012. ISBN 978-989-97939-0-3.

Finazzi F, Fassò A (2013). EM estimation of a multivariate space-time data fusion model with varying coefficients. Working papers GRASPA, 47. http://hdl.handle.net/10446/27548

Finazzi F., Scott M.E. (2013). The estimation of latent temporal patterns in multivariate geolocated time series. *SIS2013, Advances in latent variables: methods, models and applications. Brescia, June 19-21 2013.* In printing.

Finazzi F., Scott M.E., Fassò A. (2013). A model based framework for air quality indices and population risk evaluation. With an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society, series C*. **62**(2), 287-308.

Hong-di H, Wei-Zhen L (2012) Urban aerosol particulates on Hong Kong roadsides: size distribution and concentration levels with time. *Stochastic Environmental Research and Risk Assessment*. 26:177–187.

Janssen NA, Hoek G, Simic-Lawson M, Fischer P, van Bree L, ten Brink H, Keuken M, Atkinson RW, Anderson HR, Brunekreef B, Cassee FR (2011) Black Carbon as an Additional Indicator of the Adverse Health Effects of Airborne Particles Compared with PM10 and PM2.5. *Environmental Health Perspectives*. 119:1691-1699.

Beyond the threshold: the efficiency of Italian manufacturing firms

Giancarlo Ferrara, Francesco Vidoli

[1] SOSE - Soluzioni per il Sistema Economico S.p.A., Italy

E-mail for correspondence: gferrara@sose.it, fvidoli@sose.it

Abstract: Firms' size is a mainstream in the study on economic growth. Within this research area, this work analyses the presence of different marginal returns to labour with the aim of suggesting new insights for the Italian manufacturing system. We consider a two stage approach combining the order-m non-parametric frontier estimation and the 'broken-line' regression models. An homogeneous sample of manufacturing Small and Medium Enterprises (SMEs) extracted from the Database of Italian Ministry of the Economy and Finance annual survey (Studi di Settore) is used for the application of the proposed methods.

Keywords: Efficiency; Manufacturing; Frontier; Robust order-m; Segmented

1 Introduction

The research aims at evaluating the differential productive efficiency for a given set of Italian manufacturing firms in order to estimate critical nodes within which companies are structurally similar.
In order to analyze firms' efficiency, the frontier approach allows a clearer and more accurate analysis of marginal rates of substitution and of the different individual productivity than regression (Daraio and Simar, 2007). Within the nonparametric efficiency techniques, Florens and Simar (2005) proposed a two stage method for estimating the parametric model in an original way overcoming most drawbacks of the classical approaches: first, plot the cloud of data points estimating the nonparametric frontier to outline, in a flexible and robust way, the shape of boundary and, secondly, fit the stochastic linear model on these projected points.
Since the full parametric approximation in the second stage is too restrictive, we propose to evaluate only some properties of the frontier, i.e. the presence of thresholds within which companies are structurally similar, with the aim of examining the impact of company size on manufacturing firm efficiency in a "best practice" framework.

This document does not necessarily reflect the official opinion of the SOSE - Soluzioni per il Sistema Economico S.p.A. and commits only the authors.

The paper is structured as follows: the estimating algorithm is described in section 2, an application to Italian manufacturing Small and Medium Enterprises (SMEs) in section 3 and conclusions in section 4.

2 Methods

Denoting X as the single input associated to the single output Y, the Data Generating Process (DGP) of (X, Y) is supposed to be completely characterized by:

$$H_{XY}(x, y) = Prob(X \leq x, Y \geq y), \tag{1}$$

and the support of H_{XY}, noted as Ψ, is interpreted as the probability of a unit operating at the level (x, y) to be dominated. In an input oriented framework, this joint probability can be decomposed as follows:

$$\begin{aligned} H_{XY}(x, y) &= Prob(X \leq x | Y \geq y) Prob(Y \geq y) \\ &= S_{X|Y}(x|y) F_Y(y). \end{aligned} \tag{2}$$

An input oriented efficiency score $\widehat{\theta}(x, y)$ for $(x, y) \in \Psi$ is defined for all y with $F_Y(y) > 0$ as:

$$\begin{aligned} \widehat{\theta}(x, y) &= \inf\{\theta | (\theta x_0, y_0) \in \Psi\} \\ &= \inf\{\theta | H(\theta x, y) > 0\}. \end{aligned} \tag{3}$$

By considering the two stage analysis as proposed by Florens and Simar (2005), the efficient frontier and the $\widehat{\theta}(x, y)$ scores are estimated, in the first step, *via* the order-m approach in order to mitigate the bias due to the presence of influential observations (Cazals *et al.*,2002). While keeping its nonparametric nature, the expected order$-m$ frontier does not impose convexity on the production set and allows for noise (with zero expected values). As highlighted by Daraio and Simar (2007), the order-m efficiency score can be seen as the expectation of the minimal input efficiency score of the unit (x, y), when compared to m units randomly drawn from the population of firms with output levels greater than y. This is certainly a less extreme benchmark for the unit (x, y) than the 'absolute' minimal achievable level of inputs: it is compared to a set of m peers (potential competitors) producing more than its level y taken as a benchmark.
Given the estimation of the individual efficiency scores $\widehat{\theta}(x, y)$, we propose, in the second step, to study only some aspects of the frontier and especially the different marginal returns to labor in terms of breakpoints varying with inputs rather than the full stochastic parametric approximation of the production frontier as usually done. To meet this purpose, we use a piece-wise or segmented linear regression model (Muggeo, 2003) in order to bypass strong assumptions regarding the production function form and its

stochastic properties and to identify the inflection points (i.e. thresholds) in the production frontier of the efficient SMEs.
Roughly speaking, in a simple linear regression context, denoting by (x_i, y_i) the ith observation $(i = 1, 2, \ldots, n)$ associated to the response variable Y (i.e. output) and the covariate X (i.e. input), a segmented relationship between $\mu = E(Y|X = x)$ and X can be formulated as follows:

$$E(Y|X = x) = \mu = \beta_0 + \beta_1 X + \beta_2 (x - \psi)_+, \tag{4}$$

where $(x - \psi)_+ = (x - \psi) \times I(x > \psi)$ and $I(\cdot)$ the indicator function equal to one if the condition is verified; according to such model specification, β_1 is the left slope, β_2 is the difference-in-slopes and ψ is the breakpoint (Muggeo, 2008).
The estimation procedure may be easily summarized as follows:

- estimation of an efficiency score for each SMEs without assuming an *a priori* functional form of the efficient frontier by order$-m$ estimator;
- after selecting the subset of efficient firms on the basis of the previous results, search for potential thresholds.

We initially fit the segmented model (4) on SMEs with $\widehat{\theta}(x, y)$ scores lying in the [1-α, 1+α] efficiency interval, excluding from the analysis all firms that appear to be 'super' efficient ($\widehat{\theta}(x, y) > 1 + \alpha$) or inefficient ($\widehat{\theta}(x, y) < 1 - \alpha$). Since the choice of $\alpha \in (0, 1)$ is arbitrary, we suggest to evaluate the robustness of the associated $\hat{\psi}$ by varying the α value.

3 Application

The application has been carried on $n = 1053$ firms belonging to an homogeneous sample of Italian SMEs operating in the mechanical sector (Database of Italian Ministry of the Economy and Finance annual survey, Studi di Settore) given its key role in the Italian manufacturing productive system and considering that the engineering sector accounts for more than 42% of total manufacturing added value (Federmeccanica, 2006). Furthermore, its contribution to technological innovation exploited by other industries determines a greater competitiveness of the overall industrial production process.
Given the peculiarities of the sector under consideration and for sake of simplicity, the production process is here specified by considering a single input (number of Employees) and a single output (Revenues in Euro).
Figure 1 shows the presence of a strong change in slope at $\hat{\psi} = 20.83$ ($\alpha = 0.2$) suggesting a decrease in the marginal return to labor around 80%. Break-points kernel density, obtained by shrinking the efficiency interval, assesses threshold stability.

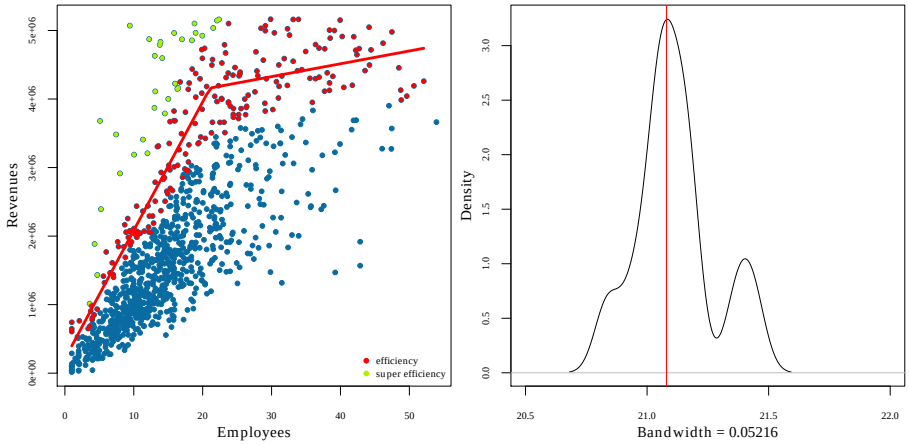


FIGURE 1. Segmented frontier (red line) and $\hat{\psi}$ kernel density.

4 Conclusions

This work provides a novel framework to evaluate efficiency of Italian manufacturing firms, in an attempt to partially fill the existing gap in the relevant literature due essentially to the lack of firm level data. We propose a flexible two step procedure finding new insights associated to the SMEs size. Results should be deeply analyzed by including in the model internal and external factors associated with tangible and intangible factors such as R&D, human capital, public infrastructure and degree of internalisation.

References

Cazals, C. and Florens, J. and Simar, L. (2002). Nonparametric frontier estimation: A robust approach. *Journal of econometrics*, **106**, 1–25.

Daraio C. and Simar L. (2007). *Advanced Robust and Nonparametric Methods in Efficiency Analysis*. Springer.

Federmeccanica (2006). *Indagine congiunturale*.

Florens J. and Simar L. (2005). Parametric approximations of nonparametric frontiers. *Journal of Econometrics*, **124**, 91–116.

Muggeo, V.M.R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, **22**, 3055–3071.

Muggeo, V.M.R. (2008). Segmented: an R package to fit regression models with broken-line relationships. *R News*, 8/1, 2025.

Handling missing data in longitudinal studies: an application to healthcare data

Álvaro J. Flórez [12], Ana Nora A. Donaldson [12], Mercedes Andrade[1], Javier Torres [1],Nairn Wilson [2]

[1] Universidad del Valle, Colombia
[2] King's College London, United Kingdom

E-mail for correspondence: alvaro.florez@univalle.edu.co

Abstract: The presence of missing data in longitudinal datasets is a very common problem. In certain cases the missingness affects not only the accuracy of the estimates, but also can introduce bias, and therefore invalid conclusion. Therefore, we revise the most widely used methods of handling missing data evaluating and comparing their performance under different simulated scenarios of missingness. The simulated dataset was obtained using a real-life study conducted at a public hospital of Cali, Colombia, South America.

Keywords: Missing data; Longitudinal data; Mixed effect model; Simulation

1 Introduction

Longitudinal studies that collect information of the same individual repeatedly in time is widely used in medical research. Unfortunately when we work with human subjects, the possibility to have missing information is really big. This is because patients can withdraw from the study for multiple reasons, like treatment inefficacy, secondary effects or unrelated with the study reasons (Daniels, M. & Hogan, J., 2008).
The analysis with missing values will necessarily have a loss of precision in the estimates, since the sample size is reduced. But the biggest problem is the potential bias that the estimates may have and this could lead to wrong conclusions. This problem depends, not only on the method to handle missing data, but on the reasons that led missing values (called *missing data mechanism*)(Philipson et al. 2008, Fielding et al.,2012).
The aim of this paper is to present a simulated based comparison of different techniques to handle missing data under different scenarios of missing data. The simulated datasets were obtained using a real-life study of low birth-weight conducted at a public hospital of Cali, Colombia, South America (Bermudez et al.,2013). A random coefficient model (Verbeke & Molenberghs, 2000) was used to fit the dataset and to create simulated

datasets. The missing values were simulated according the three missing data mechanism proposed by Rubin (1976).
We selected four methods to compare their performance in this study. Two conventional methods: complete cases analysis (CC) and the simple imputation method called last observation carried forward (LOCF), both widely used nowadays (Wood et al., 2004, Fielding et al.,2012). And two more robust techniques: multiple imputation (MI) and likelihood-based methods using all available data (we called available cases - AC).

2 Simulated Scenarios

The different datasets were simulated using the next model (Bermudez et al.,2013),

$$Y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 t_i^2 + \beta_3 PC_i + b_{0i} + b_{1i} t_j + b_{2i} t_j^2 + \varepsilon_{ij} \qquad (1)$$

Were Y_{ij} is the weight of the newborn i in the time t_j, $t_j = (0, 12, 18, 24, 36)$ (months), and PC_i is 1 if the newborn's mother had prenatal care, and 0 otherwise. $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $\boldsymbol{b_i} = (b_{0i}, b_{1i}, b_{2i})$ are the coefficient for fixed effects and random effects, respectively. ε_{ij} is the random error. The random effects and random error where simulated through these distributions, $b_i \sim N(0, \boldsymbol{D})$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$.
All the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2)$ were estimated fitting the model (1), using the real data of low-birth weight.
The missing data scenarios were chosen according to the three missing data mechanism proposed by Rubin (1976) and Little, R & Rubin, D. (2002): Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR). Therefore, the probability of a missing value of Y_{ij}, for each mechanism, is,

$$\begin{array}{ll} P(R_{ij} = 1) = \alpha_0 & \text{MCAR} \\ P(R_{ij} = 1) = \Phi(\alpha_0 + \alpha_1 y_{i,j-1}) & \text{MAR} \\ P(R_{ij} = 1) = \Phi(\alpha_0 + \alpha_1 y_{i,j}) & \text{MNAR} \end{array}$$

Were R_{ij} is the missing data indicator, $R_{ij} = 1$ if y_{ij} is missing, and $R_{ij} = 0$ if y_{ij} is observed. We chose α_0 and α_1 such that the amount of missing data in the dataset are approximately 30%

3 Methodology

According to the missing data mechanism we have three scenarios. We created 1000 simulated datasets for each scenario. In each one we fitted the model (1) using four selected techniques to handle missing data (CC-LOCF-MI-AC). The comparison of these methods were made using the relative bias of each estimate $(\bar{\hat{\theta}}_i - \theta_i)/\theta_i$. All the simulations and results were made using the statistical software R.

4 Results

Table 1 displays relative bias of estimates of $\boldsymbol{\theta}$ using each technique under each missing data mechanism scenario (except the covariances of $\boldsymbol{b_i}$ that show similar results as variance of $\boldsymbol{b_i}$).

TABLE 1. estimate relative bias for each technique of handling missing data

		Fix effects				Random Effects			
	Method	β_0	β_1	β_2	β_3	$\sigma^2_{b_0}$	$\sigma^2_{b_1}$	$\sigma^2_{b_2}$	σ^2_{ε}
MCAR	CC	0.00	0.00	0.00	-0.02	0.86	0.08	0.13	-0.03
	LOCF	0.02	-0.19	-0.08	0.01	0.08	9.05	7.32	0.07
	AC	0.00	0.00	0.00	0.00	0.63	0.08	0.13	-0.04
	MI	0.00	0.00	0.00	0.00	0.63	0.08	0.14	-0.04
MAR	CC	-0.07	-0.06	-0.05	-0.28	-0.17	-0.29	-0.03	-0.11
	LOCF	0.02	-0.23	-0.12	-0.28	1.23	10.4	9.43	0.11
	AC	0.00	0.00	0.00	0.00	0.65	0.09	0.16	-0.04
	MI	0.00	0.00	-0.01	0.00	0.64	0.09	0.15	-0.04
MNAR	CC	0.04	-0.08	-0.07	-0.13	0.61	-0.33	- 0.08	-0.07
	LOCF	0.01	-0.25	-0.16	-0.07	0.60	9.05	7.69	0.02
	AC	0.00	-0.04	-0.03	-0.03	0.73	-0.21	0.00	-0.06
	MI	0.00	-0.04	-0.03	-0.04	0.73	-0.21	0.00	-0.06

Under MCAR scenario, only LOCF provides biased estimates of fixed effects ($\boldsymbol{\beta}$), especially underestimation the slope, the other three methods provide approximately unbiased estimates of $\boldsymbol{\beta}$. Under MAR conventional methods provides biased estimates, both methods underestimated the slope, and the prenatal care effect. Meanwhile, MI and AC provide unbiased estimates of all fixed effects.
Under MNAR scenario, all methods provide biased estimates, but MI and AC with less bias in all parameters. Under all scenarios of missing data mechanism there are biased estimates of the variance and covariance of random effects and error. But the variances are greatly overestimated with LOCF.

5 Conclusions

All aspects of the estimated linear mixed model (mean and variance - covariance structure) were influenced by the particular technique used and the missing data mechanism, in some cases leading to different conclusions (especially in MNAR situations). In general, CC and LOCF showed much poorer results than MI or AC.
Although CC provides unbiased estimates under MCAR, its estimates had the highest standard error; as expected because it does not use all available information. LOCF even under MCAR situations provides biased estimates. Both show poor performance on MAR and MNAR scenarios underestimating the slope and the prenatal care effects, this may lead to wrong

conclusions about the performance of the low birth-weight program of the hospital.
Compared with traditional methods, both, AC and MI show better performance providing unbiased estimates for fixed effects on MAR situations, but with little biased estimates on MNAR scenario. this bias will be greater if the proportion of missing data is larger.
Therefore, in cases when we have doubts of MCAR or MAR assumption and MNAR is a plausible situation, is better to conduct a sensitivity analysis. Other recommendation is to collect additional information related with potential causes of missingness, and to include it in the multiple imputation analysis. Schafer & Graham (2002) assert that this extra information may effectively convert an MNAR situation to MAR.

References

Bermudez, M., Andrade, M. and Torres, J. (2013). Análisis temporal de neonatos con bajo peso nacidos en un hospital de nivel iii de Cali, Colombia a travs de un modelo de coeficientes aleatorios. *In Press*.

Daniels, M. and Hogan, J. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modelling and Sensitivity Analysis*. Chapman & Hall/CRC.

Fielding, S., Fayers, P. and Ramsay, C.R. (2012). Analysing randomised controlled trials with missing data: Choice of approach affects conclusions. *Contemporary Clinical Trials*, **33(3)**, 461 – 469.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 2 ED.

Philipson, P., Kee Ho, W. and Henderson, R. (2008). Comparative Review of Methods for Handling Drop-out in Longitudinal Studies. *Statistics in Medicine*, **27(30)**, 6276 – 6298.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, **63(3)**, 581 – 592.

Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, **7(1)**, 147 – 177.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

Wood, A.M., White, I.R., and Thompson, S.G. (2004). Are Missing Outcome Data Adequately Handled? A Review of Published Randomized Controlled Trials in Major Medical Journals. *Clinical Trials*, **4(1)**, 368 – 376.

Bayesian estimation of a discrete choice model for household labor supply in Austria

Agnes Fussl[1], Sylvia Frühwirth-Schnatter[2], Christine Zulehner[3]

[1] Department of Applied Statistics, Johannes Kepler University Linz, Austria
[2] Institute for Statistics and Mathematics, Vienna University of Economics and Business, Austria
[3] Department of Economics, Johannes Kepler University Linz, Austria

E-mail for correspondence: agnes.fussl@jku.at

Abstract: Our work considers Bayesian estimation of a discrete choice model for household labor supply in Austria which provides a basis for a microsimulation approach. We estimate a very flexible multinomial logit model allowing for fixed and random effects as well as choice and unit specific covariates. To perform Markov Chain Monte Carlo (MCMC) sampling we rewrite the multinomial logit model as an augmented model which involves some latent variables called random utilities. The parameters appearing in the regression model are estimated by using a data-augmented independence Metropolis-Hastings sampler.

Keywords: Microsimulation; Discrete choice model; Multinomial logit; Mixed effects; Bayesian estimation; Data Augmentation; Random utility model; MCMC; Panel Data; Choice and unit specific covariates.

1 Introduction

Microsimulation approaches are widely used in economics to answer the question on which factors influence labor supply of individuals or households. They allow for evaluating the consequences of a policy reform such as changes in tax system, social security contributions or social transfers (e.g. parental leave benefits) on a sample of economic units on the microdata level. The main variable of interest is labor supply and is chosen in order to find out whether a policy reform causes a change in a unit's labor supply (e.g. enlargement/reduction of labor supply, entering/dropping out of the labor market). It is assumed that units choose the category with the largest personal benefit/utility from a relatively small number of labor supply alternatives. The estimation of labor supply is performed via a discrete choice model. Although the interpretation of the results is limited since microsimulation approaches are based on many assumptions, they enable to simulate the effects of policy changes in advance instead of analysing them ex post.

Our work considers Bayesian estimation of such a discrete choice model for household labor supply in Austria. We estimate a very flexible mixed effects multinomial logit model incorporating both fixed and random effects as well as choice and unit specific covariates. To perform efficient MCMC sampling we rewrite the multinomial logit model as an augmented latent variable model called (difference) random utility model (dRUM/RUM). The parameters appearing in the regression model are estimated by using a data-augmented independence Metropolis-Hastings (MH) sampler.

2 Data

Our microsimulation approach is based on the EU-SILC (European Union Statistics on Income and Living Conditions) dataset for Austria for the years 2004 – 2007, which contains longitudinal microdata on income, poverty, housing, labor and living conditions amongst others. Typically, each household is observed over a four year period, which leads to an unbalanced panel dataset for our analysis.
To obtain the final dataset for model estimation we carry out several preparation steps (e.g. imputation of missing wages, tax/transfer calculation) using the STATA routines implemented by Grünberger (2009) and Rabethge (2009). For the current work we only keep couple households with children ($\approx$ 2400 households). Since most observed males are full-time employed or unemployed, we define 6 discrete choices of household labor supply resulting from the possible pair combinations of males (full-time/unemployed) and females (full-time/part-time/unemployed), where part-time employment denotes ≥ 1 and full-time employment ≥ 35 working hours.

3 Bayesian inference

In the present work we estimate the discrete choice model by means of a very flexible multinomial logit model allowing for fixed and random effects as well as choice and unit specific covariates.
Let $\{y_{it}\}$ be a sequence of repeated categorical data observed for N subjects i $(i = 1, \ldots, N)$ on T_i occasions t $(t = 1, \ldots, T_i)$, where each y_{it} is assumed to take a value in one of $m + 1$ unordered categories labeled by $L = \{0, \ldots, m\}$. The probability that y_{it} takes the value k for each $k \in \{1, \ldots, m\}$ in terms of covariate information is modeled by the following multinomial logit model:

$$\Pr(y_{it} = k|\boldsymbol{\alpha}, \boldsymbol{\xi}_i, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) = \frac{\exp(\mathbf{x}^f_{kit}\boldsymbol{\alpha} + \mathbf{x}_{it}\boldsymbol{\beta}_k + \mathbf{x}^r_{kit}\boldsymbol{\xi}_i)}{1 + \sum_{l=1}^{m} \exp(\mathbf{x}^f_{lit}\boldsymbol{\alpha} + \mathbf{x}_{it}\boldsymbol{\beta}_l + \mathbf{x}^r_{lit}\boldsymbol{\xi}_i)}, \quad (1)$$

where $\boldsymbol{\alpha}$ is the vector of fixed effects, $\boldsymbol{\beta}_k$ are category specific regression coefficients and $\boldsymbol{\xi}_i$ is the vector of unit specific random effects (including

random intercept). The covariates in the model consist of $\mathbf{x}_{kit}^f$, a row vector of choice specific covariates assigned to the fixed effects, $\mathbf{x}_{it}$, a row vector of unit specific covariates, and $\mathbf{x}_{kit}^r$, a row vector of choice specific covariates assigned to the random effects.

We estimate the regression coefficients by using MCMC sampling with data augmentation, where the multinomial logit model is rewritten as a latent variable model called random utility model (RUM), which was first introduced by McFadden (1974):

$$u_{0it} = \mathbf{x}_{0it}^f \boldsymbol{\alpha} + \mathbf{x}_{0it}^r \boldsymbol{\xi}_i + \mathbf{x}_{it}\boldsymbol{\beta}_0 + \epsilon_{0it}, \qquad \epsilon_{0it} \sim \mathcal{EV} \tag{2}$$

$$u_{kit} = \mathbf{x}_{kit}^f \boldsymbol{\alpha} + \mathbf{x}_{kit}^r \boldsymbol{\xi}_i + \mathbf{x}_{it}\boldsymbol{\beta}_k + \epsilon_{kit}, \qquad \epsilon_{kit} \sim \mathcal{EV} \tag{3}$$

$$y_{it} = k \Leftrightarrow u_{kit} = \max_{l \in L} u_{lit},$$

where the observed category y_{it} is equal to the category with maximal utility and $\boldsymbol{\beta}_0$ is set 0 for reasons of identifiability. The errors appearing in (2) and (3) are i.i.d extreme value distributed ($\mathcal{EV}$), so that the multinomial logit model in (1) results as the marginal distribution of y_{it}.

Following Frühwirth-Schnatter and Frühwirth (2010, 2012), we construct the more efficient difference random utility model (dRUM) by choosing a baseline category $k_0 = 0$ and taking the difference of (2) and (3):

$$z_{kit} = \left[\mathbf{x}_{kit}^f - \mathbf{x}_{0it}^f\right] \boldsymbol{\alpha} + \left[\mathbf{x}_{kit}^r - \mathbf{x}_{0it}^r\right] \boldsymbol{\xi}_i + \mathbf{x}_{it}\boldsymbol{\beta}_k + \varepsilon_{kit}, \qquad \varepsilon_{kit} \sim \mathcal{LO} \tag{4}$$

$$y_{it} = \begin{cases} 0, & \text{if } \max_{l \in L_{-0}} z_{lit} < 0, \\ k > 0, & \text{if } z_{kit} = \max_{l \in L_{-0}} z_{lit} > 0, \end{cases}$$

where $z_{kit} = u_{kit} - u_{0it}$ is defined as the difference in utility. The errors $\boldsymbol{\varepsilon}_{it} = (\varepsilon_{1it}, \ldots, \varepsilon_{mit})$ in (4) follow a multivariate logistic distribution $\mathcal{LO}_m$ with logistic marginals and are no longer independent across categories. The variance-covariance matrix $\mathbf{R}$ of $\boldsymbol{\varepsilon}_{it}$ is given by

$$\mathbf{R} = \frac{\pi^2}{6}(\mathbf{I}_m + \mathbf{e}_m \mathbf{e}_m') = \frac{\pi^2}{3} \begin{pmatrix} 1 & 0.5 & \ldots & 0.5 \\ 0.5 & 1 & \ldots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5 & 0.5 & \ldots & 1 \end{pmatrix}.$$

Reformulating the dRUM model in (4) to a multivariate regression model yields:

$$\mathbf{z}_{it} = \mathbf{X}_{it}^f \boldsymbol{\beta} + \mathbf{X}_{it}^r \boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_{it}, \tag{5}$$

where

$$\mathbf{X}_{it}^f = \begin{pmatrix} \mathbf{x}_{1it}^f - \mathbf{x}_{0it}^f & \mathbf{x}_{it} & & 0 \\ \vdots & & \ddots & \\ \mathbf{x}_{mit}^f - \mathbf{x}_{0it}^f & 0 & & \mathbf{x}_{it} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k \end{pmatrix}.$$

To perform Bayesian inference we assume that the fixed effects $\boldsymbol{\beta}$ are a priori normally distributed $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0)$ with known hyperparameters $\mathbf{b}_0$ and $\mathbf{B}_0$. The random effects $\boldsymbol{\xi}_i$ are a priori normally distributed $\boldsymbol{\xi}_i \sim \mathcal{N}(\boldsymbol{\xi}, \mathbf{Q})$ with normal and, respectively, Inverted Wishart ($\mathcal{IW}$) hyperpriors $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{c}_0, \mathbf{C}_0)$ and $\mathbf{Q} \sim \mathcal{IW}(\nu, \mathbf{Q}_0)$. The parameters are estimated by using a data-augmented independence MH-sampler in the spirit of Scott (2011). The basic idea is to construct an independence proposal density for the unknown parameters in model (5) by approximating the error distribution by a normal distribution with the same expectation and variance. Since this data-augmented independence MH-sampler has proved successfully in other applications (e.g. Fussl, Frühwirth-Schnatter and Frühwirth, 2013), we expect a good performance for the current estimation problem and will present first results.

References

Frühwirth-Schnatter, S., and Frühwirth, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. In T. Kneib and G. Tutz (Eds.): *Statistical Modelling and Regression Structures Festschrift in Honour of Ludwig Fahrmeir*, Heidelberg: Physica-Verlag, pp. 111 – 132.

Frühwirth-Schnatter, S., and Frühwirth, R. (2012). Bayesian Inference in the Multinomial Logit Model. *Austrian Journal of Statistics*, **41**, 27 – 43.

Fussl, A., Frühwirth-Schnatter, S., and Frühwirth, R. (2013). Efficient MCMC for Binomial Logit Models. *ACM Transactions on Modeling and Computer Simulation*, **23**, Article 3, 1 – 21.

Grünberger, K. (2009). *Strukturelle Modelle des Arbeitsangebots: Eine Schätzung erwerbsbezogener Präferenzen österreichischer Haushalte.* Diploma thesis, University of Vienna, 73 pages.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.): *Frontiers of Econometrics*, New York: Academic, pp. 105 – 142.

Rabethge, B. (2009). *Die Methode der Mikrosimulation am Beispiel einer Abschaffung des Alleinverdienerabsetzbertrags.* Diploma thesis, University of Vienna, 81 pages.

Scott, S.L. (2011). Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Statistical Papers*, **52**, 87 – 109.

Comparison of GAM, FDA and DFA for water quality data

Kelly Gallacher[1], Claire Miller[1], Marian Scott[1]

[1] School of Mathematics and Statistics, University of Glasgow, Scotland, UK

E-mail for correspondence: k.gallacher.3@research.gla.ac.uk

Abstract: It is often of interest to model temporal trends and seasonal patterns in environmental data sets but it is not always clear which is the best statistical method to use. Three methods (GAM, DFA and FDA) are applied here to model the seasonal pattern for nitrates in river water. We propose a comparison based on comparing the shape of the seasonal patterns using a measure of visual distance, and curvature. The results show that the shapes of the seasonal patterns differ in terms of the location and magnitude of peaks and troughs of the signal.

Keywords: additive models; functional data analysis; dynamic factor analysis; distance; curvature.

1 Introduction

The Environment Agency for England and Wales (EA) monitor river water to ensure compliance with EU regulations (WFD, 2000) by taking measurements of determinands of interest at monitoring stations located along river networks. The EA have provided data for the period 1990 to 2010 for observations of Total Oxidised Nitrogen (TON) made at approximately monthly intervals.
It is often of interest to model the seasonal pattern and temporal trends in environmental spatio-temporal data sets. It is not always clear however which is the best statistical method to use. Three statistical methods (GAM, DFA and FDA) have been applied to a subset of the EA data. Each statistical method incorporates spatial structure in a different way. We propose comparing the estimated seasonal patterns using a measure of distance between curves, and curvature.

2 Statistical Methods

Additive Models (GAM's) enable relationships between explanatory variables and the response to be modelled as smooth functions. Model (1) was fitted to the data from 28 monitoring stations using the `mgcv` package

(Wood, 2006) in R, where $s()$ indicates a smooth term fitted using thin plate penalised regression splines and $\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$. A natural log transformation is used to stabilise variability in the response and a smooth spatial surface is estimated along with a smooth temporal trend and seasonal pattern.

$$log(\text{TON}) = s_1(\text{ Easting, Northing}) + s_2(\text{ Year.day}) + s_3(\text{ Day of year}) + \varepsilon \quad (1)$$

Automatic smoothing techniques such as GCV are not appropriate here due to correlation in the errors and smoothing parameters were chosen so the estimated curve was smooth enough to capture the main shape of the seasonal pattern and flexible enough to capture any interesting features. Further work is required to refine the choice of smoothing parameters. The estimated seasonal pattern was then extracted from the model output.
In addition, GAM's were fitted to each of the 28 monitoring stations separately as in Model (2), where $\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The same smoothing parameters were used for all 28 curves and chosen as described for Model (1).

$$log(\text{TON}) = s_1(\text{Year.day}) + s_2(\text{Day of year}) + \varepsilon, \quad (2)$$

The 28 seasonal patterns were extracted and used to estimated an average seasonal pattern for the whole geographic area was estimated using two techniques where the data of interest are the estimated seasonal patterns and not the original data points.

Functional Data Analysis (FDA) is the analysis of information on curves or functions (Ramsay & Silverman, 1997). The functional mean and functional standard errors were calculated from the 28 individual seasonal curves.

Dynamic Factor Analysis (DFA) is a dimension reduction technique where n time series are modelled using a linear combination of m underlying common trends ($m < n$), and explanatory variables (Zuur et al., 2003). The number of common trends to be estimated and the form of the covariance matrix must be chosen by the user. In this work spatial correlation not already captured by the common trends is modelled with a non-diagonal covariance matrix. The 'best' model is chosen using AIC. DFA was applied to the 28 individual seasonal patterns+residuals estimated using Model (2). The estimated common trends were smoothed using cubic splines and the smoothing parameters were chosen in the same way as described for Model (1).

3 Comparing the curves

Two methods are proposed here to compare the shapes of the seasonal patterns estimated by the three methods in Section 2: (1) A hypothesis test based on a measure of visual distance between curves and (2) curvature.

Seasonal patterns $\mu^{(i)}, i = 1, .., n$ can be compared using a measure of distance, d_V:

$$d_V(\mu^{(i)}, \mu^{(j)}) \equiv \left(\int_{-1}^{1} \delta(i,j)^2 \mathrm{d}t + \int_{-1}^{1} \delta(j,i)^2 \mathrm{d}t \right)^{\frac{1}{2}}$$

For a given time-point t, $\delta(ij)$ is the minimum Euclidean distance between point $\mu^{(i)}(t)$ and all points on $\mu^{(j)}, i \neq j$, and d_V tries to capture how similarity of curves would be assessed 'by eye' (see Minas et al., (2011) for details). For each $\mu^{(i)}$, both axes are rescaled before calculating d_V thus ensuring differences are based on shape and not scale.

Minas et al. (2011) propose the distance based test statistic $\mathrm{DBF}_{\Delta_{d_V}} = \mathrm{tr}(\mathbf{B}_{\Delta_{d_V}})/\mathrm{tr}(\mathbf{W}_{\Delta_{d_V}})$ where $\mathrm{tr}(\mathbf{B}_{\Delta_{d_V}})$ and $\mathrm{tr}(\mathbf{W}_{\Delta_{d_V}})$ are the trace of the distance matrices representing between and within group variability, respectively. $\mathrm{DBF}_{\Delta_{d_V}}$ is used to test the hypotheses $H_0 : d_V(\mu^{(i)}, \mu^{(j)}) = 0$ vs. $H_1 : d_V(\mu^{(i)}, \mu^{(j)}) \neq 0$

Curvature, calculated using (3) where $f'(t)$ and $f''(t)$ are the first and second derivatives respectively of the seasonal patterns at time t, is a measure of how 'bendy' a curve is. Large values indicate turning points on $\mu^{(i)}$, and a straight line is present in $\mu^{(i)}$ where curvature = 0.

$$\kappa(t) = \frac{f''(t)}{[1 + (f'(t))^2]^{\frac{3}{2}}} \tag{3}$$

4 Results

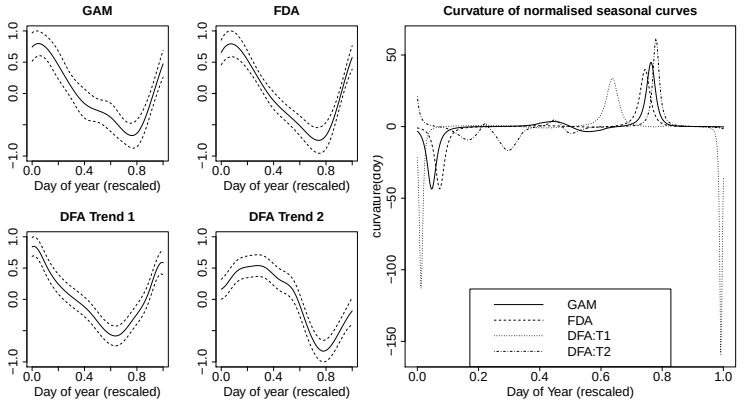


FIGURE 1. Re-scaled seasonal patterns (left and centre, solid) with error bands (dashed). Curvature of re-scaled seasonal patterns (right).

Figure 1 (left and middle) shows the estimated seasonal patterns. The FDA seasonal pattern is slightly smoother than the GAM seasonal pattern. The

best DFA model has a diagonal covariance matrix and 2 common trends suggesting the seasonal pattern is not the same across the whole LHA. The four seasonal patterns were shown to be significantly different ($p < 0.001$) using the hypothesis test described in Section 3. Inspection of curvature plotted for each seasonal pattern (Figure 1) makes it possible to identify the nature of the differences. It appears that the main differences in shape of the seasonal patterns are the time of the turning points (location of peaks and troughs) and how sharp or shallow the turning points are (magnitude of peaks and troughs). The comparison of DFA Trend1 and DFA Trend2 with the GAM and FDA seasonal patterns is particularly useful in identifying that DFA Trend1 picks up the main seasonal pattern and DFA Trend2 captures deviation from this.

5 Conclusions

The proposed comparisons illustrate that each method estimates slightly different seasonal patterns. Since each method models spatial structure in a different way, these differences could be due to the seasonal pattern changing over space. Future work will include repeating these analyses to compare the shape of the long term trend and refining the choice of smoothing parameters. This work will then be extended further to incorporate the river network structure into the models.

Acknowledgments: Special Thanks to the Environment Agency for providing the data and advice and to the EPSRC for funding this work.

References

European Parliament. (2000). Directive 2000/60/EC of the European Parliament, establishing a framework for community action in the field of water policy. *Official Journal of the European Communities*, **327**, 1 – 72.

Minas, C., Waddell, S.J., and Montana, G. (2011). Distance-based differential analysis of gene curves. *Bioinformatics*, **27**, 3135 – 3141.

Ramsay, J.O., and Silverman, B.W. (1997). *Functional Data Analysis.* New York, London: Springer.

Wood, S.N. (2006). *Generalized Linear Models: an introduction with R* (2006). Boca Raton, FL; London: Chapman & Hall/CRC.

Zuur, A.F., Fryer, R.J., Jolliffe, I.T., Dekker, R. and Beukema, J.J. (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*, **14**, 665 – 685.

Penalized Non-Linear Principal Components Analysis for Ordinal Variables

Jan Gertheiss[1] and Henk A.L. Kiers[2]

[1] Department of Animal Sciences, Georg-August-University Goettingen, Germany
[2] Faculty of Behavioural and Social Sciences, University of Groningen, The Netherlands

E-mail for correspondence: jgerthe@uni-goettingen.de

Abstract: Nonlinear principal components analysis (PCA) for categorical data constructs new variables by assigning numerical values to categories such that the proportion of variance in those new variables that is explained by a predefined number of principal components is maximized. We propose a penalized version of nonlinear PCA for ordinal variables that is an intermediate between standard PCA on category labels and nonlinear PCA as used so far. Our approach offers both better interpretability of the nonlinear transformation of the category labels as well as better performance on validation data than unpenalized nonlinear PCA.

Keywords: Categorical Variables; Non-Monotone Principal Components Analysis; Optimal Scaling; Smoothing.

1 Introduction

The objective of principal components analysis (PCA) is to reduce the dimension of the data at hand by finding uncorrelated linear combinations of the original variables. These linear combinations – called *principal components* (PCs) – should explain as much of the variability in the original data as possible. Here we consider ordinal variables, that is, categorical variables with levels that can be reasonably ordered. Though many practitioners simply treat category labels as numeric values and apply standard PCA, this way of analysis may be questionable, since ordinal variables do not have metric scale level.

The idea of nonlinear PCA for categorical data is to construct new variables by assigning numerical values to categories such that the proportion of variance in those new variables that is explained by the first, lets say m, PCs is maximized. This process is called 'optimal quantification', 'optimal scaling' or 'optimal scoring'; cf. Linting et al. (2007). However, while the found transformations – the 'quantifications' – maximize the explained variance on the data at hand, the 'training data', it is by no means clear

that they will also work well on new data. In fact, by simply maximizing the explained variance on the training data, often the found transformations rather account for random fluctuations in the data than for substantial nonlinearity. In addition, the obtained quantifications are sometimes erratic and thus hard to interpret. Here, we propose a penalized version of nonlinear PCA for ordinal variables that is an intermediate between standard PCA on category numbers and nonlinear PCA as described above.

2 Penalized Principal Components Analysis

The idea of PCA is to find uncorrelated, standardized linear combinations $y_{ir} = x_i^{\mathrm{T}} a_r$ of the original data $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$, with the variation in the original data that is explained by the first m vectors of loadings $a_1, \ldots, a_m$, $a_r = (a_{r1}, \ldots, a_{rp})^{\mathrm{T}}$, being as large as possible.
With ordinal variables, vectors x_i contain only integers $1, 2, \ldots$, with entry x_{ij} indicating the level of the jth variable that is observed at the ith subject. As numbers $1, 2, \ldots$ are just class labels, linearity in these labels as assumed by usual PCA is very restrictive and actually not appropriate for categorical data. Therefore, nonlinear PCA constructs new variables by assigning numerical values to categories in terms of $\phi_{ij} = \varphi_j(x_{ij})$, with scaling functions $\varphi_j : \mathbb{N} \to \mathbb{R}$. Then, standard PCA as described above is applied to the recoded variables. To find appropriate, or 'optimal' functions φ_j, the proportion of variance in the transformed variables that is explained by the first m PCs is maximized, with m being fixed at a certain value. For that purpose, it is useful to consider the $a_1, \ldots, a_m$ and corresponding score vectors $y_1, \ldots, y_m$, $y_r = (y_{1r}, \ldots, y_{nr})^{\mathrm{T}}$, as the solution of the least squares problem $L(Y, A) = \sum_j \sum_i (x_{ij} - \sum_r y_{ir} a_{rj})^2 \longrightarrow \min$, with $(Y)_{ir} = y_{ir}$, $(A)_{jr} = a_{rj}$. For nonlinear PCA, criterion $L(\Phi, Y, A) = \sum_j \sum_i (\phi_{ij} - \sum_r y_{ir} a_{rj})^2$ is minimized as a function of matrices A, Y and Φ, with $(\Phi)_{ij} = \phi_{ij} = \varphi_j(x_{ij})$; see Linting et al. (2007). Now, A and Y correspond to loadings and respective PC scores when using the transformed variables. Scaling function φ_j can also be represented by the vector $\theta_j = (\theta_{j1}, \ldots, \theta_{jk_j})^{\mathrm{T}}$ where θ_{jl} is the value that is assigned to category l of the jth variable, k_j denotes the highest level of variable j.
For fixed quantifications Φ, $L(\Phi, Y, A)$ is minimized by the usual PCA solution on data matrix Φ (note, we just replaced X by Φ). For fixed Y and A, minimization of $L(\Phi, Y, A)$ becomes a "regression problem"

$$L(\Phi, Y, A) = \sum_j \sum_i (u_{ij} - z_{ij}^{\mathrm{T}} \theta_j)^2 \longrightarrow \min, \tag{1}$$

with $u_{ij} = \sum_r y_{ir} a_{rj}$, and $z_{ij} = (z_{ij1}, \ldots, z_{ijk_j})^{\mathrm{T}}$ being a design vector of length k_j with entry $z_{ijl} = 1$ if at subject i variable j has value l, and zero otherwise. For finding the final solution, Φ and $\{Y, A\}$ are alternately fixed

at their current value, and it is cycled through the two optimization steps until convergence.

When $L(\Phi, Y, A)$ at (1) is considered, however, only the nominal scale level of the variables is used. In regression problems, it has been proposed to use special penalties to incorporate the covariates' ordinal scale level; see, e.g., Tutz and Gertheiss (2012), and references therein. Similar approaches can be used here. In particular, penalizing nonlinearity in the coefficients as done in Gertheiss and Oehrlein (2011) seems promising. The idea is not to minimize $L(\Phi, Y, A)$ as a function of $\theta = (\theta_1^{\mathrm{T}}, \ldots, \theta_p^{\mathrm{T}})^{\mathrm{T}}$, but its penalized version

$$L_p(\Phi, Y, A) = \sum_j \sum_i (u_{ij} - z_{ij}^{\mathrm{T}} \theta_j)^2 + \lambda J(\theta). \tag{2}$$

For penalty $J(\theta)$ we choose $J(\theta) = \sum_{j=1}^{p} \sum_{l=2}^{k_j - 1} ((\theta_{j,l+1} - \theta_{jl}) - (\theta_{jl} - \theta_{j,l-1}))^2$. By using this penalty, we penalize nonlinearity in the θ-coefficients that belong to the same variable. The strength of penalization is controlled by parameter λ. With $\lambda = 0$, optimal scaling for categorical variables as described above is obtained; with $\lambda \to \infty$, coefficients are forced to be linear, which is equivalent to usual PCA using class labels $1, 2, \ldots, k_j$ for variable j. With $0 < \lambda < \infty$, coefficients are nonlinear but smoother than with unpenalized nonlinear PCA, which makes good sense for ordinal variables, as wiggly coefficient vectors θ_j are hard to interpret.

3 Illustration and Application

We illustrate the proposed method using the `ehd` data from the R package `psy` (Falissard, 2009). The data consists of 269 observations of 20 ordinally scaled variables forming a polydimensional rating scale of depressive mood (Jouvent et al., 1988). Each item is measured on a five-point scale.

We scale the variables by using the presented technique for nonlinear PCA with $m = 5$. Figure 1 shows some of the obtained quantifications for different values of smoothing parameter λ. It is seen that with larger λ quantifications become more and more linear, which is equivalent to standard (linear) PCA using just the category labels.

For evaluating the performance of our approach, we use 5-fold cross-validation. Figure 1 (right) shows the mean proportion of variance explained as a function of penalty parameter λ (on a logarithmic scale) for both the training data as well as the validation data. On the training data, this function is of course monotonically decreasing in λ, as with smaller λ more emphasis is put on the data. For $\lambda = 0$, the original nonlinear PCA approach is obtained, where the explained variance is maximized by construction. On the validation data, however, it's a different story. Here, unpenalized nonlinear PCA (see $\lambda \to 0$) is even worse than standard (linear) PCA, which is obtained for $\lambda \to \infty$. The best results on the validation sets, however, are obtained for λ-values between $1 = 10^0$ and $10 = 10^1$. To

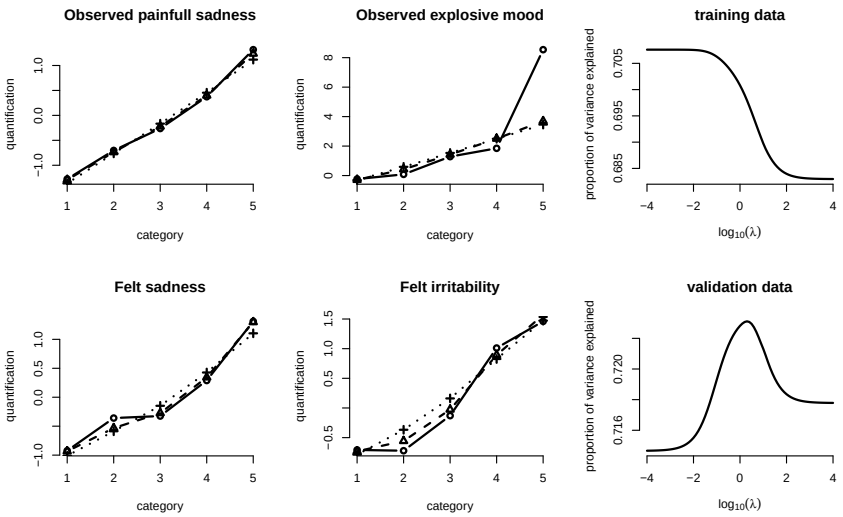


FIGURE 1. Left/middle: Category quantifications obtained by penalized nonlinear PCA for different values of penalty parameter $\lambda = 0$ (solid $\circ$), $\lambda = 1$ (dashed $\triangle$), $\lambda = 10$ (dotted $+$). Right: Mean proportion of variance explained by the first 5 PCs as a function of penalty parameter λ when 5-fold cross-validation is used for performance evaluation.

obtain the final scaling rule, we would use $\lambda = 10^{0.3}$, where the explained variance on the validation data is maximized.

References

Falissard, B (2009). *psy: Various procedures used in psychometry.* R package version 1.0.

Gertheiss, J. and Oehrlein, F. (2011). Testing linearity and relevance of ordinal predictors. *Electronic Journal of Statistics*, **5**, 1935 - 1959.

Jouvent, R., Vindreau, C., Montreuil, M., Bungender, C., and Windlocher, D. (1988). La clinique polydimensionnelle de humeur depressive: Nouvelle version echelle ehd. *Psychiatrie et Psychobiologie*, **3**, 245 - 253.

Linting, M., Meulman, J.J., van der Kooij, A.J., and Groenen, P.J.F. (2007). Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, **12**, 336358.

Tutz, G. and Gertheiss, J. (2012). Rating scales as predictors – the old question of scale level and some answers. *Psychometrika*, accepted for publication.

Selection of mixed beta regression model for longitudinal data

Viviana Giampaoli[1], Olga Usuga[2], Patricia Bertolotto[3]

[1] Departamento de Estatística, Universidad de São Paulo, Brasil ,
[2] Departamento de Ingeniería Industrial, Universidad de Antioquia, Colombia ,
[3] Departamento de Matematica, Universidad Nacional de Cordoba, Argentina

E-mail for correspondence: vivig@ime.usp.br

Abstract: Generalized Additive Models for Location, Scale and Shape (GAMLSS) is a general class of statistical models that include a beta regression model. To select predictors and covariate effects in GAMLSS are used different strategies. In this work we proposed a methodology to select mixed beta regression models for longitudinal data and we study the performance of the strategies used in GAMLSS. A simulation study comparing the strategies for model selection in relation to percentage of times the correct model was chosen and the observed efficiency is presented. According to the simulation results the proposed method seems to be a promissory method.

Keywords: Akaike information criterion; beta distribution; model selection.

1 Introduction

The GAMLSS proposed by Rigby and Stasinopoulos (2005) are an alternative when it is intended to model longitudinal data related to rates or proportions. Strategies for model selection in GAMLSS are perfomed by different functions on the package `gamlss`. In this work, we propose one strategy to select mixed beta regression models for longitudinal data and we study the performance of this strategy and the strategies for model selection in GAMLSS.

2 Mixed beta regression model

Let y_{ij} be the response value for subject i at time t_{ij}, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, n$. In the beta random intercept model, it is assumed that the conditional distribution of y_{ij} given $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2})^T$ follows a distribution beta with a density determined by $f(y; \mu, \sigma) = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}$, where $\alpha = \mu(1-\sigma^2)/\sigma^2$, $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$, $0 < \mu < 1$, and $0 < \sigma < 1$. We will assume the following model $y_{ij} \mid \gamma_{i1}, \gamma_{i2} \overset{\text{ind}}{\sim} \text{Be}(\mu_{ij}, \sigma_{ij})$, $\gamma_{i1} \overset{\text{i.i.d}}{\sim} N(0, \lambda_1^2)$,

$\gamma_{i2} \overset{\text{i.i.d}}{\sim} N(0, \lambda_2^2)$, where λ_1 and λ_2 are the standard deviation of random effects. This model is obtained by assuming that the mean and dispersion parameter of y_{ij} satisfy the following functional relations $logit(\mu_{ij}) = \boldsymbol{x}_{ij1}^T\boldsymbol{\beta}_1 + \gamma_{i1}$, $(\sigma_{ij}) = \boldsymbol{x}_{ij2}^T\boldsymbol{\beta}_2 + \gamma_{i2}$ where $\boldsymbol{x}_{ij1} = (x_{ij11}, x_{ij21}, \ldots, x_{ijp_11})^T$ and $\boldsymbol{x}_{ij2} = (x_{ij12}, x_{ij22}, \ldots, x_{ijp_22})^T$ contain values of explanatory variables, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{21}, \ldots, \beta_{p_11})^T$ and $\boldsymbol{\beta}_2 = (\beta_{12}, \beta_{22}, \ldots, \beta_{p_22})^T$ are the fixed parameter vectors, and $\boldsymbol{\gamma}_i$ is the random intercept vector.

3 Model selection

3.1 Model selection in gamlss

Rigby and Stasinopoulos (2005) discussed a variety of strategies to select predictors and covariate effects in GAMLSS. They proposed to use a generalized version of the AIC, defined as $\text{GAIC}(a) = -2\ell(\hat{\boldsymbol{\theta}}) + a\text{df}$. The GAIC consists of the negative log-likelihood and a fixed penalty factor a multiplied by the total effective degrees of freedom df. Note that $a = 2$ or $a = \log(n)$ leads to the classical AIC or Bayesian information criterion respectively. The strategies to select predictors and covariate effects are performed by the functions `stepGAIC`, `stepGAICAll.A` and `stepGAICAll.B` of the `gamlss` package. In these functions we can define or not the scope of the models, that is the range of models examined in the stepwise search.

3.2 Model selection proposed

We propose an appropriate methodology to select beta regression models with random effects for longitudinal data that combines the model selection method of the GAMLSS and the linear mixed models with longitudinal data proposed by Rigby and Stasinopoulos (2005) and Ryoo (2010), respectively. The addition and elimination the terms on the regression structure is performed by forward and backward procedures. The methodology is performed by the function `AICP` that is implemented in this work and has the following strategy:

1. Build a model for μ using the following approach I. In this approach we start with a random intercept model, then, step by step, we add the time variable, the covariates, the interactions between the time variable and covariates and finally we include the random effects.

2. Given the model for μ build a model for σ using the approach I.

3. Given the model for σ check whether the terms for μ are needed using the approach II. In this approach we start with a saturated random intercept model, then, step by step, we eliminate the terms of the polynomial time and we add the covariates, the interactions between the time variable and covariates and finally we add the random effects.

4 Simulation Study

In this section, we compare the performance of the different strategies to select predictors and covariate effects in mixed beta regression model for longitudinal data through a Monte Carlo simulation study. The strategies that we compare are performed by the functions `stepGAIC` (G), `stepGAICAll.A` (A), `stepGAICAll.B` (B) and `AICP` (P). The criterion that we use is the classical AIC. In the first three functions we can define or not the scope of the models and in the function `AICP` is not defined this argument. We consider n observations taken from each of N subjects. Observations $y_{i1}, y_{i2}, \ldots, y_{in}$ are taken at time points $t_{i1}, t_{i2}, \ldots, t_{in}$, $i = 1, 2, \ldots, N$. The time was generated from $t = (n-1)/n$ and the data were generated according to the model: $\text{logit}(\mu_{ij}) = \beta_{11} + t_{ij}\beta_{21} + \gamma_{i1}$, $\text{logit}(\sigma_{ij}) = \beta_{12} + t_{ij}\beta_{22}$, where $i = 1, 2, \ldots, N$,$j = 1, 2, \ldots, n$ and $\beta_{11} = 0.10$, $\beta_{21} = -0.10$, $\beta_{12} = -0.15$, $\beta_{22} = 0.15$ and $\gamma_{i1} \sim N(0, \lambda^2)$. The number of observations per subject is $n = 3$, the number of subjects are $N = 10, 20, 50$ and the standard deviation of random effect is $\lambda = 0.5$. Three models were fitted to the data, the regression structures for the parameter μ are given in Table 1 and the regression struture for the parameter σ is the same of the previous model. The continuous covariates x_{ij1}, x_{ij2}, and x_{ij3}, which are not time dependent, are generated from $\text{U}(0, 0.5)$. In each experiment, we run 1000 iterations and we assess the performance of the strategies through the percentage of times the correct model was chosen and the observed efficiency $\text{OE} = \|\mu_c - \mu_s\|^2$, where μ_c and μ_s are the mean vectors of the correct and selected models, respectively. The simulations were performed in R.14.3.

TABLE 1. Fitted models.

Model	Regression structure for μ	P.
M1	$\text{logit}(\mu_{ij}) = \beta_{11} + t_{ij}\beta_{21} + x_{ij1}\beta_{31} + \gamma_{i1}$	4
M2	$\text{logit}(\mu_{ij}) = \beta_{11} + t_{ij}\beta_{21} + x_{ij1}\beta_{31} + x_{ij2}\beta_{41} + \gamma_{i1}$	5
M3	$\text{logit}(\mu_{ij}) = \beta_{11} + t_{ij}\beta_{21} + x_{ij1}\beta_{31} + x_{ij2}\beta_{41} + x_{ij3}\beta_{51} + \gamma_{i1}$	6

The Table 2 display the percentage of times the correct model is chosen when three models are fitted to the data, respectively. The results show that when we not define the range of models examined in the stepwise search the `gamlss` functions exhibit weakest performance, returning the lowest percentage of correct model identification and low efficiency. It is also been that when the number of subjects N is large the percentage of times the correct model is chosen and the observed efficiency increased. The increased of the covariates on the fitted model decrease the percentage of times the correct model is chosen and the efficiency for the strategies A, B and P.

TABLE 2. Percentage of times the correct model is chosen when three models are fitted to the data.

		Without scope			With scope			
Model	N	G	A	B	G	A	B	P
M1	10	0	14	15	89	77	54	62
	20	0	20	18	100	79	64	63
	50	0	21	17	100	79	72	70
M2	10	2	6	4	80	50	30	40
	20	8	14	13	100	63	42	47
	50	9	17	11	100	65	53	53
M3	10	3	4	6	43	43	28	46
	20	9	9	10	34	34	21	47
	50	11	11	11	47	46	30	49

5 Conclusions

When we compared the strategies P and the strategies G, A and B without define the range of models examined in the stepwise search on the percentage of times the correct model is chosen the strategy P has better performance. Thus, the use of the proposed methodology is important in case there are not predetermined models. In another situation, the percentages presented by strategies G and A are slightly higher than P. However, in general, for the observed efficiency the strategy P had a performance similar to A and slightly higher than G.

Acknowledgments: The authors acknowledge the brazilian agencies INCT-FCx, CNPq and FAPESP by the financial support.

References

Rigby, R. A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society, Series C*, **54**, 507 – 544.

Ryoo, J. H. (2010). Model Selection with the Linear Mixed Effects Model for Longitudinal Data. PhD thesis, University of Minnesota, USA.

Dynamics of the Optimal Maintenance Policy under Imperfect Repair Models

Gustavo L. Gilardoni[1], Maria Luiza G. de Toledo[2], Marta A. Freitas[3], Enrico A. Colosimo[3]

[1] Universidade de Brasília, Brazil
[2] Universidade Federal de Ouro Preto, Brazil
[3] Universidade Federal de Minas Gerais, Brazil

E-mail for correspondence: gilardon@unb.br

Abstract: It is discussed both determination and practical implementation of an optimal preventive maintenance policy under imperfect repair which takes into account the information provided by observing the failure history of a repairable system.

Keywords: Poisson process; power law process; repairable systems

1 Introduction

We will assume that at any time the operator of the repairable system can decide to perform a (perfect) *preventive maintenance* (PM) action, which leaves the system in *as good as new* condition. A PM policy specifies the moments at which the system is maintained. On the other hand, every time the system fails between two successive PM actions, it is necessary to perform a repair action to bring it back to operating condition. Usually it is assumed that such action is a *minimal repair* (MR), in the sense that it leaves the system at the same condition it was immediately before failing (i.e., *as good as old*). Under MR, the optimal PM periodicity can be determined after noting that, in this case, the failure history of the system follows a nonhomogeneous Poisson process (NHPP). (Gilardoni and Colosimo, 2007) The argument goes essentially as follows. Let $N(t)$ be the number of failures in $(0, t]$ and define the mean function $H(t) = EN(t)$, the intensity $h(t) = H'(t)$, and assume that $h(\cdot)$ is increasing. If the PM and MR actions have fixed costs C_{PM} and C_{MR} and one decides to perform PMs at every τ units of time, the expected cost per unit of time would be

$$C(\tau) = [C_{PM} + C_{MR}H(\tau)]/\tau\,. \tag{1}$$

Differentiating, one finds the optimal periodicity τ_{OPT} as the solution of

$$D(\tau) = \tau h(\tau) - H(\tau) = C_{PM}/C_{MR}\,. \tag{2}$$

In practice, h and H are unknown and hence have to be estimated and plugged in (2) to obtain an estimate of τ_{OPT} (cf. Gilardoni and Colosimo, 2007, 2011 or Gilardoni et al, 2013).
Assuming MR actions at each failure may be too restrictive in practical applications. More realistic models assume that, after each failure, an *imperfect repair* (IR) action leaves the system at some point between the *as good as old* and the *as good as new* conditions. For instance, Kijima et al (1988) introduced a class of *virtual age models* to describe the operation of a system under IR. In its simplest form, the model assumes a baseline intensity $\lambda(t)$ and that, after a failure, the repair reduces the actual age of the model by a factor θ. Hence we have actual age t and virtual age $V(t) = t - t_{N(t)} + \theta t_{N(t)}$, so that the intensity of failure at time t is actually $\lambda[V(t)]$, where $0 < t_1 < \cdots < t_n < \cdots$ are the failure times. The parameter θ measures the degree of the IR; $\theta = 0$ and $\theta = 1$ correspond respectively to PM and MR.
Focusing on *optimal* PM ***periodicity***, may be somewhat shortsighted, in the sense that it does not take into account the history of the process and hence makes sense only in a scenario where $N(t)$ has independent increments. If, on the other hand, the IR model results in a process with dependent increments, then the history has to be considered when determining an optimal PM policy.
The rest of this note is organized as follows. Section 2 deals with the optimal PM policy under *any* IR model. The main finding, —see equation (6) below, states there exists a constant, independent of the history of the process, such that a PM action should be performed whenever the conditional intensity of the IR failure process given the available information attains that constant. Finally, Section 3 deals briefly with the associated inference problem (i.e., estimation of the optimal PM policy) in the virtual age model.

2 Optimal PM policy under IR

As before, supposse a failure process $N(t)$ subject exclusively to IR actions and define $H(t|s_0) = H(t|\mathcal{H}(s_0^-)) = E[N(s_0 + t) - N(s_0)|\mathcal{H}(s_0^-)]$, where $\mathcal{H}(s_0^-)$ is the failure history up to the moment immediately before s_0. In other words, $H(t|s_0)$ is the expected number of failures during the next t units of time given the history of the process up to s_0. We will consider the following assumptions: First, both PM and IR actions are performed instantaneously. Also, the operator of the system can perform a PM action at any time, meaning that there is no delay between the diagnostic of the necessity and the performance of a PM action. Second, the costs of the PM and IR actions are independent of the history of the system and have expectation C_{PM} e C_{IR} respectively. Third, for each s_0, $H(t|s_0^-)$ is differentiable for every t and its derivative $h(t|s_0^-)$ is continuous and strictly increasing.

A *maintenance policy* for the time interval (s_0, S) specifies the number of PMs n and its moments $s_0+\tau_1 < s_0+\tau_1+\tau_2 < \cdots < s_0+\tau_1+\cdots+\tau_n$, where $\tau_i > 0$ and $\sum_{i=1}^{n} \tau_i < S - s_0$. We will write $M(s_0, S) = (n; \tau_1, \ldots, \tau_n)$. Given a maintenance policy $M(s_0, S) = (n; \tau_1, \ldots, \tau_n)$, its expected cost given $\mathcal{H}(s_0^-)$ is

$$C[M(s_0, S)] = nC_{PM}+C_{IR}\left\{H(\tau_1|s_0)+\sum_{j=2}^{n} H(\tau_j|0)+H(S-s_0-\tau_1-\cdots-\tau_n|0)\right\}. \tag{3}$$

A maintenace policy $M_{OPT}(s_0, S)$ is optimal if $C[M_{OPT}(s_0, S)] \leq C[M(s_0, S)]$ for every other $M(s_0, S)$. Since the first PM action renews the system, it should be clear that, if $M_{OPT}(s_0, S) = (n, \tau_1, \ldots, \tau_n)$, then $M_{OPT}(0, S - s_0 - \tau_1) = (n - 1, \tau_2, \ldots, \tau_n)$. This shows that, to solve the general case, it is important to understand the problem with $s_0 = 0$.

The problem without information ($s_0 = 0$). In order to obtain the optimal policy we will proceed in two stages. First, we will assume the number n of PMs fixed and will obtain the optimal PMs moments. Then, we will discuss how to obtain the optimal n for an infinite horizon (i.e., for $S \to \infty$)

Considering n fixed in (3) and differentiating with respect to τ_i we get that $h(\tau_i|0) = h(S - \tau_1 - \cdots - \tau_n)$ for $i = 1, \ldots, n$. Since $h(\cdot|0)$ is strictly increasing, this implies that $\tau_i = S/(n+1)$. In other words, in this case the optimal policy specifies in fact an optimal period.

Now, to obtain the optimal n, we substitute the previous times again in (3) to obtain $c(n) = nC_{PM} + (n + 1)C_{IR}H(\frac{S}{n+1}|0) = S\,\frac{n+1}{S}\,[C_{PM} + C_{IR}H(\frac{S}{n+1}|0)] - C_{PM}$. Hence, to obtain the optimal n one should minimize $c^*(n) = \frac{n+1}{S}\,[C_{PM} + C_{IR}H(\frac{S}{n+1}|0)]$— compare with (1). Since the function $f(\tau) = [C_{PM} + C_{IR}H(\tau|0)]/\tau$ is convex, it follows that the optimal n is either $n_{OPT} = [S\tau^* - 1]$ or $n_{OPT} = [S\tau^* - 1] + 1$, where τ^* is the minimizer of $f(\tau)$ and $[a]$ is the integer part of a. Putting these considerations together and letting $S \to \infty$, it follows that the optimal PM policy calls for PM actions at every

$$\tau_{OPT} = \lim_{S\to\infty} \frac{S}{n_{OPT} + 1} = B^{-1}(C_{PM}/C_{IR}), \tag{4}$$

where we have defined $B(t) = th(t|0) - H(t|0) = \int_0^t uh'(u|0)\,du$. This is essentially the same solution given in (2).

The general case ($s_0 > 0$). Consider now an optimal policy $M(s_0, S) = (n, \tau_1, \ldots, \tau_n)$. For large S it follows from (4) that $\tau_2 = \cdots = \tau_n = B^{-1}(C_{PM}/C_{IR})$ and $n = (S - s_0 - \tau_1)/B^{-1}(C_{PM}/C_{IR})$. Hence, to obtain the optimal policy we have now to optimize with respect to the remaining variable τ_1. Substituting in (3) and differentiating with respect to τ_1 we get that the optimal policy $M(s_0, S)$ must satisfy

$$h(\tau_{1,OPT}|s_0) = h[B^{-1}(C_{PM}/C_{IR})|0], \tag{5}$$

$n_{OPT} = \frac{S - s_0 - \tau_{1,OPT}}{B^{-1}(C_{PM}/C_{IR})}$ and $\tau_{2,OPT} = \cdots = \tau_{n_{OPT},OPT} = B^{-1}(C_{PM}/C_{IR})$. Note that in a purely dynamical implementation of this solution, the only relevant equation is (5). This is because one monitors the history of the system (s_0) up to a time which solves (5), at which moment a PM action is performed and a renewal occurs. Then, one monitors again the history of the renewed system (again s_0) and so on. In other words, we never get to apply the last two equations. For this reason, we call equation (5) the *fundamental law of preventive maintenance*.

Moreover, although (5) may suggest that in order to implement the optimal policy one has to evaluate $h(t|s_0)$ for every possible s_0, if PM actions can be scheduled without delay, one only need actually to evaluate $h(t|t)$. More precisely, denote by $\tau_{OPT}(s_0)$ the solution of (5). In implementing the optimal policy, the operator of the system monitors the failure history and at each s_0 computes $\tau_{OPT}(s_0)$. In general he or she would have that $\tau_{OPT}(s_0) > s_0$ and will keep going without performing a PM. In other words, the only way that a PM action would be eventually performed is if for some s_0 one has that $\tau_{OPT}(s_0) = s_0$. In other words, a PM action would be performed if and only if $\lim_{s_0 \to \tau_1} h(\tau_1|s_0) = h(B^{-1}(C_{PM}/C_{IR})|0)$. This is quite nice, because usually $h(t|t)$ is much easier to compute than $h(t|s_0)$. For instance, for the simple virtual age model, it is easy to show that $h(t|t^-) == \lambda[t - (1-\theta)t_{N(t)}] = \lambda[V(t)]$ (see Kijima et al, 1988). Hence, (5) becomes now

$$\lambda[\tau_{1,OPT} - (1-\theta)t_{N(\tau_{1,OPT})}] = h[B^{-1}(C_{PM}/C_{IR})|0]\,, \tag{6}$$

or, remembering that $V(t) = t - t_{N(t)} + \theta t_{N(t)}$,

$$\tau_{1,OPT} - t_{N(\tau_{1,OPT})} = \lambda^{-1}\{h[B^{-1}(C_{PM}/C_{IR})|0]\} - \theta\, t_{N(\tau_{1,OPT})}\,. \tag{7}$$

In other words, a PM action will occur whenever the virtual age attains the value $\lambda^{-1}\{h[B^{-1}(C_{PM}/C_{IR})|0]\}$.

3 Statistical inference for the virtual age model

Consider the virtual age model with a baseline power law intensity $\lambda(t) = (\beta/\eta)\,(t/\eta)^{\beta-1}$ with $\beta > 1$. Suppose that the system is observed up to time T and observed failure times at times $0 < t_1 < \cdots < t_n < T$. Let $V(t_i) = \theta t_i$ and $V(t_i^-) = t_i - (1-\theta)t_{i-1}$. The likelihood is

$$L(\beta, \eta, \theta) = \left(\prod_{i=1}^{n} \lambda[V(t_i^-)] e^{\Lambda[V(t_i)] - \Lambda[V(t_i^-)]} \right) e^{\Lambda[T - (1-\theta)t_n]}\,, \tag{8}$$

where $\Lambda(t) = \int_0^t \lambda(u)\,du = (t/\eta)^{\beta}$ (see Toledo et al, 2013, for details).

In order to estimate the right hand side of (6) we proceed as follows. First, we maximize numerically the likelihood (8) to obtain the MLEs $\hat{\beta}$, $\hat{\eta}$ and

$\hat{\theta}$. Then, to estimate $H(t|0)$, we simulate many systems with the estimated parameters and use a Nelson-Aalen estimator. Following Gilardoni and Colosimo (2011), an estimate of $h(t|0)$ which takes into account the monotonicity constraint can now be obtained as the derivative of the *greatest convex minorant* (GCM) of the Nelson-Aalen estimate $\hat{H}(t|0)$. This estimates $\hat{H}(t|0)$ and $\hat{h}(t|0)$ can now be used to obtain $\hat{B}(t) = t\hat{h}(t|0) - \hat{H}(t|0)$. Inverting $\hat{B}(t)$ we get an estimate of $B^{-1}(C_{PM}/C_{IR})$. Since we already have computed $\hat{h}(t|0)$, this means that we obtain an estimate of the right hand side of (6). Likewise, the right hand side of (7) can be estimated now after noting that $\lambda^{-1}(x) = \eta[\eta x/\beta]^{1/(\beta-1)}$.

Two comments are in order here. First, the Monte Carlo simulation has to be done only once during the entire process, because the right hand sides of (6) and (7) involve only $H(t|0)$ and $h(t|0)$. Second, the size of the simulation can be taken large enough to make its precision at least an order of magnitude larger than the precision of the MLEs of the parameters, so that in practice the relevant uncertainty in the final estimates would depend only on the precision of the MLEs. Therefore, it is not too difficult to compute precision and confidence intervals of the optimal PM times. Full details about this are omitted here for reason of space but will be given in the full length paper.

Acknowledgments: This work was partially financed by CNPq, CAPES, FINATEC and UnB/DPP grants

References

Gilardoni, G. L. and E. A. Colosimo (2007). Optimal maintenance time for repairable systems. *Journal of Quality Technology*, **39**, 48 – 53.

Gilardoni, G. L. and E. A. Colosimo (2011). On the superposition of overlapping Poisson processes and nonparametric estimation of their intensity function. *Journal of Statistical Planning and Inference*, **171**, 3075 – 3083.

Gilardoni, G. L., M. D. de Oliveira and E. A. Colosimo (2013). Nonparametric estimation and bootstrap confidence intervals for the optimal maintenance time of a repairable system. *Computational Statistics & Data Analysis* (to appear).

Kijima, M., H. Morimura and Y. Suzuki (1988). Periodical replacement problem without assuming minimal repair. *European Journal of Operational Research*, **37**, 194 – 203.

Toledo, M. L. G. de, M. A. Freitas, E. A. Colosimo and G. L. Gilardoni (2013). Optimal periodic maintenance policy under imperfect repair: A case study on off-road engines. *Unpublished manuscript*.

Penalized spline smoothing for delay in Pulmonary Tuberculosis diagnosis

Dulce Gomes [1], Patrícia A. Filipe[1], Carla Nunes [2], Bruno de Sousa [2,3]

[1] Escola de Ciência e Tecnologia, Univ. de Évora, CIMA/UE, Portugal
[2] Escola Nac. de Saúde Pública, CMDT/IHMT, Univ. Nova de Lisboa, Portugal
[3] Faculdade de Psicologia e Cincias da Educao, Univ. de Coimbra, Portugal

E-mail for correspondence: dmog@uevora.pt

Abstract: In this work, Cox's proportional hazards model with penalized spline functions for time-varying covariate effects are presented in order to model the time between the onset of the first symptoms and the diagnosis of Pulmonary Tuberculosis (PTB) in Portugal. Using this approach revealed a better fitting of the model to the data than Cox's classical model. Results showed that an earlier diagnosis tends to be associated with the following: being male, no alcohol consumption, being diagnosed through active screening, having had a previous PTB treatment, being HIV positive, and being a smoker.

Keywords: Pulmonary Tuberculosis Control; Delay in Diagnosis; Survival Analysis; P-spline

1 Introduction

In the current study we intend to explore several factors that may explain the time between the onset of the first symptoms and the diagnosis of Pulmonary Tuberculosis (PTB) ("delay" in diagnosis), taking into account the dynamics of the endemic (Filipe *et al.*, 2012).

In Cox's classical model the assumption of proportional hazards often restricts it applications, since it means that covariate effects remain constant over survival time. Many techniques have been developed to overcome the proportional hazards constraint, with one such approach being to consider covariate effects as smooth functions that can be modeled by splines. Thus, our study considered an extended Cox model which incorporate a penalized spline smoothing function (P-spline). This approach was originally introduced by O'Sullivan (1986), but the procedure achieved general recognition with the paper by Eilers and Marx (1996). Our main purpose was to study the delay in the diagnosis of PTB, when several covariates such as age, sex, factors historically related to the disease (drugs consumption, smoking, HIV infection), incidence rate of the disease (*per* 10^{-5}), diagnosis type (1-passive screening resulting from having symptoms, 2-active screening resulting from contacts with children 10-13 years of age and other groups,

and 3-other) were considered. The variables age and incidence rates were treated as continuous. To be able to capture the nature of the data, our model did not established any constraints on how the continuous variables affect such delay were established. For instance, a child under 10 or 11 years of age is, in general, regularly observed by a clinician, thus it is expected to have a smaller delay time in the diagnosis of PTB. Also, it seems reasonable to admit that in geographical areas with high incidence rates, a closer supervision over suspicious symptoms occurs.

2 Penalized Spline Smoothing Function

Let $s(z)$ be a smooth function and, for convenience of notation, z a single covariate. The smooth term s is typically estimated using spline functions. The proposed model is an extension of the usual Cox model. It is used to model the *hazard function* $h(t)$ - the risk of occurring the event after time t -, and can be written as $h(t|X) = h_0(t)\exp(BX + s(z))$, where $h_0(t)$ represents the baseline hazard function, $X = (X_1, X_2, \ldots, X_k)$ are the vectors of time-constant covariates, $B = (B_1, B_2, \ldots, B_k)$ is the associated parameters vector. The interpretation of the B coefficients is similar to the original Cox model. Since the smooth function parameters do not have a direct interpretation, graphical methods can be used to describe the relationship between z and the risk.
A spline function $s(z)$ of degree k is a piecewise polynomial, where the polynomial pieces (all of degree k) join together at the knots $\psi_i, i = 1, \cdots, \Psi$, thus $s(z) = \alpha_{0i} + \alpha_{1i}z + \alpha_{2i}z^2 + \cdots + \alpha_{ki}z^k$, $\psi_i \leq z < \psi_{i+1}$. In order to ensure the continuity at the knots, it is common to define the spline of degree k, with knots $\psi_i, i = 1, \cdots, \Psi$, as a linear combination of spline basis functions, $g_i(z) = (z - \psi_i)^k$, for $z > \psi_i$ ($g_i(z) = 0$, otherwise), $i = 1, \cdots, \Psi$. Thus, the spline function $s(z)$ can be written as follows

$$s(z) = \alpha_{0i} + \alpha_{1i}z + \alpha_{2i}z^2 + \cdots + \alpha_{ki}z^k + \sum_{j=2}^{\Psi-1} \gamma_j g_j(z).$$

The main goal is to estimate the coefficients of the smooth function through a certain penalizing criterion. The most commonly used is the penalized residual sum of squares criterion (Schoenberg, 1964),

$$\sum_{i=1}^{n}(y_i - s(z))^2 + \theta \int (s^{(d)}(z))^2 dz,\ d = 1, \cdots,\ k-1,$$

where y denotes the dependent variable and θ the parameter that controls the curve smoothing. This parameter takes values on the domain $[0, \infty[$, where $\theta = 0$ means the curve passes trough all the points and $\theta \to \infty$ results in a straight line.

3 Main Results

Our main database corresponds to a total of 28,612 notified cases of PTB, from Portugal mainland between 2000 and 2010. About 15% of the initial

database was excluded due to missing values, measuring errors or severe outliers related to the delay in diagnosis. The inclusion conditions were notified PTB cases with delay in diagnosis between 1 and 285 days. About 72% are male. From the 16,630 cases to which HIV tests results are available, approximately 18% are HIV positive. From the aproximately 22,000 individuals that explicitly answered were validated, 9.7% were drug users, 18.6% admitted that were alcohol consumers, 12% were smokers, 1.9% individuals were homeless and 2% were inmates. For the variable number of previous treatments for PTB, we have observed that 88.8% never had a treatment before (new case of PTB) and the remaining individuals had at least 1 previous treatment. The percentage of individuals diagnosed through passive screening were 88.4%, with 4% being diagnosed through active screening, and the remaining due to other reasons. Based on the selected cases, the delay in diagnosis (in days) has a mean, median and standard deviation of 71.3, 57.0 and 53.3 days, respectively. Notice the extremely high standard deviation.

To check whether time-varying effects lead to an improvement of the fit when compared to time-constant effects (Cox's classical model), we have used the AIC criterion. Our analysis showed a better fit when considering time-varying effects. The assumptions and suitability of the model were confirmed.

In table 1 we show the estimated regression coefficients (for the time-constant covariates) $\hat{\beta}_i$, the estimated multiplier $\exp(\hat{\beta}_i)$ and the corresponding p-value for the Wald test. The values of $\exp(B)$ revealed an increased probability of earlier diagnosis for males (13%), non consumers of alcohol (7%), diagnosed through active screening (1%, as compared to passive screening) and for individuals that had a previous PTB treatment (6%). Being HIV positive and smoker also increase the risk of an earlier diagnosis of PTB (aprox. 11% and 21%, respectively).

TABLE 1. Estimated regression coefficients in the selected model.

Covariates	$\hat{\beta}$	$\exp(\hat{\beta})$	p-value
Sex (F*)	0.12498	1.130	< 0.001
Alcohol (Yes*)	0.06763	1.070	< 0.001
Smoker (Yes*)	-0.19309	0.824	< 0.001
HIV (Yes*)	-0.10448	0.901	< 0.001
Motif diag2	0.01313	1.010	0.770
Motif diag3	-0.10477	0.901	< 0.001
NTrt1+	0.05415	1.060	0.043

* reference class

The plots in figure 1 represent the estimates of the time-varying regression coefficients as a function of time t (in days) for the delay (solid line), together with 95% confidence intervals (dashed line). They suggest that the effects of age and incidence rates on delay are not constant throughout the

period under study. There is an increasing effect of age on delay in early ages (approx. under 5 years of age), decreasing after that, becoming below zero at around the age of 40. The effect of the incidence rates on the delay showed a slow increase from low incidence values, becoming above zero around 50. The abnormal behavior in the graph for of the highest incidence rates cases are due to a small number of observations.

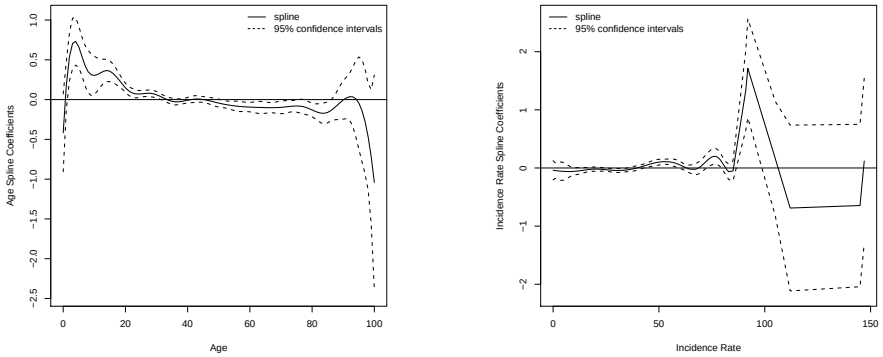


FIGURE 1. Age (left) and Incidence Rate (right) effects on delay.

The results were obtained using software R. For the main results, the functions **coxph** and **pspline** of package *survival* were applied. Regarding the **pspline** function, the AIC was the chosen criterion used to determine θ.

4 Conclusion

An extended hazard Cox model was applied to the delay in diagnosis of PTB where P-splines were used to model the effects of age and incidence rates as smooth functions. The resulted model showed a better fit to our data when compared to Cox's classical model.

Acknowledgments: This work, within the research project PTDC/SAU-SAP/116950/2010, was financed by FCT/MCTES.

References

Eilers, P. H. C., Marx, B. D. (1996). Flexible smoothing with B-splines and penalties.*Stat. Science 11(2), pp. 89 – 121.*

Filipe, P.A., Gomes, D., Nunes, C., Silva, M., de Sousa, B., and Briz, T. (2012). Delay in diagnosis of Pulmonary Tuberculosis in Portugal. *Proc. of the 27th IWSM, pp. 501 – 06.*

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r:P519-527). *Stat. Science* **1**, *pp. 502 – 18.*

Schoenberg, I. (1964). On interpolation by spline functions and its minimum properties. *Int. Ser. of Num. Analysis* ***5****, pp. 109 – 129.*

Change-point analysis for in environmental time series

A. Manuela Gonçalves[1], Marco Costa[2], Lara Teixeira[1]

[1] Departamento de Matemática e Aplicações, CMAT-Centro de Matemática, Universidade do Minho, Portugal
[2] Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, CMAF-UL, Portugal

E-mail for correspondence: mneves@math.uminho.pt

Abstract: Change-points are present in many environmental time series. Time variations in environmental data are complex and they can hinder the identification of the so-called change-points when traditional models are applied to this type of problems. In this study, it is proposed an alternative approach for the application of the change-point analysis by taking into account this data structure (seasonality and autocorrelation) based on the Schwarz Information Criterion (SIC). The approach was applied to time series of surface water quality variables measured at eight monitoring sites.

Keywords: Change-point analysis; SIC; Autocorrelation; Seasonality; Mean and variance shift.

1 Introduction

In this study is proposed the application of the Schwarz Information Criterion (SIC) to detect the change-point in mean and variance in time series of water quality variables. The data concerns the River Ave hydrological basin situated in the Northwest of Portugal, where monitoring has become a priority in water quality planning and management in this watershed. The water quality variable analyzed is Dissolved Oxygen (DO), one of the most important variables in assessing surface water quality in a river's hydrological basin (Costa and Gonçalves, 2011 and Gonçalves and Costa, 2012), measured (in milligrams per liter (mg/l)) monthly from January 1999 to December 2011 in eight monitoring sites: Cantelães (CANT), Taipas (TAI), Ferro (FER), Golães (GOL), Vizela Santo Adrião (VSA), Riba d'Ave (RAV), Santo Tirso (STI) and Ponte Trofa (PTR). In this work, the behavior study of the time series of DO water quality variable is addressed in line with the research of Gonçalves and Costa (2011), Gonçalves and Alpuim (2011), who recently studied trend alterations in environmental variables, including time series of water quality variables. By performing an

exploratory analysis, we concluded that the DO observed values over time (each time series consists at most of 156 observations) presented changes in mean and/or variance in the series (in particular between 2004 and 2006). As regards the average, it apparently increases or decreases according to the monitoring site, but it reduces the variability of the observations in all monitoring sites, more evidently on some of them. Another important feature is the indication of a seasonal component. This is due to the seasonal relationship between DO concentration with the weather patterns throughout the year, particulary temperature changes and precipitation intensity.

2 The informational approach

In order to detect changes in time series, the case of a change-point in both the mean and the variance, the aim is to test the following hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_n = \mu \quad \wedge \quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2 \tag{1}$$

versus the alternative hypothesis

$$\begin{gathered} H_1 : \mu_I = \ldots = \mu_1 = \mu_k \neq \mu_{k+1} = \ldots = \mu_n = \mu_{II} \\ \wedge \\ \sigma_I^2 = \sigma_1^2 = \ldots = \sigma_k^2 \neq \sigma_{k+1}^2 = \ldots = \sigma_n^2 = \sigma_{II}^2. \end{gathered} \tag{2}$$

Based on Akaike's work, in 1978 Schwarz proposed the Schwarz Information Criterion (SIC). The SIC is defined as following

$$\mathrm{SIC}_j = -2 \ln\, L(\hat{\Theta}_j) + p_j \ln n, \quad j = 1, 2, \ldots, M, \tag{3}$$

where n is the sample size. This criterion is based on the maximum likelihood function of a given model penalized by the number of parameters that are estimated in the model. Under H_0, the SIC is denoted by $\mathrm{SIC}(n)$ and it is obtained as

$$\mathrm{SIC}(n) = -2 \ln\, L_0(\hat{\mu}, \hat{\sigma}^2) + 2 \ln n, \tag{4}$$

$$= n \ln 2\pi + n \ln \sum_{i=1}^{n} (X_i - \overline{X})^2 + n + (2 - n) \ln n. \tag{5}$$

where $L_0(\hat{\mu}, \hat{\sigma}^2)$ is the maximum likelihood function with respect to H_0. Under H_1, the SIC is denoted by $\mathrm{SIC}(k)$ for fixed k, $2 \leq k \leq n - 2$, is obtained as

$$\mathrm{SIC}(k) = -2 \ln\, L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}_I^2, \hat{\sigma}_{II}^2) + 4 \ln n \tag{6}$$

$$= n \ln 2\pi + k \ln \hat{\sigma}_I^2 + (n - k) \ln \hat{\sigma}_{II}^2 + n + 4 \ln n, \tag{7}$$

where $L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}_I^2, \hat{\sigma}_{II}^2)$ is the maximum likelihood function under H_1. The decision to accept H_0 or H_1 is based on the principle of minimum criterion. According to the information criterion principle, we are going to estimate the position of the change-point k such that SIC(k) is the minimal. Then, the estimation of the position of the change-point by $\hat{k}$ is given by $SIC(\hat{k}) = \min_{2 \leq k \leq n-2} SIC(k)$. In order to assess significance, a critical value c_α can be included in the decision rule for a significance level α, where $c_\alpha \geq 0$. The model with a change-point SIC(k) is selected if

$$\min_{2 \leq k \leq n-2} SIC(k) + c_\alpha < SIC(n) \tag{8}$$

otherwise, the model with no change-point SIC(n) is more reasonable. The approximate critical values for different series lengths that were obtained through the asymptotic distribution are presented in Chen and Gupta (1999).

3 Change-point detection procedure

The DO time series present statistical properties as a constant mean and seasonality whose parameters must be estimated at the same time. Thus, the adjusted model is

$$X_t^{(M1)} = \mu + s_t + \epsilon_t, \quad t = 1, \ldots, n, \tag{9}$$

where μ is the global series mean, s_t is the seasonal component and ϵ_t is a white noise with $\mathrm{E}(\epsilon_t^2) = \sigma^2$. The change-points detection considers the errors series $\hat{\epsilon_t} = X_t^{(M1)} - \hat{\mu} - \hat{s}_t, t = 1, \ldots, n$.
The aim is to detect change-points in both the mean and the variance, i.e., to test the null hypothesis (1) *versus* the alternative hypothesis (2), through SIC application to the new series $\{\hat{\epsilon_t}\}_{t=1,\ldots,n}$, corresponding the SIC($n$) to the model (5) and the SIC(k) to the model (7). For a better understanding of the differences between information criterion values of the different models, will be represented SIC(k) values and the SIC(n) $-$ c_α values for two significance levels, $\alpha = 0,05$ and $\alpha = 0,01$, and they are represented in the graphics (Figure 1) by horizontal reference lines. If, statistically, a change-point is detected, a second model will be adjusted to the original data,

$$X_t^{(M2)} = \mu_t + s_t + \epsilon_t, \quad t = 1, \ldots, n, \tag{10}$$

where s_t is the seasonal component for $t = 1, \ldots, n,$,

$$\mu_t = \begin{cases} \mu_I & \text{if} \quad t \leq k \\ \mu_{II} & \text{if} \quad t > k \end{cases} \quad \text{and } \epsilon_t = \begin{cases} N(0, \sigma_I^2) & \text{if} \quad t \leq k \\ N(0, \sigma_{II}^2) & \text{if} \quad t > k \end{cases}.$$

After the adjustment of the model (10), it follows the binary segmentation process with the second change-points detection, in the two errors sequences, before and after change-point. However, the data analysis was conservative by taking into account the performed simulation study (not presented in this article) and in agreement with Beaulieu et al. (2012): the presence of autocorrelation in the observations, even weak ones ($\phi \approx 0.3$), tends to originate the detection of false change-points. Thus, in this study when SIC(n) and SIC(k) values are very close, even if the change-point is statistically significant, we decided not to consider the existence of a second change-point.

4 Results and discussion

Taking into account the previous studies about this hydrological basin (Gonçalves and Alpuim 2011, Costa and Gonçalves 2011, Gonçalves and Costa 2011) and the inspection of data series, it is reasonable to consider that the series do not present trends (for instance, a linear trend). Moreover, works that compare DO data series (and other water quality variables, Gonçalves and Alpuim 2011) in different water monitoring sites concluded that there is a common pattern in the evolution of these variables considering the same hydrological basin. Thus, it is reasonable to consider the same change-point model for all eight water monitoring sites. The linear model M1 (9) was adjusted to DO data series (original data, without any transformation).

TABLE 1. Results of change-point procedures (n_i-number of observations in site i, $\hat{k} = \text{argmin}_{2 \leq k \leq 154} \text{SIC}(k)$).

Site	n_i	SIC(n)	$\hat{k}$	SIC($\hat{k}$)	$c_{5\%}$	change-point
CANT	150	345.67	73	287.25	6.802	Jan/2005
TAI	151	321.79	70	307.96	6.791	Oct/2004
RAV	155	456.53	89	436.03	6.746	May/2006
STI	154	523.33	89	493.54	6.757	May/2006
PTR	154	482.48	83	443.22	6.757	Nov/2005
FER	152	356.58	70	341.12	6.780	Oct/2004
GOL	151	348.35	77	312.58	6.791	May/2005
VSA	151	358.44	74	321.06	6.791	Feb/2005

The SIC procedure was applied to all series according to the methodology shown above, considering the asymptotic critical values at a 5% significance level. Table 1 summarizes the results of SIC procedures. For all series was detected a change-point significant considering the respective critical value. One should notice that in all series the differences SIC(n) − SIC($\hat{k}$) are

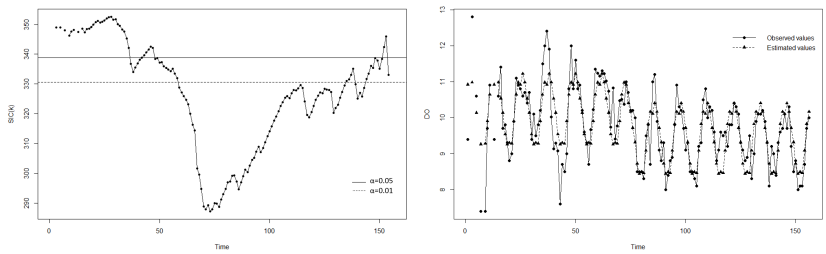


FIGURE 1. SIC(k) values for Cantelães series and adjustment of linear model considering the change-point in Cantelães.

clearly superior to the approximate critical values at a 5% significance level. Moreover, considering a 1% significance level, only the difference $\mathrm{SIC}(n) - \mathrm{SIC}(\hat{k})$ relatively to the Taipas series (TAI) is lower than the approximate critical value of $c_{1\%}$ (for instance, $c_{1\%} \approx 15.079$ when $n = 150$). Thus, change-point procedures are assertive about the existence of a change-point in both mean and variance in each series, even considering a conservative significance level. For instance, Figure 1 represents SIC(k) values, $2 \leq k \leq 154$, for Cantelães series and the values $\mathrm{SIC}(n) - c_\alpha$ with $\alpha = 1\%, 5\%$.
As the assumptions of normality and independence are not present in some time series, a simulation study was carried out (not presented in this paper) in order to evaluate the methodology's performance when applied to non-normal data series with or without time correlation.

References

Beaulieu, C., Chen, J., and Sarmiento, J.L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Phil. Trans. R. Soc. A.*, **370**, 1228 – 1249.

Costa, M. and Gonçalves, A.M. (2011). Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk Assessment*, **25**, 151 – 163.

Gonçalves, A.M. and Alpuim, T. (2011). Water quality monitoring using cluster analysis and linear models. *Environmetrics*, **22**, 933 – 945.

Gonçalves, A.M. and Costa, M. (2012). Predicting seasonal and hydrometeorological impact in environmental variables modelling via Kalman filtering. *Stochastic Environmental Research and Risk Assessment*, (doi: 10.1007/s00477-012-0640-7).

Chen, J. and Gupta, A.K. (1999). Change point analysis of a Gaussian model. *Statistical Papers*, **40**, 323 – 333.

Another View on Conditional Correlations

Radek Hendrych[1]

[1] Dept. of Probability and Math. Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Prague 8, Czech Republic

E-mail for correspondence: hendrych@karlin.mff.cuni.cz

Abstract: The aim of the paper is to introduce an innovative approach to conditional covariance and correlation modelling, which is useful e.g. in the multivariate GARCH context. The suggested two-step method is based on the LDL decomposition of the conditional covariance matrix and state space modelling with the associated Kalman recursions. Together, they provide a dynamic orthogonal transformation of observed multivariate time series. This time-varying transformation indeed simplifies further (second step) conditional variance modelling of stochastic vector data due to their simultaneously uncorrelated elements.

Keywords: conditional correlation; conditional covariance; Kalman recursion.

1 Model Framework

Consider a multivariate stochastic vector process $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ of dimension $(n \times 1)$. Denote $\mathcal{F}_t$ the σ-algebra generated by observed time series $\{\mathbf{X_t}\}$ up to and including time t. In this framework, assume the following model

$$\mathbf{X}_t = \mathbf{H}_t^{1/2}\mathbf{Z}_t, \tag{1}$$

where $\mathbf{H}_t$ is the $(n\times n)$ positive definite conditional covariance matrix of $\mathbf{X}_t$ given $\mathcal{F}_{t-1}$. Furthermore, one supposes that $\{\mathbf{Z}_t\}$ is an $(n\times 1)$ i.i.d. stochastic vector process such that it has following first two moments: $\mathsf{E}(\mathbf{Z}_t) = \mathbf{0}$ and $\mathsf{cov}(\mathbf{Z}_t) = \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of order n.
In the model (1), the conditional and the unconditional moments of $\mathbf{X}_t$ can be easily derived:

$$\mathsf{E}(\mathbf{X}_t|\mathcal{F}_{t-1}) = \mathbf{0}, \quad \mathsf{cov}(\mathbf{X}_t|\mathcal{F}_{t-1}) = \mathbf{H}_t^{1/2}(\mathbf{H}_t^{1/2})^{\mathrm{T}} = \mathbf{H}_t, \tag{2}$$

$$\mathsf{E}(\mathbf{X}_t) = \mathbf{0}, \quad \mathsf{cov}(\mathbf{X}_t) = \mathsf{E}(\mathbf{H}_t), \quad \mathsf{cov}(\mathbf{X}_t, \mathbf{X}_{t+h}) = \mathbf{0},\ h \neq 0. \tag{3}$$

Hence, from (2), it is evident that $\mathbf{H}_t^{1/2}$ is any $(n \times n)$ positive definite matrix such that $\mathbf{H}_t$ is the conditional covariance matrix of $\mathbf{X}_t$ given $\mathcal{F}_{t-1}$. From the theoretical point of view, $\mathbf{R}_t$ (the conditional correlation matrix of $\mathbf{X}_t$ given $\mathcal{F}_{t-1}$) can be obtained by the straightforward normalization of the conditional covariance matrix $\mathbf{H}_t$.

2 Conditional Covariances and Correlations via LDL Decomposition and State Space Modelling

Generally, the main task is to find out the time-varying behaviour of the conditional covariance matrix $\mathbf{H}_t$ with special regard to modelling of the conditional correlations $\mathbf{R}_t$. From the mathematical point of view, a sort of models, which work with the parameter matrices $\mathbf{H}_t$ or $\mathbf{R}_t$ directly, is indeed worthy of interest. Particularly, both $\mathbf{H}_t$ and $\mathbf{R}_t$ must be symmetric and positive (semi)definite. In addition, the conditional correlation matrix $\mathbf{R}_t$ must have unit diagonal elements. Undoubtedly, such requirements might bring really tough constraints into estimation, especially in the case of higher dimension. For this reason, it is more effective to consider some other representations of these matrices which naturally simplify or completely eliminate these restrictions (Tsay, 2005).

2.1 Orthogonal Transformation Using LDL Decomposition

Following the algebraic theory, each real symmetric positive definite matrix has a unique LDL decomposition, see e.g. Harville (1997). Namely, let the conditional covariance matrix $\mathbf{H}_t$ have the LDL reparametrization in the standard form, i.e.

$$\mathbf{H}_t = \mathbf{L}_t \mathbf{D}_t \mathbf{L}_t^{\mathrm{T}} \quad [= (\mathbf{L}_t \mathbf{D}_t^{1/2})(\mathbf{L}_t \mathbf{D}_t^{1/2})^{\mathrm{T}} = \mathbf{H}_t^{1/2}(\mathbf{H}_t^{1/2})^{\mathrm{T}}], \tag{4}$$

where $\mathbf{L}_t$ is a $(n \times n)$ lower triangular matrix with the unit diagonal and $\mathbf{D}_t$ is a $(n \times n)$ diagonal matrix with positive elements $d_{ii,t}$ on its diagonal. In particular, $\det(\mathbf{L}_t) = 1$, $\mathbf{L}_t$ is invertible and the inverted matrix $\mathbf{L}_t^{-1}$ is also a $(n \times n)$ lower triangular matrix with unit diagonal elements. Point out that the decomposition (4) requires no parameter constraints for $\mathbf{H}_t$ being symmetric and positive definite since this is guaranteed in such a structure.
The form of the matrix $\mathbf{L}_t$ provides a natural orthogonal transformation:

$$\mathbf{Y}_t = \mathbf{L}_t^{-1} \mathbf{X}_t \quad [= \mathbf{L}_t^{-1} \mathbf{H}_t^{1/2} \mathbf{Z}_t = \mathbf{D}_t^{1/2} \mathbf{Z}_t]. \tag{5}$$

The transformation $\mathbf{Y}_t$ has with respect to the declared assumptions and (2)-(5) the following conditional and unconditional moments:

$$\mathsf{E}(\mathbf{Y}_t | \mathcal{F}_{t-1}) = \mathbf{0}, \quad \mathsf{cov}(\mathbf{Y}_t | \mathcal{F}_{t-1}) = \mathbf{D}_t, \tag{6}$$

$$\mathsf{E}(\mathbf{Y}_t) = \mathbf{0}, \quad \mathsf{cov}(\mathbf{Y}_t) = \mathsf{E}(\mathbf{D}_t), \quad \mathsf{cov}(\mathbf{Y}_t, \mathbf{Y}_{t+h}) = \mathbf{0},\ h \neq 0. \tag{7}$$

2.2 Model Estimation

For dynamic estimation of some of the unknown quantities in the LDL decomposition (4) in the framework of the model (1) with given entire

sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$, state space modelling should be really useful. The issue of state space models and associated Kalman recursions is elaborated in many different publications, see e.g. Brockwell and Davis (2002) and the references given therein.

In regarding to (5), assume the following dynamic discrete-time linear state space representation (generalized analogy of the recurrent OLS estimator):

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \tag{8}$$

$$\mathbf{X}_t = \mathbf{G}_t\boldsymbol{\beta}_t + \mathbf{Y}_t,\ t = 1, \ldots, T. \tag{9}$$

Denote $\boldsymbol{\beta}_t$ the $(n(n-1)/2 \times 1)$ vector containing all unknown row elements of $\mathbf{L}_t$. The $(n \times (n(n-1)/2)$ matrix $\mathbf{G}_t$ clearly includes only zeroes and the elements of $\mathbf{X}_t$ due to (5). The state equation (8) is simply given by the multivariate random walk. Indeed, one can suppose some other (and more sophisticated) versions of the state equation (8).

Recapitulate crucial assumptions of this model: $\{\boldsymbol{\beta}_1, ((\boldsymbol{\varepsilon}_t)^{\top}, \mathbf{Y}_t^{\top})^{\top}\}_t$ is a sequence of uncorrelated random vectors with finite second moments and $\mathsf{E}\boldsymbol{\varepsilon}_t = \mathbf{0}$, $\mathsf{cov}(\boldsymbol{\varepsilon}_t) = \mathbf{Q}_t$, $\mathsf{E}\mathbf{Y}_t = \mathbf{0}$, $\mathsf{cov}(\mathbf{Y}_t) = \mathbf{R}_t$ and also $\mathsf{cov}(\boldsymbol{\varepsilon}_t, \mathbf{Y}_t) = \mathbf{0}$. The matrix $\mathbf{R}_t$ is supposed to be diagonal due to (7).

The initial state vector $\boldsymbol{\beta}_1$ is assumed to be random with the expected value $\mathsf{E}(\boldsymbol{\beta}_1) = \mathbf{0}$ and the variance $\mathsf{var}(\boldsymbol{\beta}_1) = \kappa\mathbf{I}_{n(n-1)/2}$, $\kappa \to \infty$, i.e. the so-called standard diffuse prior. The covariance matrices $\mathbf{Q}_t$ and $\mathbf{R}_t$ could be captured essentially by constant parameter matrices which are estimated via a maximum likelihood procedure (Brockwell and Davis, 2002).

In the given framework, the standard Kalman recursive formulas for filtering, predicting and smoothing can be used to obtain corresponding estimators of $\boldsymbol{\beta}_t$ and consequently also the transformed vector $\mathbf{Y}_t$ with simultaneously uncorrelated elements. The conditional variances $d_{ii,t}$ of $Y_{i,t}$ can be therefore viewed by some advanced univariate methods, e.g. by means of GARCH models (Tsay, 2005).

3 Empirical Results

To examine the empirical performance of the introduced approach to conditional covariance and correlation modelling based on the LDL decomposition and the state space representation, the following empirical application is considered. In particular, the daily correlations between log-returns on stocks and bonds are investigated. In general, there is no consensus about how stocks and long term bonds are related. Short-run correlations are obviously affected, e.g. by new announcements. The long-run correlations between these two type of assets should be state dependent, e.g. driven by macroeconomic factors. The way how the correlation links respond to these factors might be changed over time (Engle, 2009). The daily logarithmic returns on the S&P 500 index (GSPC) and 30-year bond futures

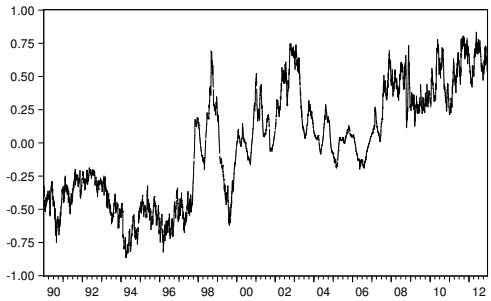


FIGURE 1. The estimated conditional correlations (S&P 500 and 30Y bonds).

(TYX) from 3 January 1990 to 14 December 2012 are observed (i.e. 5753 observations, available on http://finance.yahoo.com).
The estimated conditional correlations are presented graphically in Figure 1. Generally, the time-varying correlations are mostly negative during the 90's, rather positive after the year 2000 and positive at the end of the observed period.
The estimated model structure was additionally verified by several test criteria, e.g. the Ljung-Box test or the ARCH-LM test (Tsay, 2005), and compared with other common methods of dynamic correlation modelling. Thus, one could conclude that the proposed technique based on the LDL decomposition and the state space model (8)-(9) seems to be suitable in this framework and that it is at least competitive with other methods.

Acknowledgments: This work was supported by the Czech Science Foundation (the grant GA P402/12/G097) and by the grant SVV 2013-267315.

References

Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting*. New York: Springer.

Engle, R.F. (2009). *Anticipating Correlations: A New Paradigm for Risk Management*. Princeton, N.J.: Princeton University Press.

Harville, D.A. (1997). *Matrix Algebra from a Statistician's Perspective*. Secaucus, N.J.: Springer.

Tsay, R.S. (2005). *Analysis of Financial Time Series*. Hoboken, N.J.: Wiley-Interscience.

Posterior Singular Spectrum Analysis (PSSA)

Lasse Holmström[1], Ilkka Launonen[1]

[1] Department of Mathematical Sciences, University of Oulu, Oulu, Finland

E-mail for correspondence: Lasse.Holmstrom@oulu.fi

Abstract: We find structure in a time series using a combination of Singular Spectrum Analysis (SSA) and Bayesian modeling.

Keywords: Time series;SSA;Bayesian analysis;Credible features;Oscillations.

1 Introduction

Singular spectrum analysis (SSA) (e.g. Golyandina et al. 2001) can be used to decompose a time series into a trend, periodic and quasiperiodic components, and noise. However, as SSA is essentially an algebraic, non-statistical technique, one cannot be sure whether the suggested signal components are true features of the underlying phenomenon or just noise. We propose to extend SSA by combining it with Bayesian modeling and posterior analysis of the credibility of the suggested underlying time series features.

The basic SSA algorithm consists of four steps. First, a so-called trajectory matrix is formed from segments of the time series by applying a sliding window. The window length is a parameter specified by the user. Second, using singular value decomposition, the trajectory matrix is represented in an optimal basis as a sum of eigentriples formed from singular values and their associated eigenvectors. The eigentriples are then divided into groups that represent the underlying features of the time series. Finally, the group sums are transformed back to time series vectors which are then taken to represent the salient components of the original time series.

Harmonics in the time series produce pairs of eigentriples whose singular values are similar in magnitude. Typically, these eigentriples are not fully separated from noise. In addition, noise can produce pairs of eigentriples that resemble an oscillation and, when the noise is heavy, their singular value magnitude can exceed that of the actual signal. Therefore, statistical inference is needed to confirm the genuineness of the oscillatory patterns found by SSA.

Our approach is to combine SSA with Bayesian posterior simulation. Thus, suppose we have available a sample generated from the posterior distribution of a time series. First, fix the sliding time window length and compute

the SSA eigentriples. This can be done either by using the observed noisy time series or, in some cases, the posterior mean of the sample. Then, to find interesting signal components, project the posterior sample on the subspaces defined by the eigentriples and make inferences about the credible features in these projections, such as local maxima, minima, trends, and oscillations. Their credibility is assessed based on the posterior distribution of the slope, as represented by the slopes of the projected sample. In this way we can for example test the credibility of an apparent trend or an oscillation with some phase and frequency. We call this method Posterior SSA (PSSA) and have demonstrated its potential with examples based on artificial and real data (cf. Holmström and Launonen 2012). In Section 3 we discuss an example that involves mean Pacific sea level changes between 1992 and 2011.

2 SSA and PSSA

Consider a time series $F = (f_0, f_1, \ldots, f_{N-1})^T$ of length N and a window length $L \leq N$. The trajectory matrix is an $L \times (N-L+1)$ matrix X whose jth column is the subseries $(f_{j-1}, \ldots, f_{L+j-2})^T$. Its singular value decomposition is $X = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T$ where d is the number of positive singular values and the eigentriples $(\sqrt{\lambda_i}, U_i, V_i)$ contain the singular values $\sqrt{\lambda_i}$ in decreasing order, together with their associated left and right singular vectors. SSA partitions the eigentriples into groups that appear to correspond to meaningful features of F. The sums of the matrices $\sqrt{\lambda_i} U_i V_i^T$ within each group are finally averaged skew-diagonally to produce the trajectory matrices of the desired structural time series components. Our PSSA algorithm instead first infers the credible features corresponding to the individual eigentriples, discards those that appear to arise from noise, and then combines the remaining components into structural features according to SSA guidelines.

We start with a sample $\{G^{(1)}, \ldots, G^{(n)}\}$ from the posterior distribution of a time series G of interest, $G = (g_0, g_1, \ldots, g_{N-1})^T$. In our examples, this sample can be obtained in two different ways: 1) a Bayesian model is built for an observed noisy time series F and the resulting posterior $p(G|F)$ is sampled or, 2) data only indirectly associated with G is used for modeling and the resulting posterior is sampled. An example of the first case is demonstrated below and the second case is illustrated in Holmström and Launonen (2012) by the post Ice Age mean July temperature time series modeled on the basis of fossil abundance data from lake sediments.

The idea in PSSA is to first perform SSA on F or, in the indirect data case, on the posterior mean, and then to analyze for credible interesting features the posterior distributions of the time series components corresponding to the resulting eigentriples. The credibility of the features is summarized by maps that indicate where the slope of the component time series in question

is deemed to be negative or positive with posterior probability at least α, where α varies between 0.5 and 1 (see bottom panel of Figure 1).

3 An example

We apply PSSA to find credible features of mean sea level change in the Pacific from the end of 1992 to the end of 2011. The time series contains 699 values, averaging three or four measurements per month. It was obtained form the CU Sea Level Research Group web-site and displayed in the top panel of Figure 1.

The observed time series F is modeled as $F = G + \varepsilon$, where $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$ and $\sigma^2 \sim \mathrm{Inv} - \chi^2(\nu_0, \sigma_0^2)$, with the hyperparameters ν_0 and σ_0^2 chosen to reflect our prior belief of the value of σ^2. For the underlying time series G one uses a smoothing prior that penalizes for roughness as measured by the variability of the second differences (Holmström and Launonen 2012).

Considering the components in the middle panel of Figure 1, the first two contain seemingly the trend and the sum of the 3rd and 4th represents the seasonal sea level oscillation. Thus, the uniformly high credibility in their respective maps is not surprising (Figure 1, bottom panel). The component pairs (7,10) and (8,9) have wavelengths of 4.16 and 1.64 years, respectively, and may correspond to the ENSO-related cycles reported in Unal and Ghil (1995). The rest of the components with high credibility areas are more difficult to assess. The 12th and the 15th components could possibly form a pair with a period of approximately 2.47 years which would correspond to a third oscillation reported in Unal and Ghil (1995), although in this case they are not fully separated from noise which can be seen in their wiggliness. Their credibility maps also show some gray areas which could perhaps relate to the inherent noise. As for other potentially interesting pairs, their lack of robustness (pair (23,24)) or irregularity (pair (11,22)) make them unlikely to correspond to any true oscillation.

References

CU Sea Level Research Group. (2012). Global and regional mean sea level time series. http://sealevel.colorado.edu/. Accessed in February, 2012.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. (2001). *Analysis of Time Series Structure, SSA and related techniques*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Holmström, L., and Launonen, I. (2012). Posterior singular spectrum analysis. Submitted for publication. Available on-line at http://cc.oulu.fi/∼llh/preprints/PSSA.pdf.

Unal, Y.S., and Ghil, M. (1995). Interannual and interdecadal oscillation patterns in sea level. *Climate Dynamics*, 11:255–278, 07/1995.

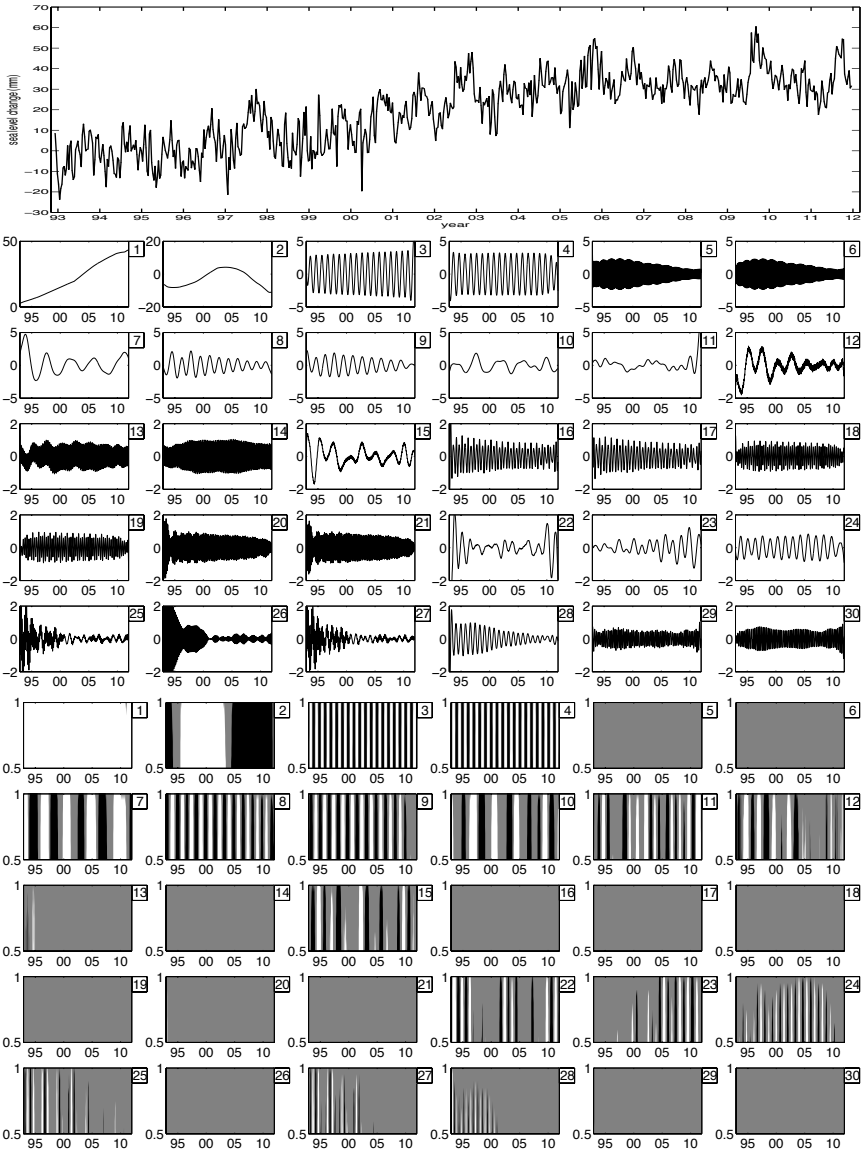


FIGURE 1. Top panel: The mean sea level change in the Pacific as estimated from satellite altimeter readings. Middle panel: The first 30 SSA components, calculated with the window length $L = 349$. Bottom panel: The PSSA credibility maps of the components in the middle panel. In each map, time is on the horizontal axis and the credibility level α is on the vertical axis. White and black indicate credibly positive and negative slopes, respectively, and gray indicates a slope which is not credibly different from zero. The inference about the credible features of the components is performed jointly over all maps.

A Zero-Inflated and Overdispersed Marginalized Model for Correlated Counts

Samuel Iddi[1], Geert Molenberghs[2,1]

[1] I-BioStat, KU Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium
[2] I-BioStat, Universiteit Hasselt, Agoralaan 1, 3500 Hasselt, Belgium

E-mail for correspondence: Samuel.Iddi@med.kuleuven.be

Abstract: Iddi and Molenberghs (2012) merged the attractive features of the so-called combined model of Molenberghs *et al* (2010) and the marginalized model of Heagerty (1999) for hierarchical non-Gaussian data with overdispersion. In this model, the fixed-effect parameters retain their marginal interpretation. Lee *et al* (2011) also developed an extension of Heagerty (1999) to handle zero-inflation from count data, using the hurdle model. To bring together all of these features, a marginalized, zero-inflated, overdispersed model for correlated count data is proposed. Using an empirical dataset, it is shown that the proposed model leads to important improvements in model fit.

Keywords: Marginal multilevel model; Random effects model; Overdispersion; Poisson model; Zero-Inflation

1 Introduction

Count data are gathered in a multitude of settings. For their univariate form, a generalized linear model (GLM) based on the Poisson distribution is regularly assumed, a member of the exponential family. Four features have called for extension. First, because empirical data generally exhibit more heterogeneity than that provided by the mean-variance relationship of the Poisson (overdispersion, but underdispersion is also possible), a collection of extensions has been proposed, such as the negative binomial (NB).
Second, the occurrence of zeros beyond what is predicted by the Poisson are often encountered. Models addressing this are, for example, the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB).
Third, assuming measurements are taken hierarchially, within-unit association is likely present. The generalized linear mixed model (GLMM) is a commonly used random-effecs model to address this. While this model is well established, further complication arises when overdispersion and zero inflation are also present. To address this, overdispersion, Molenberghs *et al* (2010) introduced the combined model (CM) that decomposes the Poisson mean into two multiplicative components, one for each phenomenon.

Fourth, by including individual-specific random effects into the predictor, the fixed effects no longer have a marginal interpretation but are interpreted conditional upon the random effects. We present a model that, while making use of the aforementioned random effects, still admits a marginal interpretation. This multilevel marginal model (MMM) approach is based on Heagerty (1999). This model further simultaneously accounts for overdispersion and zero-inflation. The model is illustrated with real data.

2 Zero-Inflated, Overdispersed, Marginalized Multilevel Model

Let Y_{ij} denote count $j = 1, \ldots, n_i$ for cluster $i = 1, \ldots, N,$, following a Poisson distribution with mean number of events λ_{ij}. We formulate a model that allows for all four issues mentioned in the introduction (Iddi and Molenberghs 2012). The proposed model is:

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij}^m + (1 - \pi_{ij}^m) f_i(0|\lambda_{ij}^m) & \text{if} \quad y_{ij} = 0, \\ (1 - \pi_{ij}^m) f_i(y_{ij}|\lambda_{ij}^m) & \text{if} \quad y_{ij} = 1, 2, \ldots \end{cases}$$

where the marginal mixing probability π_{ij}^m and marginal Poisson mean $\lambda_{ij}^m = E(Y_{ij})$ are related to covariates: $\text{logit}(\pi_{ij}^m) = x'_{1ij}\beta^m$ and $\log(\lambda_{ij}^m) = x'_{2ij}\alpha^m$. Next, a conditional specification follows:

$$P(Y_{ij} = y_{ij}|\theta_{ij}, b_i) = \begin{cases} \pi_{ij}^c + (1 - \pi_{ij}^c) f_i(0|\theta_{ij}, b_{1i}, \lambda_{ij}^c) & \text{if} \quad y_{ij} = 0, \\ (1 - \pi_{ij}^c) f_i(y_{ij}|\theta_{ij}, b_{1i}, \lambda_{ij}^c) & \text{if} \quad y_{ij} = 1, 2, \ldots \end{cases}$$

where the probit $\pi_{ij}^c = \Phi^{-1}(\Delta_{1ij} + z'_{1ij}b_{1i})$ and $\lambda_{ij}^c = \theta_{ij}\exp(\Delta_{2ij} + z'_{2ij}b_{2i})$. The overdispersion random effect, $\theta_{ij} \sim \text{Gamma}(u_{ij}, v_{ij})$ is introduced in the Poisson model. For $b_i = (b_{1i}, b_{2i})' \sim N(0, D)$ and based on

$$\lambda_{ij}^m = \int_b \int_\theta \theta_{ij}\exp(\Delta_{ij} + z'_{ij}b_i) dG_\theta dF_b = \int_b E(\theta_{ij})\exp(\Delta_{ij} + z'_{ij}b_i) dF_b \quad (1)$$

where $G_\theta(\cdot)$ and $F_b(\cdot)$ are the cumulative distribution function of θ_{ij} and b_i respectively, we derive: $\Delta_{1ij} = \sqrt{1 + z'_{1ij}Dz'_{1ij}}\Phi^{-1}\left[\text{expit}(x'_{1ij}\beta^m)\right]$ and $\Delta_{2ij} = -\log(u_{ij}v_{ij}) + x'_{2ij}\alpha^m - \frac{1}{2}z'_{2ij}Dz'_{2ij}$. Thanks to the probit link, closed forms exist. The marginal mean still uses the logit, enabling an odds-ratio interpretation.

3 Estimation

We proceed via maximum likelihood. The observed data likelihood for subject i, conditional on the overdispersion random effect is:

$$f_i(\beta, \alpha, D, \phi) = \int_b \prod_{j=1}^{n_i} f(y_{ij}|b_i) f(b_i|D) db_i,$$

TABLE 1. *Epilepsy Trial. Parameter estimates (standard errors) for the marginalized models (bottom). RE: random effect.*

Effect	Par.	MMM Est.(s.e.)	Zero-Inflated MMM Est.(s.e.)	Combined MMM Est.(s.e.)	Zero-Inflated Comb. MMM Est.(s.e.)
		Poisson Part			
Interc. placebo	α_{00}	1.396(0.189)	1.375(0.170)	1.476(0.196)	1.428(0.183)
Slope placebo	α_{01}	-0.014(0.004)	-0.004(0.005)	-0.025(0.008)	-0.012(0.007)
Interc. treatment	α_{10}	1.226(0.190)	1.378(0.172)	1.220(0.197)	1.337(0.186)
Slope treatment	α_{11}	-0.012(0.004)	-0.007(0.005)	-0.019(0.008)	-0.005(0.007)
Slope diff.	$\alpha_{01} - \alpha_{11}$	0.002(0.006)	-0.003(0.007)	0.013(0.011)	0.008(0.010)
Std. Dev. RE	σ_1	1.076(0.086)	0.973(0.082)	1.063(0.087)	1.009(0.086)
		Zero-Inflated Part			
Intercept	β_0		-2.296(0.296)		-2.428(0.321)
Slope	β_1		0.066(0.017)		0.066(0.018)
Std. Dev. of RE	σ_2		1.254(0.192)		1.292(0.208)
Overd. Par.	$v = \frac{1}{u}$			0.406(0.0348)	0.179(0.018)
Correlation	ρ		-0.138(0.1601)		-0.080(0.167)
AIC		-6810	-7222	-7664	-7682

from which the likelihood follows. The distribution of Y_i conditional on b_i and marginal over θ_{ij} is given for the zero-inflated combined model by:

$$f(y_{ij}|b_i) = I(y_{ij}=0)\pi_{ij} + (1-\pi_{ij})\begin{pmatrix} u_j + y_{ij} - 1 \\ u_j - 1 \end{pmatrix} \times \left(\frac{v_j}{1+\kappa_{ij}v_j}\right)^{y_{ij}} \left(\frac{1}{1+\kappa_{ij}v_j}\right)^{u_j} \kappa_{ij}^{y_{ij}}.$$

In fitting the MMM, the conditional distributions are specified by replacing the terms $x'_{1ij}\beta$ and $x'_{2ij}\alpha$ in the zero-inflated version of the combined model with the analytical expressions for Δ_{1ij} and Δ_{2ij}, respectively, as the mean models relate separately to these terms. Implementation is within SAS NLMIXED.

4 Analysis of Epilepsy Data

A description of the data is provided in Molenberghs *et al* (2010). The data come from a randomized, double-blinded, parallel group multi-center study aimed at comparing placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's. Weekly seizure counts are available. We fit our model and several sub-models to the data. Denote the number of epileptic seizures for patient i at week j by Y_{ij} and the occasion on which Y_{ij} was measured by t_{ij}. Assuming that Y_{ij} follows a combined model with $\lambda^c_{ij} = \theta_{ij}\kappa_{ij}$, assume $\theta_{ij} \sim \text{Gamma}(u, v)$, and

$$\ln(\kappa_{ij}) = \begin{cases} \alpha_{00} + \alpha_{01}t_{ij} + b_i & \text{if placebo,} \\ \alpha_{10} + \alpha_{11}t_{ij} + b_i & \text{if treated.} \end{cases}$$

The marginal model for the zero-inflated probabilities is given by $\ln(\pi_{ij}^m) = \beta_0 + \beta_1 t_{ij}$. The corresponding conditional models are specified by introducing a normally distributed random intercept, $b_{1i} \sim N(0, \sigma_1^2)$ in the Poisson model and $b_{2i} \sim N(0, \sigma_2^2)$ in the binomial model and the correlation between the binomial and count components is represented by ρ.
Results of these models are presented in Table 1. Generally, the fixed-effect parameters are close to each other. Their interpretations are not just subject-specific but can be extended to the whole population. Use 'CO' for combined and 'ZI' for zero inflation. Comparing the MMM and ZIMMM to the COMMM and ZICOMMM models, we see improvement in the model fit owing to the gamma random effects. Also, model fit improves if the normal random effects are supplemented with zero-inflation. Therefore, it is key that the more complex model results in a considerably improvement in the model. This is essential for inferences and for prediction.

5 Concluding Remarks

We have proposed a flexible model to simultaneously address issues of zero-inflation, overdispersion, and data hierarchies, while retaining a population-averaged interpretation of fixed effect parameters like in classical Poisson models. Through an empirical study, we have demonstrated that it is not sufficient to address either two of the three phenomena, while ignoring the remaining one. Our extension led to considerable improvement, thereby ensuring parameter interpretation is for the whole population, where a population may be defined in terms of fixed-effects profile. A marginal interpretation is often of interest to public health experts, who seek solutions or interventions for the population at large and therefore might find conditional models such as the GLMM or the combined model cumbersome.

Acknowledgments: The authors acknowledge support from the IAP research Network P7/06 of the Belgian Government (Belgian Science Policy).

References

Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688–698.

Iddi, S. and Molenberghs. G. (2012). A combined overdispersed and marginalized multilevel model. *Computational Statistics and Data Analysis*, **56**, 1944–1951.

Molenberghs, G., Verbeke, G., Demétrio, C., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.

Bayesian generalized additive models for location, scale and shape for insurance data

Nadja Klein[1], Thomas Kneib[1], Stefan Lang[2]

[1] Georg-August-University Göttingen, Germany
[2] University of Innsbruck, Austria

E-mail for correspondence: nklein@gwdg.de

Abstract: Generalized additive models for location, scale and shape define a flexible semiparametric class of regression models in which the exponential family assumption for the response is relaxed. While ordinary regression only analyzes the effects of covariates on the mean of a response, more complex parameters of the underlying distribution can be described using structured additive predictors. However, more complicated and numerically demanding likelihood functions are a consequence. An alternative to likelihood-based estimations are efficient Markov chain Monte Carlo techniques. Especially constructing adequate proposal densities which automatically deliver approximations of the full conditionals play a crucial role. In this way simultaneous estimations of nonlinear effects, spatial variations, random effects and interactions between risk factors in the data set are possible. As special cases we analyze claim frequencies and claim sizes arising in insurance data, both in due consideration of the large amount of zero observations. Therefore, zero-inflated models as an expansion of the classical Poisson regression and zero-adjusted models will be presented. For comparison of models with respect to the distribution, we consider quantile residuals as an effective graphical device and scoring rules that allow to quantify the predictive ability of the models. The deviance information criterion is used for further model specification.

1 Introduction

Calculations of car insurance premiums are based on a detailed statistical analysis of the risk structure of the policyholders. An important role is to model the claim frequencies which generally depend on characteristics of the policyholders and the car. In many applications the multitude of the policyholders do not cause any claims within the policy such that a large amount of zero observations occurs within the response. This fraction of zeros is considerably larger than expected with a Poisson distribution fitted to the data. To overcome this limitation, zero-inflated count data regression models assume data coming from a two-stage process where a binary process determines whether an observation is always zero or realized by an usual count data distribution. Furthermore, some covariate effects, e.g.

age of the car, are expected to have a nonlinear effect on the claim frequencies or spatial information should be included in the regression model. We therefore propose zero-inflated models within the framework of generalized additive models for location, scale and shape (GAMLSS), proposed by Rigby and Stasinopoulos (2005), from a Bayesian point of view.
If there is information about the claim sizes the purpose might be to price premiums correctly and to predict the risk of claims at the same time. Heller, Stasinopoulos and Rigby (2006) proposed the zero-adjusted inverse Gaussian distribution to model one particular car insurance data set. Inference was based on maximum likelihood estimations within the `gamlss` package in R, see (Stasinopoulos and Rigby 2007). We develop such models in a Bayesian framework where various candidates of continuous distributions like log-normal, inverse Gaussian or a member of the generalized beta family are conceivable. All of them could accommodate the right skewness of the claim size distribution that often occurs in applications based on fits of the data.
Compared to frequentist GAMLSS formulations our Bayesian approach has the advantage to include the choice of smoothing parameters directly in the estimation run and to provide valid credible intervals which are difficult to obtain based on asymptotic maximum likelihood theory.

2 Regression models

2.1 Zero-inflated models

We assume that zero-inflated count data y_i as well as covariate information ν_i have been collected for individuals $i = 1, \ldots, n$. The conditional distribution of y_i is then described in terms of the density $p(y_i|\nu_i) = \pi_i \mathbf{1}_{\{0\}}(y_i) + (1 - \pi_i)\tilde{p}(y_i|\nu_i)$, that arises from the hierarchical definition of the responses as $y_i = \kappa_i \tilde{y}_i$ where κ_i is a binary process, $\kappa_i \sim \mathrm{B}(1 - \pi_i)$, and $\tilde{y}_i$ follows a standard count data model, $\tilde{y}_i \sim \tilde{p}$. The amount of extra zeros introduced compared to the count data distribution of $\tilde{y}_i$ is determined by the probability π_i. We will consider two special cases for the count data part, namely the Poisson distribution $\tilde{y}_i \sim \mathrm{Po}(\lambda_i)$ and the negative binomial distribution $\tilde{y}_i \sim \mathrm{NB}(\delta_i, \delta_i/(\delta_i + \mu_i))$. The latter choice is particularly suited if the count data part of the response distribution is overdispersed. To allow maximum flexibility in the zero-inflated count data regression specifications, both the parameter for the excess of zeros as well as the parameters of the count data part of the distribution are related to regression predictors constructed from covariates via suitable link functions.

2.2 Zero-adjusted models

Many of the models in literature where distributions for analyzing claim sizes have been considered are models for the subclass of policies which

had claims within the observation period. Zero-adjusted models include discrete-continuous distributions with a probability mass at zero and an appropriate continuous component. The distribution of y_i for given covariate information ν_i can be written in terms of the mixed density $p(y_i|\nu_i) = (1-\pi_i)\mathbf{1}_{\{0\}}(y_i) + \pi_i f(y_i|\nu_i)(1-\mathbf{1}_{\{0\}}(y_i))$, where $f(y_i|\nu_i)$ is the density of a continuous distribution and π_i is the probability of a claim. As in the previous section suitable link functions are chosen for all model parameters.

2.3 Semiparametric Regression

For each predictor of the previous two sections we assume a semiparametric structured additive specification, $\eta_i = \beta_0 + f_1(\nu_i) + \ldots + f_p(\nu_i)$, where, for notational simplicity, we drop the parameter index from the predictor and the included effects. While β_0 is an intercept term representing the overall level of the predictor, the generic functions $f_j(\nu_i)$, $j = 1, \ldots, p$, relate to different types of regression effects combined in an additive fashion. In structured additive regression, each function is approximated in terms of d_j basis functions such that $f_j(\nu_i) = \sum_{k=1}^{d_j} \beta_{jk} B_k(\nu_i)$. Special cases are linear effects, P-splines for nonlinear effects, see (Lang and Brezger, 2004), Markov random fields, see (Rue and Held, 2005), or random effects.

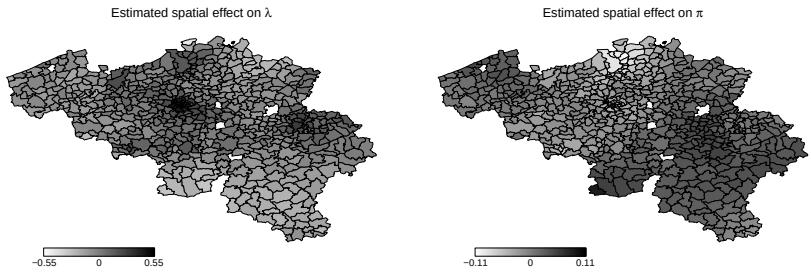
3 Inference

The full conditionals for the regression coefficients arising from the basis function expansion are not accessible analytically because of the complex structure of the likelihoods. As a consequence we will develop suitable proposal densities based on iteratively weighted least squares (IWLS) approximations. The basic result is a Gaussian proposal density with expectation and covariance matrix corresponding to the mode and the curvature of the quadratic approximation. To simplify the description we assume for the moment a model with only one predictor η but the principle idea immediately carries over to our multi-predictor framework since in the MCMC algorithm we are always working with sub-blocks of coefficients corresponding to only one predictor component. Let $l(\eta)$ be the log-likelihood depending on the predictor η. The quadratic approximation to the part of the penalized log-likelihood term depending on η leads to the working model $z^{(t)} \sim \mathrm{N}\left(\eta^{(t)}, \left(W^{(t)}\right)^{-1}\right)$ where $z = \eta + W^{-1}v$ is a vector of working observations with the predictor as expectation, $v = \partial l/\partial\eta$ is the score vector and W are working weight matrices, with $w_i = E(-\partial^2 l/\partial\eta_i^2)$ on the diagonals and zero otherwise. Based on this approximation, we obtain that the IWLS proposal distribution for β_j is $N(\mu_j, P_j^{-1})$ with expectation $\mu_j = P_j^{-1} Z_j' W(z - \eta_{-j})$ and precision matrix $P_j = Z_j' W Z_j + \frac{1}{\tau_j^2} K_j$, where

$\eta_{-j} = \eta - Z_j\beta_j$ is the predictor without the j-th component. The result is a Metropolis-Hastings algorithm for the different parameters.

4 First results

In a first step, the zero-inflated models have been applied to the claims arising in a data set of size $n = 162,548$ observations from a car insurance in Belgium of the year 1997. Typical properties are age of the policyholder and vehicle, engine power as well as previous claim experience, where all of them are modeled continuously using P-splines. In addition to various binary covariates, the data set provides spatial information about the domicile of the policyholders, which is included by a Markov random field. A raw descriptive analysis of the response gives roughly 80% of zero claims. While this summary does not take into account potential covariate effects, it already provides an indication that zero-inflation might be relevant. The use of quantile residuals and empirical calculations give no sign of overdispersion in the data so that the zero-inflated Poisson model is applied. The figure below shows the estimated spatial effect on both parameters. We estimate the rate to be higher in urban areas like Brussels, where by contrast the probability of extra zeros is supposed to be smaller in such areas.



References

Heller G., Stasinopoulos D. M., Rigby R. A. (2006). The zero-adjusted Inverse Gaussian distribution as a model for insurance data. In: *Proceedings of the 21th International Workshop on Statistical Modelling*

Lang, S., Brezger, A. (2004). Bayesian P-splines *Journal of Computational and Graphical Statistics*, **13**, 183 – 212.

Rigby, R.A., Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape (with discussion). *App. Stat.*, **53**, 507 – 554.

Rue. H., Held, L. (2005). *Gaussian Markov Random Fields*. C&H/CRC.

Stasinopoulos, D.M., Rigby, R.A. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1 – 46.

On a null variance estimator for the Mantel-Haenszel risk difference and corresponding confidence interval

Bernhard Klingenberg[1]

[1] Department of Mathematics & Statistics, Williams College, USA

E-mail for correspondence: bklingen@williams.edu

Abstract: We give a new variance estimator for the common difference of proportion in stratified 2×2 tables under the null and use it to derived a new confidence interval that is available in closed from. We present simulation results comparing our interval to various others and illustrate the interval using data from a stratified clinical trial with 56 centers.

Keywords: Meta Analysis; Stratified 2×2 tables; Difference of proportion.

1 Wald Confidence Interval for Common Risk Difference

When analyzing stratified 2×2 tables, computing pooled estimators along the lines of Cochran (1954) and Mantel and Haenszel (1959) is the established procedure. The pooled estimator for the common difference of proportion δ in $i = 1, \ldots, K$ stratified 2×2 tables is given by (Greenland and Robins, 1985)

$$\hat{\delta}_{MH} = \frac{\sum_{i=1}^{K} w_i(y_{i1}/n_{i1} - y_{i2}/n_{i2})}{\sum_{i=1}^{K} w_i} = \frac{\sum_{i=1}^{K}(n_{i2}y_{i1} - n_{i1}y_{i2})/n_{i+}}{\sum_{i=1}^{K} w_i}, \quad (1)$$

where $w_i = n_{i1}n_{i2}/n_{i+}$ are so-called Cochran weights and $n_{i+} = n_{i1}+n_{i2}$ is the total sample size in stratum i. Throughout, we assume that the y_{ij}'s are independent binomial $\text{Bin}(n_{ij}, \pi_{ij}), j = 1, 2$. Greenland and Robins (1985) plugged sample proportions into the expression for $\text{Var}[n_{i2}y_{i1} - n_{i1}y_{i2}]$ to obtain and estimate for $\text{Var}[\hat{\delta}_{MH}]$.

Under homogeneity of the risk difference one can write $\pi_{i1} = \delta + \pi_{i2}$ or $\pi_{i2} = \pi_{i1} - \delta$ for all i. Substituting these into the variance formula for $\text{Var}[n_{i2}y_{i1} - n_{i1}y_{i2}]$ leads to two different expressions, which, when averaged, yield

$$\text{Var}[n_{i2}y_{i1} - n_{i1}y_{i2}] = E[\delta P + Q], \quad (2)$$

where $P = \sum_i P_i$ and $Q = \sum_i Q_i$ with

$$P_i = \frac{n_{i1}^2 y_{i2} - n_{i2}^2 y_{i1} + n_{i1} n_{i2}(n_{i2} - n_{i1})/2}{n_{i+}^2},$$
$$Q_i = \frac{y_{i1}(n_{i2} - y_{i2}) + y_{i2}(n_{i1} - y_{i1})}{2n_{i+}}.$$

Replacing δ by $\hat{\delta}_{MH}$ in (2) and ignoring the expected value leads to the variance estimator (Sato, 1989) $\widehat{\text{Var}}[\hat{\delta}_{MH}] = \left(\hat{\delta}_{MH} P + Q\right)/W^2$, where $W = \sum_i w_i$. A Wald-type confidence interval (CI) for δ has form $\hat{\delta}_{MH} \pm z_{\alpha/2}\sqrt{\widehat{\text{Var}}[\hat{\delta}_{MH}]}$. This confidence interval is presented in survey articles (e.g., Agresti and Hartzel, 2001), while the one based on the Greenland and Robins variance estimator is presented in widely circulated epidemiology textbooks (e.g., Rothman 2002). Note that the CI above is equivalent to the acceptance region of the test $H_0 : \delta = \delta_0$ vs. $H_a : \delta \neq \delta_0$ using

$$T = \frac{(\hat{\delta}_{MH} - \delta_0)^2}{\widehat{\text{Var}}[\hat{\delta}_{MH}]}$$

as a test statistic, which is asymptotically Chi-square with $df = 1$.

1.1 A New Variance Estimator for the Mantel-Haenszel Risk Difference

We expect better asymptotic performance when estimating the variance under the null, as then the null distribution of the test statistic is closer to Chi-square. One way to obtain a null variance estimator for $\hat{\delta}_{MH}$ is to plug in δ_0 for δ in Sato's variance formula, leading to

$$\widehat{\text{Var}}_{\delta_0}[\hat{\delta}_{MH}] = (\delta_0 P + Q)/W^2. \tag{3}$$

Then,

$$T_0 = \frac{(\hat{\delta}_{MH} - \delta_0)^2}{\widehat{\text{Var}}_{\delta_0}[\hat{\delta}_{MH}]}$$

is an alternative test statistic for H_0. Inverting T_0, i.e., solving $T_0 = \chi^2_\alpha$ for δ_0, where χ^2_α is the upper α quantile of the Chi-square distribution with $df = 1$ leads to a quadratic equation. Solving it yields the following closed-form solution for the upper and lower bound of the confidence interval for δ:

$$b/2 \pm \sqrt{b^2/4 - c}, \text{ with } b = 2\hat{\delta}_{MH} + (P/W^2)\chi^2_\alpha,\ c = \hat{\delta}^2_{MH} - (Q/W^2)\chi^2_\alpha.$$

This interval, as opposed to the Wald interval, is not symmetric about $\hat{\delta}_{MH}$, which is advantageous when the distribution of $\hat{\delta}_{MH}$ is skewed. An alternative to (3) is to estimate $\text{Var}[n_{i2}y_{i1} - n_{i1}y_{i2}]$ directly under H_0 without going through (2). This leads to score-type CI that cannot be discussed here but are included in the simulations that follow.

2 Simulation Study and Example

Figure 1 shows boxplots of the actual coverage probability (estimated via 7600 simulations) over 500 random parameter settings when $K = 3, 5$ or 10, equal and constant (across strata) sample sizes $n_{i1} = n_{i2} = 25$ and true common risk difference $\delta = 0, 0.1$ or 0.2.

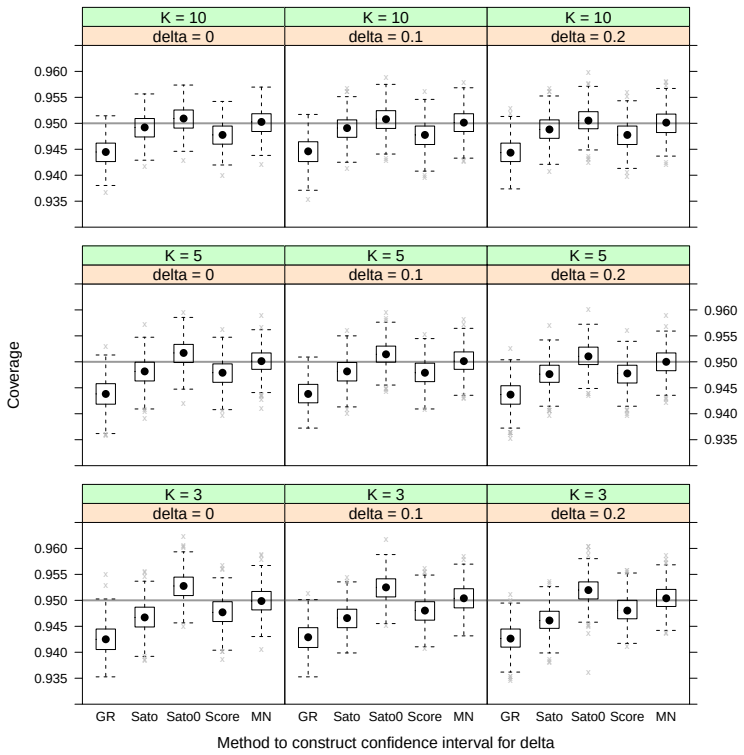


FIGURE 1. Coverage Probability for the common risk difference under various setting for the number K of centers/studies and the true effect δ. The interval estimators investigated are: Greenland-Robins (GR), Sato, the new approach proposed in Section 1.1 ("Sato0"), the score interval for a linear combination of proportions ("Score"), and the score interval similar to one proposed by Miettinen and Nurminen ("MN").

The new interval proposed in Section 1.1 (called "Sato0")and the Miettinen-Nurminen score type interval outperform all others. Note that, unlike the Miettinen-Nurminen interval, the Sato0 interval is available in closed form and does not need iteration.

In a recent vaccine trial, a vaccine was compared to placebo in $K = 56$ dif-

ferent centers. The sample sizes in the treatment group varied from around 20 to around 50 per center. Since allocation to treatment or placebo followed a 3:1 ratio, sample sizes in the placebo group are by about 1/3 smaller. Table 1 shows data for a few selected centers.

TABLE 1. Data excerpt from stratified vaccine trial in 56 centers.

Center	y_{i1}	n_{i1}	y_{i2}	n_{i2}	$\hat{\pi}_{i1} - \hat{\pi}_{i2}$
1	10	46	2	15	0.08
38	4	18	0	6	0.22
45	3	18	1	5	-0.03

A 95% confidence interval for the common risk difference between the vaccinated group and the placebo group, using the new interval from Section 1.1 equals [-0.71%, 6.49%].

References

Agresti, A. and Hartzel, J. (2000). Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine* **19**, 1115 – 1139.

Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10**, 417 – 451.

Gart, J. and Nam, J. (1990). Approximate interval estimation of the difference in binomial parameters: Correction for skewness and extension to multiple tables. *Biometrics* **46**, 637 – 643.

Greenland, S. and Robins, J.M. (1985). Estimation of a common effect parameter from sparse follow-up data. *Biometrics* **41**, 55 – 68.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* **22**, 719 – 748.

Rothman, K. (2002) *Epidemiology: an introduction.* New York: Oxford University Press.

Sato, T (1989). On the Variance Estimator for the Mantel-Haenszel Risk Difference. *Biometrics*, **45**, 1323 – 1324.

Model based segmentation of TV advertising scheduling patterns

Arnošt Komárek[1], Tomáš Kincl[2], Lenka Komárková[2]

[1] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
[2] Faculty of Management, University of Economics in Prague, Czech Republic

E-mail for correspondence: komarek@karlin.mff.cuni.cz

Abstract: This paper employs a model based classification for longitudinal data to identify typical scheduling patterns of TV sponsorship spots which is a type of TV advertisement.

Keywords: Classification; Longitudinal data; Segmentation.

1 Introduction

The advertising campaign is set according its goals and objectives. To ensure the highest efficiency of the campaign, the companies use different approaches to scheduling and timing the advertisements. There are different scheduling patterns identified to adjust the campaign timing according to the communication goals. The volume of advertising during the campaign may be continuous with steady (i.e. reminder advertising for matured products or building brand awareness), rising (i.e. to concentrate attention around a particular event) or falling (i.e. fade after initial launch of a new product) trend during the campaign. There are more scheduling pattern identified (i.e. flighting or pulsing) used for short and heavy advertising periods. The campaign length also reflects the nature of the communicated message and the goals of the campaign. For example longer campaigns (weeks or years) are often directed towards building the longer term effects of favorable brand image and strong brand loyalty

2 Data and research questions

In this paper, we concentrate on analysis of scheduling patters of so called TV sponsorship spots broadcasted in the Czech TV channels during 2011. The TV sponsorship is one possible type of the advertising campaign which can take many forms, i.e., TV billboards, sponsored trailers, injections, identifications, sponsorship reminders, break bumpers. Very often, the TV

sponsorship comes before/after the broadcast (or within as a break bumper) and is mostly 10 or 15 seconds long. Other often used types are injections of various lengths (5 – 60 seconds).
Data for our analysis were gathered by the Mediaresearch company (`http:// www.mediaresearch.eu`) which is the research agency conducting electronic monitoring of TV viewership in the Czech Republic. Data contain the broadcasting history of more than 5 000 unique TV sponsorship spots (*unique commercials*) that appeared during 2011 on one of 13 Czech TV channels that offer this type of advertisement.
Let us now introduce some notation. Let $Y_{i,t}$ be the number of broadcasting occurrences of the ith commercial ($i = 1, \ldots, N$) during week t ($t = 0, \ldots, T_i$) since its prime. Since only rarely (in less than 5% of cases), a particular commercial is being broadcasted longer than 16 weeks (4 months), we limit our analysis to data with $t \leq 16$. Our goal is to use the observed values of $\mathbf{Y}_i = \left(Y_{i,0}, \ldots, Y_{i,T_i}\right)^\top$ which characterize the scheduling history of the ith commercial to identify typical scheduling patterns of the TV sponsorship spots. This problem being often referred to as a problem of segmentation.

3 Model based segmentation

The observed values of the scheduling histories $\mathbf{Y}_i$ ($i = 1, \ldots, N$) might be viewed as longitudinal data and the problem of segmentation as a problem of classification based on the observed longitudinal profiles. To this end, a model based classification method of Komárek and Komárková (2013b) and a related contributed `R` (R Core Team, 2013) package `mixAK` (Komárek and Komárková, 2013a) might be exploited for this purpose.
As it is usual with model based classification, it is assumed that the scheduling history $\mathbf{Y}_i$ of the ith commercial is generated according to one of K models where K is the number of segments (groups). History generation according to the kth model ($k = 1, \ldots, K$) happens with an unknown probability w_k, where $0 < w_k < 1$, $\sum_{k=1}^{K} w_k$. In our particular application, the following (linear mixed) model is assumed for the scheduling history $\mathbf{Y}_i$ of the ith commercial provided it belongs to the kth segment:

$$\log(Y_{i,t}) = b_{i,0} + b_{i,1}\, t + b_{i,2}\, t^2 + \varepsilon_{i,t}, \qquad t = 0, \ldots, T_i,$$

where the random effect vector $\mathbf{b}_i = \left(b_{i,0}, b_{i,1}, b_{i,2}\right)^\top$ is assumed to follow a normal distribution with an unknown mean vector $\boldsymbol{\mu}_k = \left(\mu_{k,0}, \mu_{k,1}, \mu_{k,2}\right)^\top$ and an unknown covariance matrix $\mathbb{D}_k$. The random error vector $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i,1}, \ldots, \varepsilon_{i,T_i}\right)^\top$ is assumed to be independent of $\mathbf{b}_i$ and following a zero mean normal distribution with a diagonal covariance matrix with an unknown residual variance σ^2 being the same for all segments.

TABLE 1. Estimated characteristics of the scheduling patter segments.

Segment k	**Weight** w_k	**Intercept** $\mu_{k,0}$	**Linear term** $\mu_{k,1}$	**Quadratic term** $\mu_{k,2}$
1	0.192	1.297	−0.0277	0.00102
2	0.470	0.311	0.0497	−0.00267
3	0.282	2.231	−0.1065	0.00321
4	0.056	2.574	−0.6740	0.06130

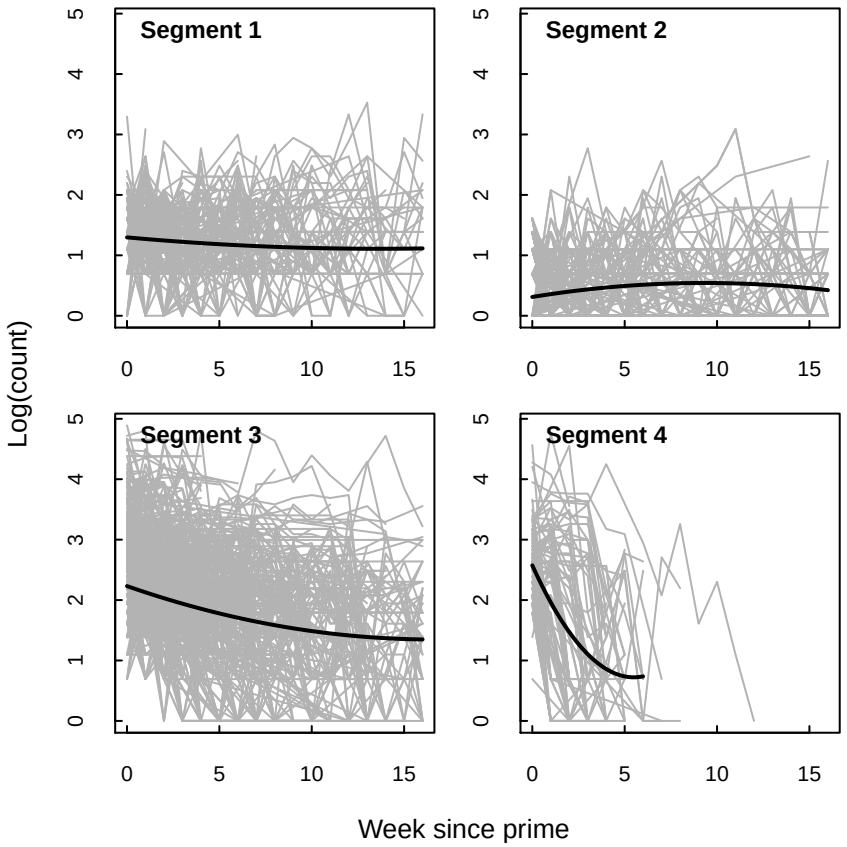


FIGURE 1. Observed (grey) and mean (black) evolution of the logarithmic number of broadcasted TV sponsorship in each of four segments.

In summary, the kth segment ($k = 1, \ldots, K$) is characterized by the kth mean vector $\boldsymbol{\mu}_k$ and the kth covariance matrix $\mathbb{D}_k$. The mean vector determines the mean evolution of the logarithmic number of weekly broadcasting occurences of the kth segment whereas the covariance matrix the variability of the individual scheduling patterns around the mean pattern given by $\boldsymbol{\mu}_k$.

Komárek and Komárková (2013a) describe a Bayesian approach to estimation of unknown parameters and subsequent classification. They also suggest to use an approach based on penalized expected deviance (PED, Plummer, 2008) for selection of an optimal number of segments. Their methods have been applied to our application.

4 Results and discussion

The optimal number of segments according to PED is four, i.e., $K = 4$. Estimated segment weights and parameters of the segment specific patterns represented by the mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_4$ are given in Table 1. Graphically, the segment specific patterns together with the observed histories for the individual creativities being classified in each patters are shown on Figure 1. More detailed discussion of results and results of more advanced analyses exploting also information on additional characteristics of each commercial (length, type of channel where broadcasted, ...) shall be postponed to a journal paper being in progress.

Acknowledgments: The second and the third author have been supported by the Czech Science Foundation Grants GAČR P403/12/2175 and P403/12/1557, respectively.

References

Komárek, A. and Komárková, L. (2013a). Capabilities of `R` package `mixAK` for clustering based on multivariate continuous and discrete longitudinal data. *Submitted for publication,* preprint available from `http://msekce.karlin.mff.cuni.cz/~komarek/publication.html#mixAKclust`.

Komárek, A. and Komárková, L. (2013b). Clustering for multivariate continuous and discrete longitudinal data. *Annals of Applied Statistics*, **7**, 177 – 200.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523 – 539.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. `http://www.R-project.org/`.

Statistical Models to Density-dependence Detection in Mediterranean Deer Populations

Antonio J. López-Montoya[1,2], Concepción Azorit[1], Irene García-Garrido[2], Ramón Gutiérrez[3], Javier Moro[3]

[1] Dep. Animal and Vegetal Biology and Ecology, University of Jaén, Spain,
[2] Dep. of Statistics and Operations Research, University of Granada, Spain,
[3] Autonome Organism of National Parks, Ministry of the Environment and Rural and Marine Affairs, Spain

E-mail for correspondence: ajlm@correo.ugr.es

Abstract: The detection of density dependent as an intrinsic factor influencing dynamic deer populations is an interesting focus in ecological research and a very useful information for a wiser management herd. In this work several statistical models such as linear models, Ricker and Gompertz models, and an autoregressive state-space model (SSM) have been tested in order to detect direct Density-Dependence (DD) in red deer (*Cervus elaphus hispanicus*) populations of Southern Spain. We use hunting data temporal series as population abundance estimates. We found that the Gompertz model and the SSM lead us to parallel conclusions that prove direct DD in our populations. Monitoring red deer in two separate populations, but in the same climatic and Mediterranean environment, allow us a better assessment of the effect of DD in contrast to extrinsic factor such as hunting pressure.

Keywords: Density-dependence; Gompertz model; Kalman filter; Population growth rate; State-space model.

1 Introduction

In this work, we investigated the effect of intrinsic factor (DD) on population growth rate fluctuation of free-living red deer (*Cervus elaphus hispanicus*) from two separated reserves of Southern Spain. Using hunting data temporal series from 2001 to 2011 we explore for the first time in these red deer populations direct DD through lineal mathematical models and autoregressive SSM.

We have a database with the annual hunting extractions from years 2001 to 2011, and taking advantage of this data, we decided to use the hunting extractions as an abundance index based on studies of Simard et al. (2012), to detect the existence of DD in our reserves.

We expect a consistent DD because a strong overabundance causing chronic browsing and summer-autumn deer mortality on the study area before 1997.

2 Statistical Models

Our goal is to detect if abundance affects the population growth rate and the population fluctuations. We assume that the discrete growth rate is defined of the form:

$$r_t = \log\left(\frac{H_t}{H_{t-1}}\right)$$

where r_t is the realized per capita rate, H_t is the hunting size (abundance index) at the year t and H_{t-1} is the hunting size at the year $t-1$.

The first model used to estimate direct DD is a simple linear model, direct DD is defined by the slope of the regression coefficient β_1 of the regression line between the natural logarithm of hunting size at the year t, and the natural logarithm of hunting size at the year $t-1$, this model is given by:

$$\log(H_t) = \beta_0 + \beta_1 \log(H_{t-1}) + \varepsilon_t \tag{1}$$

where the residuals ε_t are assumed to be Gaussian and uncorrelated. If we have a negative slope of β_1, this indicate direct DD. Turchin (2003) suggests that facing H_t against H_{t-1}, can mask the presence of direct DD and that to avoid this, it is more convenient to deal with r_t against H_{t-1}.

Two of the easier and more popular models in which we test direct DD by r_t against H_{t-1} are the stochastic Ricker and Gompertz models. The Ricker model assumes an exponential DD and takes the form:

$$H_t = H_{t-1} \exp\left(a + bH_{t-1} + \varepsilon_t\right) \tag{2}$$

likewise, the Gompertz model is written as:

$$H_t = H_{t-1} \exp\left(a + b \log(H_{t-1}) + \varepsilon_t\right). \tag{3}$$

In both models, the residuals ε_t are assumed to be Gaussian and uncorrelated, we say that direct DD exists if the value of b is negative and significantly different from zero. In Turchin (2003), we have further description of these models.

Another model to test direct DD is the autoregressive SSM, we will use the Kalman filter to apply the likelihood function and restricted likelihood function of Dennis et al. (2006) and adjust the stochastic Gompertz population growth as a SSM:

$$\begin{aligned} \log(H_t) &= a + c \log(H_{t-1}) + E_t \\ \log(Y_t) &= \log(H_t) + F_t \end{aligned} \tag{4}$$

in this case, H_t is supposedly the true population size, Y_t is the observed population size, $E_t \sim \mathcal{N}(0, \sigma^2)$ and $F_t \sim \mathcal{N}(0, \tau^2)$ are the error terms with variance σ^2 and sampling variance τ^2 . When $c = 1$, the process is density independent and on the log scale is a Gaussian random walk with drift given by a. When $|c| < 1$, the model has a stationary distribution and smaller values of c imply greater DD (see Staples et al.(2004),(2005); Dennis et al.(2006); Seavy et al.(2009)).

3 Results

In the next two tables we can see the values of the fitted models to our data to prove the existence of direct DD. The fitted models are the Ricker model, the Gompertz model and the autoregressive SSM.

TABLE 1. Density dependence test for the red deer population between the years 2001-2011. From left to right: reserve, type of model (Ricker or Gompertz), slope (b), probability value, coefficient of determination (R^2) and Shapiro-Wilk test.

Reserve	Model	b	$p-value$	R^2	$S-Wtest$
R1	Ricker	-0.0018	0.1352	0.2303	0.8507
	Gompertz	-0.7102	0.0532	0.3547	0.1532
R2	Ricker	-0.0017	0.1163	0.2512	0.0738
	Gompertz	-0.5428	0.1023	0.2688	0.0335

TABLE 2. Maximum likelihood (ML) and restricted maximum likelihood (REML) parameters estimates in the Gompertz SSM. From left to right: reserve, method, intercept (a), regression coefficient (c) of the Gompertz model on the logarithmic scale, variance of real population densities time-series (σ^2), variance of the observation error (τ^2) and Akaike information criterion (AIC).

Reserve	Method	a	c	σ^2	τ^2	AIC
R1	ML	0.3738	0.9355	0.0179	0.1276	18.0155
	REML	0.3363	0.9419	0.0197	0.1403	20.4560
R2	ML	0.3888	0.9344	0.0687	0.4889	32.7929
	REML	0.3248	0.9452	0.0756	0.5378	33.8900

The results of the direct DD test are reported in Table 1 for the two reserves in study. In the R1 reserve, the Gompertz model accounted for direct DD better than the Ricker model, with a higher value of b that was significantly different from zero at the 95% level. In the R2 reserve, both Gompertz and

Ricker models had a p-value higher than 95% level. For both models, the R^2 values were always low but we admit Gompertz and Ricker models. The Shapiro-Wilk test was significant for all cases (except Gompertz model in R2 reserve), suggesting a Gaussian distribution of residuals.

From Table 2, we can see that according to AIC the ML method is better than the REML one. According to the values of the parameter c, we conclude that, in both cases there exist direct DD.

Due to the analysis performed by previous models, we can conclude the presence of direct DD in the two reserves. Monitoring red deer in two separate populations but in a same climatic and Mediterranean environment allow us a better assessment of the effect of DD in contrast to extrinsic factor such as hunting pressure.

Acknowledgments: This study was approved by the Bioethical Committee of University of Jaén and supported by Ministry of Agriculture, Food and Environment (National Parks, Spain), the projects P07-RNM-03087 and CGL-2011-23919 and by the European Fund for Regional, Development (FEDER).

References

Dennis, B., Ponciano, J.M., Lele, S.R., Taper, M.L. and Staples, D.F. (2006). Estimating density dependence, process noise, and observation error. *Ecological Monographs*, **76**, 323 – 341.

Seavy N.E. and Reynolds, M.H. (2009). Seabird nest counts: a test of monitoring metrics using Red-tailed Tropicbirds. *Journal of Field Ornithology* , **80**, 297 – 302.

Simard, M.A., Côté, S.D., Gingras, A. and Coulson, T. (2012). Test of density dependence using indices of relative abundance in a deer population. *Oikos*, **121**, 1351 – 1363.

Staples, D.F., Taper, M. L. and Dennis, B. (2004). Estimating population trend and process variation for PVA in the presence of sampling error. *Ecology*, **85**, 923 – 929.

Staples, D.F., Taper, M. L. and Shepard, B.B. (2005). Risk-based viable population monitoring. *Conservation Biology*, **19**, 1908 – 1916.

Turchin, P. (2003). *Complex population dynamics*. Princenton Univ. Press.

GsymPoint: An R Package for estimating the Generalized Symmetry Point as the optimal cutpoint in continuous diagnostic tests

Mónica López-Ratón[1], Carmen Cadarso-Suárez[1], Elisa M. Molanes-López[2], Emilio Letón[3]

[1] Biostatistics Unit, Department of Statistics and Operations Research, Universidad de Santiago de Compostela, Santiago de Compostela, Spain
[2] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain
[3] Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia, Madrid, Spain

E-mail for correspondence: `monica.lopez.raton@usc.es`

Abstract: The selection of optimal cutpoints in continuous diagnostic tests is an important issue for classifying individuals in two groups (healthy and diseased). Additionally, the incorporation of costs for the misclassification rates is crucial although not taken into account most of the times. In the literature, several criteria for choosing the optimal cutpoint have been studied depending on the ultimate goal. One of them is the Generalized Symmetry Point that has been recently introduced using two approaches: one based on the General Pivotal Quantity under the assumption of normality and the other based on Empirical Likelihood without any parametric assumptions. This work introduces the R package `GsymPoint`, for estimating the Generalized Symmetry Point and the corresponding cost based Sensitivity and Specificity accuracy measures. The use of this package is illustrated with a real biomedical dataset.

Keywords: Empirical Likelihood; Generalized Pivotal Quantity; optimal cutpoint; R package.

1 Introduction

The classification of individuals in the healthy and diseased groups in continuous diagnostic tests is usually based on a cutoff value c, such that in general, individuals with a diagnostic test value equal to or higher than c are classified as diseased (positive test) and as healthy otherwise (negative test). Several strategies for selecting optimal cutpoints in diagnostic tests have been proposed depending on the ultimate goal (see for example, Pepe 2003). One of the best-known methods is based on the Symmetry Point c_S (Greiner et al. 1995), defined as the point where $p(c_S) = q(c_S)$, with p denoting the Specificity and q the Sensitivity. Taking into account the

prevalence of disease π and the costs associated to the False Positives and False Negatives misclassifications, c_{F-} and c_{F+}, respectively, we have defined the Generalized Symmetry Point in López-Ratón et al. 2012, c_{GS}, as follows :

$$r(1 - p(c_{GS})) = 1 - q(c_{GS}),$$

where $r = \frac{1-\pi}{\ell\pi}$ and $\ell = \frac{C_{F-}}{C_{F+}}$. In this work, we introduce `GsymPoint`, a package written in R (R Development Core Team 2012) for estimating the Generalized Symmetry Point in continuous diagnostic tests. In Section 2, we briefly review two methods included in this package for obtaining point estimates and confidence intervals for c_{GS}, $p(c_{GS})$ and $q(c_{GS})$: one method based on the Generalized Pivotal Quantity (GPQ) and the other on Empirical Likelihood (EL). In Section 3, we describe the `GsymPoint` package. Finally, in Section 4 we give an illustration of the practical application of the package using a real biomedical dataset.

2 Inference methods

In this section, we briefly present the GPQ and EL methods. The GPQ method was introduced by Weerahandi (1993). The Generalized Symmetry Point, c_{GS}, and the corresponding Sensitivity and Specificity measures are computed following the same reasoning as in Lai et al. (2011), assuming that the diagnostic test or a monotone transformation of Box-Cox type follows a Normal distribution. The EL method was first introduced by Thomas and Grunkemeier (1975). As the parameter of interest c_{GS} can be seen as two specific quantiles, the $p(c_{GS})$-th quantile of the healthy population and the $r(1 - p(c_{GS}))$-th quantile of the diseased population, the same reasoning as in Mólanes-López and Letón (2011) is followed to make inference on c_{GS}, $p(c_{GS})$ and $q(c_{GS})$.

3 The `GsymPoint` Package

This section introduces the R-based `GsymPoint` package where the inference methods described in Section 2 have been implemented for practical applications. This package only requires a data-entry file, which must, at minimum, contain the variables that indicate the diagnostic marker, the disease status (diseased/healthy) and whether the Generalized Symmetry Point is computed according to the levels of a categorical covariate, the variable that indicates such levels.

The main function of the package is the `gsym.point()` function, which uses the selected method(s) to compute the Generalized Symmetry Point, with its Sensitivity and Specificity accuracy measures, and creates a class gsym.point object. The call to this function is as follows:

```
gsym.point(methods, data, marker, status, tag.healthy,
+ categorical.cov = NULL, pop.prev = NULL, control =
+ control.gsym.point(), CFN = 1, CFP = 1, conf.level = 0.95,
+ trace = TRUE).
```

The `methods` argument is a character vector specifying the method/s used for estimating the Generalized Symmetry Point ("EL", "GPQ" or both). The `data` argument is the data frame containing the needed variables; `marker` and `status` arguments are character strings with the names of the diagnostic test variable and the variable that distinguishes healthy from diseased individuals, respectively. The value codifying healthy individuals in this last variable is indicated in the `tag.healthy` argument.
The `categorical.cov` argument is a character string with the name of the categorical covariate according to which optimal cutpoints are to be computed. By default it is NULL (no categorical covariate is considered).
The `pop.prev` argument is the value of the disease's prevalence. By default it is NULL, i.e., the prevalence is estimated by the sample prevalence, appropriate for cross-sectional studies. However, when other type of studies are considered, a given value for the prevalence can also be specified.
The `CFN` and `CFP` arguments are the costs of False Negative and False Positive classifications, respectively. The default value is 1 for both.
The `conf.level` argument is the value of the confidence level (1-α), and by default is equal to 0.95.
Moreover, there are some extra arguments, specific to each method. They are included in the `control` argument, a list of control values for the estimating process specified by means of the `control.gsym.point()` function. When no arguments are given to this function, the default values are used.

4 Biomedical application

In this section, we describe the application of the R-based `GsymPoint` package, considering a study conducted on 141 patients admitted to the Cardiology Department of a Teaching Hospital in Galicia (northwest Spain) for evaluation of chest pain or cardiovascular disease. The aim of the study was to investigate the clinical usefulness of leukocyte elastase determination in the diagnosis of coronary artery disease (CAD). All patients underwent coronary angiography during the investigation: 96 had coronary lesions (diseased patients) and 45 had non-stenotic coronaries (non-diseased patients). More details of this dataset can be found in Amaro et al.(1995).
The main objective here is to select the optimal cutpoint of elastase concentrations, given by the Generalized Symmetry Point, to diagnose CAD. The first step consists on downloading the `GsymPoint` package and the dataset `elas` in R (included in the package):

```
R> library("GsymPoint")
R> data("elas")
```

```
Call:
gsym.point(methods = "EL", data = data, marker = "elas",
 status = "status", tag.healthy = 0, categorical.cov = NULL,
 pop.prev = NULL, CFN = 2, CFP = 1,
 control = control.gsym.point(), conf.level = 0.95,
 trace = TRUE)

Healthy:  45 individuals
   Min       1Q   Median     Mean       3Q      Max Std. dev
  5.00    15.00    31.00    29.52    41.00    56.00    14.62
-------------------------------------------------
Diseased:  96 individuals
   Min       1Q   Median     Mean       3Q      Max Std. dev
 13.00    32.00    43.00    49.73    60.25   163.00    27.43

Sample prevalence: 68.085%

************************************************
OPTIMAL CUTOFF: GENERALIZED SYMMETRY POINT
************************************************

Area under the ROC curve (AUC):  0.744 (0.659, 0.828)

---------------------------------
METHOD: EL
---------------------------------
             Estimate 95% CI lower limit 95% CI upper limit
cutoff      26.3629472         23.3672584         29.4842178
Specificity  0.4076066          0.3034827          0.5477526
Sensitivity  0.8611578          0.8367538          0.8940045
```

FIGURE 1. Output of `GsymPoint` package.

To compute the Generalized Symmetry Point using the `elas` dataset, simply use the syntax shown below. In this case, we consider the sample prevalence (`pop.prev` = 0.68), `CFN` = 2 and `CFP` = 1. Since these data do not follow the parametric assumption needed by the GPQ method, we only show the 95%-confidence intervals obtained with the EL method.

```
R> cutpoint <- gsym.point(methods = "EL", data = elas,
marker = + "elas", status = "status", tag.healthy = 0,
categorical.cov = + NULL, pop.prev = NULL, CFN = 2, CFP =
1, control =
+ control.gsym.point(), conf.level = 0.95, trace = TRUE)
```

A numerical summary of the results can be obtained by calling up the `summary.gsym.point()` function, which can be abbreviated by `summary()`:

```
R> summary(cutpoint)
```

In Figure 1, the output of `GsymPoint` package is shown. In this case, the `summary.gsym.point()` function displays: firstly, a summary of leukocyte elastase values in healthy and diseased populations; and secondly, the point

estimates and the EL based 95%-confidence intervals for the Generalized Symmetry Point and its corresponding Sensitivity and Specificity measures.

Acknowledgments: This research has been supported by several Grants from the Spanish Ministry of Science & Innovation. M. López-Ratón and C. Cadarso-Suárez acknowledge support to MTM2010-09213-E and MTM2011-28285-C02-00. E.M. Molanes-López acknowledges support to MTM2010-09213-E, ECO2011-25706 and MTM2011-28285-C02-02. E. Letón acknowledges support to MTM2010-09213-E and MTM2011-28285-C02-02.

References

Amaro, A., et al. (1995) Plasma leukocyte elastase concentration in angiographically diagnosed coronary-artery disease. *European Heart Journal*, **16**, 615 – 622.

Lai, C.Y., Tian, L. and Schisterman, E.F. (2011). Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point. *Comput. Stat. Data Anal.*, DOI:10.1016 /j.csda.2010.11.023.

López-Ratón, M, Cadarso-Suárez, C, Molanes-López, E.M, and Letón, E. (2012) Inference of the symmetry point with different costs for the Specificity and Sensitivity. In: *Proceedings of 27th International Workshop on Statistical Modelling*, Prague, Czech Republic, pp. 191.

Molanes-López, E.M. and Letón, E. (2011). Inference of the Youden index and associated threshold using empirical likelihood for quantiles. *Stat. Med.*, **30**, 2467 – 2480.

Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

Thomas, D.R. and Grunkemeier, G.L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Am. Stat. Assoc.*, **70**, 865 – 871.

Weerahandi, S. (1993). Generalized confidence intervals. *J. Am. Stat. Assoc.*, **88**, 899 – 905.

The Birnbaum–Saunders survival model with cure fraction under different of activation mechanisms

Francisco Louzada[1], Vicente G. Cancho[1], Glagys D.C. Barriga[2], Dipak K. Dey[3]

[1] ICMC, Universidade de São Paulo, Brazil
[2] Universidade Estadual Paulista, Brazil
[3] University of Connecticut, USA

E-mail for correspondence: louzada@icmc.usp.br

Abstract: In this paper we propose a new cure rate survival model based on a Birnbaum-Saunders distribution. The model is conceived inside a scenario of latent competing causes with the presence of a cure fraction, where the occurrence of the event of interest may be activated by different kinds of mechanisms. We explore the use of Markov chain Monte Carlo methods to develop a Bayesian analysis for the proposed model. Case deletion influence diagnostics are developed based on the ψ-divergence, which includes the Kullback-Leibler, J-distance, L_1 norm and χ^2-square divergence measures. Simulation studies are performed and proposed methodology is illustrated on a real malignant melanoma data.

Keywords: Birnbaum–Saunders distribution; Cure fraction models; Geometric distribution; lifetime data.

1 Introduction

In many medical problems, such as chronic cardiac diseases and various different types of cancer, a cumulative individual damage may be caused by various unknown causes. This degradation leads to a fatigue process, whose propagation lifetimes can be suitably modeled by a Birnbaum-Saunders (BS) distribution (Balakrishnan *et al*, 2007; Leiva *et al*, 2008). The survival function of the BS model is given by $S_{BS}(t) = \Phi[-\frac{1}{\alpha}(\sqrt{t/\lambda} - \sqrt{\lambda/t})]$, for $t > 0$, where $\Phi(.)$ is the standard normal cumulative distribution function, $\alpha > 0$ and $\lambda > 0$ are respectively, shape and scale parameters.
The main goal of this paper is to present a generalization of the BS model, hereafter the GBS cure rate (GBScr) model, conceived inside a scenario of latent competing causes with the presence of a cure fraction, where the occurrence of the event of interest may be activated by different kinds of mechanisms (Cooner *et al*, 2007).

2 Model Formulation

The GBScr distribution is derived as follows. For an individual in the population, let M denote the unobservable number of causes of the event of interest for this individual. Assume that M follows a geometric distribution with parameter θ and probability mass function $P(M = m) = \theta(1-\theta)^m$, $m = 0, 1, \ldots$. The time for the j^{th} cause to produce the event of interest is denoted by Z_j, $j = 1, \ldots, M$. We assume that, conditional on M, the Z_j are i.i.d. with BS distribution with survival function given by $S_{BS}(.)$. Also, we assume that $Z_1, Z_2, \ldots$ are independent of M. The observable time to event is defined by the random variable $Y = Z_{(R)}$, where R depends on M, $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(R)} \leq \cdots \leq Z_{(M)}$ are the order statistics and $Y = \infty$ if $M = 0$. In many biological processes R can be interpreted as a resistance factor of the immune system of the individual. If the event of interest occurs (e.g., cancer relapse), then the random variable Y takes the value of the R^{th} order statistics $Z_{(R)}$. In other words, as in Cooner *et al* (2007), R out of M causes are required to produce the event of interest. The resistance factor can be a fixed constant, a function of M or a random variable specified through a conditional distribution on M.

In this paper, we deal with three specifications for R, $R = 1$ directing to a first activation mechanism, random R directing to a random activation mechanism and $R = M$ directing to a last activation mechanism. Thus we scan all possible mechanisms of activation.

Assuming that given $M \geq 1$, the conditional distribution of R is uniform on $\{1, \ldots, M\}$ (random activation mechanism). Under this setup, the surviving function for the population is given by

$$\begin{aligned} S_{\text{ran}}(y) &= P(Y > y) \\ &= P(M = 0) + \sum_{k=1}^{\infty} \sum_{R=1}^{k} P(Z_{(R)} > y | R, M = k) P(R | M = k) P(M = k), \quad (1) \end{aligned}$$

where $P(Z_{(R)} > y | R, M = k) = \sum_{i=0}^{R-1} \binom{k}{i} (F_{\text{BS}}(y))^i (S_{\text{BS}}(y))^{k-i}$, which is the cumulative distribution function of a binomial distribution, with k trials and success probability $F_{\text{BS}}(y) = 1 - S_{BS}(y)$. Then, considering a geometric distribution, the survival function of Y in (1) under random activation mechanism is given by

$$S_{\text{ran}}(y) = \theta + (1-\theta) S_{BS}(y), \quad (2)$$

where $B(x; k, F_{BS}(y)) = P(X = x)$ and $X \sim \mathit{Binomial}(k, F(y))$. We observe that the (2) is a mixture cure model with cured fraction $p_0 = P(M = 0) = \lim_{y \to \infty} S_{\text{ran}}(y) = \theta$.

As a second setup, the so-called first activation mechanism, we suppose that the event of interest happens due to any one of the possible causes.

Therefore, for $R = 1$, the time to event is $Y = Z_{(1)} = \min\{Z_1, \ldots, Z_M\}$, with survival function given by

$$S_{\min}(y) = \theta/(1 - (1 - \theta)S_{\mathrm{BS}}(y)). \tag{3}$$

The cured fraction is given by $p_0 = \theta$.
In our third scenario, also known as the last activation mechanism, the event of interest only takes place after all the M causes have been occurred, so that $R = M$ and the observed failure time is $Y = Z_{(M)} = \max\{Z_1, \ldots, Z_M\}$, with survival function given by

$$S_{\max}(y) = 1 + \theta - \theta/(1 - (1 - \theta)F_{\mathrm{BS}}(y)), \tag{4}$$

so that the cured fraction is $p_0 = \theta$.
It his easy to proof that under conditions of models in (2), (3) and (4) we have that, $S_{min}(y)$ in (3) $\leq S_{ran}(y)$ in (2) and $S_{max}(y)$ in (4) $\geq S_{min}(y)$ in (2).
Completing our model, we propose to relate the cured fraction to the covariates by the logistic link $\theta_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ encapsulates the vector of regression coefficients, so that for each group of individuals represented by $\boldsymbol{x}_i$, we have a different cured fraction.

3 Inference

Let us consider the situation where the failure time Y not completely observed and is subject to right censoring. Let C_i denote the censoring time. In a sample of size n, we then observe $T_i = \min\{Y_i, C_i\}$ and $\delta_i = \mathrm{I}(Y_i \leq C_i)$, where $\delta_i = 1$ if T_i is a failure time and $\delta_i = 0$ if it is right censored, for $i = 1, \ldots, n$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ denote the vector of covariates for the i^{th} individual. Then, under non-informative censoring, we write the likelihood function as $L(\boldsymbol{\vartheta}; \boldsymbol{\mathcal{D}})$, where $\boldsymbol{\vartheta} = (\alpha, \lambda, \boldsymbol{\beta}^\top)^\top$ and $\boldsymbol{\mathcal{D}} = (\boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{x})$, $\boldsymbol{t} = (t_1, \ldots, t_n)^\top$. Moreover, we assume that $\pi(\boldsymbol{\vartheta}) = \pi(\beta)\pi(\alpha)\pi(\lambda)$, with all the hyper-parameters specified in order to express non-informativeness. The joint posterior density $\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}}) \propto L(\boldsymbol{\vartheta}; \boldsymbol{\mathcal{D}})\pi(\boldsymbol{\vartheta})$ is analytically intractable. So, we based our inference on the MCMC simulation methods. Particularly, we resort to the Metropolis–Hastings algorithm.
Model comparison is made via deviance information criterion (DIC), the expected Akaike information criterion (EAIC) and the expected Bayesian (or Schwarz) information criterion (EBIC) can be used.
Bayesian case influence diagnostics is based on the ψ-divergence between P and $P_{(-i)}$, where P denotes the posterior distribution of $\boldsymbol{\vartheta}$ for full data, and $P_{(-i)}$ denotes the posterior distribution of $\boldsymbol{\vartheta}$ without the ith case. Specifically, $D_\psi(P, P_{(-i)}) = \int_{\vartheta \in \Theta} \psi\left(\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}}^{(-i)})/\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})\right) \pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})\, d\boldsymbol{\vartheta}$, where ψ is a convex function with $\psi(1) = 0$. With $\psi(z) = -\log(z)$ defining the

Kullback-Leibler (K-L) divergence, $\psi(z) = (z-1)\log(z)$ defining the J-distance, $\psi(z) = 0.5|z-1|$ defines the variational distance or L_1 norm, and $\psi(z) = (z-1)^2$ defines the χ^2-square divergence.

TABLE 1. Bayesian criteria (DIC/EAIC/EBIC) for the fitted models.

Activation	First	Last	Random
Criteria	423.3/430.9/450.8	434.1/440.3/460.3	428.9/435.3/451.9

4 Malignant melanoma data

In this section we work out an example employing our modeling. The data set includes 205 patients observed after operation for removal of malignant melanoma in a period of following up of 15 years (Scheike, 2009). The observed time (T) ranges from 10 to 5565 days and refers to the time until the patient's death or the censoring. Patient dead from other causes, as well as patients still alive at the end of the study are assumed to be censored observations (72%). We take tumor thickness, ulceration status and sex as covariates.

We fitted the GBScr models according to (2), (3) and (4). According to the DIC, $EAIC$ and $EBIC$ criteria (Table 1), the GBScr model under the first activation mechanism stands out as the best one, which we then select as our working model. Considering the of the GBScr model under the first activation mechanism the ψ-divergence measures were computed. For all ψ-divergence measures, the case 5 was identified as the most influential.

Acknowledgments: Special Thanks to CNPq and FAPESP, Brazil.

References

Balakrishnan, N., Leiva, V. & Lopez, J. (2007). Acceptance sampling plans from truncated life tests based on the generalized Birnbaum-Saunders distribution. *Communications in Statistics: Simulation and Computa- tion*, **36**, 643 – 656.

Cooner, F., Banerjee, S., Carlin, B. P. & Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, **102**, 560 – 572.

Leiva, V., Riquelme, M., Balakrishnan, N. & Sanhueza, A. (2008). Lifetime analysis based on the generalized Birnbaum-Saunders distribution *Computational Statistics and Data Analysis*, **52**, 2079 – 2097.

Scheike, T. (2009). *timereg package*. R package version 1.1-0. With contributions from T. Martinussen and J. Silver. R package version 1.1-6.

Inference in complex biological systems with Gaussian processes and parallel tempering.

Benn Macdonald[1], Frank Dondelinger[2], Dirk Husmeier[1]

[1] University of Glasgow, Department of Mathematics and Statistics, Scotland
[2] The Netherlands Cancer Institute, Netherlands

E-mail for correspondence: b.macdonald.1@research.gla.ac.uk

Abstract: Parameter inference in mathematical models of complex biological systems, expressed as coupled ordinary differential equations (ODEs), is a challenging problem. These depend on kinetic parameters, which cannot all be measured and have to be ascertained a different way. However, the computational costs associated with repeatedly solving the ODEs are often staggering, making many techniques impractical. Therefore, aimed at reducing this cost, new concepts using gradient matching have been proposed. This paper combines current adaptive gradient matching approaches, using Gaussian processes, with a parallel tempering scheme, in order to compare 2 different paradigms using the same nonlinear regression method. We use 2 ODE systems to assess our technique, showing an improvement over the recent method in Calderhead et al. (2008).

Keywords: Parameter inference; Ordinary differential equations; Adaptive gradient matching; Gaussian processes; Parallel tempering.

1 Introduction

Ordinary differential equations (ODEs) have many applications in modelling the behaviours of systems, from fluid mechanics to systems biology. Often, there is enough knowledge of a system to model it through mathematical equations, but there is intrinsic uncertainty in the kinetic parameters governing these. Conventional methods involving Markov Chain Monte Carlo (MCMC) tend to involve integrating the system of ODEs at each iterative step, to compare how well the sampled parameters match the data. However, the computational cost can be overbearing, making these methods impractical for larger systems, and more modern methods have sought an alternative to the explicit solution. The work by Calderhead et al. (2008), Campbell and Steele (2012) and Dondelinger et al. (2013), involves fitting an interpolant to the data, then comparing the gradients from the interpolant to those from the ODEs (known as gradient matching). The original method proposed by Calderhead et al. (2008) uses a methodological simplification, which effectively ignores the posterior correlation between the

ODE parameters and the Gaussian process (GP) hyperparameters in the sampling scheme, whereas Dondelinger et al. (2013) sample all the parameters from the posterior distribution (adaptive gradient matching (AGM)). Both Dondelinger et al. (2013) and Campbell and Steele (2012) temper towards the posterior (β-tempering, Section 2.), but Campbell and Steele (2012) differs with regards to the mismatch parameter (the difference between the gradients). Whereas Dondelinger et al. (2013) infer the mismatch parameter, Campbell and Steele (2012) temper this mismatch towards zero (γ-tempering). Since this is gradual, it avoids convergence problems. We combine both methods to create an adaptive gradient matching technique, using Gaussian processes and parallel tempering (both the β and γ variety).

2 Methodology

Consider a set of T arbitrary time points $t_1 < ... < t_T$, and a set of noisy observations $\mathbf{Y} = (\mathbf{y}(t_1), ..., \mathbf{y}(t_T))$, where $\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t)$, N = $\dim(\mathbf{x}(t))$, $\mathbf{X} = (\mathbf{x}(t_1), ..., \mathbf{x}(t_T))$. The signals of the system are described by ordinary differential equations (ODEs), of the form

$$\mathbf{x}' = \frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}, t); \qquad \mathbf{x}(t_1) = \mathbf{x}_1 \tag{1}$$

where $\boldsymbol{\theta}$ is a parameter vector of length p, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_n^2 \mathbf{I})$. Then,

$$P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\sigma}) = \prod_n \prod_t P(y_n(t)|x_n(t), \sigma_n) = \prod_n \prod_t N(y_n(t)|x_n(t), \sigma_n) \tag{2}$$

Now let $\mathbf{x}_n$ and $\mathbf{y}_n$ be T dimensional column vectors containing the n^{th} row of $\mathbf{X}$ and $\mathbf{Y}$. Following Calderhead et al. (2008), we place a GP prior on $\mathbf{x}_n$, $p(\mathbf{x}_n|\boldsymbol{\phi}) = N(\mathbf{x}_n|\mathbf{0}, \mathbf{C}_{\phi_n})$, where $\mathbf{C}_{\phi_n}$ is a positive definite matrix of covariance functions with hyperparamters ϕ_n. As the derivative of a GP is itself a GP, the conditional distribution for the state derivatives is

$$p(\mathbf{x}'|\mathbf{x}, \boldsymbol{\phi}) = N(\mathbf{m}_n, \mathbf{K}_n) \tag{3}$$

(analytical solutions to $\mathbf{m}_n$ and $\mathbf{K}_n$ in Dondelinger et al. (2013)). Assuming additive Gaussian noise with state-specific variance γ_n, from (1) we get

$$p(\mathbf{x'}_n|\mathbf{X}, \boldsymbol{\theta}, \gamma_n) = N(\mathbf{f}_n(\mathbf{X}, \boldsymbol{\theta}), \gamma_n \mathbf{I}) \tag{4}$$

Dondelinger et al. (2013) link the interpolant in (3) with the ODE model in (4) using a products of experts approach, obtaining a joint distribution for $p(\mathbf{X}', \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma})$. This can then be marginalised over in closed form (see Dondelinger et al. (2013) for details), to obtain $p(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma})$.
Following Dondelinger et al. (2013), we sampled $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ from the posterior distribution with MCMC. However, we did not sample $\boldsymbol{\gamma}$ directly, but instead followed Campbell and Steele (2012) to set up a ladder of fixed values associated with the "temperatures" of a parallel tempering scheme, choosing a Log_{10} scale. For details see the online supplementary material at http://www.stats.gla.ac.uk/~dhusmeier/MyPapers/IWSM2013Macd.pdf

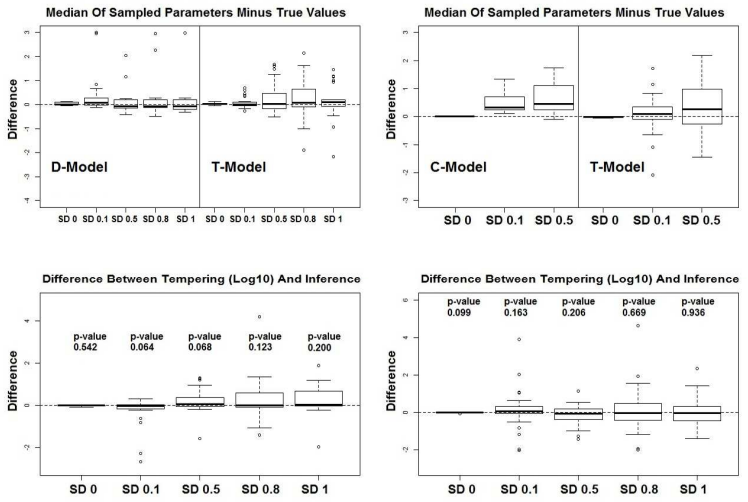


FIGURE 1. Parameter estimation accuracy of $\boldsymbol{\theta}$ over noise instantiations, for the Fitz-Hugh Nagumo (left) and Lotka-Volterra (right) systems. Some outliers in the plots have been removed for scalability. The dashed lines show zero difference. Top Row: Boxplots, over the 10 datasets, of differences between the median of sampled parameters and true values. The solid line splits the D-Model/C-Model (left) from the T-Model (right). Bottom Row: Boxplots, over the 10 datasets, of the differences in parameter estimation accuracy for the D-Model and T-Model. The p-values for a paired t-test are shown above the corresponding boxplot.

3 Results

We tested our method on the Fitz-Hugh Nagumo (FitzHugh (1961) and Nagumo et al. (1962)) and Lotka-Volterra (Lotka (1932)) ODE models. For space restrictions, details of the equations and parameters have been relegated to the online supplementary material.

We introduce the abridged notation used in this section: The method described in Calderhead et al. (2008) shall be denoted, C-Model, the method described in Dondelinger et al. (2013), D-Model, and the new combined method proposed in this paper, T-Model. For each system, method and added observational noise level, 10 datasets were generated. By averaging over these, we are able to remove specific characteristics of a dataset and observe more clearly our method's performance. The median was used as an estimator of the parameters and the true values were subtracted from the sampled parameter estimates. The distributions (of estimate minus true value) over the 10 datasets were compared.

The first row of FIGURE 1. shows the distribution of the estimate to the true parameter for the D-Model, C-Model and T-Model (Log_{10}), for the

FhN and LV systems. For zero noise, both the C-Model and T-Model have boxplots centred very close to zero, displaying good performance. However, when increasing the noise, the C-Model no longer has a distribution centred around zero (no part of the distribution for noise = 0.1 and only a small part of the lower tail for noise = 0.5). For all noise instantiations, the T-Model (and D-Model) has most of its mass centred around zero. Therefore, if averaging over all datasets, for the T-Model, the true parameters are close to the estimates i.e. this technique is unbiased. The second row of FIGURE 1. allows us to check how robust our technique is. The plots show the distributions of the differences between the absolute distance of the estimator to the true parameter for the T-Model and D-Model. These distributions are centred around zero, indicating that there is no noticeable difference between the parameter estimation accuracy of these two techniques. We can therefore see that our technique is robust to noise.

4 Conclusion

We have carried out a comparative evaluation of two schemes for adaptive gradient matching: posterior inference vs. parallel tempering of the gradient mismatch hyperparameter. The tempering scheme was originally proposed in the context of splines-based regression, which we have adapted to non-parametric Bayesian modelling, with Gaussian processes. An application to data, generated from two different systems of ODEs, shows no significant difference between the parallel tempering and posterior inference. We found that both methods outperform a related method by Calderhead et al. (2008), considered the current state of the art.

References

Calderhead, B. et al. (2008). Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *NIPS*, **22**,

Campbell, D. and Steele., R.J. (2012). Smooth functional tempering for nonlinear differential equation models. *Stat Comput*, **22**, 429 – 443.

Dondelinger, F. et al. (2013). ODE parameter inference using adaptive gradient matching with Gaussian processes. *In Press: 16th AISTATS*

FitzHugh, R. (1961). Impulses and physiological states in models of nerve membrane. *Biophys. J.*, **1**, 445 – 466.

Lotka, A. (1932). The growth of mixed populations: two species competing for a common food supply. *Journal of the Washington Academy of Sciences*, **22**, 461 – 469.

Nagumo, J.S. et al. (1962). An active pulse transmission line simulating a nerve axon. *Proc. Inst. Radio Eng.*, **50**, 2061 – 2070.

Modelling of the distribution of incomes with the use of finite mixtures of distributions

Ivana Malá[1]

[1] University of Economics, Prague, Czech Republic

E-mail for correspondence: malai@vse.cz

Abstract: In the contribution the distribution of the net yearly incomes of the Czech households in 2005-2010 is modelled with the use of finite mixtures of lognormal and gamma distributions with unknown component membership. The net yearly income per equivalised unit according to the methodology of the European Union are studied together with the development of the household size and number of equivalised units (according to EU and OECD definitions). Finite mixture models are useful for the description of distributions of random variables in non-homogeneous populations as the incomes of the households are. The models with 3 artificial components was chosen from mixtures with 2-4 components. The EM algorithm is used to obtain estimates of parameters, all computations are made in R.

Keywords: finite mixture of distributions; lognormal distribution; gamma distribution; income distributions; EM algorithm

1 Introduction

The modelling of the distribution of incomes is a frequently treated problem as the results are of interest of a large spectrum of analysts and researchers. Estimated characteristics as the mean or median income are followed up by wide range of people. In this contribution data of incomes of the Czech households are analyzed. The development of incomes per capita in the Czech Republic can be found for example in Bílková, Malá (2012) or Bartošová, Bína (2009). The goal of this text is to construct a finite mixture models to fit net yearly equivalised incomes of the Czech households in 2004-2010. Income data are usually very non-homogeneous and mixture models are the suitable approach how to treat it. As a result of the modelling we obtain information about components and its distribution as well as about structure of the mixture (proportions of components in the mixture) and the distribution in the whole population of households. Equivalised incomes are defined as a total net yearly income divided by the equalised size of household (equalised number of units). Number of units reflects the structure of the household and the possibility to share spend-

ings and usually are evaluated according to two methodologies, European Union (EU) and OECD:

OECD: first adult 1, members above 13 years 0.7, members below 13 0.5
EU: first adult 1, members above 13 years 0.5, members below 13 0.3.

All units (number of members, equivalised units given by EU and OECD methodology) are equal for single member households, otherwise it follows

number of members > units OECD>units EU.

For the equivalised incomes the inequality is reversed. In this text equivalised net year income in the the Czech Republic (in Czech koruna, CZK) is treated according to EU methodology. All presented models are acceptable (however chi-square test rejects the distribution), the Akaike information criterion could be used to find the best model (in each year only). The EM algorithm (McLachlan, Peel (2000)), implemented in the program R, is used to find maximum likelihood estimates of unknown parameters.

2 Results

Suppose the distribution of equivalised income of the Czech households is a mixture of K lognormal and gamma components. The mixture of probability distributions is given as

$$f(x, \boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j f(x, \boldsymbol{\theta}_j),$$

where for $j = 1, ..., K$ $f(x, \boldsymbol{\theta}_j)$ are probability densities of lognormal or gamma distributions and π_j denotes weights of the components in the mixture. Unknown parameters in the model (if number of components K is selected) are parameters of component distributions $\boldsymbol{\theta}_j$ and $K-1$ parameters π_j. For two-parametric distributions the vector parameter $\boldsymbol{\psi}$ contains $2K+(K-1)$ parameters to be estimated, Kleiber, Kotz (2003), Wiper et al (2001). For the estimation of unknown parameters data from the Living Conditions Survey (a national module of the European Union Statistics on Income and Living Conditions (EU-SILC)) dealing with the Czech households in 2005-2011, that cover incomes from 2004 to 2010, are used. The survey has been carried by the Czech Statistical Office annually since 2005 (CZSO). From this survey, data about net incomes, number of members of the household and number of equivalised units are used together with the weights of households reflecting the two-stage samplings scheme. The sample sizes in the analysed years are 4,351, 7,483, 9,675, 11,294, 9,911, 9,098 and 8,066 households. Large samples properties of estimates can be used to obtain standard errors of estimates.

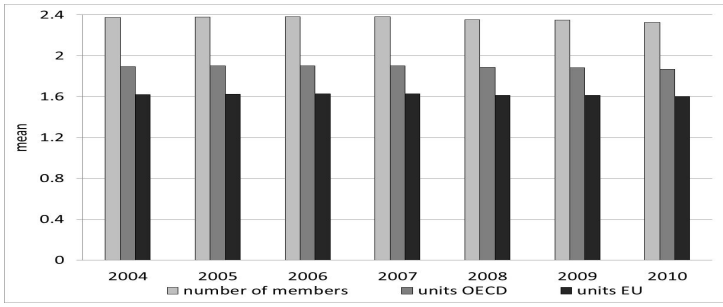


FIGURE 1. Mean number of members and equivalised units 2005-2010.

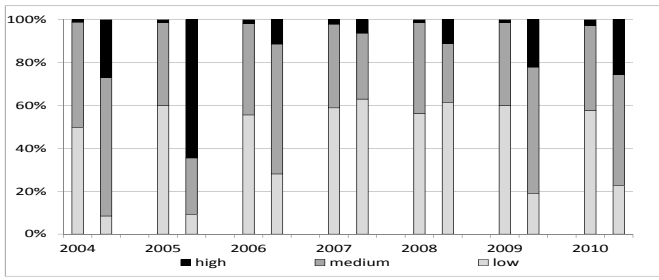


FIGURE 2. Estimated proportions of components 2005-2010, lognormal distribution left column, gamma distribution right column.

From the Figure 1 the slow decrease in all three characteristics of the size of the Czech households is observable, the number of members of the households declines slowly from the mean value 2.37 in 2005 to 2.26 in 2010. Percentage of single member households increased from 22.5 per cent by one percent.

The model with three components was chosen from the models with 2 to 5 components (according to the Akaike criterion, numeric results, interpretation). The three component model distinguishes three artificial subgroups that can represent households with low, medium and high incomes. In the Figure 2 the estimated component proportions are shown for lognormal and gamma distributions. It is obvious that the percentage strongly depends on the component distribution and they really differ for analysed distributions. In the Figure 3 estimated expected values are shown for both distributions (columns), values are similar for low and medium income households, there

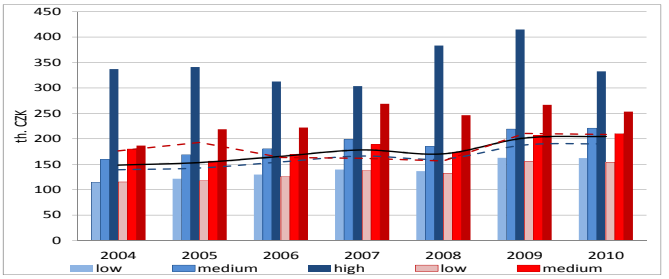


FIGURE 3. Estimated component expected values for both models 2005-2010.

is large difference in the high incomes component. The gamma model found this component with high level and small proportion, the lognormal model lower value with higher proportions. Estimated expected values from the mixtures are given (dashed lines) in the Figure 3 together with the mean values (solid line) evaluated from the samples.

Acknowledgments: Research was supported by the grant IG 410062 from the Faculty of Informatics and Statistics, University of Economics, Prague.

References

Bartošová J., Bína V. (2009). Modelling of Income Distribution of Czech Households in Years 1996-2005. *Acta Oeconomica Pragensia.*, **17**, 3 – 18.

Bílková D., Malá I. (2012). Application of the L-Moment Method when Modelling the Income Distribution in the Czech Republic. *Austrian Journal of Statistics.*, **41**, 125 – 132.

Kleiber, C., Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences.* Wiley-Interscience, New York.

McLachlan, G. J., Peel, D. (2000). *Finite Mixture Models.* Wiley series in Probability and Mathematical Statistics: Applied Probability and Statistics Section, New York.

Wiper, M., Rios Insua, D., Ruggeri, F. (2001). Mixtures of Gamma Distributions with Applications. *Journal of Computational and Graphical Statistics*, **10**, 440 – 454.

CZSO (08.03.2013). *Czech Statistical Office.* URL: http://www.czso.cz

Study of the longevity in Sardinia: an application of the Beta skew-normal regression

Valentina Mameli[1], Monica Musio[2], Luca Deiana[3]

[1] Dipartimento di Scienze Statistiche, Università di Padova, Italy
[2] Dipartimento di Matematica ed Informatica, Università di Cagliari, Italy
[3] Dipartimento di Scienze Biomediche, Università di Sassari, Italy

E-mail for correspondence: mameli@stat.unipd.it

Abstract: In real applications normality of the errors is a routine assumption for the linear model, but it may be unrealistic. In fact often residuals exhibit non-normal shape, with an heavy right or left tail. In this work, we relax the normality assumption by considering that the errors follow a Beta skew-normal distribution. The new regression model includes as special cases the skew-normal and the normal one. We apply such model to study the longevity in Sardinia.

Keywords: Centenarians; Beta skew-normal; regression; longevity; Sardinia.

1 Introduction

Sardinia has been called "the Centenarian island"(Deiana and Vaupel (2006)). In fact, it turns out to be one of the regions with more alive centenarians in the world. In recent years, several projects have started with the objective to understand and analyse which factors may be related to the longevity in Sardinia. In particular, the project AKEA (see for example Deiana and Vaupel (2006), Poulain et al. (2004) and the references therein) was focused on a census of all centenarians living in Sardinia. This project has highlighted the presence in the island of geographical areas in which the phenomenon of longevity is particularly important. In this work we have analysed data from the AKEA project, collected and validated in two villages of Sardinia. Our aim is to address the following question: Do members of families in which there are centenarians live longer in average? This is achieved by comparing the mean age at death for individuals belonging to a centenarians' family to that for individuals from families having not centenarians. Our response variable, the age of death, is strongly asymmetric. Recent statistical literature has seen an increasing interest in the construction of flexible parametric families of distributions that exhibit skewness and kurtosis different from the normal distribution. Azzalini (1985) defined

the skew-normal distribution and studied its properties. Subsequently, Azzalini and Capitanio (1999) introduced the skew-normal regression model. Recently, Mameli and Musio (2013) proposed a new distribution, called Beta skew-normal (BSN), which generalizes the SN distribution. In this work we propose an extension of the SN regression model, in which the errors follow a BSN distribution. We apply this model to study the longevity in Sardinia. The paper unfolds as follows: the definition of the BSN distribution and the BSN regression model are presented in section 2. Section 3 is devoted to data and results of the statistical analysis. A brief discussion is given in section 4.

2 Statistical model

2.1 The Beta skew-normal distribution

A random variable Z is said to have a Beta skew-normal distribution with parameters λ, a and b ($BSN(\lambda, a, b)$), if its density is given by

$$g^{B}_{\Phi_\lambda(z)}(z;\lambda,a,b) = \frac{1}{B(a,b)}(\Phi_\lambda(z))^{a-1}(1-\Phi_\lambda(z))^{b-1}\phi_\lambda(z),\ z \in R,$$

with $a > 0,\ b > 0,\ \lambda \in R$. The functions $\Phi_\lambda(\cdot)$ and $\phi_\lambda(\cdot)$ are the cdf and the pdf of the skew-normal distribution, respectively. Here, $B(a,b)$ is the beta function. The BSN distribution can be generalized by the inclusion of the location and scale parameters which we identify as μ and $\sigma > 0$. Thus if $Z \sim BSN(\lambda,\ a,\ b)$ then $X = \mu + \sigma Z$ is a $BSN(\mu, \sigma, \lambda, a, b)$.

2.2 The Beta skew-normal regression model

The Beta skew-normal regression model is defined as

$$y_i = x_i\beta + \sigma\epsilon_i, \quad \text{for} \quad i = 1, \ldots, n$$

where $\beta = (\beta_1, \ldots, \beta_k)$ is a $k \times 1$ vector of unknown regression parameters ($k < n$), $x_i = (x_{i1}, \ldots, x_{ik})$ is the vector of k covariates. Under the assumption that each error ϵ_i is distributed as a $BSN(\lambda, a, b)$, the response variable y_i is distributed as a $BSN(x_i\beta, \sigma, \lambda, a, b)$. The new model contains as special cases the skew-normal and the normal regression models. The log-likelihood function based on a sample of N independent observations is

$$l(\theta) = \sum_{i=1}^{N}\left\{\log(t_i) + \log(v_i)^{(a-1)} + \log(1-v_i)^{(b-1)} - \log(\sigma Beta(a,b))\right\},$$

where $t_i = \phi_\lambda(z_i)$ and $v_i = \Phi_\lambda(z_i)$, with $z_i = \frac{y_i - x_i\beta}{\sigma}$. The score vector $U(\theta)$, obtained by differentiating $l(\theta)$ with respect to $\theta = (\beta, \sigma, \lambda, a, b)$, has

the following components

$$\begin{aligned} U_\beta(\theta) &= \sum_{i=1}^N \left[\frac{x_i(z_i - w_i t_i - \lambda h_i)}{\sigma}\right]; \\ U_\sigma(\theta) &= \sum_{i=1}^N \left[\frac{z_i^2 - 1 - z_i(w_i t_i + \lambda h_i)}{\sigma}\right]; \\ U_\lambda(\theta) &= \sum_{i=1}^N \left(z_i h_i + w_i \frac{\partial v_i}{\partial \lambda}\right); \\ U_a(\theta) &= \sum_{i=1}^N \{\log v_i - [\psi(a) - \psi(a+b)]\}; \\ U_b(\theta) &= \sum_{i=1}^N \{\log(1 - v_i) - [\psi(b) - \psi(a+b)]\}; \end{aligned}$$

where $\psi(t)$ is the di-gamma function. Furthermore, $w_i = \frac{(a-1)+(2-a-b)v_i}{(1-v_i)v_i}$, and $h_i = \frac{\phi(\lambda(z_i))}{\Phi(\lambda(z_i))}$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and the pdf of the normal distribution, respectively. Maximum likelihood estimates of the parameters can be found by setting the above expressions equal to zero and solving them simultaneously. Since analytical solutions are not available in closed form, the R package `maxLik` is used to find estimates numerically.

3 Data description and statistical analysis

Data were collected in the two villages of Ovodda and Tiana in Sardinia, located 5 km apart in the province of Nuoro, in an area characterized by exceptional longevity. The data were taken from the project AKEA, and cover the time period from 1860 to 2009. The experimental design was as follows: we first identified the most recent centenarians who die; then we included in the study all the present and previous members of their family, as identified by the family name. In this way, 8 families were identified and compared with 7 families in which, in the same period, there were no centenarians. The centenarians used to identify their families were removed from the study. As the causes of death in children are very different in nature from those in adults, we have restricted our analysis to individuals who died in adulthood (at 30 years or more). The dataset contains 932 cases, 698 of which coming from families of centenarians. We consider as response variable the age at death, which is strongly asymmetric. For each individual sampled we also know the date of birth, the sex and the year of death, that we consider as potential predictors. Since this area is characterized by an exceptional male longevity as well as a low female/male centenarian ratio (see Poulain et al. (2004)), we have excluded the variable sex from the model. As predictors we consider the year of death of each individual and the dummy variable $Cent1$ (which is 1 if there is at least one centenarian in the family and 0 otherwise). Then we assume the following model

$$y_i = \beta_0 + \beta_1 year\_death_i + \beta_2 Cent1_i + \sigma\epsilon_i, \quad \text{for} \quad i = 1, \ldots, 932$$

where the error ϵ_i follows a $BSN(\lambda, a, b)$ density function. We use the likelihood ratio (LR) test statistic for comparing the SN model against

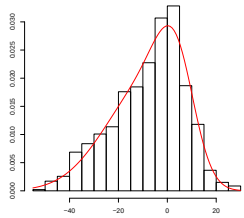


FIGURE 1. Histogram of the residuals with superimposed the BSN density

the BSN one. Based on the values of the LR test statistic, the BSN model provides a better fit than the SN ($LR = 9.604$, p-value $= 0.008$). The estimates of the regression parameters are $\hat{\beta}_0 = 73.999$, $\hat{\beta}_1 = 8.211$ and $\hat{\beta}_2 = 2.394$. The estimated density function is plotted in figure 1.

4 Discussion

In this paper we have introduced the BSN regression model and we have used it to study the longevity in Sardinia. Our analysis shows a significant effect on longevity for persons belonging to a centenarians' family. Indeed, the effect is probably underestimated, since we have taken no account of the mother's family. Even so, it seems clear that longevity has a genetic cause. This finding should motivate more detailed studies to investigate the genetic characteristics of centenarians' families.

References

Azzalini, A. (1985). A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, **12(2)**, 171 – 178.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society, Series B*, **61**, 579 – 602.

Deiana, L. and Vaupel, J. (2006). Longevity in Sardinia, The Centenarian Island. *Biochimica Clinica*, **30(1)**, ISSN 0393 3504.

Mameli, V. and Musio, M. (2013). A new generalization of the Skew-normal distribution: the Beta skew-normal. *Communications in Statics: Theory and Methods*, to appear.

Poulain, M., Pes, G., Grasland, C., Carru, C., Ferrucci, L., Baggio, G., Franceschi, C., and Deiana, L. (2004). Identification of a geographic area characterized by extreme longevity in Sardinia Island: the AKEA study. *Experimental Gerontology*, **39**, 1423 – 1429.

An Application of The Vector AutoRegression (VAR) Model to The Analysis of The Sun–Earth's Climate Connection

Elizabeth Martínez-Gómez[1,2], Victor Guerrero[1], Francisco Estrada[3,4]

[1] Departamento de Estadística, Instituto Tecnológico Autónomo de México, Río Hondo 1, Col Progreso Tizapán, 01080, México D.F., México
[2] Center for Astrostatistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA, 16802, USA
[3] Centro de Ciencias de la Atmósfera, Universidad Nacional Autónoma de México, Circuito Exterior s/n, Ciudad Universitaria, 04510, México D.F., México
[4] Institute for Environmental Studies, De Boelelaan 1087, 1081 HV Amsterdam, The Netherlands

E-mail for correspondence: elizabeth.martinez@itam.mx

Abstract: It is well-known that the Sun has an obvious effect on terrestrial climate since its electromagnetic radiation is the main energy for the outer envelopes of Earth. This is one of the so-called *solar-terrestrial physics* problems. It has been the subject of speculation and research by scientists for many years. Understanding the behaviour of natural fluctuations in the climate is especially important because of the possibility of man-induced climate changes. Many studies have been conducted to show correlations between solar activity and various meteorological parameters, nevertheless historical observations of solar activity were restricted to sunspot numbers and it was not clear how these could be physically related to meteorological factors. The theoretical models proposed to explain these relationships are not conclusive since the complexity of the system, that is, many correlated variables, lack of data, and an incomplete understanding of the physical mechanisms responsible for such an interaction. In this work we study the problem through the application of the Vector AutoRegression Model (VAR). The advantage of this approach is that we can analyse several variables simultaneously and search for the causality among them. The results indicate that the sun (described only by the number of sunspots and its total irradiance) has a weak connection to Earth, at least for the major climate phenomena. Further investigation must be conducted by including another proxy variables for the solar activity and studying the solar cycles separately.

Keywords: time series; Vector Autoregression; climate; Sun.

1 The solar and terrestrial connection

The Sun varies over a broad span of timescales, from its brightening over its lifetime to the fluctuations commonly associated with magnetic activity over days to years. The latter activity includes most prominently the 11-year sunspot cycle and its modulations. Variations in the total solar irradiance (TSI) incident on Earth's atmosphere can cause imbalances in Earth's radiation budget that can induce temperature shifts near the surface. The temperature of Earth can be understood to a first approximation as controlled by the balance between the radiative energy received from the Sun and Earth's thermal emission of radiative energy to space. Thermal emission increases with increasing temperature, and Earth can be thought of as settling into an equilibrium by adjusting its temperature so that this thermal radiation balances the solar energy absorbed by the planet. An increase or a decrease in the TSI is expected on this basis to increase or decrease the temperature of Earth. The observed sunspot number has been demonstrated to be negatively correlated with the cosmic ray flux. The cosmic-ray flux reaching Earth's surface is modulated by the strength of the solar wind. It is now understood that this decrease in cosmic rays is due to changes in the magnetic field geometry in the heliosphere, the bubble blown in the interstellar medium by the solar wind. Higher levels of solar activity lead to a decrease in the cosmic-ray flux at Earth. Cosmic rays are potentially implicated in climate change on Earth because as they penetrate Earth's atmosphere they leave behind an ionized path that could serve as a source of condensation centres that in turn affect cloudiness and Earth's albedo (reflectivity of solar radiation). Research is being conducted on these potential mechanisms and their possible relevance as a climate-forcing agent.

2 The Vector AutoRegression (VAR) model

The VAR model is one of the most successful, flexible, and easy to use models for the analysis of multivariate time series. It is a natural extension of the univariate autoregressive model (AR) to dynamic multivariate time series (Sims 1972, 1980, 1982). It describes the evolution of a set of k variables (called *endogenous variables*) over the same sample period ($t = 1, 2, \ldots, T$) as a linear function of only their past evolution

$$y_t = c + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \epsilon_t \tag{1}$$

where c is a $k \times 1$ vector of constants (intercept), α_i is a $k \times k$ matrix (for every $i = 1, \ldots, p$), p is the number of lags (that is, the number of periods back), and ϵ_t is a $k \times 1$ vector of error terms. In addition there are some assumptions about the error terms: a) the expected value is zero, that is,

$E[\epsilon_{it}] = 0$ with $t = 1, \ldots, T$, and b) the errors are not autocorrelated, $E[\epsilon_{it}\epsilon_{j\tau}] = 0$ with $\tau \neq t$.
The determination of the number of lags p is a trade-off between the dimensionality and simpler models. To find the optimal lag length we can apply a Log-Likelihood Ratio test (LR) test or an information criterion (Lütkepohl 1993).
After the parameters in the $VAR(p)$ model (Eq. 1) have been estimated through Ordinary Least Squares (OLS), we need to interpret the dynamic relationship between the indicated variables using the Granger causality.

3 Application of the VAR model to the Sun–Earth's Climate connection

Our purpose is to analyse the relationship between the Sun (solar activity) and the major climate phenomena through a time series analysis. The data are taken from the National Geophysical Data Center (NGDC) and the National Climatic Data Center (NCDC).

3.1 Description of the data

Any study of a connection between solar variability and Earth's climate must involve historical observations, not only for developing empirical models but also for testing physically-based hypotheses. It is therefore important to have long records of high-quality observations of both the Sun and Earth's climate and to be able to formulae a physical model explaining the causality. However, there are few long-term observations that really are suitable for such studies. The most complete long-term data sets that exist are *surface measurements of temperature*, *sea-level pressure*, and *sea surface temperature*. Similarly, there are few high-quality, long-term solar observations, and one of the few direct solar quantities that has been recorded for more than 100 years is the *sunspot activity*. There also have been measurements of the geomagnetic field for more than 100 years, and these records hold some information about solar activity. Thus we have,

1. **Solar activity**. Comprises photospheric and chromospheric phenomena such as sunspots, prominences and coronal disturbances. We select two indicators: *Sunspot number* (1700–today) and *Total Solar Irradiation (TSI)* (1610–1978, reconstructed and 1978–today, satellites).

2. **Earth's global climate**. We are interested in a description of Earth's climate as a whole, with all the regional differences averaged. We select the following variables: *Pacific Decadal Oscillation (PDO)* (1900–today), *North Atlantic Oscillation (NAO)* (1821–today), *Hemispheric temperature anomalies* (1850–today).

4 Summary

The possible relation between the solar activity and the terrestrial climate has been addressed in many works. Most of them search for periodicities or correlations among the set of variables that characterize the solar activity and the major climate parameters. In this work we have proposed and estimated a VAR model to explain such a "possible" connection. To apply the VAR model we have analysed the time series for the most remarkable features of both the solar activity and climate. From our statistical analysis we find that the Sun can be modelled by a VAR(8) where the chosen variables are good proxies for the solar activity. A second model which accounts for the relation between the Sun and terrestrial climate is given by a VAR(4) where the solar variables are taken as exogenous. This latter model is a first evidence that the solar activity *does not strongly affect the major climate variables*, in other words, the long-term variability. In a forthcoming work we will include a term related to the cloudiness (albedo) and we will separate certain cycles in the solar activity to determine the epochs in which the connection with the terrestrial climate could be stronger.

Acknowledgments: E. Martínez-Gómez acknowledges CONACyT Postdoctoral Fellowship, the Schlumberger Foundation and the Center for Astrostatistics at Penn State University for the financial and academic support. V. Guerrero thanks the support from Asociación Mexicana de Cultura, A. C.

References

Lean, J., Beer, J. and Bradley, R. (1995). Reconstruction of solar irradiance since 1610: Implications for climate change. *Geophysical Research Letters*, **22**, 23, 3195 – 3198.

Lütkepohl, H. (1993). *Introduction to multiple time series analysis*. Germany: Springer Verlag.

Sims, C. A. (1972). Money, Income, and Causality. *American Economic Review*, **62**, 540 – 552.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, **48**, 1 – 48.

Sims, C. A. (1982). Policy Analysis with Econometric Models. *Brooking Papers on Economic Activity*, **1**, 107 – 152.

Influence diagnostics in mixed-effects models with censored data using the multivariate-t distribution

Larissa A. Matos[1] and Victor H. Lachos[1]

[1] Department of Statistics, Campinas State University, Brazil

E-mail for correspondence: larissam@ime.unicamp.br

Abstract: HIV RNA viral load measures are often subjected to some upper and lower detection limits depending on the quantication assays, and consequently the responses are either left or right censored. Linear and nonlinear mixed-effects models with censored response (LMEC and NLMEC), are routinely used to analyze this type of data. Although normal distributions are commonly assumed for random effects and residual errors, such assumptions make inferences vulnerable to outliers. The sensitivity to outliers and the need for heavy tailed distributions for random effects and residual errors motivate us to use a likelihood-based inference for linear and nonlinear mixed effects models with censored response (NLMEC/LMEC) based on the multivariate Student-t distribution. In order to examine the performance of the proposed model, some simulation studies are presented in order to show the robust aspect of it against outlying and influential observations. The sensitivity of the EM estimates under some usual perturbation schemes in the model or data, the local influence curvatures are derived and some diagnostic graphics are proposed. The one-step approximations of the estimates in the case-deletion model are also obtained.

Keywords: Censored data; HIV viral load; EM Algorithm; Influential observations; Linear and nonlinear mixed models.

1 Introduction

Studies of HIV viral dynamics, often considered to be a key issue in AIDS research, considers repeated/longitudinal measures over a period of treatment routinely analyzed using linear and non-linear mixed effects models (LME/NLME) to assess rates of changes in HIV-1 RNA level or viral load (Wu, 2005, 2010). Viral load measures the amount of actively replicating virus and its reduction is frequently used as a primary endpoint in clinical trials of anti-retroviral (ARV) theraphy. However, depending on the diagnostic assays used, its measurement may be subjected to some upper and lower detection limits, below or above which they are not quantifiable (resulting in left or right censoring). The proportion of censored data in these

studies may not be small (Hughes, 1999) and so the use of crude/adhoc methods, such as substituting a threshold value or some arbitrary point like mid-point between zero and cut-off for detection (Vaida & Liu, 2009), might lead to biased estimates of fixed effects and variance components (Wu, 2010).
Recently, Matos, Prates, Chen & Lachos (2012) proposed an algorithm for linear and nonlinear mixed effects models with censored response based on the multivariate Student-t distribution, using this models we developed and presented diagnostic measures for assessing the local influence.
The study of influence analysis is an important and key step in data analysis subsequent to parameter estimation. This can be carried out by conducting an influence analysis for detecting influential observations. There are two primary approaches for detecting influential observations. The first approach is the case-deletion approach (Cook, 1977) and it is an intuitively appealing method (see also Cook and Weisberg, 1982). Deletion diagnostics such as Cooks distance or the likelihood distance have been applied to many statistical models. The second approach, which is a general statistical technique used to assess the stability of the estimation outputs with respect to the model inputs, is the local influence approach of Cook (1986). Following the pioneering work of Cook (1986), this method has received considerable attention recently in the statistical literature on mixed effects models (LME/NLME); see, for example, Lesaffre & Verbeke (1998), Zhu & Lee (2001), Lee & Xu (2004), Osorio et al. (2007) and Russo et al. (2009), among others.
Zhu & Lee (2001) developed an approach for performing local influence analysis for general statistical models with missing data, and it is based on the Q-displacement function that is closely related to the conditional expectation of the complete-data log-likelihood in the E-step of the EM algorithm. This approach produces results very similar to those obtained from Cooks method. Moreover, the case-deletion can be studied by Q-displacement function following the approach of Zhu et al. (2001). So, using the same methods as in Matos et al. (2013) we develop here methods to obtain case-deletion measures and local influence measures by using the method of Zhu et al. (2001) (see also Lee & Xu, 2004; Zhu & Lee, 2001) in the context of mixed effects models with censored response based on the multivariate Student-t distribution.

2 Methodology

Based on Matos, Prates, Chen & Lachos (2012) where proposed an algorithm for linear and nonlinear mixed effects models with censored response based on the multivariate Student-t distribution, we developed and presented diagnostic measures for assessing the local influence.

Diagnostic analysis

Influence diagnostic techniques are used to identify anomalous observations that impact on model fitting or statistical inference for the assumed statistical model. There are primarily two approaches for detecting influential observations. The case- deletion approach (Cook, 1977) is the most popular one for identifying influential observations. To assess the impact of influential observations on parameter estimates some metrics have been used for measuring the distance between $\widehat{\theta}_{[i]}$ and $\widehat{\theta}$, such as the likelihood distance and Cooks distance. The second approach is a general statistical technique used to assess the stability of the estimation outputs with respect to the model inputs (Cook, 1986). By using the results of Zhu et al. (2001), we introduce here the case-deletion measures and the local influence diagnostics for the censored data on the basis of Q-function; see Zhu & Lee (2001). We first consider the case- deletion measures, then the local influence measures, and finally the perturbation schemes employed.

Numerical illustrations

We illustrate the performance of the proposed methods with the analysis of two HIV datasets, previously analyzed by Vaida & Liu (2009), and the analysis of a simulated example.
This work provides a new insight into the classical diagnostics methods for censored linear and nonlinear mixed effects models, typically used for analyzing censored HIV viral load outcomes, and also presents an useful expectation conditional maximization (EMC) algorithm, which enable the development of diagnostic influence measures. Explicit expressions are obtained for the Hessian matrix $\ddot{Q}$ and for the matrix Δ under different perturbation schemes. For NLMEC, the analysis is mathematically (and computationally) feasible through a linearization procedure. The proposed methodology has been applied to two recent (left and right-censored) AIDS studies, which is freely downloadable from R.

Acknowledgments: L. A. Matos acknowledges support of the FAPESP-Brazil.

References

Cook, R. (1977). Detection of influential observation in linear regression. *Technometrics*, 15 – 18.

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133 – 169.

Matos, L. A., Lachos, V. H, Balakrishnan, N. and Vilca-Labra, F.

(2013). Inluence Diagnostics in Linear and Nonlinear Mixed-Effects Models with Censored Data. *Computational Statistics and Data Analysis*, **57**, 450 – 464.

Matos, L. A., Prates, M. O., Chen, M-H. and Lachos, V. H. (2012). Likelihood Based Inference for Linear and Nonlinear Mixed-Effects Models with Censored Response Using the Multivariate-t Distribution. Accepted for publication in Statistica Sinica. doi:10.5705/ss.2012.043

Meng, X. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **81**, 633 – 648.

Meng, X. and Van Dyk, D. (1997). The EM algorithm: An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **59**, 511 – 567.

Vaida, F. and Liu, L. (2009). Fast implementation for normal mixed Effects models with censored response. *Journal of Computational and Graphical Statistics*, **18**, 797 – 817.

Bayesian estimation with INLA for logistic multilevel models

Silvia Metelli[1], Leonardo Grilli[1], Carla Rampichini[1]

[1] Department of Statistics, Computer Science, Applications - University of Florence, Italy

E-mail for correspondence: grilli@disia.unifi.it

Abstract: In multilevel models for binary responses, estimation is computationally challenging due to the need to evaluate intractable integrals. In this paper, we investigate the performance of a recently proposed Bayesian method for deterministic fast approximate inference, called Integrated Nested Laplace Approximation (INLA). In particular, we conducted a simulation study, comparing the results obtained via INLA with the results obtained via MCMC, i.e. the traditional estimation method in the Bayesian context, and via maximum likelihood with adaptive quadrature. Particular attention is devoted to the case of small sample size and to the specification of the prior distribution for the variance.

Keywords: Integrated Nested Laplace Approximations; Logistic multilevel models; MCMC estimation; Prior specification.

1 Introduction

Multilevel models are a standard tool for the analysis of hierarchically structured data. However, in most non linear case the estimation process is complicated by the fact that the likelihood cannot be written in closed form, requiring numerical or Monte Carlo techniques. Two popular approaches are maximum likelihood with Adaptive Gaussian Quadrature (AGQ) and Bayesian inference via MCMC methods. The latter approach gives more accurate results at the cost of increasing the computational effort and adding a dependency of the results on the specification of the prior distributions and in large samples the computational time of MCMC can be prohibitive. Therefore, it is worthwhile to evaluate the performance of alternative estimation methods, such as the Integrated Nested Laplace Approximation (INLA) proposed by Rue et al. (2009). This method performs approximate Bayesian inference based on multiple Laplace approximations combined with numerical integration. Fong et al. (2010) considered the application of INLA to generalized linear mixed models. However, they did not carry out any simulation study to assess the properties of the estimators; moreover, the comparison was limited to maximum likelihood via PQL, which in

general is less accurate than adaptive quadrature. In this paper we assess the performance of Bayesian inference via INLA for a random intercept logit model, making comparisons with Bayesian MCMC Gibbs sampling and maximum likelihood with Adaptive Gaussian Quadrature.

2 The simulation study

The simulation study refers to the following two-level random intercept logit model, for a dicothomous response y_{ij}, where level 2 units are indexed by $j = 1, ..., J$ and level 1 units by $i = 1, ..., n_j$:

$$logit(\pi_{ij}) = \alpha + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \gamma_1 z_{1j} + \gamma_2 z_{2j} + u_{0j}, \quad u_{0j} \sim N(0, \sigma^2_{u0}) \quad (1)$$

where $\pi_{ij} = P(y_{ij} = 1 \mid x_{1ij}, x_{2ij}, z_{1j}, z_{2j}, u_{0j})$, x_{1ij} is a continuous level 1 variable, x_{2ij} a binary level 1 variable, z_{1j} a continuous level 2 variable and z_{2j} a binary level 2 variable.

We consider several balanced designs with different number of clusters (10, 40, 70, 100) and cluster size (10, 30, 50). For the level 2 variance σ^2_{u0}, we consider the values 0.25, 1.44 and 6.25. The number of scenarios is thus $4 \times 3 \times 3 = 36$. For each configuration, 1000 datasets are generated.

The three examined estimation methods are: Bayesian inference via INLA, using the `R` function `inla`; via MCMC (Gibbs sampling), using the JAGS program; maximum likelihood through AGQ, using the `R` package `lme4`.

Moreover, for Bayesian methods we use flat prior distributions N(0,1000) for the fixed effects and three different specifications for the precision (the inverse of the level 2 variance): (*i*) $\Gamma(1, 0.0005)$, the default choice of the `inla` function; (*ii*) $\Gamma(0.001, 0.001)$, the default choice of the BUGS software (Lunn et al., 2000); (*iii*) $\Gamma(0.5, 0.0164)$, recently suggested by Fong et al. (2010). The last specification yields random effects u_{0j} with a marginal Cauchy distribution such that $e^{u_{0j}} \in [0.1, 10]$ with probability 0.95.

We focus on three measures of performance: the relative bias for the fixed effects and for the variance component, the coverage of confidence intervals for the fixed effects and the computational time.,

The estimates of the fixed effects are accurate regardless of the sample size, whereas the estimates of the variance component are quite sensitive to the number of clusters. An extensive comparison of different prior specifications is extremely time consuming with MCMC, but it is relatively fast with INLA. Figure 1 shows, for the three considered prior specifications, the plots of the relative bias on the level 2 variance in the case $\sigma^2_{u0} = 0.25$ (for the other values of σ^2_{u0} the findings are analogous).

The INLA method always converged quickly. When the number of clusters is large all the considered prior specifications yield satisfactory results. However, in the case of 10 clusters, only the prior $\Gamma(0.5, 0.0164)$ proposed by Fong et al. (2010) gives an acceptable bias on the variance component,

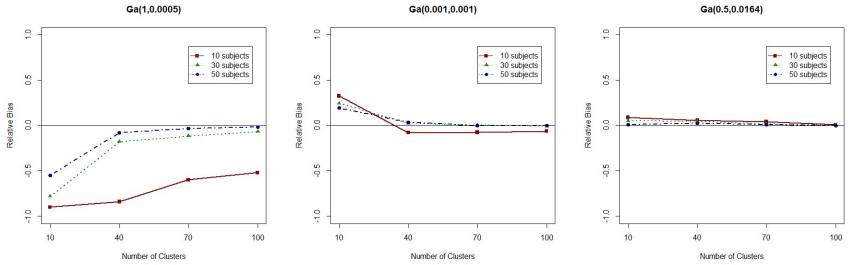


FIGURE 1. Relative bias for the level 2 variance $\sigma^2_{u0} = 0.25$ for the three considered priors. INLA estimates.

whereas the prior $\Gamma(1, 0.0005)$ (`inla` function default) yields a severe downward bias and the $\Gamma(0.001, 0.001)$ (BUGS default) yields a severe upward bias. The prior $\Gamma(0.5, 0.0164)$ also gives better coverage of the intervals for the fixed effects.

For the simulations with MCMC, we considered the priors with the best and worst performances (Γ(0.5,0.0164) and Γ(1,0.0005), respectively). We also tested maximum likelihood with AGQ. Table 1 reports the relative bias on the level 2 variance with $\sigma^2_{u0} = 0.25$ for INLA, MCMC and AGQ, where the first two methods are based on the prior distribution $\Gamma(0.5, 0.0164)$.

TABLE 1. Relative bias for level 2 variance $\sigma^2_{u0} = 0.25$ by number of clusters and cluster size. $\Gamma(0.5, 0.0164)$ prior for the precision $1/\sigma^2_{u0}$

J	n_j	INLA	MCMC	AGQ
10	10	0.102	0.208	-0.444
10	30	0.056	0.088	-0.432
10	50	0.012	0.064	-0.256
40	10	0.060	0.058	-0.164
40	30	0.047	0.034	-0.124
40	50	0.038	0.028	-0.092
70	10	0.034	0.024	-0.104
70	30	0.020	0.022	-0.068
70	50	0.011	0.022	-0.064
100	10	0.011	-0.020	-0.088
100	30	0.004	-0.016	-0.036
100	50	0.000	0.004	-0.028

For both bias and coverage, the patterns of the three methods are similar. Overall, INLA is more accurate than MCMC, which is more accurate than AGQ, even if the differences vanish as the number of clusters increases.

The simulations for MCMC confirm the key role of the prior specification

of the level 2 variance, so that a careful choice is far more important than the estimation method. It is worth to note that, for Bayesian methods, the sign of the bias on the level 2 variance depends on the prior specification, whereas AGQ is certain to yield a downward bias consistently with the findings of other simulation studies (Moineddin et al., 2007).
As for the computational load, the advantage of INLA over MCMC is considerable and it rapidly increases with the sample size, for example the computational time is 1 over 50 in the design with 100 clusters of size 50. Moreover, in the same design, the computational times of INLA and AGQ are nearly identical, thus closing the traditional gap in computational times between Bayesian and maximum likelihood methods.

3 Conclusions

Our simulation study has shown that the INLA method is fast and accurate for fitting random intercept logit models, even if, with few clusters, the estimates of the level 2 variance are not robust to the choice of the prior distribution for the precision. In the Bayesian context, the choice of the prior distribution is far more relevant than the choice of the estimation method. A noteworthy result is that the prior proposed by Fong et al. (2010) yields satisfactory results. Finally, INLA is considerably faster than MCMC and it has computational times similar to AGQ.

Acknowledgments: This research has been supported by the Italian government FIRB 2012 project n. RBFR12SHVV_003: *Mixture and latent variable models for causal inference and analysis of socio-economic data.*

References

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319 – 392.

Fong Y., Rue H. and Wakefield, J. (2010). Bayesian inference for Generalized Linear Mixed Models. *Biostatistics*, **11**, 397 – 412.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325 – 337.

Moineddin, R., M ,Matheson, F. and Glazier, R. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, **7**:34.

Modeling Credit Spreads Using Nonlinear Regression

Radoslava Mirkov[1], Thomas Maul[1], Ronald Hochreiter[2], Holger Thomae[1]

[1] TriSolutions GmbH, Germany
[2] WU Vienna University of Economics and Business, Austria

E-mail for correspondence: radoslava.mirkov@trisolutions.de

Abstract: The term structure of credit spreads is studied with an aim to predict its future movements. A completely new approach to tackle this problem is presented, which utilizes nonlinear parametric models. The Brain-Cousens regression model with five parameters is chosen to describe the term structure of credit spreads. Further, we investigate the dependence of the parameter changes over time and the determinants of credit spreads.

Keywords: nonlinear regression; random starting values; credit spreads.

1 Introduction and Motivation

We study the historical development of the credit spread curves, and are interested in forcasting future movements of credit spreads. In economic sciences, credit spreads represent the premium paid for specific risks embedded in a bond. The risk factors include geopolitical and macroeconomic variables. For details, see e.g. Schlecker (2009). The existing methods used in the banking industry proved unsatisfactory in times of financial crisis, as the relationship between issuer and reference curves has changed.
We scrutinize the behavior of credit spreads from a completely different perspective. We model the credit spread curve not by the common layer-factor approach, but we approximate the curve by a nonlinear parametric function with several parameters. Then we concentrate on finding the dependencies between these parameters and timely available and observable indicators and market data. This is motivated by the fact that the complexity of the parametric curve can be reduced to a small number of parameters so that changing patterns of the curve structure can be understood in terms of changes in these parameters. Also, each nonlinear parametric curve may be summarized by its parameter estimates as a single low-dimensional multivariate observation, which then may be subject to a regression or a correlation analysis.

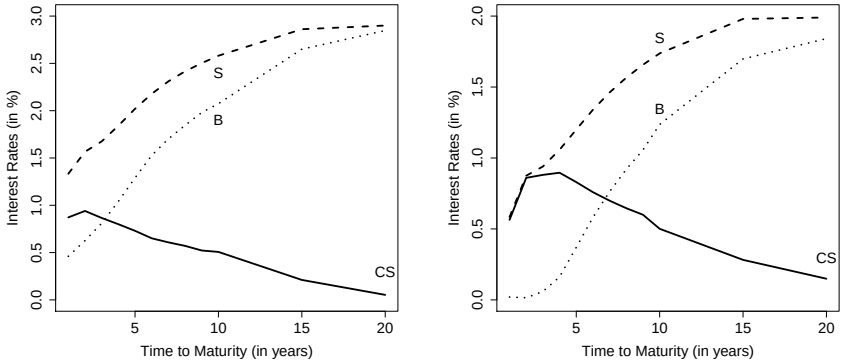


FIGURE 1. The term structure of ISDA fixings (S, dashed line), German government bonds (B, dotted line) and credit spreads (CS, solid line) on October 21, 2011 (left) and May 31, 2012 (right).

The following model which describes the structure of credit spreads y_i for given times to maturity x_j is studied:

$$y_i = y(x_j) + \varepsilon_i,$$

where x_j, $j = 1, \ldots, 12$, denotes the time to maturity in years of quoted credit spreads, usually $x = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20)$, and $\varepsilon_i \sim \mathcal{F}(0, \sigma^2)$ are error terms with zero mean and constant variance σ^2, for $i = 1, \ldots, n$.

2 Data Description

The analyzed data set is based on an excerpt from the Bloomberg data base, and contains daily quotation of ISDA fixings for Euro and German government bonds for all maturities for the period between June 2011 and June 2012, i.e. $n = 258$. The credit spreads y_i are obtained by subtracting the issuer curve from the reference curve, as Figure 1 shows graphically. Comparing credit spread curves in Figure 1 (left) and (right), it becomes evident that the structure of credit spread curves is changing, and we need a model that will be flexible enough to capture possible developments and different shapes of credit spread curves.

3 Nonlinear Regression Model

We fit a parametric nonlinear logistic regression model and analyze the term structure of credit spreads. The generalization of the log-logistic regression, the so-called Brain-Cousens model (BC-model) is proposed by

Ritz and Streibig (2008) for this kind of dependency. Some authors propose some variant of spline regression for similar problems, see e.g. Jarrow et al. (2004).
The BC-model is defined by

$$y(x_j) = c + \frac{d + f\,x_j - c}{1 + \exp(b\,[\ln(x_j) - \ln(e)])}, \tag{1}$$

and the parameters in model (1) have the following meaning: c and d define the upper and lower horizontal asymptotes, f is the slope of the upper asymptote, while b and e describe the shape of the decrease of the curve, i.e. e is the inflection point of the curve, and b is proportional to the slope at $x_j = e$. A similar approach for describing nonlinear dependancies is proposed in Friedl et al. (2012).
The starting values for the iteration necessary to calculate the least squares estimates of parameters are obtained either by using the parameter estimates of the previous day or by generating random starting values. If the parameter estimates from the previous day used as the starting values for the next day's estimation did not lead to convergence, random starting values from the interval $[-2, 2]$ are used. This approach is introduced in order to obtain convergence also in cases when the behavior of credit spreads changes dramatically from one day to another. It also enables the identification of days when market fluctuations influence the development of credit spreads strongly.
The results of the parameter estimation for both curves shown in Figure 1 are given in Table 1. We refer to Figure 2 for a graphical representation of the fitted models. We note that for the model fit shown in Figure 2 (left), nine random iterations of starting values were necessary to obtain convergence. This method yields convergence for 231 out of 258 days. Without random starting values, the convergence is obtained for only 180 days.

TABLE 1. MLEs (std. errors) of the BC-model (left) and (right).

b	c	d	e	f
4.9129	0.0772	0.9300	12.7469	-0.0209
(1.3833)	(0.0648)	(0.0307)	(1.0744)	(0.0094)
1.3550	-1.5464	-0.2754	2.4529	1.4679
(0.1951)	(1.1451)	(0.3344)	(0.8448)	(1.0474)

4 Conclusion

We study the term structure of credit spreads with an aim to predict their future movements. We suggest a completely new approach to tackle this

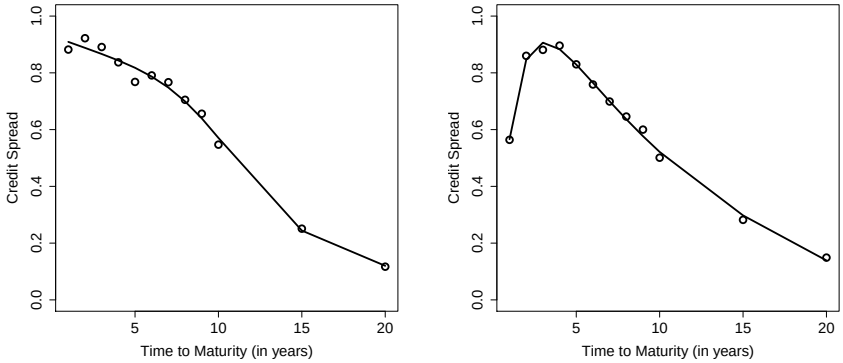


FIGURE 2. Fitted BC model given by (1) (solid line) and quoted credit spread values at time to maturity on October 21, 2011 (left) and May 31, 2012 (right).

problem, and instead of modeling credit spreads by the means of the layer-factor model, we utilize a nonlinear parametric model and concentrate on its parameters. The Brain-Cousens regression model with five parameters is chosen to describe the term structure of credit spreads. Random starting values are introduced in order to obtain convergence of parameter estimates also in cases when the behavior of credit spreads changes dramatically. Eventually, the dependence of the parameter values and given microeconomic factors over time is to be analyzed.

References

Jarrow, R., Ruppert, D., and Yu, Y. (2004). Estimating the Interest Rate Term Structure of Corporate Debt With a Semiparametric Penalized Spline Model. *Journal of the American Statistical Association*, **99(465)**, 57 – 66.

Friedl, H., Mirkov, R., and Steinkamp, A. (2012). Modeling and Forecasting Gas Flow on Exits of Gas Transmission Networks. *International Statistical Review*, bf 80(1), 24 – 39.

Ritz, C., and Streibig, J.C. (2008). *Nonlinear Regression with R.* New York: Springer.

Schlecker, M. (2009) *Credit Spreads Einflussfaktoren, Berechnung und langfristige Gleichgewichtsmodellierung.* Lohmar/Köln: Eul Verlag.

Estimation of the conditional survival function for successive survival times

Ana Moreira[1], Luís Machado[1]

[1] University of Minho, Department of Mathematics and Applications, Portugal

E-mail for correspondence: a.moreira.cris@gmail.com

Abstract: In longitudinal studies of disease, patients may experience several events through a follow-up period. In these studies, the sequentially ordered events (and the gap times) are often of interest and lead to problems that have received much attention recently. In recent years significant contributions have been made regarding these topics. Issues of interest include the estimation of bivariate survival, marginal distributions and the conditional distribution of the second gap time given the first gap time. In this work we consider the estimation for the survival of second gap time given the first gap time. Different approaches will be considered for estimating these quantities, all based on the Kaplan-Meier estimator of the survival function. Real data illustration based on a bladder study is included. In this dataset one major goal is to characterize the number of months that a patient have without second recurrence given he/she has spent a certain number of months without first recurrence.

Keywords: conditional survival; gap times; Kaplan-Meier.

1 Introduction

In many medical studies individuals can experience several events across a follow-up study. In these studies, the times between two consecutive events are often of interest and lead to problems that have received much attention. The events of concern may be of the same nature (e.g. cancer patients may experience recurrent disease episodes) or represent different states in the disease process (e.g. alive and disease-free, alive with recurrence and dead). If the events are of the same nature this are usually referred as recurrent event, whereas if they represent different states (i.e. multi-state models) they are usually modeled thought their intensity functions (Meira-Machado et al. 2009).

Let (T_1, T_2) be a pair of gap times of successive events, which are observed subjected to random right-censoring. Let C be the right-censoring variable, assumed to be independent of (T_1, T_2) and let $T = T_1 + T_2$ be the total time. Because of this, we only observe $(\tilde{T}_{1i}, \tilde{T}_{2i}, \Delta_{1i}, \Delta_{2i})$, $1 \leq i \leq n$, which are n independent replications of $(\tilde{T}_1, \tilde{T}_2, \Delta_1, \Delta_2)$, where

$\tilde{T}_1 = T_1 \wedge C$, $\Delta_1 = I(T_1 \leq C)$, and $\tilde{T}_2 = T_2 \wedge C_2$, $\Delta_2 = I(T_2 \leq C_2)$ with $C_2 = (C - T_1)I(T_1 \leq C)$ the censoring variable of the second gap time. Define $\tilde{T} = T \wedge C$ and let F_1 and G denote the distribution functions of T_1 and C, respectively. Since T_1 and C are independent, the Kaplan-Meier product-limit estimator based on the pairs $(\tilde{T}_{1i}, \Delta_{1i})$'s , consistently estimates the distribution F_1. Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on $(\tilde{T}_{1i} + \tilde{T}_{2i}, \Delta_{2i})$'s. Because T_2 and C_2 will be in general dependent, the estimation of the marginal distribution for the second gap time is not a simple issue. The same applies to the bivariate distribution function $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$. This issue have received much attention recently. In this work we present four estimators for the conditional survival function of second gap time given the first gap time. Different approaches will be considered, all based on the Kaplan-Meier estimator of the survival function. We have conducted extensive simulation studies to compare all four methods regarding its bias, mse and its variance. The performance of the estimators was also investigated for different sample sizes.

2 Estimators

In this section we will present four different approaches for estimating the conditional distribution function of the second gap time given the first gap time $S_{2|1}(x, y) = P(T_2 > y | T_1 > x)$, all using the Kaplan-Meier estimator of survival. A simple estimator for the conditional distribution function (CKM) is given by

$$\hat{S}_{2|1}(x, y) = P(T_2 > y | T_1 > x) = \hat{S}_{KM}(y | T_1 > x, \Delta_1 = 1)$$

where $\hat{S}_{KM}(y)$ the Kaplan-Meier estimator based on the pairs $(\tilde{T}_{2i}, \Delta_{2i})$'s. The $\hat{S}_{KM}(y | T_1 > x, \Delta_1 = 1)$ is the conditional survival function for the subset of $T_1 > x$ and $\Delta_1 = 1$ (the Kaplan-Meier estimator based on the pairs $(\tilde{T}_{2i}, \Delta_{2i})$'s such that $\tilde{T}_{1i} > x$ and $\Delta_{1i} = 1$). The idea behind the next estimator is to use the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data (KMW) and it is given by

$$\tilde{S}_{2|1}(x, y) = P(\tilde{T}_2 > y | \tilde{T}_1 > x) = \frac{P(\tilde{T}_1 > x, \tilde{T}_2 > y)}{P(\tilde{T}_1 > x)}$$

$$= \frac{P(\tilde{T}_1 > x) - P(\tilde{T}_1 > x, \tilde{T}_2 \leq y)}{1 - P(\tilde{T}_1 \leq x)}$$

The $P(\tilde{T}_1 > x, \tilde{T}_2 \leq y) = \frac{1}{n} \sum_{i=1}^{n} W_i I(\tilde{T}_{1i} > x, \tilde{T}_{2i} \leq y)$, where $W_i = \frac{\Delta_{2i}}{n - R_i + 1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{2j}}{n - R_j + 1}\right]$ is the Kaplan-Meier weight attached to $\tilde{T}_i$ when estimating the marginal distribution of T from $(\tilde{T}_i, \Delta_{2i})$'s, and for

which the ranks of the censored $\tilde{T}_i$'s, R_i, are higher than those for uncensored values in the case of ties. One estimator related to estimator above, in which they assume a presmoothed version of the Kaplan-Meier estimator (KMPW) (Dikta, 1998).

$$\tilde{\tilde{S}}_{2|1}(x,y) = \sum_{i=1}^{n} W_i^{\star} I(\tilde{T}_{1i} > x, \tilde{T}_{2i} \leq y)$$

where $W_i^{\star} = \frac{m(\tilde{T}_{1i},\tilde{T}_i)}{n-R_i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(\tilde{T}_{1j},\tilde{T}_j)}{n-R_j+1}\right]$ is the presmoothed Kaplan-Meier weight, where m stands for a parametric binary regression model.
Another estimator for the conditional survival function is obtained using Inverse Probability of Censoring Weighted (IPCW). Further details can be seen in the paper by Meira-Machado et al. (2011).

3 Simulation study

In this section, we compare by simulations the four estimators, for the conditional survival function. We consider one simulated scenario that described in Lin's paper (see their Section 3). In this scenario, the gap times were generated from Gumbel's bivariate distribution function, the so-called Fairlie-Gumbel-Morgenstern families of bivariate cdf's, $F(x,y) = F_1(x)F_2(y)[1 + \delta(1 - F_1(x)((1 - F_2(y))]$ where $|\delta| \leq 1$ for a bivariate density to exist. The marginal distributions, F_1 and F_2 are exponential with rate parameter 1. The case of independence is obtained for $\delta = 0$ while the maximum of correlation (between T_1 and T_2) for the bivariate exponential distribution is obtained for $\delta = 1$ with bound equal to 0.25 (Table 1).
For scenario we have considered two sample sizes, $n = 100$ and $n = 1000$ (Table 2) and for each simulation, 5000 samples were generated. For each setting we computed the mean and standard deviations for the conditional survival estimators at pairs of time points (x, y), where x and y takes values corresponding to: marginal survival probabilities of 0.8, 0.6, 0.4 and 0.2 for the bivariate exponential scenario.

TABLE 1. True values for conditional survival function with maximum correlation.

	0.2231	0.5108	0.9163	1.6094
0.2231	0.83203	0.64801	0.44799	0.23200
0.5108	0.86402	0.69601	0.49599	0.26401
0.9163	0.89602	0.74401	0.54400	0.29601
1.6094	0.92802	0.79201	0.59199	0.32801

TABLE 2. MSE of the estimated conditional survival function along 5000 trials for sample size $n = 1000$ with standard deviation.

	y	0.2231	0.5108	0.9163	1.6094
	x				
	0.2231	0.00033	0.00098	0.00296	0.01185
		(0.0177)	(0.0276)	(0.0373)	(0.0465)
	0.5108	0.00046	0.00155	0.00511	0.02103
		(0.0209)	(0.0339)	(0.0480)	(0.0609)
KMW	0.9163	0.00075	0.00300	0.01089	0.04666
		(0.0263)	(0.0456)	(0.0680)	(0.0892)
	1.6094	0.00181	0.00966	0.04029	0.16962
		(0.0401)	(0.0776)	(0.1213)	(0.1463)
	0.2231	0.00033	0.00065	0.00163	0.00475
		(0.0167)	(0.0254)	(0.0330)	(0.0390)
	0.5108	0.00044	0.00097	0.00284	0.00844
		(0.0196)	(0.0310)	(0.0423)	(0.0515)
KMPW	0.9163	0.00069	0.00171	0.00556	0.01714
		(0.0245)	(0.0407)	(0.0577)	(0.0717)
	1.6094	0.00169	0.00417	0.01292	0.06591
		(0.0383)	(0.0639)	(0.0901)	(0.1062)
	0.2231	0.00135	0.00316	0.00369	0.00245
		(0.0183)	(0.0243)	(0.0263)	(0.0274)
	0.5108	0.00118	0.00269	0.00331	0.00278
		(0.0215)	(0.0298)	(0.0347)	(0.0400)
CKM	0.9163	0.00118	0.00273	0.00404	0.00701
		(0.0274)	(0.0403)	(0.0513)	(0.0791)
	1.6094	0.00242	0.00712	0.02108	0.05922
		(0.0470)	(0.0811)	(0.1422)	(0.2098)
	0.2231	0.00033	0.00098	0.00296	0.01185
		(0.0177)	(0.0276)	(0.0373)	(0.0465)
	0.5108	0.00046	0.00155	0.00511	0.02103
		(0.0209)	(0.0339)	(0.0480)	(0.0609)
IPCW	0.9163	0.00075	0.00300	0.01089	0.04666
		(0.0263)	(0.0456)	(0.0680)	(0.0892)
	1.6094	0.00181	0.00966	0.04029	0.16962
		(0.0401)	(0.0776)	(0.1213)	(0.1463)

4 Conclusion

In this paper we present several nonparametric estimators of the conditional survival function for the second gap time given the first gap time. Simulations showed that the estimator KMPW present better results.

Acknowledgments: The authors acknowledge receiving financial support from the Portuguese Ministry of Science, Technology and Higher Education in the form of grants , PTDC/MAT/104879/2008 and SFRH/BD/62284/2009. This research was financed by FEDER Funds through Programa Operacional Factores de Competitividade COMPETE and by Portuguese Funds through FCT - Fundaçãoo para a Ciência e a Tecnologia, within the Project Est-C/MAT/UI0013/2011 and CMAT.

References

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *The Annals of Statistics*, **16**, 1475 – 1489.

Dikta, G. (1998). On semiparametric random censorship models. *Journal of Statistical Planning and Inference*, **66**, 253 – 279.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457 – 481.

Lin, D. Y., Sun, W. and Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika*, **86**, 59 – 70.

Meira-Machado, L., Uña-Álvarez, J., Cadarso-Suárez, C. and Andersen,P.K. (2009): Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research*, **18**, 195 – 222.

Meira-Machado, L., Uña-Álvarez, J., Datta, S. (2011). Conditional Transition Probabilities in a non-Markov Illness-death Model. *Discussion Papers in Statistics and Operation Research*, **n 11/03**, ISSN: 1888-5756, Depsito Legal VG 1402 - 2007.

Moreira, A. and Machado, L. (2012). survivalBIV: Estimation of the Bivariate Distribution Function for Sequentially Ordered Events Under Univariate Censoring. *Journal of Statistical Software*, **46**, 1 – 16.

Serrat, C. and Gómez, G. (2007). Nonparametric bivariate estimation for successive survival times. *SORT*, **31**, 75 – 96.

Wang, W. and Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika*, **85**, 561 – 572

A Semiparametric Approach for Multivariate Longitudinal Count Data

Darcy Steeg Morris[1]

[1] U.S. Census Bureau, Washington, DC U.S.A.

E-mail for correspondence: darcy.steeg.morris@census.gov

Abstract: This paper presents a semiparametric method for estimating the marginal response and association parameters in a random effects multivariate longitudinal count model. In the context of the generalized estimating equations (GEE) framework, we use a specific form of the covariance matrix of the response vector based on a model that induces dependence over time and outcomes using random effects. This moment based method is robust to distributional misspecification and reduces the computational burden associated with a high-dimensional joint distribution by avoiding parametric assumptions on the response and unobserved effects. Through a simulation study, we compare finite sample robustness properties of this semiparametric method with a pseudo-likelihood approach that imposes distributional assumptions. Both of these methods are then used to analyze a dataset of insurance claim counts for three types of coverage over time.

Keywords: Generalized Linear Mixed Models; Correlated Count Data; Longitudinal Data; Generalized Estimating Equations; Unobserved Heterogeneity.

1 Introduction

Correlated count data commonly arise in fields such as business, economics and demography through longitudinal studies. Methods for accounting for correlation in a single longitudinal count outcome are well-established and straightforward to implement. However, often times the researcher is presented with multiple longitudinal outcomes with an underlying relationship that should not be ignored. For example, we are interested in how unobserved individual-specific risk characteristics are related across three types of personal insurance coverages: home, auto collision and auto comprehensive. Separate generalized linear mixed models for each of the claim counts can be fit, but a joint model for the multiple claim counts properly addresses our research question. This research focuses on generalized linear mixed models (GLMMs) with correlated random effects that induce

marginal association between the multiple claim rates through the joint dependence on the random effects. In such a model, the covariance structure of the random effects tells us about the relationship between the unobserved heterogeneity in the multiple claim counts.

Maximum likelihood estimation of the multivariate longitudinal GLMM requires distributional specification of the unobserved heterogeneity and is also computationally prohibitive. Specifically, assuming the count outcomes y_{itk} are conditionally Poisson distributed with mean λ_{itk}, it involves maximizing the following marginal likelihood:

$$\prod_{i=1}^{N} \int_{u_{iK}} ... \int_{u_{i1}} \left\{ \prod_{k=1}^{K} \prod_{t=1}^{T_i} e^{-u_{ik}\lambda_{itk}} \frac{(u_{ik}\lambda_{itk})^{y_{itk}}}{y_{itk}!} \right\} g(u_{i1}, ..., u_{iK}) du_{i1} ... du_{iK}$$

where $g(u_{i1}, ..., u_{iK})$ is the multivariate density of the random effects, K is the number of dependent count variables, T_i is the time dimension for subject i, and N is the number of subjects. Pairwise likelihood is one approach to reduce the computational complexity of evaluating and maximizing the above integral (Fieuws and Verbeke 2006). This method reduces the full likelihood to a composite likelihood that involves fitting all pairwise GLMMs, but this can still be computationally prohibitive as the dimension of the random effect vector increases even in seemingly simple cases and is not robust to misspecification. We present a semiparametric approach that uses the moments implied by the GLMM with correlated random effects in the framework of generalized estimating equations (Liang and Zeger 1986, Prentice 1988, Gourieroux et. al. 1984) to estimate the joint model. In simulation, we show that the robustness property of the semiparametric approach is maintained in finite samples. In simulation and empirical analysis, we find that the semiparametric method runs about 25 times faster than the pairwise likelihood method.

2 Methodology

2.1 Assumptions

The following formalizes the minimal assumptions associated with this class of multivariate longitudinal count models with multiplicative correlated random effects. These are the assumptions maintained in this research.

Conditional Moments and Model Assumptions

$$(i)\ E(y_{itk}|\mathbf{x}_{ik}, u_{ik}) = u_{ik} exp(\mathbf{x}_{itk}^T \beta_k) = u_{ik}\lambda_{itk}$$
$$(ii)\ E(y_{itk}|\mathbf{x}_{ik}, u_{ik}) = Var(y_{itk}|\mathbf{x}_{ik}, u_{ik})$$
$$(iii)\ E(u_{ik}|\mathbf{x}_{ik}) = E(u_{ik}) = 1 \text{ and } Var(u_{i1}, \ldots, u_{iK}|\mathbf{x}_{ik}) = \Sigma$$
$$(iv)\ y_{itk} \perp y_{isl}|(u_{ik}, u_{il})$$

By the law of iterated expectations, the first two marginal moments for this class of multivariate longitudinal count models can be derived.

Marginal Moments

$$(i)\ E(y_{itk}|\mathbf{x}_{itk}) = exp\left(\mathbf{x}_{itk}^T\beta_k\right) = \lambda_{itk}$$
$$(ii)\ V_i \equiv Var(\mathbf{y}_i|\mathbf{x}_i) = diag(\lambda_i^T) + \Sigma \otimes \mathbf{1}_{T_i}\mathbf{1}_{T_i}^T \circ \lambda_i\lambda_i^T$$

where $\circ$ is element-wise multiplication, $\otimes$ is the Kronecker product, and $\mathbf{1}_{T_i}$ is a T_i-dimensional vector of ones.

2.2 Semiparametric Estimation

The procedure for the semiparametric approach relies on the moment conditions implied by the marginal mean and variance and involves iterating between moment-based estimation of the covariance parameters of the random effect, Σ, and moment-based estimation of the regression parameters, β. This approach is an extension of quasi-generalized pseudo maximum likelihood (QGPML) estimators developed by Gourieroux et. al. (1984) and the extended GEE approach developed by Prentice (1988). The estimator $(\hat{\beta}, \hat{\Sigma})$ for β and Σ is defined as the solution to:

$$U(\beta,\Sigma) = \sum_{i=1}^{N} \begin{pmatrix} D_i^T & 0 \\ 0 & E_i^T \end{pmatrix} \begin{pmatrix} V_i & 0 \\ 0 & W_i \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y}_i - \lambda_i \\ \mathbf{R}_i^* - \mathbf{V}_i^* \end{pmatrix} = 0$$

where $D_i = \frac{\partial \lambda_i^T}{\partial \beta}$, V_i is the model based variance matrix, $E_i = \frac{\partial \mathbf{V}_i^{*T}}{\partial \mathbf{\Sigma}^*}$, $W_i = I_{T_iK}$ and R_i is the cross product of residuals. Define $\mathbf{R}_i^*$, $\mathbf{V}_i^*$ and $\mathbf{\Sigma}^*$ to be the vector of unique elements of R_i, V_i and Σ, respectively.

The semiparametric estimator involves a two-step iterative procedure. After finding initial estimates for β via a procedure such as nonlinear least squares (i.e. ignoring any dependence), $Var(\mathbf{y}_i)$ can be consistently estimated by the cross product of the residuals, R_i. Define the vector of residuals $\hat{\mathbf{r}}_{ik} = \begin{bmatrix} \hat{r}_{i1k} & \dots & \hat{r}_{iT_ik} \end{bmatrix}^T$ with $\hat{r}_{itk} = y_{itk} - e^{\mathbf{x}_{itk}^T\hat{\beta}_k}$, so that:

$$R_i = \widehat{Var}(\mathbf{y}_i) \quad = \quad \begin{bmatrix} \hat{\mathbf{r}}_{i1} \\ \vdots \\ \hat{\mathbf{r}}_{iK} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{r}}_{i1}^T & \dots & \hat{\mathbf{r}}_{iK}^T \end{bmatrix}$$

Next the empirical variance estimate R_i and the model defined variance structure V_i are used to estimate Σ. Specifically, the relation between R_i and V_i implies $\frac{KT_i(KT_i+1)}{2}$ estimating equations for Σ from the distinct elements of the two matrices. These estimating equations define the estimator $\hat{\mathbf{\Sigma}}^*$ for $\mathbf{\Sigma}^*$:

$$U(\mathbf{\Sigma}^*) = \sum_{i=1}^{N} E_i^T(\hat{\beta}) W_i^{-1} \left(\mathbf{R}_i^*(\hat{\beta}) - \mathbf{V}_i^*(\hat{\beta}, \mathbf{\Sigma}^*)\right) = 0$$

An estimate for $\mathbf{V}_\mathbf{i}^*$ can now be obtained by plugging $\hat{\mathbf{\Sigma}}^*$ into the model defined variance structure. The estimator $\hat{\beta}$ for β is then found as the solution to:

$$U(\beta) = \sum_{i=1}^{N} D_i^T(\beta) \hat{V}_i^{-1}(\hat{\beta}, \hat{\mathbf{\Sigma}}^*) \left(\mathbf{y}_i - \lambda_i(\beta)\right) = 0$$

The roots of the estimating equations $U(\beta, \Sigma)$ are solved for via an iterative procedure, updated at each iteration by the previous value of the $\sqrt{N}$-consistent estimator of β given Σ and the $\sqrt{N}$-consistent estimator of Σ given β, until convergence. Consistency results follow from the work of Liang and Zeger (1986), Gourieroux et. al. (1984) and properties of two-step M-estimators. Standard errors are based on the joint asymptotic distribution of $\sqrt{N}(\beta - \hat{\beta})$ and $\sqrt{N}(\mathbf{\Sigma}^* - \hat{\mathbf{\Sigma}}^*)$ specific to the GEE framework developed by Prentice (1988). Please see Morris (2012) for details.

3 Monte Carlo Simulation Studies

We conduct Monte Carlo simulation studies to assess the finite sample performance of the semiparametric and the pairwise likelihood approach. The following data generating process is considered:

$$y_{itk}|u_{ik}, \beta_k \sim Poisson(u_{ik}\lambda_{itk}) \text{ where } \lambda_{itk} = e^{\beta_{k0} + x_{itk}\beta_{k1}}$$
$$\mathbf{u}_i = (u_{i1}, u_{i2}, u_{i3}) \sim g(1, \Sigma)$$

Each simulation study varies the distributional assumptions for the true random effect distribution g: (1) the base case where g is multivariate lognormal, (2) the skewed case where g is a multivariate gamma distribution constructed via a Gaussian copula, and (3) the bimodal case where g is a mixture of multivariate lognormal distributions. Table 1 presents and Figure 1 graphically depicts results of the simulation studies. Please see Morris (2012) for a complete presentation and discussion of the results.

TABLE 1. Simulation Study Results: Semiparametric and Pairwise Likelihood

		MV Log Normal				MV Gamma				MV Log Normal Mixture			
		Semipar.		Pairwise		Semipar.		Pairwise		Semipar.		Pairwise	
	True	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
σ_{11}	.47	.001	.047	-.003	.023	-.001	.029	.129	.132	-.003	.087	.313	.317
σ_{22}	.16	.000	.010	-.001	.007	.000	.009	.010	.013	.000	.007	.046	.047
σ_{33}	.56	.001	.058	-.001	.033	.000	.035	.131	.136	.007	.169	.077	.088
σ_{12}	.14	.000	.013	-.001	.008	-.005	.012	.013	.016	.000	.009	.104	.105
σ_{13}	.22	.001	.028	-.001	.016	-.012	.024	.032	.037	.001	.029	.197	.199
σ_{23}	.11	.000	.012	.000	.009	-.005	.012	.006	.012	.000	.011	.063	.064

The pairwise likelihood approach results in a potentially seriously biased estimator that, while more precise than the semiparametric estimator in the correctly specified case, can lead to incorrect conclusions about Σ. Our simulation studies indicate that under the chosen simulated settings, which are empirically motivated by the insurance data, the semiparametric approach offers an advantage in computing time (a 25-fold improvement - .5 vs. 12 minutes) and robustness.

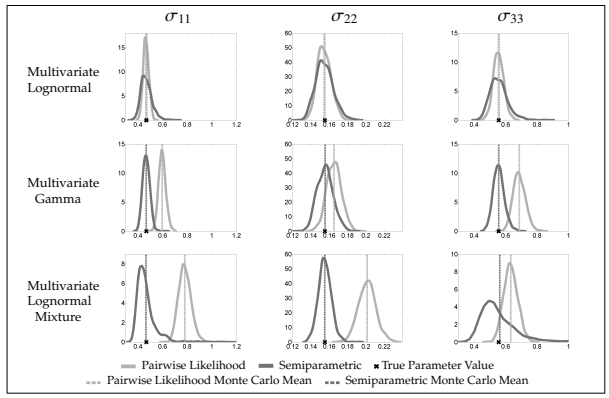


FIGURE 1. Kernel Density of Variance Parameter Estimates from Simulation

4 An Empirical Application: Insurance Data

This research is motivated by interest in how unobserved heterogeneity is associated between different types of insurance coverages. The unobserved heterogeneity in an insurance claim rate model represents the intrinsic riskiness of the policyholder that affects the claim propensity, after accounting for observable characteristics of the policyholder. The joint model of insurance claim counts is estimated using data containing nine years of policy and policyholder characteristics. Please see Barseghyan et. al. (2012) for details of this data, a full version of the claim count model, and a thorough discussion of the empirical results.

TABLE 2. Association Parameter Estimates & Standard Errors: Insurance Data

	Balanced Panel 8,731 Policies, 78,579 Obs.						Unbalanced Panel 62,425 Policies, 294,917 Obs.			
	Semipar.		Pairwise		Uni. GLMM		Semipar.		Uni. GLMM	
σ_{11}	.467	(.072)	.396	(.035)	.395	(.036)	.320	(.077)	.531	(.025)
σ_{22}	.155	(.036)	.181	(.021)	.181	(.022)	.123	(.023)	.204	(.013)
σ_{33}	.551	(.182)	.540	(.077)	.537	(.080)	.571	(.109)	.745	(.055)
σ_{12}	.136	(.021)	.137	(.019)	.	(.)	.078	(.014)	.	(.)
σ_{13}	.224	(.038)	.223	(.035)	.	(.)	.227	(.024)	.	(.)
σ_{23}	.104	(.029)	.120	(.027)	.	(.)	.122	(.019)	.	(.)

Generally, we find comparable estimates in the balanced panel analysis, albeit with larger standard errors for the variance estimates in the semiparametric approach consistent with the simulation study results. While the pairwise likelihood is computationally prohibitive in the more complete unbalanced panel analysis, the univariate GLMM and semiparametric methods result in significant differences in the estimates of the vari-

ance parameters. This result, in conjunction with simulation study results, suggests that the log normality assumption may lead to an overestimation of the association parameters. The semiparametric method is robust to this potential bias. Just as in the simulation study, the semiparametric approach offers a computational advantage over the pairwise likelihood approach (balanced: 2.5 hours vs. 60 hours, unbalanced: 22 hours vs. computationally prohibitive).

5 Conclusion

We propose a semiparametric method for multivariate longitudinal data based on a correlated random effects model and GEE. Joint modeling allows for estimation and inference on association parameters for unobserved heterogeneity, which is of economic interest in the insurance data as it provides insight into underlying risk related characteristics of the policyholders. Compared to the pairwise likelihood method, the semiparametric method shows a significant advantage in (1) computing time and (2) finite sample properties when the random effect distribution is misspecified.

Acknowledgments: This work was completed during my doctoral studies at Cornell University and benefited from guidance from Francesca Molinari, Levon Barseghyan, Jim Booth, Rob Strawderman and Josh Teitelbaum.

References

Barseghyan, L., Molinari, F., Morris, D.S., and Teitelbaum, J. (2012). Unlucky or Risky? Unobserved Heterogeneity and Experience Rating in Insurance Markets. At SSRN: http://ssrn.com/abstract=2176295.

Fieuws, S., and Verbeke, G. (2006). Pairwise Fitting of Mixed Models for Joint Modeling of Multivariate Longitudinal Profiles. *Biometrics*, **62**, 424 – 431.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo Maximum Likelihood Methods: Applications to Poisson Models. *Econometrica*, **52**, 701 – 720.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal Data Analysis using Generalized Linear Models. *Biometrika*, **35**, 13 – 22.

Morris, D.S. (2012). *Methods for Multivariate Longitudinal Count and Duration Models with Applications in Economics*. Ph.D. Dissertation, Cornell University.

Prentice, R. (1988). Correlated Binary Regression with Covariates Specific to each Binary Observation. *Biometrics*, **44**, 1033 – 1048.

A Quaternion Widely Linear Series Expansion

Jesús Navarro-Moreno[1], Rosa M. Fernández-Alcalá[1], Juan Carlos Ruiz-Molina[1], Antonia Oya[1]

[1] Department of Statistics and Operations Research, University of Jaén, 23071 Jaén, Spain

E-mail for correspondence: jcruiz@ujaen.es

Abstract: A series representation for continuous-time quaternion random signals is given. The series expansion is based on augmented statistics and provides uncorrelated scalar real-valued random variables. The proposed technique implies a dimension reduction of the four-dimensional original problem to a one-dimensional problem. As a particular case, the quaternion Karhunen-Loève expansion is obtained. Finally, an illustrative application to the quaternion widely linear detection problem is presented.

Keywords: Quaternion Random Signal; Series Expansion; Widely Linear Detection.

1 Introduction

Series expansions for stochastic processes comprise an essential tool for providing a solution to different types of problems in statistical signal processing. Its ability to split the time and random components of the signal have enabled estimation, detection and simulation problems, among others, to be solved. Specially interesting are those series representations that provide uncorrelated variables, the Karhunen-Loève (KL) expansion being the most widely used because of its optimal properties in information compression.

On the other hand, the mathematical theory of quaternions has recently aroused a great interest due to its multiple applications in signal processing such as in aerospace, computer graphics, image processing, vector-sensor signals, etc. Quaternion domain has become a tool of increasing importance in statistical multichannel processing since it accounts naturally for the correlated nature of the signal components. The suitable statistical processing for quaternion requires the augmented statistics to be considered, i.e., requires the operation on the quaternion and its involutions over the three pure unit quaternions in an orthogonal basis. This approach, called quater-

nion widely linear (QWL) processing, has been shown to outperform the traditional quaternion linear processing (Cheong Took and Mandic (2011)). In this paper we give a series expansion for the quaternion in continuous-time by using augmented statistics. As a particular case, the KL expansion for quaternionic signals is provided, thus extending the orthogonality property to the time component. Such representations are obtained from the definition of a real-valued univariate stochastic signal whose second-order statistics match that of quaternion. This strategy avoids addressing a four-dimensional vectorial problem which notably simplifies the obtaining of series expansions. Finally, two potential applications in quaternion signal detection and estimation problems are showed.

2 The WL Series Representations

We use boldfaced uppercase letters to denote matrices, boldfaced lowercase letters for column vector, and lightfaced lowercase letters for scalar quantities. Superscripts $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ represent quaternion (or complex) conjugate, transpose, and Hermitian (i.e., transpose and quaternion conjugate), respectively. E is the expectation operator and r_x is the correlation function of the signal x. The Hilbert space spanned by the variables of the stochastic signal x is denoted by $\mathcal{H}(x)$.

Let $\{q(t) = q_1(t) + q_2(t)\mathrm{i} + q_3(t)\mathrm{j} + q_4(t)\mathrm{k}, t \in [0,T]\}$ be a quaternion random signal with $q_n(t)$, $n = 1,\ldots,4$, zero-mean mean-square continuous stochastic signals. Define the augmented quaternion vector $\boldsymbol{q}(t) = [q(t), q^{\mathrm{i}}(t), q^{\mathrm{j}}(t), q^{\mathrm{k}}(t)]^{\mathrm{T}}$ where $q^{\eta}(t) = -\eta q(t)\eta$, $\eta = \mathrm{i},\mathrm{j},\mathrm{k}$, are the three perpendicular quaternion involutions. The complete description of the second-order statistics of $q(t)$ in the quaternion domain is given by the augmented quaternion vector $\boldsymbol{q}(t)$ or by its (augmented) correlation function, $\boldsymbol{R}_{\boldsymbol{q}}(t,s) = E[\boldsymbol{q}(t)\boldsymbol{q}^{\mathrm{H}}(s)]$.

We aim to find a series representation which accounts for the complete second-order description of $q(t)$. For that, we define the following real-valued stochastic signal

$$x(t) = \begin{cases} q_1(t), & t \in [0,T] = I_1; \\ q_2(-t), & t \in [-T,0) = I_2; \\ q_3(-t-T), & t \in [-2T,-T) = I_3; \\ q_4(-t-2T), & t \in [-3T,-2T) = I_4. \end{cases} \tag{1}$$

Let $\{f_n\}_n$ be a basis of functions in $L_2[-3T,T]$ and consider the Hilbert space, $\mathcal{H}(\zeta_n)$, spanned by the variables $\zeta_n = \int_{-3T}^{T} x(t) f_n(t)dt$, $n = 1,2,\ldots$ Thus, $\mathcal{H}(x)$ is separable, i.e., it follows that $\mathcal{H}(x) = \mathcal{H}(\zeta_n)$ and we have that $x(t)$ admits the representation

$$x(t) = \sum_{n=1}^{\infty} l_n(t)\xi_n$$

where the series converges in quadratic mean (q.m.) uniformly in $t \in [-3T, T]$, with ξ_n orthonormal random variables defined by

$$\xi_n = \int_{-3T}^{T} x(t) g_n(t) dt$$

and $g_n(t) = \sum_{m=1}^{n} c_{nm} f_m(t)$ where the coefficients c_{nm} are obtained from applying the Gram-Schmidt method to the random variables $\{\zeta_n\}_{n=1}^{\infty}$. Also we have

$$\int_{-3T}^{T} r_x(t,s) g_n(s) ds = l_n(t), \quad t \in [-3T, T] \tag{2}$$

Now we have conditions to find a series representation for $\boldsymbol{q}(t)$ depending on uncorrelated variables, and hence for $\boldsymbol{R}_{\boldsymbol{q}}(t,s)$. The augmented quaternion vector has the series expansion

$$\boldsymbol{q}(t) = \sum_{n=1}^{\infty} \boldsymbol{\phi}_n(t) \xi_n \tag{3}$$

in q.m. uniformly in $t \in [0, T]$, where $\boldsymbol{\phi}_n(t) = [\phi_n(t), \phi_n^{\mathrm{i}}(t), \phi_n^{\mathrm{j}}(t), \phi_n^{\mathrm{k}}(t)]^{\mathrm{T}}$ with

$$\phi_n(t) = l_n(t) + l_n(-t)\mathrm{i} + l_n(-t-T)\mathrm{j} + l_n(-t-2T)\mathrm{k},$$

and the random variables ξ_n are real-valued and orthonormal of the form

$$\xi_n = \int_{0}^{T} \boldsymbol{\psi}_n^{\mathtt{H}}(t) \boldsymbol{q}(t) dt \tag{4}$$

with $\boldsymbol{\psi}_n(t) = [\psi_n(t), \psi_n^{\mathrm{i}}(t), \psi_n^{\mathrm{j}}(t), \psi_n^{\mathrm{k}}(t)]^{\mathrm{T}}$ being

$$\psi_n(t) = \frac{1}{4}(g_n(t) + g_n(-t)\mathrm{i} + g_n(-t-T)\mathrm{j} + g_n(-t-2T)\mathrm{k})$$

and where the set $\{\boldsymbol{\psi}_n\}_n$ is biorthogonal to the set $\{\boldsymbol{\phi}_n\}_n$. Also $\boldsymbol{R}_{\boldsymbol{q}}(t,s)$ admits the representation

$$\boldsymbol{R}_{\boldsymbol{q}}(t,s) = \sum_{n=1}^{\infty} \boldsymbol{\phi}_n(t) \boldsymbol{\phi}_n^{\mathtt{H}}(s) \tag{5}$$

uniformly in $t, s \in [0, T] \times [0, T]$.
The proof of (3) and (5) can be found in Navarro-Moreno et al. (2012).
As a particular case, we have the following extension of the KL expansion to the quaternion domain

$$\boldsymbol{q}(t) = \sum_{n=1}^{\infty} \boldsymbol{\varphi}_n(t) \varepsilon_n \tag{6}$$

in q.m. uniformly in $t \in [0,T]$, where $\boldsymbol{\varphi}_n(t) = [\varphi_n(t), \varphi_n^{\mathrm{i}}(t), \varphi_n^{\mathrm{j}}(t), \varphi_n^{\mathrm{k}}(t)]^{\mathrm{T}}$ with

$$\varphi_n(t) = \frac{1}{2}(a_n(t) + a_n(-t)\mathrm{i} + a_n(-t-T)\mathrm{j} + a_n(-t-2T)\mathrm{k})$$

and λ_n and $a_n(t)$ are the eigenvalues and eigenfunctions, respectively, of the kernel $r_x(t,s)$ on $[-3T,T]\times[-3T,T]$. Likewise, $\varepsilon_n = \int_0^T \boldsymbol{\varphi}_n^{\mathrm{H}}(t)\boldsymbol{q}(t)dt$ are real-valued random variables such that $E[\varepsilon_n\varepsilon_m] = \beta_n\delta_{nm}$, with $\beta_n = 4\lambda_n$. Hence, $\{\boldsymbol{\varphi}_n\}_n$ constitute a particular choice of both $\{\boldsymbol{\psi}_n\}_n$ and $\{\boldsymbol{\phi}_n\}_n$.

3 Application: QWL Detection Problem

In order to illustrate a potential application, we consider the detection of quaternion deterministic signals in additive quaternion Gaussian noise. This problem can be modeled by the following hypothesis pair

$$\begin{aligned} H_0 &: y(t) = v(t) \qquad 0 \le t \le T \\ H_1 &: y(t) = q(t) + v(t) \qquad 0 \le t \le T \end{aligned} \tag{7}$$

where $\{q(t), t \in [0,T]\}$ is a continuous completely known quaternion signal and the noise $\{v(t), t \in [0,T]\}$ is a quaternion mean-square continuous Gaussian noise. We also assume that the augmented quaternion noise $\boldsymbol{v}(t)$ has the representation (3) and that $\boldsymbol{q}(t)$ belongs to the space spanned by the set $\{\boldsymbol{\phi}_n\}_n$. To study the problem of (7) we can consider the equivalent hypothesis pair

$$\begin{aligned} H_0 &: y_n = v_n \qquad n = 1,2,\ldots \\ H_1 &: y_n = \xi_n + v_n \qquad n = 1,2,\ldots \end{aligned}$$

where $v_n = \int_0^T \boldsymbol{\psi}_n^{\mathrm{H}}(t)\boldsymbol{v}(t)dt$, $n = 1,2,\ldots$ Thus, assuming that the detection problem (7) is not singular and according to Grenander's theorem, the log-likelihood ratio test is

$$L_{QWL}(y) = \sum_{n=1}^{\infty} \xi_n y_n - \frac{1}{2}\sum_{n=1}^{\infty} \xi_n^2$$

References

Cheong Took, C. and Mandic, D.P. (2011). Augmented second-order statistics of quaternion random signals. *Signal Processing*, **91**, 214 – 224.

Navarro-Moreno, J., Fernandez-Alcala, R.M. and Ruiz-Molina, J.C. (2012). A quaternion widely linear series expansion and its applications. *IEEE Signal Processing Letters*, **19**, 868 – 871.

Smooth Graphical models of type II: link with marginal models

Federica Nicolussi[1]

[1] Università degli studi Milano Bicocca, Italy

E-mail for correspondence: f.nicolussi@campus.unimib.it

Abstract: The graphical models (**GM**) for categorical data are models useful to represent conditional independences through graphs. In this work we propose a subclass of **GM**s proposed by Andersson, Madigan and Perlman (2001), having interesting properties for the asymptotic theory. As we will show these models can be parametrized with Marginal Models for categorical data, proposed by Bergsma and Rudas (2002). In order to show the main results on **GM***s* **II**, we analyse the data from the European Values Study (EVS), (2008). The work will be structured in two sections. In the first we will give basic concepts about the methodology, furthermore graphical models for chain graph, marginal models and the subclass of **GM***s* that will be used. In the second section we will introduce the different datasets and will be shown the applications on the different data, with the main aspects.

Keywords: Categorical data; chain graphs; EVS; hierarchical and complete marginal parametrization; log-linear parameters; Markov properties; smoothness.

1 Conditional Independence Models

1.1 Graphical Models

Graphical models for categorical variables represent the relationships among variables with the help of graphs: mathematical objects where the vertices act for the variables and the possible edges act for dependence relationships. In chain graphs can be both directed and undirected edges and there are no cycle or semi-cycles. So we can grouping the vertices in s components $T_1, T_2, ..., T_s$ in such way that within the components there are only undirected arcs and between the components only directed arcs all pointing in the same direction. When two vertices are connected by an undirected edge, we can guess that the two linked variables are "symmetrically" dependently. In this case the two vertices are called neighbours (nb). On the other hand, when there is a directed arc, we presume an unilateral dependence relationship among the variable linked by a direct arc. In this case we call parent (pa) the vertex from which the arrow starts and child (ch) the vertex where the arc ends. Anderson, Madigan and Perlman (2001) introduced

Graphical models of type II (**GM II**, proposed by Andersson, Madigan and Perlman) as generalization of both graphical models for undirected graphs (**UG**) and graphical models for directed acyclic graphs (**DAG**) (for details see Lauritzen). This choice is supported by different reasons. In the first, the grouping of variables in components allows to split the variables in "purely explicative" variables, "purely response" variables and "intervening" variables. Secondly, in the **GM*s* II**, the relationship among a variable and its explicative variables is considered marginally regarding the variables in the same component. Finally, the **GM*s* II** model the association between the variables within the same component using a log-linear approach. All these topics make the **GM*s* II** one of the easiest interpretable models. The rules to read a list of k conditional independences as $A_i \perp B_i | C_i$ by a graph are called Markov properties and, for the **GM II** are the three following:

$$\begin{array}{lll} \textbf{C1}) & T_h \perp \cup_{i<h} T_i \setminus pa_D(T_h) | pa_D(T_h) & \\ \textbf{C2a}) & A \perp T_h \setminus Nb(A) | pa_D(T_h) \cup nb(A) & \forall A \subseteq T_h \\ \textbf{C3b}) & A \perp pa_D(T_h) \setminus pa_G(A) | pa_G(A) & \forall A \subseteq T_h \end{array} \tag{1}$$

$\forall h = 1, ..., s$. Where T_h denotes the h-th component, $nb(A)$ denotes the union of all vertices that are neighbours of A end $Nb(A) = nb(A) \cup A$. Finally $pa_D(T_h)$ and $pa_G(T_h)$ are respectively the set of parents of the component T_h and the set of parents of the set A. Unlikely, as Drton (2009) showed, these three Markov properties do not always correspond to smooth models. The property of smoothness guarantees the existence of the Max Likelihood Estimation of the model. As the parametric marginal models for categorical data have useful properties for the asymptotic theory of the ML estimators, showed by Bergsma and Rudas (2002), we are interested to study which **GM** of type II can be parametrized as marginal models. At this purpose the next theorem defines a subclass of smooth **GM*s* II**.

1.2 Marginal Models

Bergsma and Rudas (2002), proposed a generalization of log-linear models where the log-linear parameters are defined even in marginal distributions. We consider q categorical variables $V = \{V_1, ..., V_q\}$, collected in a contingency table. At first it is necessary to define a hierarchical class of marginal sets $\mathcal{H} = \{\mathcal{M}_1, ..., \mathcal{M}_r = V\}$. Secondly, we define the vectors of parameters

$$\boldsymbol{\eta}_{\mathcal{L}}^{\mathcal{M}} = \boldsymbol{C}_{\mathcal{L}}^{\mathcal{M}} \log \boldsymbol{M}_{\mathcal{L}} \pi \tag{2}$$

Where, $\boldsymbol{C}$ and $\boldsymbol{M}$ are respectively a contrast and marginalization matrices. Finally, the vector of all parameters is drawn stacking the previous vectors respecting the properties of completeness and hierarchy, that is any log-linear interaction must be defined one time in the first marginal where it appears. These models present several advantages. First of all, Bergsma and Rudas (2002, Theorem 1) proved that these models are always smooth.

In addition, these models are also used to represent and to shape a list of conditional independences constricting to zero certain parameters. We used these properties to finding a subclass of **GM II** representable with marginal models, that therefore is a smooth subclass of **GM II** . Thanks to the Theorem 1 of Bergsma, Rudas and Neméth (2011), we can define this smooth class with the following theorem.

Theorem A **GM*s* II** is a marginal model if, $\forall T_h$ the set of vertices $V_j \in CH_h$, such that $Nb(V_j) \notin \mathcal{C}_h$, $\{K : K \in \mathcal{K}_h; K \cap nb(V_j) \neq \emptyset\} \subseteq J_h$.

where CH_h is the set of all vertices V_j in the component T_h having at least one parent: $pa_G(V_j) \neq \emptyset$; $\mathcal{C}_h$ is the class of all complete subsets of T_h; $\mathcal{K}$ is the class of all connected subsets of T_h; finally J_h is the class such that $A \in J_h$ iff $pa_G(A) = pa_D(T_h)$.

2 Application on European Values Study (EVS), (2008)

In order to show the main results on **GM*s* II**, we analyse the data from the European Values Study (EVS), (2008). The EVS is a research project on human values in Europe. In particular, the research involves how Europeans think about family, work, religion, politics and society. In order to investigate trends on opinions about the previous themes, we taken into account different variables. In particular, we divided the variables in three groups. In the first group we placed the variables concerning the personal data of the respondents (*i.e. sex, range of age, country,...*). In the second group there are variables about the achievements of the respondents (*i.e. education level, house owner, employed, children...*). Finally, the last group regards the variables that consider the opinion of the respondents about the main topics cited above (*i.e. family, work, religion, politics and society*). We represented each group of variables with a component in the chain graphs. We divided the Europe dataset in 5 subsets concerning the big geographical areas (Northern, Central, Southern, Occidental and Oriental). For all datasets we proposed certain graphical models in order to find the most representative model. In particular, the graphs in figure 1 and 2 represent the best conditional independence models for different areas of Europe.

The variables involved are **A**: *Range of Age* ($20 \dashv 40$, $40 \dashv 60$, > 60), **G**: *Gender* (Male, Female), **C**: *Children*(Yes, No), **E**: *Employed* (Yes, No), **T**: *Trust in the people*(Yes, No), **LS**: *Life satisfaction*(High, Low), **OS**: *Opinion on Society* (High, Mean, Low), **ID**: *Ideology* (Freedom, Either, Equality). We will highlight some interesting trends in the opinion of the European citizens of different areas.

The statistical software R-project is used with the help of the package "hmmm", (that is available from the comprehensive R Archive Network

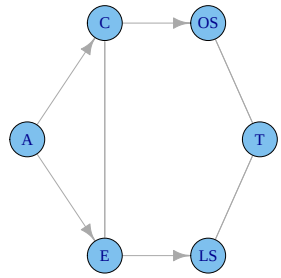


FIGURE 1. $\{A \perp LS, OS, T|C, E;\ OS \perp LS|T, C, E;\ LS, OS \perp E|C;\ T, OS \perp C|E\}$.

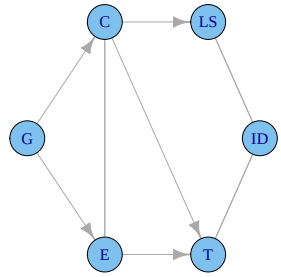


FIGURE 2. $\{G \perp LS, ID, T|C, E;\ LS \perp T|ID, C, E;\ LS, ID \perp E|C;\ ID \perp C|E\}$.

out http://cran.r-project.org/web/packages/hmmm) for the test of the marginal models and the estimation of the parameters and the packages "gRbase" (http://cran.r-project.org/web/packages/gRbase) and "RBGL" (http://www.bioconductor.org/packages/release/bioc/html/RBGL.html) to the part concerning the graphs.

References

Andersson, S.A., Madigan, D. and Perlman, M.D. (2001). *Alternative Markov properties for chain graphs*. Scand. J. Statist. 28 33-85.

Bergsma, W.P., Rudas, T., (2002). *Marginal models for categorical data*. Ann. Statist. 30, 140-159.

Bergsma, W.P., Rudas, T., Neméth, R., (2009). *Marginal conditional independence models with application to graphical modeling*. Biometrika, 97, 4, pp 1006-1012.

Carey, V., Long, L.,Gentleman, R., *RBGL: An interface to the BOOST graph library*. R package version 1.32.0.

Colombi, R., Giordano, S., Cazzaro, M., R Development Core Team (2012). *hmmm: Hierachical Multinomial Marginal Models*. R package version 1.0.0.

Dethlefsen, C., Hjsgaard,S. (2005). *A Common Platform for Graphical Models in R: The gRbase Package*. Journal of Statistical Software, 14(17), 1-12.

Drton, M. (2009). *Discrete chain graph models*. Bernoulli, 15(3), 736-753.

EVS (2010). *European Values Study 2008, 4th wave*. Data File Version 1.0.0 (2010-11-30), doi: 10.4232/1.10031.

Lauritzen, S.L. (1996). *Graphical Models*. New York: Oxford Univ. Press.

Adjusted pseudo composite likelihood ratios

Luigi Pace[1], Alessandra Salvan[2], Nicola Sartori[2]

[1] Department of Economics and Statistics, University of Udine, Italy
[2] Department of Statistical Sciences, University of Padova, Italy

E-mail for correspondence: sartori@stat.unipd.it

Abstract: For inference about a parameter of interest, in the presence of nuisance parameters, we consider a pseudo likelihood obtained from a composite likelihood by replacing the nuisance component with an estimate based on a generic estimating equation. Suitable adjustments are developed for the resulting pseudo composite likelihood ratio statistic, taking into account both nuisance estimation procedure and misspecification.

Keywords: Composite likelihood; Misspecification; Nuisance parameter.

1 Introduction

Complex model structures may give rise to full likelihoods that are intractable. It is then natural to consider, as a basis for inference, a simplified pseudo likelihood. One instance is composite likelihood (Lindsay, 1988; Varin *et al.*, 2011), useful when the fully specified likelihood is computationally cumbersome as well as when a fully specified model is out of reach. Another instance is the pseudo likelihood of Gong and Samaniego (1981), that eliminates nuisance parameters in a simple way. These two types of pseudo likelihood may be combined, as is illustrated here.
When using a pseudo likelihood we have often to content ourselves with maintaining only some of the properties of a full likelihood. We will consider pseudo likelihoods whose score function remains, at least to first order, an unbiased estimating function for the parameter of interest, i.e. the first Bartlett identity still holds. But we will admit failure of the second Bartlett identity, so that the variance of the pseudo score may differ from the expected value of minus the pseudo log-likelihood Hessian by $O(n)$, where n is proportional to a measure of information, conventionally the sample size. Under the usual regularity conditions, the first Bartlett identity allows to preserve consistency of the maximum pseudo likelihood estimator of the parameter of interest. On the contrary, a serious failure of the second Bartlett identity causes the asymptotic null distribution of the likelihood ratio statistic to take the form of a linear combination of independent chi squared random variables on one degree of freedom. The purpose here is

to illustrate a rescaled pseudo likelihood ratio statistic which recovers the usual asymptotic null chi squared distribution. In particular, we concentrate on inference for a parameter of interest based on composite likelihoods, when nuisance parameters are estimated through general estimating equations. The resulting adjusted pseudo composite likelihood ratio is developed in Pace *et al.* (2013), generalizing the proposal of Pace *et al.* (2011) for composite likelihoods.

2 Pseudo composite likelihood

Consider inference about a p-dimensional parameter of interest θ, in the presence of a q-dimensional nuisance parameter ϕ. Suppose that $y_1, \ldots, y_n$ are independent observations of m-dimensional random variables $Y_1, \ldots, Y_n$.

Difficulties in modelling interdependencies between components of Y_i may suggest the consideration of a composite likelihood (Lindsay, 1988; Varin *et al.*, 2011) in place of the full likelihood. A composite log likelihood for (θ, ϕ) has the form

$$c\ell(\theta, \phi) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_k \ell_k(\theta, \phi; A_k(y_i)) \,, \tag{1}$$

where $A_k(y_i)$, $k = 1, \ldots, K$, are K marginal or conditional events on the sample space of Y_i, giving log likelihoods $\ell(\theta, \phi; A_k(y_i))$, and where w_k, $k = 1, \ldots, K$, are positive weights.

Let us consider inference about θ based on $c\ell(\theta, \phi)$ when ϕ is replaced by $\tilde{\phi}$, solution of the estimating equation $\sum_{i=1}^{n} g_i(\phi; y_i) = 0$, such that $E_{\theta,\phi}\left(g_i(\phi; Y_i)\right) = 0$, for $i = 1, \ldots, n$, and for every θ and ϕ, or, more generally, $E_{\theta,\phi}\left(\sum_{i=1}^{n} g_i(\phi; Y_i)\right) = O(1)$. Therefore, $\tilde{\phi} \dot{\sim} N_q(\phi, \Sigma)$, where

$$\begin{aligned} \Sigma &= Q^{-1} S (Q^{-1})^{\mathrm{T}} \,, \\ Q &= E_{\theta,\phi}\left(-\frac{\partial}{\partial \phi^{\mathrm{T}}} g(\phi; Y) \right) \,, \\ S &= \mathrm{V}_{\theta,\phi}\left(g(\phi; Y)\right) = E_{\theta,\phi}\left(g(\phi; Y) g(\phi; Y)^{\mathrm{T}}\right) \,. \end{aligned}$$

The latter equality holds with error $O(1)$ if $E_{\theta,\phi}\left(\sum_{i=1}^{n} g_i(\phi; Y_i)\right) = O(1)$. The substitution of ϕ with $\tilde{\phi}$ gives the pseudo composite log likelihood $c\ell_{PS}(\theta) = c\ell(\theta, \tilde{\phi})$ and the pseudo composite log likelihood ratio

$$cW_{PS}(\theta) = 2\left\{ c\ell_{PS}(\tilde{\theta}_c) - c\ell_{PS}(\theta) \right\} \,,$$

where $\tilde{\theta}_c$ is the maximizer of $c\ell_{PS}(\theta)$. As happens for the profile composite likelihood ratio, $cW_P(\theta)$, the asymptotic null distribution of $cW_{PS}(\theta)$ is a

linear combination of p independent chi squared random variables on one degree of freedom.
The adjusted version of $cW_{PS}(\theta)$ proposed in Pace *et al.* (2013) is

$$cW_{PS}^{adj}(\theta) = A_c(\theta) cW_{PS}(\theta) \dot{\sim} \chi_p^2 ,$$

where

$$A_c(\theta) = \frac{cU_{PS}(\theta)^{\mathrm{T}} K^{-1} cU_{PS}(\theta)}{cU_{PS}(\theta)^{\mathrm{T}} H_{\theta\theta}^{-1} cU_{PS}(\theta)} , \tag{2}$$

with $cU_{PS}(\theta) = (\partial/\partial\theta) c\ell_{PS}(\theta)$ and matrices K and $H_{\theta\theta}$ calculated at $(\theta, \tilde{\phi})$ and given by

$$\begin{aligned} K &= J_{\theta\theta} + H_{\theta\phi}\Sigma(H_{\theta\phi})^{\mathrm{T}} - \Omega_{\theta\phi}(Q^{-1})^{\mathrm{T}} H_{\theta\phi}^{\mathrm{T}} - H_{\theta\phi} Q^{-1} \Omega_{\theta\phi}^{\mathrm{T}} , \\ H_{\theta\theta} &= \mathrm{E}_{\theta,\phi}\{-\partial cU_\theta(\theta,\phi)/\partial\theta^{\mathrm{T}}\} . \end{aligned}$$

Above, $cU_\theta(\theta,\phi) = (\partial/\partial\theta) c\ell(\theta,\phi)$ and

$$\begin{aligned} J_{\theta\theta} &= \mathrm{E}_{\theta,\phi}\{cU_\theta(\theta,\phi) cU_\theta(\theta,\phi)^{\mathrm{T}}\} , \\ \Omega_{\theta\phi} &= \mathrm{Cov}(cU_\theta(\theta,\phi), g(\phi;Y)) = \mathrm{E}_{\theta,\phi}(cU_\theta(\theta,\phi)\, g(\phi;Y)^{\mathrm{T}}) , \\ H_{\theta\phi} &= \mathrm{E}_{\theta,\phi}\{-\partial cU_\theta(\theta,\phi)/\partial\phi^{\mathrm{T}}\} . \end{aligned}$$

Wald and score statistics from $c\ell_{PS}(\theta)$ are

$$\begin{aligned} cW_{PS}^{e}(\theta) &= (\tilde{\theta}_c - \theta)^{\mathrm{T}} H_{\theta\theta}^{\mathrm{T}} K^{-1} H_{\theta\theta} (\tilde{\theta}_c - \theta) , \\ cW_{PS}^{u}(\theta) &= cU_{PS}(\theta)^{\mathrm{T}} K^{-1} cU_{PS}(\theta) , \end{aligned}$$

both with asymptotic χ_p^2 null distribution.

3 Equicorrelated multivariate normal data

As an illustration, let us consider Y_i as a multivariate normal with components having mean μ, variance σ^2, and correlation ρ between any two components of Y_i. This is a reparameterization of a one-way normal-theory random effects model.
We focus on ρ as the parameter of interest with $\phi = (\mu, \sigma^2)$. A simple composite log likelihood is the pairwise log likelihood. Denoting by y_{ir}, $r = 1, \ldots, m$, a generic element of y_i, $i = 1, \ldots, n$, the pairwise log likelihood is given by (1) where the events $A_k(y_i)$ are defined in terms of pairs of observations (y_{ir}, y_{is}). Here, we will set all weights equal to 1.
Using for ϕ the moment estimate $\tilde{\phi} = (\tilde{\mu}, \tilde{\sigma}^2)$, with

$$\tilde{\mu} = \bar{y} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{r=1}^{m} y_{ir}, \quad \tilde{\sigma}^2 = \sum_{i=1}^{n}\sum_{r=1}^{m}(y_{ir}-\bar{y}_i)^2/(nm) + \sum_{i=1}^{n}(\bar{y}_i-\bar{y})^2/(n-1) ,$$

we obtain the pseudo pairwise log likelihood $c\ell_{PS}(\rho) = c\ell(\rho, \tilde{\phi})$. Being the parameter of interest one-dimensional, the modifying factor (2) is simply $A_c(\rho) = H_{\rho\rho}/K$, which is equal to

$$\frac{(n-1)\, m\, (\rho+1)^2\, \left(\rho^2+1\right)}{(m\,\rho-\rho+1)^2\, \left\{(n\,m\,\rho^2-\rho^2+2\,m\,\rho-2\,\rho+2\,n\,m-m-1)\,\rho^2+n\,m-m\right\}}.$$

Table 1 reports the empirical coverage probabilities of confidence intervals for ρ in a simulation study with $n = 5$, $m = 30$, $\mu = 0$, $\sigma^2 = 1$ and $\rho = 0.2, 0.5, 0.9$. Among the statistics based on $c\ell_{PS}(\theta)$, $cW^{adj}_{PS}(\rho)$ gives empirical coverage probabilities reasonably close to the nominal levels, and very close to those of the adjusted profile composite likelihood ratio of Pace *et al.* (2011), cW^{adj}_{P}. Note also that both are often outperforming even the profile likelihood ratio from the full model, W_P.

TABLE 1. Percentage empirical coverage of confidence intervals based on different statistics in three simulations, 10,000 replications, $n = 5$, $m = 30$, $\mu = 0$, $\sigma^2 = 1$ and $\rho = 0.2, 0.5, 0.9$.

	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.9$		
	90	95	99	90	95	99	90	95	99
W_P	82.9	89.9	96.9	82.8	90.1	97.0	82.1	89.4	96.8
cW^{adj}_{P}	93.7	98.8	99.7	86.5	94.4	99.9	79.5	86.5	94.5
cW_{PS}	18.9	22.4	29.3	9.4	11.2	14.7	6.0	7.0	9.1
cW^{adj}_{PS}	93.5	98.3	99.6	89.0	95.6	99.9	83.7	88.4	93.8
cW^{e}_{PS}	69.7	73.9	79.9	74.5	79.8	86.1	86.6	91.0	95.9
cW^{u}_{PS}	94.0	99.2	99.9	86.9	92.5	98.5	78.3	81.7	87.1

References

Gong, G. and Samaniego, F. J. (1981). Pseudo Maximum Likelihood Estimation: Theory and Applications. *Annals of Statistics*, **9**, 861 – 869.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, **80**, 220 – 241.

Pace, L., Salvan, A. and Sartori, N. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica*, **21**, 129 – 148.

Pace, L., Salvan, A. and Sartori, N. (2013). Calibrating pseudo likelihood ratios. Manuscript.

Varin, C., Reid, N. and Firth. D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5 – 42.

On the use of the characteristic function of the residuals to test for the equality of regression curves

Juan Carlos Pardo-Fernández[1], M. Dolores Jiménez-Gamero[2], Anouar El Ghouch[3]

[1] Departamento de Estatística e I.O., Universidade de Vigo, Spain
[2] Departamento de Estadística e I.O., Universidad de Sevilla, Spain
[3] Institut de statistique, biostatistique et sciences actuarielles (ISBA), Université catholique de Louvain, Belgium

E-mail for correspondence: juancp@uvigo.es

Abstract: We study a new procedure to test for the equality of k regression curves in a fully nonparametric context. The test is based on the comparison of empirical estimators of the characteristic functions of the regression residuals in each population. The asymptotic behaviour of the test statistic is studied. Under the null hypothesis the distribution of the test statistic converges to a combination of independent χ^2_1 random variables.

Keywords: Comparison of regression curves; empirical characteristic function; regression residuals.

1 Introduction

Testing for the equality population characteristics is a classical problem in Statistics. In this paper we consider the comparison of k regression functions in a general nonparametric setting. Let (X_j, Y_j), $1 \leq j \leq k$, be k independent random vectors satisfying general nonparametric regression models $Y_j = m_j(X_j) + \sigma_j(X_j)\varepsilon_j$, where $m_j(x) = E(Y_j \mid X_j = x)$ is the regression function, $\sigma_j^2(x) = Var(Y_j \mid X_j = x)$ is the conditional variance function and ε_j is the regression error, which is assumed to be independent of X_j. The regression functions, the variance functions, the distribution of the errors and the distribution of the covariates are unknown and no parametric models are assumed for them. Under this framework our approach is fully nonparametric. In this conditional setting, we are interested in testing for the null hypothesis of the conditional means or regression functions

$$H_0 : m_1 = m_2 = \ldots = m_k,$$

or, in other words, the mean effect of the covariates over the responses is equal in the k populations. The alternative hypothesis is H_1 :

H_0 is not true.

The problem of testing for the equality of regression curves in nonparametric settings has been previously treated in the statistical literature. Particularly, Pardo-Fernández et al. (2007) studied an approach based on comparing the distribution functions of the regression errors. More in detail, let $\varepsilon_j = \{Y_j - m_j(X_j)\}/\sigma_j(X_j)$ be the regression error in population j. Let m_0 be the common regression curve under the null hypothesis, and define

$$\varepsilon_{0j} = \{Y_j - m_0(X_j)\}/\sigma_j(X_j) = \varepsilon_j + \{m_j(X_j) - m_0(X_j)\}/\sigma_j(X_j), \quad (1)$$

$1 \leq j \leq k$. It turns out that the null hypothesis H_0 is true if and only if, for all $1 \leq j \leq k$, the random variables ε_j and ε_{0j} have the same distribution (see Theorem 1 in Pardo-Fernández et al., 2007). This assessment can be interpreted in terms of the cumulative distribution function (cdf) or in terms of any other function characterizing the probability law of the errors. Pardo-Fernández et al. (2007) restricted their attention to the cdf.

The probability law of any random variable X is also characterized by its characteristic function (cf), $\varphi(t) = E\{\exp(\mathrm{i}tX)\}$. Recent years have witnessed an increasing number of proposals for hypothesis testing whose test statistics measure deviations between the empirical characteristic function (ecf) of the available data and an estimator of the cf under the null hypothesis. See, for example, Jiménez-Gamero et al. (2005) or Hušková and Meintanis (2009). An advantage of the cf approach over the one based on the cdf, is that the former usually requires less stringent assumptions for its validity. In addition, simulation results for finite sample sizes in these and other related papers show that the tests based on the ecf compete very satisfactorily with those based on the empirical cdf (ecdf). Having in mind the reasons above, the purpose of the present paper is to test H_0 by comparing estimators of the cfs of the random variables ε_j and ε_{0j}.

2 Testing procedure and asymptotic results

Let (X_{jl}, Y_{jl}), $1 \leq l \leq n_j$, be iid observations from (X_j, Y_j), $1 \leq j \leq k$. Let $n = \sum_{j=1}^{k} n_j$, and let $f_{mix}(x) = \sum_{j=1}^{k} p_j f_j(x)$ be the density of the mixture of covariates according to the weights $p_1, \ldots, p_k$, where $p_j = \lim n_j/n$. In order to estimate the regression errors, we first need estimators of the regression functions, $\hat{m}_j(x)$, the scale functions, $\hat{\sigma}_j(x)$, and the common regression function under H_0, $\hat{m}_0(x) = \sum_{j=1}^{k} p_j\{\hat{f}_j(x)/\hat{f}_{mix}(x)\}\hat{m}_j(x)$. With this aim we use nonparametric estimators based on kernel smoothing techniques (Nadaraya-Watson or local-linear estimators). Based on these estimators, for each population j, $1 \leq j \leq k$, we construct two samples of residuals:

$$\hat{\varepsilon}_{jl} = \frac{Y_{jl} - \hat{m}_j(X_{jl})}{\hat{\sigma}_j(X_{jl})} \qquad \text{and} \qquad \hat{\varepsilon}_{0jl} = \frac{Y_{jl} - \hat{m}_0(X_{jl})}{\hat{\sigma}_j(X_{jl})},$$

$1 \leq l \leq n_j$, whose ecfs are

$$\hat{\varphi}_j(t) = \frac{1}{n_j}\sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{jl}) \qquad \text{and} \qquad \hat{\varphi}_{0j}(t) = \frac{1}{n_j}\sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{0jl}),$$

respectively. These ecfs are nothing but consistent kernel-based nonparametric estimators of the population cfs $\varphi_j(t) = E\{\exp(\mathrm{i}t\varepsilon_j)\}$ and $\varphi_{0j}(t) = E\{\exp(\mathrm{i}t\varepsilon_{0j})\}$, respectively. The testing procedure consists of comparing $\hat{\varphi}_j(t)$ and $\hat{\varphi}_{0j}(t)$, $1 \leq j \leq k$, using a weighted L_2-distance of the form

$$T_n = \sum_{j=1}^{k} \frac{n_j}{n} \int |\hat{\varphi}_j(t) - \hat{\varphi}_{0j}(t)|^2 \, w(t)dt,$$

where w is any given non-negative weight function ($|z|$ denotes the modulus of the complex number z). Under H_0, $\varphi_j(t) = \varphi_{0j}(t)$, then T_n should be close to zero. On the other hand, large values of T_n should lead to the rejection of H_0. In practice, given a significance level, α, a threshold value above which H_0 is rejected needs to be established. To this end we need to study the null (asymptotic) distribution of T_n.
A summary of the theoretical results obtained are listed below (due to the lack of space, we do not show the details of the statements):

- Under H_0, nT_n converges in distribution to $W = Z'\mathcal{A}Z$, where Z a certain k-dimensional zero-mean Normal vector with covariance matrix Σ and $\mathcal{A} = diag(a_1, \ldots, a_k)$, where $a_j = \int t^2 |\varphi_j(t)|^2 w(t)dt$.
- In other words, the limiting distribution of nT_n under H_0 is a linear combination of independent chi-square variables, $\sum_{j=1}^{k} \beta_j \chi^2_{1,j}$, where $\chi^2_{1,1}, \ldots, \chi^2_{1,k}$ are independent chi-square random variates with one degree of freedom and $\beta_1, \ldots, \beta_k$ are the eigenvalues of $\mathcal{A}\Sigma$.
- The quantities β_j are unknown. Estimators for them, say $\hat{\beta}_j$, can be obtained by plug-in methods.
- The distribution of $\sum_{j=1}^{k} \hat{\beta}_j \chi^2_{1,j}$ can be approximated via Monte Carlo methods. This approximation allows us to obtain critical values and/or the p-value for the test based on T_n.
- It can be proved that if all the covariates have the same distribution, the errors also have the same distribution and the variance functions are equal, then the asymptotic distribution of nT_n is a re-scaled χ^2_{k-1}.
- Although the test based on T_n is fully nonparametric, it can detect local alternatives converging to the null hypothesis at a rate $n^{-1/2}$.

TABLE 1. Observed rejection proportions of the test based on the asymptotic distribution of T_n.

		α:	0.100	0.100	0.050	0.050	0.010	0.010
model	(n_1, n_2, n_3)	C:	2.00	3.00	2.00	3.00	2.00	3.00
(i)	(50,50,50)		0.125	0.112	0.062	0.062	0.008	0.007
	(100,50,50)		0.116	0.113	0.065	0.068	0.008	0.011
	(100,100,50)		0.127	0.121	0.076	0.070	0.018	0.016
	(100,100,100)		0.109	0.099	0.055	0.054	0.012	0.009
(ii)	(50,50,50)		0.909	0.899	0.839	0.815	0.620	0.588
	(100,50,50)		0.971	0.960	0.925	0.918	0.782	0.757
	(100,100,50)		0.975	0.968	0.952	0.946	0.837	0.818
	(100,100,100)		0.996	0.996	0.995	0.990	0.956	0.953

3 Simulations

We have performed an intensive simulation study to check the finite sample performance of the test based on the critical values obtained from the approximation of the asymptotic null distribution of T_n. Here we only show a small portion of the results. Consider the case of three populations ($k = 3$). The regression functions are (i) $m_1(x) = m_2(x) = m_3(x) = x$ (null hypothesis) and (ii) $m_1(x) = x$, $m_2(x) = x + 0.2$, $m_3(x) = x + 0.4$ (alternative hypothesis). The variance functions are given by $\sigma_1(x) = \sqrt{0.25}$, $\sigma_2(x) = \sqrt{0.25}$ and $\sigma_3(x) = \sqrt{0.50}$. The covariates X_1, X_2 and X_3 are distributed according to $Beta(1.5, 2)$, $Beta(2, 1.5)$ and $Beta(2, 2)$, respectively, and all regression errors are $N(0, 1)$. Nonparametric estimation of the regression functions is performed by the local-linear estimator. Smoothing parameters of the form $h = Cn^{-0.375}$ are taken (cases $C = 2$ and $C = 3$ are displayed). As weighting function $w(t)$ we take the density of a $N(0, 1)$. Table 1 displays the observed proportion of rejections in 1000 simulated data sets. The level is well approximated (model i), specially for large sample sizes, and the behaviour in terms of power (model ii) is correct.

References

Hušková, M. and Meintanis, S.G. (2009). Goodness-of-fit tests for parametric regression models based on empirical characteristic functions. *Kybernetika*, **45**, 960 – 971.

Jiménez-Gamero, M.D., Muñoz-García, J. and Pino-Mejías, R. (2005). Testing goodness of fit for the distribution of errors in multivariate linear models. *Journal of Multivariate Analalysis*, **95**, 301 – 322.

Pardo-Fernández, J.C., Van Keilegom, I. and González-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statistica Sinica*, **17**, 1115 – 1137.

Analysing Formalisation of Management Accounting by Bayesian Variable Selection in a Cumulative Logit Model

Daniela Pauger[1], Christine Duller[1], Helga Wagner[1]

[1] Department of Applied Statistics and Econometrics, Johannes Kepler University Linz, Austria

E-mail for correspondence: daniela.pauger@jku.at

Abstract:
In this paper we apply Bayesian variable selection to a cumulative logit model. We analyse the extent of formalisation of management accounting in Austrian and German family and non-family firms. Data augmentation and MCMC methods are used for posterior sampling.

Keywords: Bayesian Estimation; Ordinal Logit Model; Data Augmentation; Variable Selection; Management Accounting; Family Firm.

1 Introduction

Formalisation of management accounting, measured as the extent of written documentations of long-run concepts and strategies, is presumed as determining factor for business success. Therefore the degree of formalisation is an interesting topic for research in business administration. The level of formalisation itself is said to be influenced by several contingency factors as enterprise size or structure (family or non-family firm). The extent of written documentations is measured on a three-point scale ranging from "less or not recorded" to "fully recorded". The goal of our analysis is to find out which of various interesting aspects (size, structure, generation, state and industry) are associated with the extent of formalisation (Duller C., Feldbauer-Durstmüller B., Mitter C., 2011). The application is based on survey data from Austrian and German enterprises.
We apply a Bayesian cumulative logit model and use variable selection with spike and slab priors to determine relevant predictors. Inference is based on sampling from the posterior distribution by MCMC methods and data augmentation.

2 Model

Let y denote the ordinal response variable taking a value in one of m ordered categories and $\mathbf{x}_i$ the vector of covariates of dimension $1 \times p$. To link the covariates to the response we use a cumulative logit (proportional odds) model, where

$$\log \frac{P(y_i \leq k|\mathbf{x}_i)}{P(y_i > k|\mathbf{x}_i)} = \theta_k - \mathbf{x}_i\boldsymbol{\beta}$$

Here $\boldsymbol{\beta}$ is the $p\times 1$ vector of regression coefficients and $\theta_1, \ldots \theta_m$ are category specific parameters.

2.1 Data Augmentation for the Ordinal Logit Model

The proportional odds model has a representation via a latent utility

$$z_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i \sim \text{Log } (0,\, 1)$$

where

$$y_i = k \Longleftrightarrow \theta_{k-1} < z_i \leq \theta_k$$

and $-\infty = \theta_0 < \theta_1 < \cdots < \theta_m = \infty$. Usually θ_1 is fixed to $\theta_1 = 0$ and only $\theta_2, \ldots, \theta_{m-1}$ are estimated. We use a very accurate approximation of the standard logistic distribution with a finite scale mixture of normal distributions (Frühwirth-Schnatter and Frühwirth, 2010) to obtain an auxiliary normal model with heteroscedastic errors, given as

$$z_i = \eta_i^b + \tilde{\epsilon}_i \qquad \tilde{\epsilon}_i \sim \mathcal{N}(0, s^2_{r_i}),$$

where r_i denotes the component indicator for the mixture component. Hence additionally to the model parameters the auxiliary variables $\mathbf{z} = (z_1, \ldots, z_n)$ and $\mathbf{r} = (r_1, \ldots, r_n)$ will be sampled using the MCMC scheme.

2.2 Priors

For $\boldsymbol{\theta}$ we use a flat prior $0 < \theta_2 < \cdots < \theta_{m-1} < \infty$. To implement variable selection, we specify a spike and slab prior for $\boldsymbol{\beta}$ with a Dirac spike and a normal independence slab. This prior can be specified hierarchically by introducing an indicator vector $\boldsymbol{\delta}$, where $\delta_j = 1$ if β_j belongs to the slab component, see Malsiner-Walli and Wagner (2011) for more details.

2.3 MCMC for the Ordinal Logit Model

Bayesian inference is based on the posterior distribution, formally given by Bayes theorem as

$$p(\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{r}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\mathbf{z}, \mathbf{r}|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\theta}) p(\boldsymbol{\beta}, \boldsymbol{\delta}) p(\boldsymbol{\theta})$$

This posterior distribution is not tractable analytically, but the parameters and auxiliary variables $(\mathbf{z}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\theta})$ can be sampled from the posterior distribution using the following Gibbs sampling scheme:

Step (I). Sample $(\boldsymbol{\theta}, \mathbf{z}, \mathbf{r})$ from $p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{r}|\boldsymbol{\beta}, \mathbf{y})$

Step (Ia). Sample $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\boldsymbol{\beta}, \mathbf{y})$ marginalizing over $(\mathbf{z}, \mathbf{r})$.

Step (Ib). For $i = 1, \ldots, n$ sample z_i from $p(z_i|\boldsymbol{\beta}, y_i)$.

Step (Ic). For $i = 1, \ldots, n$ sample r_i from $p(r_i|\boldsymbol{\beta}, z_i, y_i)$.

Step (II). Sample $(\boldsymbol{\delta}, \boldsymbol{\beta})$ from $p(\boldsymbol{\delta}, \boldsymbol{\beta}|\mathbf{z})$
The full conditional posterior distribution does not depend on $\mathbf{y}$ and $\boldsymbol{\theta}$.

3 Analysis of Data on Management Accounting in Austrian and German Family Firms

3.1 Data

We use data from a survey on family and non-family firms in Austria and Germany collected between July 2009 and March 2010. This survey concentrated on the topics controlling, financial management and organisational development of the companies. 1,052 of the questionnaires are the basis of this study and 454 of them were removed because of missing values in the possible regressors.

As potential covariates which might affect the extent of formalisation of management accounting we use annual sales (categories: below 10 million Euros, between 10 and 50 million Euros, at least 50 million Euros), number of employees (dichotomous: below/at least 250 employees), business sector (dichotomous: industrial/other), state (categories: Austria, German federal states of Bavaria, Baden-Württemberg, North Rhine-Westphalia and Lower Saxony), generation (categories: first or second, third, fourth, fifth or later generation) and structure (dichotomous: no family firm and family firm).

3.2 Results

The posterior mean of the sampled threshold θ_2 is 1.48. Table 1 shows the posterior means and inclusion probability of the regressors based on 5,000 MCMC iterations after burn-in of 1,000. Estimated posterior inclusion probabilities are higher than 0.5 only for two covariates: the indicator for family firms and that for large firms. Family firms are less likely than others to have more formalised plans, whereas firms with at least 250 employees are more likely to formalise their plans than those with less than 250 employees. Annual sales, the second indicator for the size of a company, is not selected into the model. Also the effects of business sector, state and generation in family firms are not selected.

TABLE 1. Posterior means and inclusion probability of the regressors

Variable	posterior mean	inclusion prob.
intercept	1.791	1
business sector (base: not industrial)	0.005	0.04
annual sales (base: up to 10 million)		
10 to 50 million	0.041	0.12
more than 50 million	0.144	0.25
number of employees (base: 50 to 249)		
250 and more	**0.486**	**0.76**
structure (base: no family firm)		
family firm	**-0.514**	**0.85**
state (base: Austria)		
Baden-Württemberg	0.069	0.14
Bavaria	-0.002	0.04
Lower Saxony	-0.013	0.06
North Rhine-Westphalia	-0.003	0.04
generation (base: first or second generation)		
third generation	-0.032	0.10
fourth generation	-0.008	0.06
fifth generation or later	0.023	0.07

References

Duller C., Feldbauer-Durstmüller B., Mitter C. (2011). Corporate Governance and Management Accountig in Family Firms: Does Generation matter. *International Journal of Business Research*, Volume 11, Number 1, 2011, pp. 29-46.

Frühwirth-Schnatter, S. and R. Frühwirth (2010). Data augmentation and MCMC for binary and multinomial logit models. In T. Kneib and G. Tutz (Eds.), *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, Heidelberg, pp. 111 – 132. Physica-Verlag.

Malsiner-Walli, G. and H. Wagner (2011). Comparing Spike and Slab priors for Bayesian Variable Selection. *Austrian Journal of Statistics*, Volume 40, pp. 241-264.

A New Weibull Family of Hazard Models for Breast Cancer Survivals

Gleici Castro Perdona[1], Francisco Louzada[2], Cleyton Zanardo[3], Hayala Cavenague[1]

[1] FMRP, Universidade de São Paulo, Brazil
[2] ICMC, Universidade de São Paulo, Brazil
[3] NAP, Hospital de Câncer de Barretos, Brazil

E-mail for correspondence: pgleici@fmrp.usp.br

Abstract: In this paper, we discuss a family Weibull Modified of hazard model to breast cancer problematic. The breast cancer is addressed here by the high incidence and lack of knowledge in survival among women worldwide. The model is very flexible, and accommodate several particular cases. Inference procedure is based on maximum likelihood. A simulation study is performed in order to verify the frequentists properties of the maximum likelihood estimation procedure. A real example on breast cancer is addressed.

Keywords: Hazard modeling; Cure Rate Modeling; Breast cancer.

1 Introduction

Mixture models for long-term survivors have been widely used for fitting time-to-event data in which some individuals may never suffer the cause of failure under study. The modeling was introduced by Berkson and Gage (1952) and several authors have been considering them for survival modeling. Interested readers can refer to Maller and Zhou (1996)) and Perdona and Louzada-Neto (2011) for more information and literature review.
In this paper we propose a general cure rate model. The formulation is hazard-function-based since it describes the way in which the instantaneous probability of failure for a component changes with time (Lawless, 2003). The general hazard model embeds several existing lifetime models in a more general and flexible framework. The model formulation allows the accommodation of non monotone hazard function shapes, such as bathtub and unimodal and also presents a parameter denoting the presence of long-term survivors. Besides yielding a better fit to the data, the model allows the determination of the most appropriate model for a particular dataset.

2 Model Formulation

Let T be a nonnegative random variable representing the lifetime of an individual in some population. Following Lawless (2003), the hazard function at time t, is defined as $h(t) = \lim_{\Delta t \to 0} Pr(t \leq T < t + \Delta t \mid T \geq t)/(\Delta t) = f(t)/S(t)$. The class based in hazard functions Perdona and Louzada-Neto (2011), is given by

$$h(t; p, \theta, \nu) = \frac{p\theta \frac{\partial[1-g(t;\nu)]}{\partial t}}{1 - p[1 - g(t;\nu)]^{\theta}}[1 - g(t;\nu)]^{\theta-1}, \tag{1}$$

where the function $g(t;\nu)$ is the positive, monotone decrease, ν is a parameter vector of $g(.)$ θ is shape parameter and the p parameter, $0 < p < 1$, denotes the proportion of long-term survivals. The advantage of the (1) is the characterization of the modeling by the generic $g(.)$ covering a wide spectrum of hazard models.

Considering the case in which $p = 1$ and $\lambda = 0$ we have a simple Weibull model. Consequently, for $\beta = 1$ we have the exponential model. For $\lambda > 0$, we have the hazard function of a modified Weibull model (Lai, Xie and Murthy, 2003). Considering $0 < p < 1$ and $\lambda = 0$, following Berkson e Gage (1952), we obtain the Weibull-type long-term mixture model. Consequently, the long-term exponential case is obtained for $\beta = 1$.

Model (1) is very general, but it provides an easy way to test whether or not particular cases fit a data set by fitting sub models and then comparing the fit to the full model fitting. Moreover, a very important characteristic of the proposed general hazard model (1) is related to its possible shapes. It can accommodate increasing, decreasing, unimodal and bathtub hazard function shapes basically depending on the values of the parameters λ and β.

3 Inference

The estimation procedure is maximum-likelihood-based. Let $T_1^0, T_2^0, ..., T_n^0$ be the true survival times of a sample of size n. Assuming that they are independent identically distributed random variables with hazard function $h^0(t)$, for $i = 1, ..., n$, with observations subject to arbitrary right censoring, the period of follow-up for the i^{th} individual is limited to a value C_i. Subsequently, the observed survival time of the i^{th} individual is given by $T_i = min(T_i^0, C_i)$. Let $\delta_i = 1$ if $T_i = T_i^0$ (that is, if T_i is an observed death) and $\delta_i = 0$ if $T_i < T_i^0$ (that is, if T_i is a censored observation). The maximum likelihood estimates (MLEs) of the parameters of model (1), can be obtained by direct numerical maximization of the log-likelihood function $\log L = \sum_{i=1}^{n} \delta_i \log h(t_i) - H(t_i)$,where $H(t|z) = \int_0^t h(u)du$ is the cumulative hazard function (Lawless, 2003). The advantage of this procedure is that it runs immediately using existing statistical packages. Although the

maximization procedure can be performed by solving the system of non-linear equations given by the partial derivatives of $l(p, \alpha, \beta, \lambda)$ with respect to the parameters, in our experience pure Newton-Raphson schemes are extremely susceptible to failure to converge. Then, we consider the BFGS algorithm to compute the MLEs implemented in the software R via the optim function.

Large sample inference for the parameters can be based, in principle, on the MLEs and their estimated standard errors. An simulation study for examining the coverage probabilities of asymptotic confidence intervals for the parameters reveled that, overall, the coverage probabilities of the 95% confidence intervals for the parameters are close to the nominal coverage for moderate-sized samples, decreasing to about 90% when the number of units is small.

For comparison of nested models, which is the case when comparing the general hazard model (1) to some of its special sub-models, we can compute the maximum values of the unrestricted and restricted log-likelihoods to obtain the likelihood ratio statistics (LRS) (Lawless, 2003). Large positive values of the LRS give favorable evidence to the full model. However, for testing $H_0 : \lambda = 0$ versus $H_1 : \lambda > 0$ and $H_0 : p = 0$ versus $H_1 : p > 0$ (for testing the presence of long-term survivors), the test is performed on the boundary of the parameter space. In this case, following Ghitany and Maller (1992), the LRS, is assumed to be asymptotically distributed as a symmetric mixture of a chi-squared distribution with one degree of freedom and a point-mass at zero, where the reference distribution is a chi-square distribution with one degreee of freedom. Again, large positive values of the LRS give favorable evidence to the full model.

4 Breast Cancer Data

We apply the proposed methodology to a set of data on breast cancer extract from Cobre et al (2012), which consist of 40 women diagnosed with locally advanced invasive ductal carcinoma, treated at the Ribeirao Preto School of Medicine Clinic Hospital, Brazil, where all patients were submitted to neo-adjuvant (pre-operatory) chemotherapy between 2003 and 2006. The mean patient age was equals to 50.6 years old, with standard deviation (SD) equals to 6.0 years. Chemotherapy scheme consisted of two drugs whose action mechanisms were different within the cell cycle: docetaxel at mean dosage of 121.9 mg (SD=8.1 mg) per cycle and epirubicin at mean dosage of 88.8 mg (SD = 9.7 mg) per cycle, with number of cycles ranging from 4 to 6. We define the disease-free time as being the interval between surgery and relapse, termed as disease-free survival. In the sample, 75% of the patients had their follow-up time censored and we do not faced patients who die without prior relapse. After completing the neo-adjuvant chemotherapy, the patients were submitted to surgical treatment of the

breast affected by the neoplasia. The associated covariate was tumor size (in cm) where the size were 0 to 2 cm, 2 to 5 cm and greater than 5 cm. Accordingly, our general hazard model and some of its particular cases were fitted to the data. We consider here only the fit of those models with long term survivals since their presence is clear from the data. Table 1 shows the MLEs (and their corresponding standard errors in parentheses) of the parameters and the log-likelihood values, for the distributions: long term Modified Weibull (LTMW) and long term Weibull (LTW). The LRS value for testing the the LTMW and the LTW equals to 7.31 is significant, which is evidence in favor of the LTMW distribution. Such a result is corroborated by the plot of the empirical survival function which is superimposed by the estimated survival function of the LTMW distribution.

Acknowledgments: Special Thanks to FAPESP and CNPq, Brazil.

References

Berkson, J. and Gage, R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **52**, 501 – 515.

Cobre, J., Perdona, G.S.C., Louzada, F. and Peria, F. (2012). A Mechanistic Breast Cancer Survival Modeling Through The Axillary Lymph Node Chain. *Statistics in Medicine – IFirst*.

Ghitany, M., Maller, R. (1992). Asymptotic results for exponential mixture models with long-term survivors. *Statistics*, **23**, 321 – 336.

Lai, C.D., Min X. and Murthy, N.P. (2003). A Modified Weibull Distribution. *IEEE Transactions on Reliability*, **52**, 33 – 37.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley.

Maller, R.A., Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. New York: John Wiley.

Perdona, G.S.C. and Louzada, F. (2011). A general hazard model for lifetime data in the presence of cure rate. *Journal of Applied Statistics*, **38**, 1395 – 1405.

Approximate Bayesian Computation for Model Choice

Anthony N. Pettitt[1], Xing Ju Lee[1]

[1] Queensland University of Technology, Brisbane, Australia

E-mail for correspondence: a.pettitt@qut.edu.au

Abstract: Approximate Bayesian Computation (ABC) is rapidly becoming a popular and powerful tool for the Bayesian analysis of models with computationally intractable likelihood functions but from which it is possible to simulate data. Here we present an ABC algorithm for model choice problems in a Bayesian context. The method was applied to transmission modeling of methicillin-resistant *Staphylococcus aureus* (MRSA) within a hospital ward.

Keywords: disease transmission model; Monte Carlo; MRSA; summary statistics; model choice.

1 Introduction

Approximate Bayesian Computation (ABC) (or "likelihood-free" Bayesian inference) methods are having a major impact in applied science across a broad spectrum of disciplines for the Bayesian analysis of complex stochastic models with computationally intractable likelihood functions. However, it is assumed that it is possible to simulate data. The popularity of ABC is, in part, due to its conceptual simplicity and ease in implementation. We present here an extension of well-studied ABC algorithms for parameter estimation in order to incorporate model choice problems.

2 Methodology

In essence, an ABC algorithm simulates parameter values from a prior distribution and corresponding data from the model, and only accepts parameter values which generate simulated data similar to the observed data as measured by a discrepancy function. It has been shown that straightforward implementation of ABC algorithms for model choice problems can fail to produce sensible results (Robert et al., 2011). We have produced here an ABC algorithm developed to handle model choice problems by extending the regression approach outlined in Fearnhead & Prangle (2012) through the use of multinomial logistic regression to estimate model probabilities from the data. Hence the discrepancy measure now involves both

the model probabilities and posterior mean estimates. A similar approach was also recently proposed by Prangle *et al.* (2013) although they advocate the use of pairwise logistic regression for cases with more than 2 models whereas our approach handles such cases straightforwardly.
We used the SMC ABC replenishment algorithm (Drovandi & Pettitt, 2010) coupled with the regression approach introduced on a disease transmission model choice problem.

3 Example: MRSA Transmission in an intensive care unit (ICU)

We considered two different modes of cross-transmission, namely the direct contact transmission (standard model) and indirect contact transmission (alternate model) between healthcare workers (HCWs) and patients, to arrive at a model choice problem to describe the transmission of MRSA in an ICU. Data used were the weekly MRSA incidence as shown in Drovandi & Pettitt (2011). The variables and parameters used in the models are defined in Table 1. Stochastic simulation from each model were generated using the relative rates shown in Table 2 (derivation outlined in Drovandi & Pettitt (2008)).
It was assumed that there is no patient-patient or HCW-HCW transmission, perfect detection of colonisation, a constant number of patients in the ICU, and the number of colonised HCWs are at their equilibrium value given by the expression

$$\bar{Y_h} = \frac{f\left(Y_p\right) N_w}{f\left(Y_p\right) + \left[\frac{hN_w}{p_{ph}(1-h)}\right]}$$

where $f\left(Y_p\right) = Y_p$ for the standard model and $f\left(Y_p\right) = \mathbf{1}\left(Y_p > 0\right)$ for the alternate model (with $\mathbf{1}(.)$ denoting the indicator function).
The summary statistics used in the regression steps of Fearnhead & Prangle (2012) were the mean, variance, median absolute deviance, maximum, autocovariances and autocorrelations of lag 1 up to lag 5, number of zeros, ones and twos recorded, the AR(1) regression coefficient and the coefficients of categorical regression

$$x_t = \beta_0 + \beta_1 \mathbf{1}(x_{t-1} = 0) + \beta_2 \mathbf{1}(x_{t-1} = 1) + \beta_3 \mathbf{1}(x_{t-1} = 2) + \beta_4 \mathbf{1}(x_{t-1} > 2).$$

where x_t is the weekly incidence for week t.
The final estimated model probability obtained from the ABC algorithm was 0.7840 in favour of the standard model (Figure 1). This was well in agreement with the posterior model probability of 0.7864 for the standard model estimated using numerical integration.
The point and interval estimates obtained for ϕ_p (median of 0.0408 and 95% empirical credible interval $[0.031, 0.053]$) were similar to previous estimates reported in Drovandi & Pettitt (2011).

4 Discussion

The model choice results obtained provide evidence that direct transmission better explains the observed incidence of MRSA in the ICU compared with indirect transmission between patients and HCWs. Some caution should be taken in generalising this finding as the models considered are relatively simplistic. However, they suffice for the illustration of the utility of the ABC model choice algorithm presented here and enable the comparison of the results obtained with the model probabilities computed using numerical integration.

Acknowledgments: The authors are grateful to the Australian Research Council for financial support.

TABLE 1. Definition of parameters and variables used in the MRSA transmission models.

Symbol	Description	Value
Variables		
Y_h	Number of colonised HCWs	
Y_p	Number of colonised patients	
N	Incidence	
Parameters of interest		
ϕ	Transmission rate for direct contact	
ϕ'	Transmission rate for indirect contact	
Fixed parameters		
N_w	Ward size	15
μ	Discharge rate of uncolonised patients	1/4
μ'	Discharge rate of colonised patients	1/10.6
σ	Proportion of patients already colonised upon admission	0.03
h	Hand hygiene compliance parameter	0.39
p_{ph}	Probability of colonisation of a HCW	0.13

TABLE 2. Relative rates of events for the MRSA transmission models.

Event	Transition	Relative rate
Colonisation upon arrival	$(Y_p, N) \to (Y_p + 1, N)$	$\mu\sigma(N_p - Y_p)$
Colonisation within ward	$(Y_p, N) \to (Y_p + 1, N + 1)$	$\phi f(\bar{Y_h})(N_p - Y_p)$
Recovery	$(Y_p, N) \to (Y_p - 1, N)$	$\mu'(1 - \sigma)Y_p$

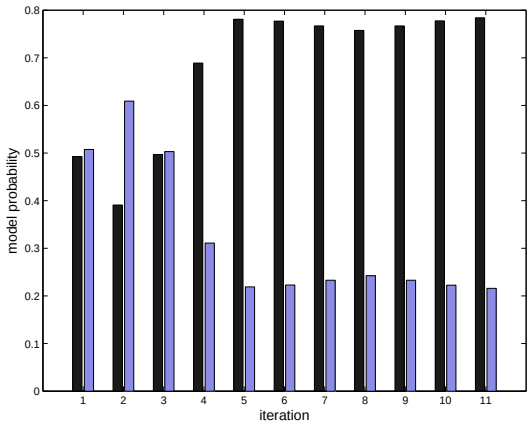


FIGURE 1. Model probabilities across SMC iterations for the standard model (dark bars) and alternate model (light bars).

References

Drovandi, C.C. and Pettitt, A.N. (2008). Multivariate Markov process models for the transmission of methicillin-resistant *Staphylococcus aureus* in a hospital ward. *Biometrics*, **64**:3, 851 – 859

Drovandi, C.C. and Pettitt, A.N. (2011). Estimating transmission rates of nosocomial pathogens using approximate Bayesian computation. *Statistical Communications in Infectious Diseases*, **3**:1.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**,1 – 28.

McBryde, E.S., Pettitt, A.N. and McElwain, D.L.S (2007). A stochastic mathematical model of methicillin resistant *Staphylococcus aureus* transmission in an intensive care unit: Predicting the impact of interventions. *Journal of Theoretical Biology*, **245**:3, 470 – 481.

Prangle, D., Fearnhead, P., Cox, M.P., Biggs, P.J. and French, N.P. (2013). Semi-automatic selection of summary statistics for ABC model choice. *ArXiv e-prints*.

Robert, C.P., Cornuet, J.-M., Marin, J.-M., and Pillai, N.S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, **108**:37, 15112 – 15117.

Trend and regional analysis of fatal off-piste and backcountry avalanche accidents in Austria within the years 1968 and 2011

Christian Pfeifer[1], Achim Zeileis[1], Peter Höller[2]

[1] Institut fr Statistik, Universitt Innsbruck, A–6020 Innsbruck
[2] Bundesamt und Forschungszentrum fr Wald, Institut fr Lawinenforschung, A–6020 Innsbruck

E-mail for correspondence: christian.pfeifer@uibk.ac.at

Abstract: In this article we analyze trend and regional patterns of fatal avalanche accidents caused by alpine skiers and snowboarders.

Keywords: Fatal snow avalanche events, time series, regional distribution.

1 Introduction and Data

In the Alps, backcountry skiing has become very popular in the last 50 years. Unfortunately, there are a lot of fatal accidents due to snow avalanches caused by skiers and snowboarders. In Austria, about 25 fatalities caused by snow avalanches every year are expected. Furthermore it is reported that the number of fatalities is more or less constant over the time (see for example Brugger et al., 2001).
In this paper our focus is on accidents caused by backcountry skiers keeping in mind that accidents due to backcountry skiing is by far the most common way to be involved in avalanche accidents. Until now there has not been an investigation for this special group of avalanche incidents in Austria. For our study we built a data base of fatal avalanche accidents recording the

- date
- municipal area where the accident took place
- federal state of the municipality
- number of persons involved
- number of fatalities
- type of activity (on/off-piste, backcountry skiing etc.)

of fatal accident events in Austria within the years 1980-2011 which is available from the annual reports of the Kuratorium fr alpine Sicherheit (1973–2011) and the annual reports of the information services of the federal states (see Amt der Tiroler Landesregierung, 1994–2010). For years before that time (back to 1968) we used aggregated information published in the annual reports of the Kuratorium fr alpine Sicherheit.

2 Methods

We propose the following model for capturing the:

$$\log(y_t) = f(t) + x_t$$

where y_t denotes the number of annual avalanche fatalities over time t. The logarithms of these count data are modeled as the sum of potentially nonlinear trend function $f(t)$ and a stationary remainder x_t. To account for potential serial correlation and periodic variation in the remainder, we consider autoregressive moving-average (ARMA) effects. In order to estimate the nonlinear function $f(t)$ we use the `R` package `mgcv` (see R Development Core Team, 2012; Wood, 2006).
Further on, for looking at the regional distribution of avalanche fatalities we build small area maps based on Austrian municipalities using the `GIS`-software `ArcMap`. We use Markov random field smoothing which helps us to identify regional hot spots of avalanche fatalities.

3 Results

In Figure 1, we give the plot of the nonlinear trend function of avalanche fatalities in Austria
If we look at the estimated function we take notice of an increasing trend (1970: approx 12 fatalities, 2010 approx 22 fatalities). Further on we point out that there is a lot of variation of the observed counts around the estimated function. Additionally, we notice a peak in the 1980s (1982-1988). We did not find any substantial serial correlation (contrary to the results of Pfeifer, Zeileis (2012)) or any sort of periodicity in the remainder x_t (see Höller, 2009; Tschirky et al., 2000).
Figure 2 gives the spatial distribution of avalanche fatalities within the years 1981 and 2011 based on Austrian municipalities. The coloring, however, is based on smoothed Markov random field estimates of avalanche fatalities (the number corresponding with each spatial unit in the plot is equal to the original count) .

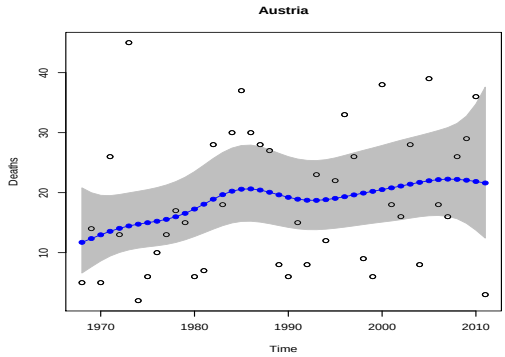


FIGURE 1. Observed and estimated annual avalanche fatalities in Austria within 1968–2011 including the 90% confidence band.

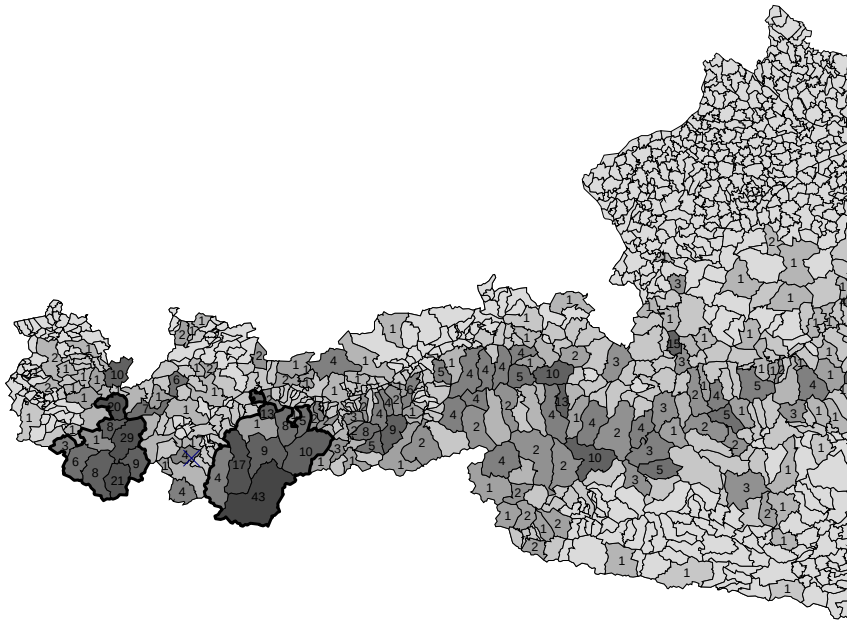


FIGURE 2. Regional distribution of avalanche fatalities in Austria within 1981–2010 including the observed numbers of fatalities within each spatial unit.

4 Conclusion

As the result of the trend analysis we notice an increasing trend of avalanche fatalities within the years 1968 and 2011. Additionally we take notice of a peak in the 1980s (an unusual period of increased snowfall/covering of snow in the 80's is discussed as a reason for the higher avalanche frequency).
As the result of the regional analysis we notice two hot spots of avalanche fatalities in Figure 2: 'St. Anton a. Arlberg (29)' (Arlberg-Silvretta) and 'Slden (43)' (southern part of tztal, Stubai-Khtai).
Because of the increasing trend and the rather 'narrow' regional distribution of the fatalities consequences on prevention of avalanche accidents are discussed.

References

Amt der Tiroler Landesregierung (1994–2010). *Schnee und Lawinen, Jahresberichte*. Innbruck.

Brugger, H., Durrer, B., Adler-Kastner, L., Falk, M., and Tschirky, F. (2001). Field management of avalanche victims. *Resuscitation* **51**, 7 – 15.

Höller, P. (2009). Avalanche cycles in Austria: an analysis of the major events in the last 50 year. *Natural Hazards*, **48**, 399 – 424.

Kuratorium fr alpine Sicherheit (1973–2011). *Sicherheit im Bergland, Jahrbcher des Kuratoriums fr alpine Sicherheit*. Innsbruck.

Pfeifer, C., and Rothart, V. (2004). On probabilities of avalanches triggered by alpine skiers. An application of models for counts with extra zeros. In: *Proceedings International Workshop of Statistical Modelling 2004*. Florence.

Pfeifer, C. (2010). On probabilities of avalanches triggered by alpine skiers. An empirically driven decision strategy for backcountry skiers based on these probabilities. In: *Proceedings International Workshop of Statistical Modelling 2010*. Glasgow.

Pfeifer, C. (2011). On probabilities of avalanches triggered by alpine skiers. Models with random effects. In: *Proceedings International Workshop of Statistical Modelling 2011*. Valencia.

Pfeifer, C., and Zeileis, A. (2012). Trend analysis of snow avalanche accidents in Tyrol within the years 1989-2010. In: *Proceedings International Workshop of Statistical Modelling 2012*. Prague.

R Development Core Team (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.*Vienna.

Tschirky, F., Brabec, B., and Kern, M. (2000). *Lawinenunfälle in den Schweizer Alpen – eine statistische Zusammenstellung mit den Schwerpunkten Verschüttung, Rettungsmethoden und Rettungsgeräte.* http://www.slf.ch/praevention/lawinenunfaelle/unfallstatistik-de.pdf

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R.* Boca Raton: Chapman and Hall/CRC.

Functional Data Analysis via Quasi U-Statistics Based Tests

Aluísio Pinheiro[1], Pranab Kumar Sen[2]

[1] University of Campinas, Brazil
[2] University of North Carolina, Chapel Hill, USA

E-mail for correspondence: apinheiro.unicamp@gmail.com

Abstract: In this paper we study the testing for the grouping of individuals in FDA. The response is a high-dimensional discrete observation of an underlying functional, and groups can be previously defined or data-driven. The approach is based in quasi U-statistics decomposition of wavelet diversity measures. The procedure is illustrated in movement temporal curves of *Gracilianus microtarsus* specimens. The numerical studies include a detailed analysis of the data set and the importance of the choice of the wavelet filter and thresholding policies.

Keywords: within-populations diversity measures, between-populations diversity measures, asymptotic normality, U-statistics, non-standard asymptotics

1 The Functional Model

Consider G groups of individuals for which iid discretized curves can be drawn. Each observation is composed of K time-point evaluations of a function of interest f such that

$$dX_{gl}(t) = f_g(t)dt + \epsilon dW_{gl}(t), \quad t \in [0,1], \tag{1}$$

where ϵ is the diffusion parameter, $f_g(t)$ is a deterministic unknown function and $\{W_{gl}(t) : t \in \mathbb{R}\}$ are either: independent standard Brownian motions (Abramowich et al., 2004; Abramowich and Angelini, 2006) - **case i.i.d.**, or independent CTARMA processes (Kist and Pinheiro, 2012) -**case dep.** , for $g = 1, \ldots, G$, $l = 1, \ldots, n_g$. We call $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}, \ldots, \mathbf{X}_{G1}, \ldots, \mathbf{X}_{Gn_G}$ the n_g K-dimensional observations in group $g = 1, \ldots, G$, respectively. Suppose f belongs to a convenient Besov space. We write the nonlinear wavelet estimators as $\hat{f}_{gl}(\cdot) = \sum_{j,k} \hat{c}^{thr}_{gl,jk}\psi_{jk}(\cdot)$. One can write the L_2 norm as $\|\hat{f}_{gl} - \hat{f}_{g\prime l\prime}\|^2_{L_2} = \sum_{j,k}(\hat{c}^{thr}_{gl,jk} - \hat{c}^{thr}_{g\prime l\prime,jk})^2$. The hypotheses of interest are

$$H_0 : f_1(\cdot) = \cdots = f_G(\cdot) \;\; vs \;\; H_1 : \exists 1 \leq g \neq g\prime \leq G : f_g(\cdot) \neq f_{g\prime}(\cdot).$$

Suppose: G (≥ 2) independent groups with respective distribution functions (K-dimensional or even infinite dimensional) and sample sizes $n_1, \ldots, n_G$,

so that all F's are defined on a common probability space; and a non-negative kernel function $\phi(\mathbf{x}, \mathbf{y})$ which can be expressed as a convex linear combination of one-dimensional convex kernels. Let

$$U_{n,g} = \binom{n_g}{2}^{-1} \sum_{1 \le i < j \le n_g} \phi(\mathbf{X}_{gi}, \mathbf{X}_{gj}), \quad g = 1, \ldots, G. \tag{2}$$

Note that the $U_{n,g}$ are U-statistics and, thus, they are unbiased estimators for $\delta(F_g) = E\phi(\mathbf{X}, \mathbf{Y})$, whenever $\mathbf{X}, \mathbf{Y}$ *i.d.* F_g. Similarly, let

$$U_{n,gg\prime} = \frac{1}{n_g n_{g\prime}} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g\prime}} \phi(\mathbf{X}_{gi}, \mathbf{X}_{g\prime j}), \quad 1 \le g < g\prime \le G. \tag{3}$$

two-sample U-statistics estimators for $\delta(F_g, F_{g\prime})$ so that $2\delta(F_g, F_{g\prime}) \ge \delta(F_g) + \delta(F_{g\prime})$. Take $n = n_1 + \cdots + n_G$ and the pooled sample U-statistics which can be decomposed as within- (W_n) and between-groups (B_n):

$$\begin{aligned} U_n &= \sum_{g=1}^{G} \frac{n_g}{n} U_{n,g} + \sum_{1 \le g < g\prime \le G} \frac{n_g n_{g\prime}}{n(n-1)} \{2U_{n,gg\prime} - U_{n,g} - U_{n,g\prime}\} \\ &= W_n + B_n, \end{aligned} \tag{4}$$

$W_n > 0$ *a.s.* under either H_0 or H_1. Under the latter, B_n may assume both positive and negative values, but $E(B_n) > 0$, and large positive values of B_n are statistically expected. Moreover, the classical Hoeffding non-degeneracy and CLT hold, i.e., $n^{1/2}(B_n - E(B_n))$ is asymptotically normal. On the other hand, under H_0, $E(B_n) = 0$ and the degeneracy on the kernel of B_n leads to a nonstandard asymptotic situation, as follows. We use the squared L_2 distance as an example, but the analogous decomposability is true for a larger class of kernel functionals. Let $\delta_{F_g} = E\int(dX_{g1} - dX_{g2})^2 = 2\epsilon^2$, $\delta_{F_g, F_{g\prime}} = E\int(dX_{g1} - dX_{g\prime 1})^2 = \int(f_g - f_{g\prime})^2 + 2\epsilon^2$, $g \ne g\prime = 1, \ldots, G$. We have $EB_n = 2\sum_{1 \le g < g\prime \le G} \int(f_g - f_{g\prime})^2$. Hence, under H_0, $EB_n = 0$ and, otherwise, $EB_n > 0$. B_n can be written as: $B_n = \sum_{i,j}^{1,n} \eta_{nij} \phi(\mathbf{X}_i, \mathbf{X}_j)$, where $\eta_{nij} = 1$ if i and j come from different groups, and $-(n - n_g)/((n_g - 1)$, if i and j are both from group g, for $1 \le g \le G$. The first term of Hoeffding's decomposition of ϕ is given by $\phi_{1,g}^{g\prime}(X_{g1}) = \int(f_g - f_{g\prime})^2 + \epsilon^2(1 + \int(dW_{g1})^2 + 2\epsilon\int(f_g - f_{g\prime})dW_g$, which means that $\phi_{1,g}^{g\prime}(X_{g1}) + \phi_{1,g\prime}^{g}(X_{g\prime 1}) - \phi_{1,g}^{g}(X_{g1}) - \phi_{1,g\prime}^{g\prime}(X_{g\prime 1}) = 2\int(f_g - f_{g\prime})^2 + 2\epsilon[\int(f_g - f_{g\prime})dW_g - dW_{g\prime}]$. Under H_0, the first order Hoeffding term is zero a.s., and is a nondegenerate r.v. otherwise (Pinheiro et al., 2009; 2011). We apply the sub-group decomposition for the estimators of the f_g's (Kist and Pinheiro; 2012), thus generating a quasi U-statistics for differences of f_g for which the asymptotic normality can be proven following Pinheiro et al. (2011), leading to nB_n being asymptotically normal. We then have a asymptotically normal test statistic under either H_0 and H_1, which is unusual for dissimilarity cases. However, the test statistic converges to a normal distribution under different rates: $n^{1/2}$ for H_1, and n for H_0.

2 Application

The procedure is illustrated in a data set composed by the movement curves during 12 hours taken every 10 seconds for 8 males and 7 females from the *Gracilianus microtarsus* species. We have $G = 2$ groups: females and males. We then estimate f_1 and f_2 as seen in Figure 1. The diversity measures are apportioned into their intra- and between-groups diversities components, and the quasi U-statistic B_n is computed, and compared to the critical value for testing H_0 vs H_1. The kernel was the squared L_2-distance between functionals.

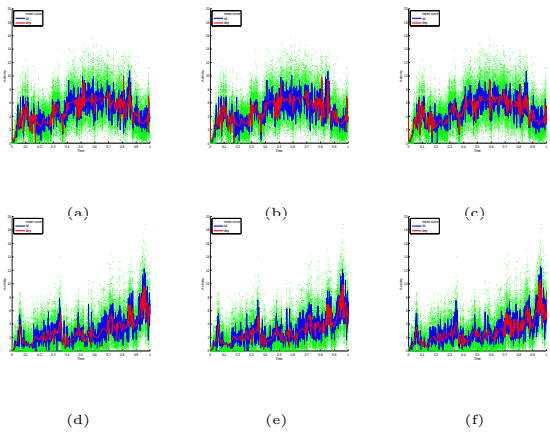


FIGURE 1. Regularized Wavelet Mean 12-hour Curve for the Movements of the Female ((a)-(c)) and Male (d)-(f)) *Gracilianus microtarsus*. The wavelet filter is the Symmlet8 ((a),(d)), Coiflets3 ((b),(e)), Daubechies6 ((c),(f)), and the regularizing parameters are $j_s = 5$ and $J_\eta = 9$.

Figure 1 shows three estimators for each gender, based on three different wavelet bases: Symmlets 8, Coiflets 3, and Daubechies 6. The data is shown in green. The independent FANOVA model estimator is shown in blue, and the dependent FANOVA model estimator is shown in red. One can notice the differences between $\hat{f}_1$ and $\hat{f}_2$. Moreover, the proposed 'dep' estimators are much more regularized then the previously available wavelet estimators. The inferential conclusion is that there are statistically significant differences between males and females in their temporal movement curves. The computed B_n bootstrap p-value is 0.0000 with 10000 bootstrap replications.

The numerical studies involves several choices. Besides the three wavelet bases, one could choose the dep-i.i.d. estimator, as well as the regularizing

parameters for the thresholding procedure. Finally, *robust* MAD or the standard deviation are employed for the estimates of the measure of the noise variability.
Standard deviation is in general superior to MAD, since the latter yields usually less regular estimated curves. The choice of the wavelet basis is quite unimportant. The use of anyone leads to the same inferential results. Some local characteristics of the estimated curves are highlighted or shadowed by each basis, but the test results are the same.
The choice of the smoothing parameters is by far the most important issue of all. The automatic procedures is not always successful, but in general values within a reasonable range, far from both 0 and the maximum level, yields the same inferential results.

3 Discussion

We present a novel test for functional data which can be easily computed for even large sets ($n \uparrow \infty$) of very finely sampled curves ($K \uparrow \infty$). Pinheiro et al. (2009; 2011) prove that this procedure provides one with good test statistics under mild regularity conditions on both the kernel function and the grouping (and/or sampling) characteristics. We embed our proposed test statistic within the quasi U-statistics framework, and illustrate that its use together with non-linear wavelet estimation yields very good results, which are usable for noisy data, and under a variety of wavelet bases, and thresholding policies.

References

Abramovich, F., Antoniadis, A., Sapatinas, T., and Vidakovic, B. (2004). Optimal testing in a fixed-effects functional analysis of variance model, *Int. J. Wavelets Multiresolut. Inf. Process.*, **2**(4), 323 - 349.

Abramowich, F. and Angelini, C. (2006). Testing in mixed-effects FANOVA models, *J. Statist. Plann. Inference*, 136 (12), 4326:4348.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution, *Ann. Math. Statist.*, 19(3), 293–325.

Kist, A. and Pinheiro, A. (2012). Wavelet Functional Data Models for Dependent Errors. *submitted for publication.*

Pinheiro, A., Sen, P.K., Pinheiro, H.P. (2009). Decomposability of high-dimensional diversity measures: quasi U-statistics, martingales and nonstandard asymptotics. *J. Multiv. Anal.*, 100, 1645-1656.

Pinheiro, A., Sen, P.K., Pinheiro, H.P. (2011). A class of asymptotically normal degenerate quasi U-statistics, *Ann. I. Stat. Math.*, 63, 1165–1182.

Boosting Multi State Models

Holger Reulen[1], Thomas Kneib[1]

[1] University of Goettingen, Germany

E-mail for correspondence: hreulen@uni-goettingen.de

Abstract: This abstract describes the ideas behind estimating multi state models using the functional gradient descent (FGD) boosting algorithm and is structured as follows: Multi state models are motivated in section 1, followed by a brief description of the FGD boosting algorithm in section 2, and a description of the integration of multi state models in FGD boosting in section 3.

Keywords: Multi state models; Functional gradient descent boosting.

1 Multi state models

The analysis of categorical quantities changing their state over the course of time is often needed in practice and therefore has a great tradition in statistical research. The furthest developed model class is present for single terminal event processes (called *duration* or *survival analysis*). Such a single terminal event model is the simplest type of multi state models and is characterized by a transition from the initial state at time origin to the absorbing state at random time T.
Competing risk models generalize single event models by analyzing the timing of various terminal events: as long as an individual is in its initial state, several risks are concurrently active, leading to transitions into the respective absorbing states. If one or more of the states is not absorbing, another single or competing risk setting emerges instantaneously with the actual transition into the non-absorbing state. This sequence of arising single or competing risk models is called *multi state model*.

2 Functional gradient descent boosting

The functional gradient descent (FGD) boosting algorithm, in its componentwise definition, is a convenient estimation approach and a well established answer to variable selection and model choice questions for regression analyses in a lot of different response settings, e.g. linear, generalized linear, structured (geo)additive, survival or quantile regression models (Kneib et al., 2009).

In general, boosting is a functional gradient descent method that seeks the solution of the optimization problem

$$\eta^* (\mathbf{x}) = \arg\min_{\eta(\mathbf{x})} \mathbb{E} \left(\rho \left(y, \eta \left(\mathbf{x} \right) \right) \right),$$

where $\rho (\cdot, \cdot)$ is a suitable loss function. Further on, y denotes the response and $\eta (\mathbf{x})$ is the linear predictor depending on a covariate vector $\mathbf{x}$, e.g.

$$\eta (\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} \mathrm{f}_j (x_j).$$

The single parts $\mathrm{f}_j (x_j)$ of $\eta (\mathbf{x})$ are refered to as model components and may be of any type that is known from structured additive regression models, e.g. linear terms $\beta_j x_j$, (geo)additive terms $\mathrm{f}(x_j)$ using penalized spline approaches, interaction terms $\beta_j x_k x_l$, varying coefficients $\mathrm{f}_j (x_k) \cdot \mathrm{I}_{\{x_l \neq \text{Ref.}\}}$ or effect surfaces $\mathrm{f}_j (x_k, x_l)$.

In practice the loss function $\rho (y, \eta)$ is replaced by the empirical risk

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(y_i, \eta \left(\mathbf{x}_i \right) \right).$$

We then aim to minimize this quantity iteratively with respect to η. Therefore, we fit base learners $\mathrm{g}_j (x_j)$ that correspond to the model components $\mathrm{f}_j (x_j)$ and reflect the desired properties, e.g. by penalized linear regression aproaches

$$\mathrm{g}_j (x_j) = \mathbf{X} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} + \lambda \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{z},$$

where $\mathbf{X}$ is a suitable design matrix, $\mathbf{K}$ is a penalty matrix with λ as corresponding smoothing parameter, and $\mathbf{z}$ denotes the working response, which is defined as the negative derivative of the loss function with respect to η and is recalculated in each of the $m = 1, \ldots, m_{\text{stop}}$ boosting iterations by plugging in the current linear predictor $\hat{\eta}^{[m]} (\mathbf{x}_i)$. This current linear predictor version $\hat{\eta}^{[m]} (\mathbf{x}_i)$ is formed by the step- and componentwise definition of the FGD boosting algorithm: In each boosting iteration, the best base learner, i.e. the one that leads to the largest decrease of the (empirical) loss, is selected and the respective parameter(s) is/are updated by a certain amount $\nu \in (0, 1]$ of the respective base learner parameter estimate:

$$\hat{\beta}_j^{[m+1]} = \hat{\beta}_j^{[m]} + \nu \cdot \hat{\mathrm{g}}_j (x_j), \qquad \hat{\eta}^{[m+1]} (\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_j^{[m+1]}.$$

This strategy is followed until any further iteration leads to only neglectible improvements of the likelihood (this *early stopping* is crucial for variable selection and several possible strategies can be applied to find the number m_{stop} of performed iterations).
For detailed description of the FGD boosting algorithm, see Buehlmann and Hothorn (2007).

3 Partial likelihood & working response in multi state models

3.1 Survival analysis

The construction of the loss function and the working response for FGD boosting in survival models is based on Cox's proportional hazards model formulation (Cox, 1972), i.e. Cox's log partial likelihood for observed event times t_i and censoring indicators D_i, $i = 1, \ldots, n$:

$$\log\left(\mathrm{PL}\left(\boldsymbol{\beta}|\mathbf{t}, \mathbf{D}, \mathbf{X}\right)\right) = \sum_{i=1}^{N} D_i \left[\eta\left(\mathbf{x}_i\right) - \log\left(\sum_{j=1}^{N} \mathrm{I}_{\{t_j \geq t_i\}} \cdot \exp\left(\eta\left(\mathbf{x}_j\right)\right)\right)\right],$$

The working response for Cox's proportional hazards regression model is then the partial first derivative of the log partial likelihood with respect to $\eta\left(\mathbf{x}_i\right) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, evaluated at $\eta\left(\mathbf{x}_i\right) = \hat{\eta}^{[m]}\left(\mathbf{x}_i\right)$:

$$z_i^{[m]} = D_i - \sum_{j=1}^{N} D_j \mathrm{I}_{\{t_i \geq t_j\}} \cdot \frac{\exp\left(\hat{\eta}^{[m]}\left(\mathbf{x}_i\right)\right)}{\sum\limits_{k=1}^{N} \mathrm{I}_{\{t_k \geq t_j\}} \exp\left(\hat{\eta}^{[m]}\left(\mathbf{x}_k\right)\right)}.$$

3.2 Multi state models

A Cox type log partial likelihood for general multi state models with transitions $q = 1, \ldots, Q$ is given in Andersen et al. (1993) by

$$\log\left(\mathrm{PL}\left(\boldsymbol{\beta}|\mathbf{t}, \mathbf{D}, \mathbf{X}\right)\right) = \sum_{i=1}^{n} \sum_{q=1}^{Q} \left[\int_0^{\infty} \eta\left(\mathbf{X}_{qi}\left(t\right)\right) \mathrm{dN}_{qi}\left(t\right) - \right.$$
$$\left. - \frac{1}{n} \int_0^{\infty} \log\left(\sum_{i=1}^{n} \mathrm{Y}_{qi}\left(t\right) \exp\left(\eta\left(\mathbf{X}_{qi}\left(t\right)\right)\right)\right) \mathrm{dN}_q\left(t\right)\right].$$

Here, $\mathrm{N}_{qi}\left(t\right)$, $\mathrm{N}_q\left(t\right)$ and $\mathbf{X}_{qi}$ denote transition (and individual event history specific) counting processes and design matrices, respectively.
We regard all observed transitions $i = 1, \ldots, n$ individually as if each of them was originated by one unique individual i, with observed transition q_i at time t_i and covariate observation $\mathbf{x}_i$. The $\mathbf{x}_i$ are constructed by merging products of transition indicators D_{qi}, $q = 1, \ldots, Q$, and covariate observations $x_{i1}, \ldots, x_{ip}$. Since the integral of a function with respect to a counting process is a convenient notation for a sum with a finite number of terms, we may furthermore rewrite the log partial likelihood as sum of the function evaluated at the jumping times

$$\log\left(\mathrm{PL}\left(\boldsymbol{\beta}|\mathbf{t}, \mathbf{D}, \mathbf{X}\right)\right) = \sum_{q=1}^{Q} \sum_{i=1}^{n} D_{qi} \left[\eta\left(\mathbf{x}_i\right) - \log\left(\sum_{j=1}^{n} Y_{qj}\left(t_i\right) \exp\left(\eta\left(\mathbf{x}_j\right)\right)\right)\right].$$

This likelihood is a pretty straightforward generalization of the single event log partial likelihood formulation, with key changes by summing up potentially multiple events of type q and more general risk sets that are no longer strictly decreasing with time.

Calculating the first-order partial derivative with respect to $\eta(\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, evaluated at $\eta(\mathbf{x}_i) = \hat{\eta}^{[m]}(\mathbf{x}_i)$, $i = 1, \ldots, n$, leads to the working response

$$z_i^{[m]} = \sum_{q=1}^{Q} \left[D_{qi} - \sum_{j=1}^{n} \left(D_{qj} \cdot Y_{qi}(t_j) \cdot \frac{\exp\left(\hat{\eta}^{[m]}(\mathbf{x}_i)\right)}{\sum_{k=1}^{n} Y_{qk}(t_j) \exp\left(\hat{\eta}^{[m]}(\mathbf{x}_k)\right)} \right) \right].$$

The indicator product $D_{qj} \cdot Y_{qi}(t_j)$ takes the value one if the left states of i and j are equal and j transitions into the corresponding right state of transition q. Since every observation is composed of only one unique transition q_i, we can rewrite the working response as:

$$z_i^{[m]} = 1 - \sum_{q=1}^{Q} \left[\sum_{j=1}^{n} \left(D_{qj} \cdot Y_{qi}(t_j) \cdot \frac{\exp\left(\hat{\eta}^{[m]}(\mathbf{x}_i)\right)}{\sum_{k=1}^{n} Y_{qk}(t_j) \exp\left(\hat{\eta}^{[m]}(\mathbf{x}_k)\right)} \right) \right].$$

References

Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993) *Statistical Models Based on Counting Processes.* Springer Series in Statistics, Springer-Verlag New York.

Buehlmann, P. and Hothorn, T. (2007) *Boosting Algorithms: Regularization, Prediction and Model Fitting (with Discussion).* Statistical Science, 22(4), p. 477-505.

Cox, D.R. (1972) *Regression Models and Life-Tables.* Journal of the Royal Statistical Society. Series B (Methodological) Vol. 34, No. 2, p. 187-220.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2012) *mboost: Model-Based Boosting, R package version 2.2-0.* http://CRAN.R-project.org/package=mboost..

Kneib, T., Hothorn, T., and Tutz, G. (2009) *Variable Selection and Model Choice in Geoadditive Regression Models.* BIOMETRICS 65, p. 626-634.

Ridgeway, G. (1999) *The state of boosting.* Computing Science and Statistics 31, p. 172-181.

Efficient ungrouping of coarse histograms with the penalized composite link model

Silvia Rizzi[1], Jutta Gampe[1], Paul H.C. Eilers[2]

[1] Max Planck Institute for Demographic Research, Rostock, Germany
[2] Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: rizzi@demogr.mpg.de

Abstract: Data grouped in relatively wide histogram bins are frequently encountered in practice. Here, in particular, grouped data by age are considered. We propose a penalized composite link model to estimate smooth densities by single years of age when data are provided in coarse age classes only. The performance of the model is examined in two applications: the first one illustrates mortality trajectories by age and the second studies simulated patterns of age at first marriage.

Keywords: grouped data; penalized composite link model; smoothing

1 Introduction

Grouped data are common. In demographic and epidemiological contexts age distributions are habitually provided in 5-year age intervals, often with a wide open age class at the highest ages. Typical examples are so called abridged life tables, as provided by the WHO or EUROSTAT. Figure 1 shows an example taken from the Human Mortality Database.
However, often more detailed information is needed when age-specific patterns are to be compared across time or across countries. In particular, wide age groups for the elderly can pose an obstacle. Therefore efficient methods of ungrouping coarse histograms are needed.
Here we present an approach based on the penalized composite link model (Eilers, 2007). This model implements the idea that the observed counts are indirect observations of a latent sequence that represents the true distribution. This distribution has to be estimated from the composite observed data. The only assumption made about the true distribution is smoothness, which is enforced by a penalty.
We will demonstrate the efficiency of the approach by two applications where we can assess its performance. The first example is an age at death distribution, the second example studies the age-specific incidence of first marriages.

2 Penalized composite link model for ungrouping

We denote the actually observed counts in the I bins by $y = (y_1, ..., y_I)'$. The y_i are realizations from Poisson distributions with $E(y_i) = \mu_i$. However, the vector $\mu \in \mathbb{R}^I$ results from grouping the $J > I$ original (expected) counts $\gamma = (\gamma_1, ..., \gamma_J)'$ into the histogram bins: $\mu = C\gamma$ with

$$C = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & \dots & & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 \\ & \vdots & & & \ddots & & & \vdots & \\ 0 & \dots & & & \dots & 0 & 1 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{I \times J}.$$

C is a matrix that 'composes' μ from γ, and it describes how the latent distribution γ was mixed before generating the data. The number of '1' in each row of C corresponds to the number of elements of γ that are grouped into one bin. The aim of ungrouping a histogram is to estimate the original distribution, i.e., the vector γ. As $I < J$, this problem is ill-defined and additional constraints need to be imposed. We assume that the latent distribution is smooth and implement this assumption by a difference penalty on the elements of γ. This is the penalized composite link model (PCLM).

To guarantee non-negative values of γ we write $\gamma = \exp(X\beta)$. Here X is the identity matrix, and smoothness of γ results if β is smooth. If the number of elements in γ is large, we can replace the design matrix X by a B-spline basis; again a smooth coefficient vector β produces a smooth distribution γ. Eilers (2007) showed that the PCLM can be estimated by an appropriately modified version of the interatively reweighted least squares (IRWLS) algorithm. The system of equations becomes

$$(\breve{X}'\tilde{W}\breve{X} + \lambda D_d' D_d)\beta = \breve{X}'\tilde{W}[\tilde{W}^{-1}(y - \tilde{\mu}) + \breve{X}\tilde{\beta}]. \tag{1}$$

where D_d is the matrix that computes the dth differences of the components of β. In the practical applications we chose a difference order of $d = 2$. The matrix $\breve{X}$ has elements $\breve{x}_{ik} = \sum_j c_{ij} x_{jk} \gamma_j / \mu_i$, and can be considered a 'working X' in the IRWLS algorithm. For a given value of λ the system can be easily solved.

The parameter λ modulates the smoothness of β, as forced by the difference penalty, against the fit to the observed data. We choose the value of λ that minimizes the Akaike's Information Criterion (AIC).

3 Applications

We now study the performance of the approach in two examples where we can analyze both the original data before grouping and the PCLM-ungrouped distribution. This allows us to compare how well the proposed method works.

3.1 Age at death distribution

We first apply the PCLM to mortality data. We consider period life table death counts by age in Sweden for the year 2011. The data are taken from the Human Mortality Database (HMD; available at www.mortality.org), and they are available by single years of age from 0 to 109, and then a final group 110+.

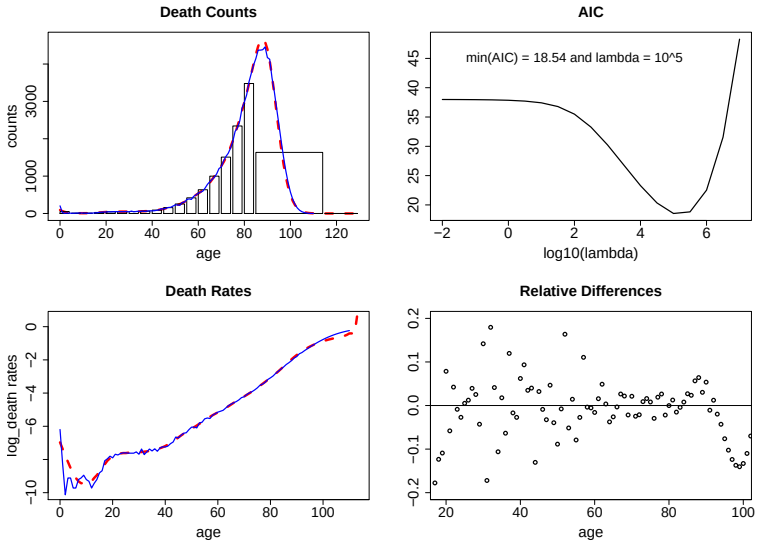


FIGURE 1. Top left: Age at death distribution, Sweden in 2011. Original data by single years of age (blue solid line), grouped counts and PCLM estimates (red dashed line). Top right: AIC over $\log_{10} \lambda$. Bottom left: Death rates (on log scale) from empirical data (blue solid line) and from PCLM estimates (red dashed line). Bottom right: Relative differences between original death counts by single year of age and PCLM estimates.

We artificially grouped the death counts in 5-year age classes plus an open-ended age interval starting at 85. This age interval of 85+ years has been common in many epidemiological databases, at least in the past, which is why we chose this particular grouping. Furthermore, we added an extra bin from 115 to 120 with 0 counts. This was done after realizing that without this information the PCLM approach had a tendency to 'stretch out' too far into the right-hand tail. The assumption of practically no deaths after age 115 can safely be made.

Figure 1, top left, shows the age at death distribution derived directly from the HMD as well as the ungrouped PCLM estimate. The concordance is striking. Additionally, we calculated the corresponding hazard (death rates) from the density because very often this characteristic of the distribution is of particular interest. The result is shown in Figure 1, bottom left.

3.2 Age at first marriage

To check whether the method is able to reproduce hazards with a completely different shape equally well, we considered the distribution of ages at first marriage. These data are usually not available by single years of age. However, there is a widely used model for this age distribution (Coale and McNeil, 1972), which has been shown to match the distribution in human populations very closely.
Using the Coale-McNeil model we simulated $s = 2000$ ages at first marriage, rounded to full years. These data were grouped in 5-year age classes from which the original distribution was estimated by the PCLM.

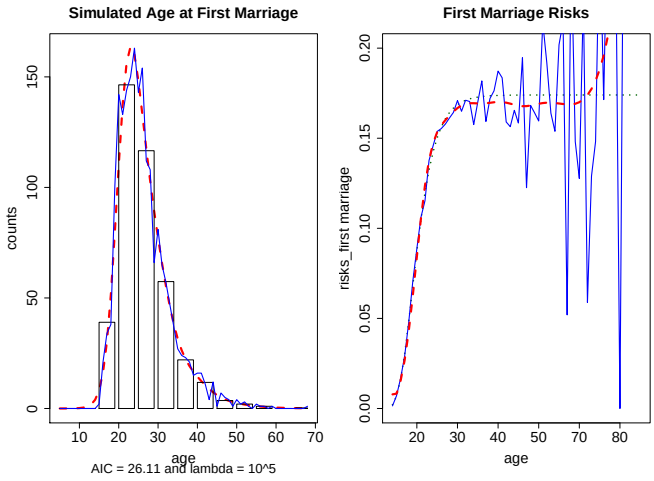


FIGURE 2. Left: Simulated ages at first marriage. Single-year counts (blue solid line), histogram and PCLM estimates (red dashed line). Right: Age-specific first marriage rates. Theoretical rate from Coale-McNeil model (green dotted line), empirical rates by single ages (blue solid line) and hazard derived from the PCLM estimates (red dashed line), average over 50 repetitions.

Again one bin with 0 counts was added, this time at the left end of the distribution, i.e., for ages younger than 15. Like in the mortality example this additional empty bin considerably improved the PCLM estimate. The result can be seen in Figure 2 and again the model performs very well.

References

Coale, A.J. and McNeil, D.R. (1972). The Distribution by Age of the Frequency of First Marriage in a Female Cohort. *Journal of the American Statistical Association*, **67**, 743 – 749.

Eilers, P.H.C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, **7**, 239 – 254.

A hyper-Poisson regression model for zero-truncated count data

Antonio José Sáez-Castillo[1], Antonio Conde-Sánchez[1], Ana Maria Martínez-Rodríguez[1], José Olmo-Jiménez, M.J.[1], José Rodríguez-Avi[1]

[1] Department of Statistics and OR, University of Jaén, Spain

E-mail for correspondence: ammartin@ujaen.es

Abstract: A regression model for zero-truncated count data based on the hyper-Poisson distribution is developed. This regression model generalizes the zero-truncated Poisson regression model. The hyper-Poisson regression model introduces the regressors in the equation of the mean and additionally, regressors can also be introduced in the equation of the dispersion parameter in order to model under- and over- dispersion.

Keywords: Regression model; Zero-truncated count data; Hyper-Poisson distribution; Overdispersion; Underdispersion.

1 Introduction

In count data modelling there are situations for which the response variable cannot obtain a value of zero. A typical example is the length of hospital stay, in days. In the literature common models for this situation are zero-truncated Poisson model and zero-truncated negative binomial model. The zero-truncated Poisson model displays underdispersion (Winkelmann, 2008; Cameron and Trivedi, 1998), when overdispersion is present, the zero-truncated negative binomial model is used (Hilbe, 2011).
In order to cope with the presence of over- and under- dispersion Sáez-Castillo and Conde-Sánchez (2013) introduced a regression model for count data in which the Crow-Bardwell or hyper-Poisson distribution (Johnson et al., 2005) is used for the response variable. The main advantage of the hyper-Poisson regression model is that the regressors may be introduced in the mean at the same time that they can influence the over- or under-dispersed behavior of the distribution. This allows us to model data which present both, overdispersion and underdispersion, in different levels of the observations, determined by different combinations of the values of the covariates. In this paper a zero-truncated regression model based in the hyper-Poisson distribution is developed.

2 The hyper-Poisson distribution

The hyper-Poisson distribution, from now in, hP, (Johnson et al., 2005) has probability mass function (p.m.f.)

$$f_y = \frac{1}{{}_1F_1\left(1;\gamma;\lambda\right)}\frac{\lambda^y}{\left(\gamma\right)_y}, \quad y = 0,\ 1,\ 2, ...,$$

where $\gamma, \lambda > 0$, $\left(a\right)_r = a\left(a+1\right)...\left(a+r-1\right) = \frac{\Gamma(a+r)}{\Gamma(a)}$ for $a > 0$ and r a positive integer, and

$${}_1F_1\left(a;c;z\right) = \sum_{r=0}^{\infty}\frac{\left(a\right)_r}{\left(c\right)_r}\frac{z^r}{r!},$$

is the confluent hypergeometric series (Johnson et al. 2005). The main characteristic of this distribution is that it is overdispersed if $\gamma > 1$, is the Poisson distribution if $\gamma = 1$ and is underdispersed if $\gamma < 1$; that is because γ is interpreted as a dispersion parameter.
It can be proved (Sáez-Castillo and Conde-Sánchez, 2013) that the mean (μ) is given by

$$\mu = \lambda - \left(\gamma - 1\right)\left(1 - f_0\right) = \lambda - \left(\gamma - 1\right)\frac{{}_1F_1\left(1;\gamma;\lambda\right) - 1}{{}_1F_1\left(1;\gamma;\lambda\right)}.$$

It is clear that when $\gamma = 1$ (the Poisson distribution case), the mean matches up with λ, but, in general, λ is not always near the mean.
When $\gamma > 1$ the hP distribution may be seen as a Poisson compound distribution with a confluent hypergeometric distribution which, in general, is given by a density function

$$f\left(p\right) = \frac{p^{u-1}\left(1-p\right)^{v-1}e^{-\lambda p}}{Beta\left(u,v\right){}_1F_1\left(u;u+v;-\lambda\right)}.$$

Thus, in this overdispersion context, the hP distribution may be compared with other compound Poisson distributions, such as the negative binomial distribution.

3 Zero-truncated hyper-Poisson distribution

In this section the main properties of the zero-truncated hP distribution are developed. First, the p.m.f. is given by

$$f_y = \frac{1}{{}_1F_1\left(1;\gamma;\lambda\right) - 1}\frac{\lambda^y}{\left(\gamma\right)_y}, \quad y = 1,\ 2, ..., \tag{1}$$

where $\gamma, \lambda > 0$.

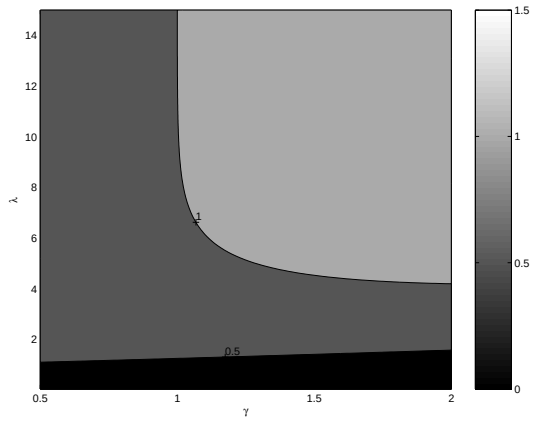


FIGURE 1. Contour plots for the variance-mean ration in terms of λ and γ.

It can be verified that when $\gamma = 1$ it is the zero-truncated Poisson distribution.
It is easy to prove that (1) verifies the recurrence equation

$$(\gamma + y)\, f_{y+1} = \lambda f_y \quad y = 1,\, 2, ...,\, . \tag{2}$$

An expression of the mean can be obtained if we sum (2) from $y = 1$ to ∞, obtaining

$$\mu = \lambda + f_1 - (\gamma - 1)\,(1 - f_1)\,. \tag{3}$$

From (1), multiplying both members by $y + 1$ and adding from $y = 1$ to ∞, a relation between mean and variance, σ^2, can be deduced,

$$\sigma^2 = \lambda\,(\mu + 1) + f_1 - (\gamma - 1)\,(\mu - f_1) - \mu^2.$$

The probability generating function (p.g.f.) is given by

$$g\,(z) = \frac{{}_1F_1\,(1; \gamma; \lambda z) - 1}{{}_1F_1\,(1; \gamma; \lambda) - 1}.$$

An alternative expression of the mean may be obtained from the derivative of the probability generating function evaluated in one, given by

$$\mu = g'(1) = \frac{\lambda}{\gamma}\frac{{}_1F_1\,(2; \gamma + 1; \lambda)}{{}_1F_1\,(1; \gamma; \lambda) - 1}.$$

In order to display the dispersion properties of the zero-truncated hP distribution, the variance-mean ratio has been analyzed in terms of λ and γ. Thus, Figure 1 shows a contour plot for the variance-mean ratio for a wide range of values of λ and γ.

4 The zero truncated hyper-Poisson regression model

Let be y_i the value of the response variable of the $i-th$ individual of the sample and $\mathbf{x}_i^{\mathrm{T}} = (1,\ x_{i1},\ x_{i2}, ...,\ x_{ik})$ the observed covariates in this $i-th$ individual. Let us consider that y_i follows a zero-truncated hP distribution with mean

$$\mu_i = e^{\mathbf{x}_i^{\mathrm{T}}\beta}$$

Two possible models have being considered depending on the parameters that are estimated: on one hand, the parameter λ will be determined by (3) from the values of μ_i and γ; from now on, we will name this model zero-truncated $hP(\lambda_i, \gamma)$. On the other hand, to obtain μ_i and λ estimates and γ will be determined by (3). In this case it is obtained a zero-truncated $hP(\lambda, \gamma_i)$ model.

The estimation of the regression coefficients β is carried out maximizing the log-likelihood function. If we have a sample $y_1,\ ...,\ y_n$, it is given by

$$\log L(\gamma, \lambda) = -\sum_{i=1}^{n} \log \Gamma(\gamma + y_i) + \log(\lambda) n\bar{y} + n\left(\log(\gamma) - \log\left({}_1F_1(1; \gamma; \lambda) - 1\right)\right).$$

It is very important to highlight that this function depends on γ and λ, while we are modelling μ, so we must replace the parameter λ or the parameter γ with its expression in terms of the mean μ, which can be deduced from (3), in each step of the optimization process. The problem is that there is not a closed expression of λ in terms of μ and γ nor of γ en terms of μ and λ, but we have overcome this difficulty solving the resulting equation by numerical methods in each evaluation of the log-likelihood function within the optimization process.

This model has been applied to real data for illustrative purposes.

References

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Hilbe, J. (2011) *Negative Binomial Regression*. Cambridge University Press.

Johnson, N. L., Kotz, S., Kemp, A. W (2005). *Univariate Discrete Distributions*. Third edition. New York: Wiley.

Sáez-Castillo, A.J. and Conde-Sánchez, A. (2013). A hyper-Poisson regression model for overdispersed and underdispersed count data *Computational Statistics and Data Analysis*, **61**, 148 – 157.

Winkelmann, R. (2008) *Econometric analysis of count data*. Springer-Verlag.

Predicting future offending in adolescents from a longitudinal survey with missing responses

Stuart J. Sharples[1], Deborah A. Costain[1], Chris Sherlock[1]

[1] Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom

E-mail for correspondence: s.sharples@lancaster.ac.uk

Abstract: It is common in longitudinal surveys for a number of the participants to dropout or simply omit responses upon follow up. The Edinburgh Study of Youth Transitions and Crime is a longitudinal survey of 4,300 young people. Whilst both attrition and omission at each sweep are small, the impact of this on the number of complete-cases is large. By multiply imputing the missing data, standard analysis methods can then be used in a way that reflects the uncertainty arising from the imputation process. A sensitivity analysis is performed by building models to predict joyriding on both the complete-case data and the multiply imputed data. The results show that an analysis of only the complete-case data would be misleading and that there are major differences in key parameter estimates as well as the variables selected during model selection.

Keywords: missing data; multiple imputation; model building.

1 Introduction

The Edinburgh Study of Youth Transitions and Crime (ESYTC) is a longitudinal cohort study of 4,300 young people concerned with personal changes during adolescence that explain why some, among all who engage in delinquent behaviour go on to become serious offenders and why others do not (Smith and McVie, 2003). The ESYTC commenced study in 1998, targeting all adolescents who were aged between $11\frac{1}{2}$ and $12\frac{1}{2}$ and thus due to start secondary school. Information was collected via self-completion questionnaires. Of particular interest are the delinquent behaviour items that were asked at every sweep. In total there are 15 items, ranging from fare dodging, being noisy in public to carrying a weapon and joyriding. Each delinquency item consisted of an initially "yes/no" response as to whether or not the participant had engaged in the behaviour, and positive responses were followed by a question on the volume (frequency of occurrence) of the behaviour. The initial sweep collected information regarding behaviour over the participants lifetime up to this point. Subsequent sweeps occurred at 12-

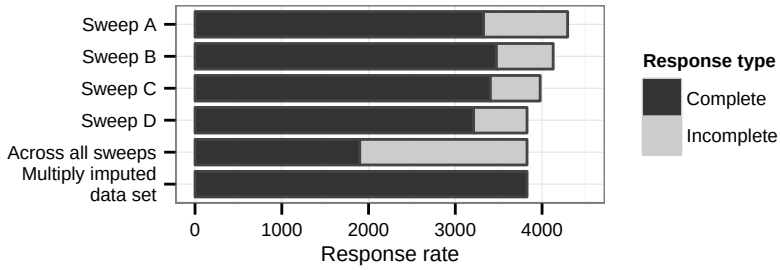


FIGURE 1. Edinburgh Study response rates at each sweep, how many of those where complete-cases, and the size of the multiply imputed data set.

monthly intervals, and were concerned with acts and changes in behaviour since the last sweep.

Shown in Figure 1 are the response rates to the individual sweeps, as well as the survey as a whole and the size of the multiply imputed data. While the attrition at each sweep is small, only 50% of the respondents who respond to all four sweeps have complete-cases and a further 49% have less than 5% missing. The priority is to multiply impute responses for analysis.

From the perspective of the individual items all but one item (gender) had missing responses. The largest amount of missing data was found to be 4.8%, but 90% of the items had less than 1% missing.

2 Model-based multiple imputation

A complete-case analysis on incomplete data makes the Missing Completely At Random (MCAR) assumption (Rubin, 1976), that those with incomplete observations do not differ in any systematic way from those with complete observations. This is often unrealistic, and if it is incorrect then parameter estimates can be biased. A more realistic assumption, Missing At Random (MAR) (Rubin, 1976), assumes that the omission of responses depends to some extent on the observed data. By taking a modelling approach to imputation it is possible to assume MAR, and by creating multiple imputed data sets the uncertainty about the imputed values themselves can be maintained. The method used to multiply impute is the Fully Conditional Specification (FCS) of Van Buuren et al. (2006) which imputes multivariate missing data on a variable-by-variable basis. FCS requires the specification of an 'imputation model' for each incomplete variable which, in part, then maintains the relationships that will be examined in the 'analysis model', as well factors predictive of non-response. In this respect, three delinquency items were associated with future non-response: the volume responses for fighting and skiving, and the binary response for stealing from home.

An important aspect of the questionnaire that needed to be taken into account was whilst some participants simply did not respond for a given delinquency item others responded "yes" but omitted how many times.

For example, at Sweep A, 21 participants failed to provide any information regarding "have you ever stolen from home?", while 18 answered "yes" but failed to specify a volume. Missing volume response should therefore be imputed conditional on the binary response being observed. By using the FCS approach, it was possible to implement this through a custom imputation function for the delinquency volume responses.
Ten imputed data sets were created in R 2.15.2 using the mice 2.13 package (Van Buuren and Groothuis-Oudshoorn, 2011); the algorithm was ran for 20 iterations. In order to assess convergence, the associations between the delinquency items and the background characteristics were monitored. Convergence was satisfactory with statistics from each of the ten data sets being equally varied and freely mixed.

3 Predicting future joyriding: a sensitivity analysis

Compared to the other delinquent acts in the ESYTC, joyriding is one of the few that is also a criminal offence. Though it has a very low prevalence (2% at sweep A), it increases over time (8% at sweep D). It is therefore of interest to identify factors that are associated, positively or negatively, with engaging in this behaviour. To predict future occurrences of joyriding, the binary responses from sweeps B, C and D were stacked and a logistic transition model fitted with the explanatory variables being the previous sweep's background and delinquent behaviour responses. Since each participant has three sets of responses, previous joyriding was automatically included in order to control for this lack of independence in the data.
In order to identify a parsimonious model, a two stage routine was devised. Starting with a saturated model, all binary delinquency items other than joyriding and background characteristics were considered for removal via a backwards elimination procedure using the pooled Wald test. The second stage then tested if the model could be improved by swapping the remaining binary delinquency items for their volume counterpart. This routine was applied to the complete-case data and the final results recorded (M1).
In order to apply this routine to the ten imputed data sets a strategy suggested by Wood et al. (2008) was used. The routine ran on each of the imputed data sets, and the items included in each were recorded. Those that appeared in more than half the models were then used to create a so called *supermodel*. This supermodel was then fitted to each of the imputed data sets (M2), with parameter estimates and standard errors being pooled using Rubin's rules (Rubin, 1987). Finally, the pooled Wald test was used to determine if any items could be dropped. A model based upon the covariate structure of M2 was fitted to the CC data (M3) in order to highlight any differences in parameter estimates.

3.1 Results

All the items selected by M1 were also selected by M2. In addition, M2 contained four more binary delinquency items and three more background characteristics, one of which was gender. From the covariates included in M2, only arson (7/10) and noisy in public (7/10) were not selected by all ten models. Due to limited space the parameter estimates for M2 and M3 have not been included here but will be shown on the poster. The parameter estimates show that the baseline odds ratio of an individual joyriding is very low, however M2 suggests a higher rate than M3. Another key difference is the effect of current joyriding, M2 suggest that the link between current and future joyriding incidences is slightly weaker than that suggested by M3. An association with the participant not living with either of their birth parents (reference category: living with birth parents) and joyriding was found in M3 but not M2, and therefore maybe spurious. The other major difference is the obvious increase in confidence in estimates for M2.

4 Discussion

When a longitudinal study contains missing data, analyses assuming MCAR such as analyses performed on only CC data should be avoided. In the case of the ESYTC, despite the amount of missing data being small, a CC analysis would only consider half of the participants. A practical alternative is to multiply impute missing responses and make the more plausible assumption of MAR. FCS, a model-based approach to imputation, was used which maintained the relationships that were later examined, and included predictive factors of non-response. By FCS requiring an 'imputation model' to be specified for each incomplete variable, it was possible to accommodate the difference between providing no information regarding a delinquent behaviour and only omitting a volume response. The sensitivity analysis showed that the CC model (M1) selection process was inferior to that performed on the MI data (M2), with a difference of several delinquency and background items being missed. In addition, fitting a model with the same covariate structure as M2 to the CC data (M3) showed key differences in parameter estimates, namely M3 underestimates the rate of joyriding compared to M2, slightly overstated the link between current and future joyriding, and potentially has a spurious association with who the participant lives with in terms of parents.

References

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Smith, D and McVie, S (2003). Theory and method in the Edinburgh study of youth transitions and crime. *British Journal Of Criminology*, **1**, 169 – 195

Van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman & Hall.

Comparison of estimation methods for variance components in elliptical mixed effects models

Danilo A. Silva[1], Cibele M. Russo[1]

[1] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brazil

E-mail for correspondence: danilo@icmc.usp.br

Abstract: Mixed effects models are usually applicable to problems with longitudinal data, repeated measures and other correlated data situations. Two commonly considered approaches for estimation in this theme are the hierarchical and marginal model, and the most common assumption for the distribution of random effects and errors is the normality. Our proposal is to compare constrained and unconstrained estimation methods for variance components in elliptical mixed effects models. The proposed methods are applied to a rat growth data set and a brief simulation study is presented.

Keywords: Constrained estimation; maximum likelihood estimation; variance components; mixed effects model.

1 Introduction

Mixed effects models are useful tools for analyzing correlated data. Although Verbeke and Lesaffre (1997) showed that the distribution of random effects does not have much influence in the parameters estimates, Osorio et al. (2007) concluded that assuming elliptical distributions for the random effects and errors may provide more robust estimates against outlying observations in linear and nonlinear mixed models. For nonlinear mixed elliptical models, Russo et al. (2012) also showed that score-type tests for variance components may be slightly more sensitive to perturbations under normality. However, no studies have been developed about estimating methods for variance components under the elliptical assumption.

2 The model

A linear mixed effects models for a vector of response variable $\mathbf{Y}_i$ $(m \times 1)$ may be written as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are matrices of known constants, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta)^\top$ a vector of unknown parameters and $\mathbf{b}_i = (b_{i1}, \ldots, b_{ir})^\top$ the random-effects coefficients. It is usual to assume $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ to be independent and to follow a normal distribution. Osorio et al. (2007) extended this model by considering a multivariate elliptical distribution for $(\mathbf{Y}_i, \mathbf{b}_i)$, such that

$$\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{b}_i \end{bmatrix} \sim \mathrm{El}_{m_i+r}\left\{ \begin{pmatrix} \mathbf{X}_i\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}; \begin{bmatrix} \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \sigma^2\mathbf{I}_{m_i} & \mathbf{Z}_i\mathbf{D} \\ \mathbf{D}\mathbf{Z}_i^T & \mathbf{D} \end{bmatrix} \right\}. \tag{2}$$

The matrices $\boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \sigma^2\mathbf{I}_{m_i}$, $\mathbf{D}$, and $\mathbf{Z}_i\mathbf{D}$ are proportional to the variance-covariance matrices $\mathrm{Var}(\mathbf{Y}_i)$, $\mathrm{Var}(\mathbf{b}_i)$ and $\mathrm{Cov}(\mathbf{Y}_i, \mathbf{b}_i)$. In this paper we assume $\mathbf{D}$ is unstructured and $\boldsymbol{\tau}$ are the vector of elements of $\mathbf{D}$.
The notation $\mathbf{W} \sim \mathrm{El}_m$ indicates that a vector $\mathbf{W}$ $(m \times 1)$ is elliptically distributed with mean $\boldsymbol{\mu} \in \mathbb{R}^m$ and scale matrix $\boldsymbol{\Sigma}$ (positive definite), with p. d. f. given by $f(\mathbf{w}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} g[(\mathbf{Y}-\boldsymbol{\mu}_W)^T {\boldsymbol{\Sigma}_W}^{-1}(\mathbf{Y}-\boldsymbol{\mu}_W)]$, where $g : \mathbb{R} \to [0, \infty)$, continuous and twice differentiable, is known as density generating function, with $\int_0^\infty u^{\frac{m}{2}-1} g(u)du < \infty$. Examples of elliptical distributions are the normal and Student-t distributions which can be obtained when $g(u) = (2\pi)^{-m/2}\exp(-u/2)$ and $g(u) = q\,(1 + u/\nu)^{-(\nu+m)/2}$, respectively, with q a constant and $u = (\mathbf{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu})$. It is usual to work with the quantities $W_g(u_i)$ and $W_g'(u_i)$ such that $W_g(u_i) = [d\log g(u_i)]/du_i$, $W_g'(u_i) = [dW_g(u_i)]/du_i$ and $u_i = [(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})]$ (see, for instance, Osorio et al., 2007).
The marginal model is given by $\mathbf{Y}_i \sim \mathrm{El}_{m_i}(X_i\boldsymbol{\beta}; \boldsymbol{\Sigma}_i)$, which preserves the mean of the hierarchical model without requiring numerical integration.

3 Unconstrained estimation method

An example of an unconstrained estimation method would be the maximum likelihood estimation of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})^\top = (\boldsymbol{\beta}, \sigma^2, \tau_{11}, \tau_{12}, \tau_{22})^\top$. One possible algorithm would be the Fisher scoring method, given by

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} + \mathbf{K}_{\beta\beta}^{-1(m)}\mathbf{U}_\beta^{(m)},$$

$$\widehat{\boldsymbol{\gamma}}^{(m+1)} = \widehat{\boldsymbol{\gamma}}^{(m)} + \mathbf{K}_{\gamma\gamma}^{-1(m)}\mathbf{U}_\gamma^{(m)}, \text{ for } m = 0, 1, 2, ...$$

The score functions $\mathbf{U}_\beta$ and $\mathbf{U}_\gamma$ and the Fisher information matrix elements, $\mathbf{K}_{\beta\beta}$ and $\mathbf{K}_{\gamma\gamma}$, are presented, for example, in Osorio et al. (2007).

4 Constrained estimation methods

Under the marginal model, the variance components of the random effects do not need to be positive, as long as the matrix Σ_i is. In the context of elliptical mixed models, the most used estimation method is a (unconstrained) iterative process, which could theoretically produce a solution where some elements of the main diagonal of D could be negative, so as $|\Sigma_i|$. There are some ways to restrain the log-likelihood function and work with a constrained algorithm. One possibility would be to include a penalty $a \log |\Sigma_i|$ to the likelihood function or to consider a decomposition of the matrix D, for instance the Cholesky decomposition or the spectral decomposition. In the last case, an unconstrained solution would be sought. Here we search for a solution preserving the link to the hierarchical model, which means to impose that the elements of the main diagonal of D are positive.

5 Application

The rats data presented by Verbeke and Lesaffre (1999) is an example where one variance component is negative. As expected, the maximum likelihood reached by the unconstrained method is greater than in the constrained method. In this application, we fit a model of type (1) for a response measured in different ages of rats, with an intercept and three covariates indicating the time effects in three groups, divided by the administrated doses of a substance. Results are presented in Table 1.

TABLE 1. Maximum likelihood estimates for normal model obtained to the rats data.

Parameter	Unconstrained Method Estimate	Constrained Method Estimate
β_0	68.6210	68.6071
β_1	7.2748	7.3157
β_2	7.4750	7.5080
β_3	6.8865	6.8697
σ^2	1.5312	1.4349
τ_{11}	2.8122	3.4159
τ_{12}	0.4838	0.0185
τ_{22}	-0.3325	0.0000
MaxLogLike	-462.9178	-464.3272

6 Simulation results

We consider different scenarios for a simulation study with $n = 120$ and $m_i = 4$ for $i = 1, \ldots, 4$, and similar results were reached, showing that, even if the sample comes from a population where the variance elements of D are positive, the unconstrained estimation method may provide negative

estimates to some of the terms in the main diagonal of D, which disconnects the hierarchical and the marginal model and provides unrealistic estimates for D. The constrained method pointed lower bias and mean squared error of the variance components estimators, although larger bias of $\widehat{\sigma}^2$. For fixed parameters, no differences were observed. It is worth noting that the differences were minimal, meaning that both methods reach practically the same results.

TABLE 2. Simulation results in one scenario, with measures along time points 8, 10, 12 and 14.

		Unconstrained Method			**Constrained Method**		
Param.	**Real**	**Estim.**	**Bias**	**MSE**	**Estim.**	**Bias**	**MSE**
β_0	5	4.9953	-0.0047	0.1592	4.9953	-0.0047	0.1592
β_1	2.5	2.4992	-0.0008	0.0032	2.4992	-0.0008	0.0032
σ^2	1.5	1.4908	-0.0092	0.0764	1.4503	-0.0497	0.0627
τ_{11}	3	2.8640	-0.1360	1.6599	2.9425	-0.0575	1.6235
τ_{12}	0.1	0.1099	0.0099	0.0218	0.0911	-0.0089	0.0185
τ_{22}	0.01	0.0100	0.0000	0.0007	0.0164	0.0064	0.0004
MaxLogLike			-229.8647			-230.0404	

7 Discussion

In this paper we propose a comparison between unconstrained and constrained estimation methods for variance components in elliptical mixed-effects models. Our preliminary results show that the (unconstrained) maximum likelihood estimation method may provide unrealistic and with greater biased estimates for the variance components, but still good estimates.

Acknowledgments: The authors thank to Fundação de Amparo à Pesquisa do Estado de São Paulo for supporting this research.

References

Osorio, F., Paula, G. A. and Galea, M. (2007). Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics and Data Analysis*, **51**, 4354–4368.

Russo, C. M., Aoki, R. and Paula, G. A. (2012). Assessment of variance components in nonlinear mixed-effects elliptical models. *Test* 21, 519-545.

Verbeke, G., and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data, *Computational Statistics & Data Analysis*, **23**, 541-556.

Verbeke, G., and Lesaffre, E. (1999). The impact of dropout on the efficiency of longitudinal experiments. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, **48**, 363-375.

Childhood Leukaemia incidence around the Belgian nuclear sites: Surrogate exposure modelling

Koen Simons [1 2], Kaatje Bollaerts [1], Michel Sonck [2 3], Sébastien Fierens [1], André Poffijn [3], Lodewijk Van Bladel [3], David Geraerts [1], Pol Gosselin [1], Herman Van Oyen [1], Julie Francart [4], An Van Nieuwenhuyse [1]

[1] Scientific Institute of Public Health (WIV-ISP), Belgium
[2] Free University Brussels (VUB), Belgium
[3] Federal Agency for Nuclear Control (FANC), Belgium
[4] Belgian Cancer Registry (BCR), Belgium

E-mail for correspondence: An.VanNieuwenhuyse@wiv-isp.be

Abstract: Health effects among populations living in the vicinity of nuclear installations have been a major area of concern for several decades already. The main focus is on childhood leukaemia. The dominant approach is an ecological study using residential proximity to the nuclear site, however complex radioactive discharge models have also been used. This paper describes an ecological study on leukaemia incidence in children living in the vicinity of nuclear installations in Belgium. Each nuclear site was treated as a point-source and single-site focussed hypothesis tests were used to test for a gradient in childhood leukaemia incidence with residential proximity to the site. In addition, two other surrogate exposures were used: prevailing wind direction and simulated radioactive discharges. The use of multiple measures of surrogate exposures and multiple statistical methods has added value when investigating a priori defined point-sources.

Keywords: Childhood Leukaemia; Nuclear Sites; Surrogate-exposure Modelling

1 Introduction

Health effects among populations living in the vicinity of nuclear installations have been a major area of public concern for several decades already. Positive findings in recent peer reviewed articles, as well as incidents with nuclear installations, have further boosted this concern. In particular Kaatsch et al (2008) has reported a 2.2 fold $[1.51;+\infty]$ increase in leukaemia incidence in children living within 5 km of a German nuclear power plant (NPP). Furthermore, in the year 2008 an INES 3 incident has

occurred at the Institute for Radio-Elements in Fleurus, Belgium, accompanied by gaseous release of I-131 to the environment (Degueldre et al (2011)). Following these events, the Belgian Minister of Social Affairs and Public Health had commissioned a national epidemiological study of the health effects of living in the vicinity of nuclear sites. A multi-disciplinary group decided that in first instance, an exploratory ecological study should be undertaken, focussing on childhood leukaemia and thyroid cancer incidence and making use of readily available data. In this paper we present the surrogate-exposure modelling approach and results for childhood leukaemia incidence. The corresponding findings for thyroid cancer and for near versus reference analyses, are described elsewhere (Bollaerts et al (2012)).

2 Data

Childhood Leukaemia incidence (0-14 years) data by age-sex, year and commune, were obtained from the Belgian Cancer Registry. Data were available for the years 2000-2008 for the Flemish Region and 2004-2008 for the Walloon-Brussels-Capital Region. Population data were obtained from the Belgian Directorate-general Statistics and Economic Information. Population counts by age-sex, year and commune were obtained for Januari 1st for each incidence year. For each commune, a wealth index was obtained, consisting in the ratio between the commune-specific average income per inhabitant and the national average income per inhabitant. Because this index shows large temporal variation, an average wealth index was calculated for each agegroup (0-4, 5-9, 10-14 years) by averaging over the wealth indices experienced during the whole lifetime, assuming no migration between communes. The study considered all sites with nuclear installations of Class I as defined by the Royal Decree of July 20th 2001 on protection of the population, the workers, and the environment against the hazard of ionising radiations. This includes two NPPs, Doel and Tihange; as well as the sites of Mol-Dessel and Fleurus. The latter are host to a combination of research and industrial activities in the nuclear sector. Of the nuclear sites outside the Belgian territory, two reside within 20 km of the border: Chooz (France) and Borsele (Netherlands). No cases of childhood leukaemia incidence were detected within the 20 km proximity zones for both sites, during the study period. For the remainder of the paper we will consider only the sites in Belgium.

3 Surrogate Exposure

Three surrogate exposures were constructed: (inverse) residential proximity to the nuclear site, prevailing winds and radioactive discharge fraction. Residential proximity from a site was calculated as the Euclidian distance between the commune and the nuclear site. In order to calculate this, each

nuclear site was treated as a point source. For Mol-Dessel and Fleurus, the highest potential source was determined, whereas for both classical NPPs a central point between the multiple facilities on the site was georeferenced using Google Maps. For each commune a centroid was obtained. Wind speed and direction were obtained from on-site measurement stations for each site, for the period 2003-2008. 16 sector wind roses were calculated, omitting wind speeds below 0.2m/s: such low wind speeds are associated with frequently changing directions. Thus for each site was calculated the percentage of time the wind blows from the site into a specific sector. The communes within the 20 km proximity area around each nuclear site, were categorised into the 16 sectors by use of their centroids and the coordinates of the nuclear sites, as obtained for the residential proximity. For a continuous discharge, the prevailing winds accurately represents the fractions of the gaseous discharge that are blown towards the different communes in the proximity of a nuclear site. Both prevailing winds and residential proximity are surrogates created from the relative positions of site and commune. Both surrogates are complementary and the next step would be to combine them into a single variable that better reflects the true exposure. Noting that the deposition of gaseous discharges is not linearly dependent on residential proximity, this requires transforming the residential proximity into the fractions of radionuclide that reach different distances from a given site. The Hotspot Model *ref* was used by the FANC. Hotspot is a biguassian dispersion model that takes into account the effective release height, the (mean) rainfall, the (mean) windspeed and the deposition velocity in wet and in dry weather of one or more radionuclides. It is assumed that the concentration profiles are approximately normal on the axis of the wind direction, as well as on the plane perpendicular to this axis. Multiplying this site- and distance-specific fraction with the site-specific prevailing winds fraction, yields an estimate for the radioactive discharge fraction. For the site of Mol-Dessel the nuclides Ar-41 and I-131 were modelled, whereas for Fleurus the nuclides I-131 and Xe-133 were modelled. These radionuclides were chosen because they are the most relevant ones for exposure to these sites.

4 Methods

We used focussed hypothesis tests, treating each nuclear site as a point source. Under the null-hypothesis, there is no relationship between surrogate exposure and cancer incidence. The alternative hypotheses are positive association between inverse distance and incidence, between prevailing winds fraction and incidence, and between radioactive discharge fraction and incidence. For each surrogate exposure and site, three tests were performed: Bithell linear risk score test (Bithell (1995)), Bithell's test using ranks and Stone's likelihood ratio test (Stone (1988)). P-values were obtained by Monte Carlo simulation from the multinomial distribution with

5000 iterations. The simultaneous use of the three tests is an approach is often justified because each test is powerful only with respect to a more specific alternative hypotheses: Bithell linear risk score test test is optimal when the relation is a linear increase, whereas Stone's test assumes only a non-descending function. Noting that the main goal is to improve the overall power and that these tests are not independent, adjustment for multiple testing seems both tedious and counter-productive. For each of the four nuclear sites, up to three surrogate exposures were used, with three different alternative hypotheses. The multitude of tests allows for outcomes that are seemingly at odds. For visual guidance, Generalized Additive Models (Hastie and Tibshirani, 1990) were used to estimate the surrogate exposure-response relations. GAMs allow for any smooth relationship and are therefore very flexible. In particular, we used overdispersed Poisson model with surrogate exposure, age, sex and average wealth index as covariates. We used P-splines and choose the smoothing parameter by optimizing the AIC.

5 Results

Before using the surrogate exposures, a classical near versus reference analysis was made using overdispersed Poison Generalized Linear Models including a dummy variable for proximity, age, sex and average wealth index. For the classical NPPs the results showed no excess risk and were not sensitive to the size of the proximity area. For Mol-Dessel and Fleurus, results were dependent on the definition of the proximity area. Table 1 shows the results of the focussed hypothesis tests. For the classical NPPs these results are again consistent and show no relation between surrogate exposures and childhood leukaemia cancer incidence. For the site of Fleurus, some results are significant or borderline significant, however the majority of the results is not. For the site of Mol-Dessel, most of the results are significant.
Figure 1 shows the relative risk as a smooth function of estimated I-131 exposure. For Fleurus, the curve does not suggest a monotone increasing relation. This is consistent with the non-significant results of Stone's likelihood ratio test and Bithell's linear risk score test. The single significant result obtained when applying Bithell's test to the ranks of surrogate exposure, is not confirmed in the image. Furthermore the other surrogate exposures do not show an association with childhood leukaemia incidence. For Mol-Dessel the majority of the focussed hypothesis tests provide significant results, however not all of them. Figure 1 provides some clues as to why Stone's test and Bithell's test have very dissimilar results: for the latter the non-uniform distribution of surrogate exposure leads to a much higher influence of the most exposed commune.

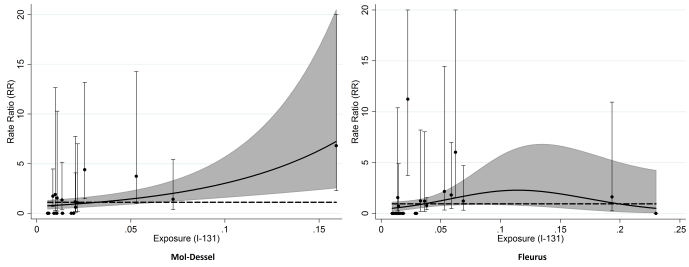


FIGURE 1. Rate Ratios (RR) and 95% point-wise confidence band of acute leukaemia incidence 0-14 years within the 20km proximity area as a smooth function of estimated gaseous I-131 discharge fraction. A scatter of the RRs for the different communes within the 20km proximity area is given on top, together with their 95% CIs. The dotted line represents the standardised incidence ratio obtained from a 20km near versus reference comparison.

6 Discussion

The current study has an ecological design, which has many limitations, most notably ecological bias: the average risk of a commune with an average exposure X is not necessarily equal to the individual risk at individual exposure X. This bias can be both towards or away from the null and the size of the bias depends on the within-area variation of exposure: with zero within-area variation, there is no ecological bias. Because all surrogate exposures are constructed from geographical coordinates, the within-area

TABLE 1. The results (p-values) of Stone's test, Bithells Linear Risk Score test (LRS), and Bithells Linear Risk Score test using ranks (LRS^2) for the three surrogate measures of exposure: (i) the (inverse) residential proximity to the nuclear site, (ii) prevailing winds fraction, and (iii) radioactive discharge fraction.

Nuclear site	Distance			Prevailing winds		
	Stone	LRS	LRS^2	Stone	LRS	LRS^2
Mol-Dessel	$< .01^+$	$.01^+$	$.01^+$	.42	$.01^+$	$.03^+$
Doel	.72	.84	.86	.56	.35	.53
Fleurus	.35	.14	.12	.11	.13	.41
Tihange	.24	.89	.91	.18	.62	.47

	I-131			Xe-133/Ar-41		
	Stone	LRS	LRS^2	Stone	LRS	LRS^2
Fleurus	.44	.08	$.02^+$	.43	.07	$.03^+$
Mol-Dessel	.43	$< .01^+$	$< .01^+$	.7	$< .01^+$	$< .01^+$

variation in surrogate exposure is a function of the size of the communes, which is not a constant for all communes. For inverse distance and radionuclide discharge fraction, the within-area variation is also dependent on the distance between the commune and the nuclear site: communes that are further away have less within-area variation than those that are close to the nuclear site. These facts further emphasize the influence of the commune that has the the highest surrogate exposure of all communes near the site of Mol-Dessel, because that commune also has a very high within-area variation. The prevailing winds and inverse distance surrogate exposures are less representative of the true exposure at the centroid of a commune, however they exhibit less within-area variation than radionuclide discharge fraction, in essence trading one source of bias for another. Because consistent results reinforce their credibility and inconsistent results diminish it, we argue that both the use of multiple surrogate exposures and the combination of multiple test types with graphical methods allows for a better qualitative judgement. For the classical NPPs, the results are consistently showing no relation with surrogate exposure. For Fleurus the majority of the results also indicates no relation, however for Mol-Dessel most of the outcomes suggest a potential link between the site and increased childhood leukaemia incidence. Taking into account the weaknesses of the study, it is recommended to investigate this further.

References

Bithell, J.F. (1995) The choice of test for detecting raised disease risk near a point source. *Stat Med* **14** 2309 – 2322.

Bollaerts, K., Fierens, S., Simons, K., Francart, J., Poffijn, A., Sonk, M., et al. (2012) *Monitoring of possible health effects of living in the vicinity of nuclear sites in Belgium* Report No.: D/2012/2505/01

Degueldre, D., Sonck, M., Vandecasteele, C.M. (2011) Rejet accidentel d'iode-131 par l'IRE sur le site de Fleurus : retour d'expérience de l'autorité de sûreté belge. *Radioprotection* **46** 159 – 173

Hastie, T., Tibshirani, R. (1990) *Generalized Linear Models* New York: Chapman & Hall/CRC.

Kaatsch, P., Spix, C., Schulze-Rath, R., Schmiedel, S., Blettner, M. (2008) Leukaemia in young children living in the vicinity of German nuclear power plants. *Int J Cancer*, **122** 721 – 726.

Stone, R.A. (1988) Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Stat Med* **7** 649 – 660.

Exploratory data analysis of energy security in the EU member countries in the period 2000-2010

Slawomir Smiech[1], Monika Papiez[1]

[1] Cracow University of Economics, Department of Statistics, Rakowicka 27, Cracow, Poland

E-mail for correspondence: smiechs@uek.krakow.pl

Keywords: Energy security; Spars PCA; PAM

1 Introduction

Fossil fuels form the foundation of energy balance in the European Union member countries. Their share in the total primary energy supply in 2010 amounted to respectively: oil (33.3%), gas (25.5%) and coal (16.2%). Net import constituted 55.5% of the total primary energy supply in 2010 and increased in comparison with 2000, when it constituted 49%. A growing dependence of the EU on imported energy and diminishing deposits of its own resources make the issues connected with energy security and energy policy of the EU one of the most important topics of debates. Additionally, energy balance of the EU does not correspond to energy balance of its particular member countries due to their diversification, which results in difficulties with developing a single energy policy. The EU member countries differ in their energy balance structure, the level of dependence on import and the level of diversification of energy suppliers. Those aspects, Papiez (2013), make the issue of energy security of the EU and the EU member countries worth considering.

There are numerous definitions of energy security developed by countries and international organizations (e.g. the International Energy Agency (IEA, 2009), the Asia Pacic Energy Research Centre (APREC, 2009) or the World Energy Council (WEC, 2009), since energy security depends on national priorities and their national concerns. Thus, it is difficult to define the term energy security precisely. Generally, most definitions of energy security make reference to its three aspects (energy, economy and ecology) and describe it as the availability of 'uninterrupted energy supplies at acceptable prices with respect to the natural environment'. Energy security is not directly measurable, although it can be approximated by multivariate set of variables. That is why in order to evaluate energy security in

quantitative terms the authors have developed indicators describing the relations between energy consumption and economic development, natural environment and social issues. The aim of the paper is to analyse energy security in the EU member countries in the period 2000 - 2010.

2 Empirical results

Taking into account a great variety of objects analysed with regard to the indicators describing energy security, which results in high volatility and the occurrence of outliers, partitioning among medoids (PAM) procedure developed by Kaufman and Rousseeuw (1990) was used. Similarly to a traditional k-means method, it assumes partitioning n observations into k clusters. PAM operates on the dissimilarity matrix, is less sensitive to outliers because it is based on the most centrally located object in a cluster (i.e. medoids), provides the silhouette which allows to determine which objects lie well within their clusters and which do not, and also shows how good is the quality of the clustering obtained. They suggested that silhouettes, i.e. average silhouette width, can be used for the selection of the best number of clusters in PAM (or in k-means methods).
In the second part of the analysis the authors focused on interpretations of the differences between the countries and clusters of countries. The classical principal components (PC) analysis is the most popular extraction and dimension reduction tool. It seeks the linear combinations of the original variables which capture maximal variance. Each PC is a linear combination of all variables and the loadings are usually non zero, which makes the interpretation difficult. Zou et al. (2004) proposed a new method called a sparse principal component analysis (sPCA). They used the lasso (elastic net) to generate modified principal component with sparse loadings. The idea is to formulate PCA as a regression-type optimization problem and obtain sparse loadings by imposing the lasso constraint on the regression coefficients.
The analysis of the level of energy security in EU member countries in the period 2000 - 2010 was conducted on the basis of variables used to obtain the Aggregated Energy Security Performance Indicator (AESPI) in (Martchamadol, Kumar 2013). As not all the values of the variables were accessible, only 15 out of 25 were selected for the analysis. The first stage of the analysis focused on the assessment of the quality of clustering based on average silhouette width. The value of average silhouette width (0.27) indicates a poor quality of clustering, which may indicate an artificial division of countries into clusters. The poor quality of clustering may also result from geographic, political and economic factors significantly differentiating the EU member countries.
The application of sPCA indicated four main components of energy security, which explained 78% of total variance. The components were named

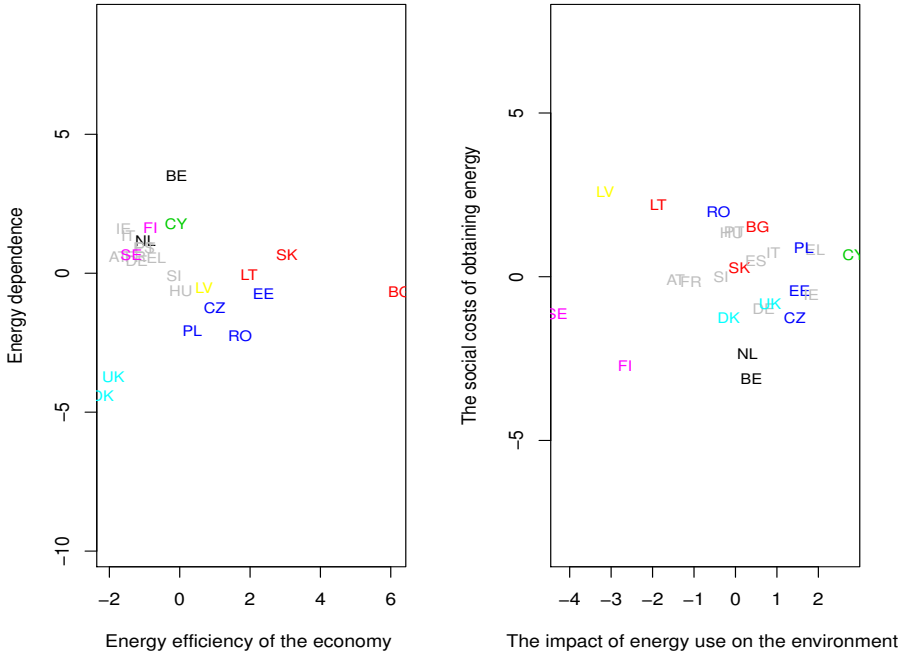


FIGURE 1. The situation in EU countries with respect to the new dimensions of energy security

according to their economic interpretation. The first sPCA component represents energy efficiency of the economy. The higher its value, the worse the economic situation of a given country with regards to energy efficiency. The highest energy efficiency in 2000 was found in Denmark and the United Kingdom, and the lowest in Bulgaria. It can be noticed that 'old' EU countries are characterised by higher energy efficiency and 'new' ones - by lower energy efficiency.

The second sPCA component represents energy dependence, i.e. dependence on energy source and type. The higher the component is, the more energy dependent a country is. In 2000 Belgium was the country with the highest dependence, and Denmark and the United Kingdom were the countries with the lowest energy dependence.

The third sPCA component represents the impact of energy use on the environment. The higher the component is, the more negative impact of energy use on the environment in a given country can be noticed. The lowest negative impact was observed in 2000 in Sweden, Finland, Lithuania and

Latvia, and the highest in Cyprus.
The fourth sPCA component represents the social costs of obtaining energy. The higher the component is, the less it costs a society of a given country to obtain energy. The lowest value of this component in 2000 was noticed in Belgium, Finland and the Netherlands and the highest in Lithuania, Latvia and Romania.

3 Conclusion

The results obtained indicate that the greatest improvement of energy efficiency took place in Romania, Bulgaria and Slovakia, and deterioration in Lithuania, Belgium and Estonia. The highest increase of energy dependency was noted in the United Kingdom and Lithuania, and the greatest decrease of energy dependency in Estonia. The negative impact of energy use on the environment decreased in Denmark and Portugal, and increased in Finland and Estonia. The social costs of obtaining energy decreased most in the United Kingdom and Ireland, and increased in Estonia.

Acknowledgments: This study benefited from a grant by the Polish National Science Centre (project DEC-2011/03/B/HS4/01134).

References

Asia Pacic Energy Research Centre (2007). A quest for energy security in the 21st century resources and constraints.

IEA (2009). World Energy Outlook 2009. *International Energy Agency*

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley series in probability and mathematical statistics.John Wiley and Sons Inc., New York.

Martchamadol, J., Kumar, S. (2013). An aggregated energy security performance indicator. *Applied Energy*, **103**, 653 – 670.

Papiez, M. (2013) *Convergence of energy security level in the EU member countries* In: Papiez, M., Smiech, S., (eds.), Proceedings of the 7th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Cracow: Foundation of the Cracow University of Economics, 107-114.

WEC (2009). World Energy and Climate Policy: 2009 Assessment. *World Energy Council Publication.*

Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**, 265 – 286.

Statistical model for multi-environment trials: Accounting for variety by environment interaction

Katia Stefanova[1]

[1] University of Western Australia, Australia

E-mail for correspondence: katia.stefanova@uwa.edu.au

Abstract: The aim of the paper is to present a statistical model and explore variety by environment ($V \times E$) interaction in the context of modelling genetic effects in different environments by using the factor analytic model.

Keywords: Factor-analytic (FA) model; Multi-environment trial (MET); Variety by environment ($V \times E$) interaction.

1 Introduction

Assessing and interpreting the $V \times E$ interaction, also referred to as genotype by environment interaction ($G \times E$), is a major challenge in any plant breeding/crop evaluation program. Widely used statistical techniques in the analysis of large sets of MET data involve two stage approach (Smith *et al* 2001). The first stage comprises single site analyses from which the spatially adjusted variety means and weights are obtained. At the second stage, a linear mixed model is used for the analysis of the combined set. Cullis *et al* (2010) presented an approach where factor-analytic techniques were applied for exploring $V \times E$ interaction. This paper presents a statistical model for the analysis of MET data and illustrates it on example of lupin variety testing data.

2 Statistical Model

Presented here model for MET data considers a series of t trials with total of m varieties grown, not necessarily at all trials. Trials performed at the same location but different years are considered as different environments. Let $\mathbf{y}^{[n\times 1]}$ denote the vector of individual plot yields combined across trials and $n = \sum_{j=1}^{t} n_j$ where n_j is the number of plots in the j^{th} trial. We assume a linear model for yield written as

$$\mathbf{y} = \mathbf{M}\eta + \epsilon, \tag{1}$$

The vector $\eta^{[mt\times 1]} = (\eta_{11}, \eta_{21}, \cdots, \eta_{m1}, \cdots, \eta_{1t}, \cdots, \eta_{mt})$ is the vector of the variety by environment ($V \times E$) effects, ordered as varieties within the environments and $\mathbf{M}^{[n\times mt]}$ is a design matrix that assigns variety by environment combinations to the vector of yields. The combined vector of residual effects from all trials is $\epsilon^{[n\times 1]}$.
The model for η, used in the context of MET data, is

$$\eta = \mathbf{1_{mt}}\mu + (\mathbf{1_t} \otimes \mathbf{I_m})\alpha + (\mathbf{I_t} \otimes \mathbf{1_m})\theta + \rho \tag{2}$$

where μ is the overall mean, $\alpha = (\alpha_{\mathbf{1}}, \alpha_{\mathbf{2}}, ..., \alpha_{\mathbf{m}})'$ is the $m \times 1$ vector of variety main effects, $\theta = (\theta_{\mathbf{1}}, \theta_{\mathbf{2}}, ..., \theta_{\mathbf{t}})'$ is the $t \times 1$ vector of environment main effects and ρ is the $mt \times 1$ vector of $V \times E$ interaction effects, $\mathbf{1_{mt}}$ and $\mathbf{1_m}$ are unit vectors, and $\mathbf{I_m}$ is the identity matrix.
A mixed model analogue of Principal Components Analysis (PCA) for the analysis of MET data was first used by Piepho (1997) and Smith *et al* (2001). Basically, in their approach the matrix of estimated $V \times E$ interaction effects from the model of equation (2) is subjected to PCA. The $V \times E$ interaction term is decomposed into a number of multiplicative terms. A random variety effect and fixed environment effect is assumed for the following model

$$\eta = \mathbf{1_{mt}}\mu + (\mathbf{1_t} \otimes \mathbf{I_m})\alpha + (\mathbf{I_t} \otimes \mathbf{1_m})\theta + (\mathbf{\Lambda_e} \otimes \mathbf{I_m})\mathbf{f_v} + \delta \tag{3}$$

where $\mathbf{\Lambda_e}^{[\mathbf{t}\times\mathbf{k}]}$ is the matrix of environmental loadings, $\mathbf{f_v}^{[\mathbf{mk}\times\mathbf{1}]}$ is the associated vector of variety scores and k is the number of the multiplicative terms (components) included in the analysis. The variance structure for the $V \times E$ interaction effects, namely $(\mathbf{\Lambda_e}\mathbf{\Lambda_e}' + \mathbf{\Psi_e}) \otimes \mathbf{I_m}$ is known as factor analytic structure of order k. It is also assumed that the vector θ comprises fixed effects and $\alpha, \mathbf{f_v}$ and δ are random effects with joint distribution

$$\begin{pmatrix} \alpha \\ \mathbf{f_v} \\ \delta \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2\mathbf{I_m} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I_k} \otimes \mathbf{I_m} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Psi_e} \otimes \mathbf{I_m} \end{pmatrix}\right],$$

where $\mathbf{\Psi_e}^{[\mathbf{t}\times\mathbf{t}]}$ is a diagonal matrix with elements commonly referred to as specific variances. Therefore,

$$E(\eta) = \mathbf{1_{mt}}\mu + (\mathbf{I_t} \otimes \mathbf{1_m})\theta$$

$$var(\eta) = \sigma_\alpha^{\mathbf{2}}(\mathbf{J_t} \otimes \mathbf{I_m}) + (\mathbf{\Lambda_e}\mathbf{\Lambda_e}' + \mathbf{\Psi_e}) \otimes \mathbf{I_m} \tag{4}$$

The vector of residual effects ϵ in (1) is presented as $\epsilon = (\epsilon_{\mathbf{1}}', \epsilon_{\mathbf{2}}', \cdots, \epsilon_{\mathbf{t}}')'$, where $\epsilon_{\mathbf{j}}^{[\mathbf{n_j}\times\mathbf{1}]}$ is the vector of residuals effects for the j^{th} trial, $j = 1, 2, ..., t$ and $\sum_j n_j = n$. Then the following model for the residual effects is considered:

$$\epsilon_{\mathbf{j}} = \mathbf{X_{p_j}}\tau_{\mathbf{p_j}} + \mathbf{Z_{p_j}}\mathbf{u_{p_j}} + \mathbf{e_j} \tag{5}$$

where $\tau_{\mathbf{p_j}}$ and $\mathbf{u_{p_j}}$ are vectors of fixed and random effects, respectively with design matrices $\mathbf{X_{p_j}}$ and $\mathbf{Z_{p_j}}$ and $\mathbf{e_j}$ is the vector of plot error effects for the j^{th} trial. It is assumed that

$$\begin{pmatrix} \mathbf{u_{p_j}} \\ \mathbf{e_j} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{G_{p_j}} & \mathbf{0} \\ \mathbf{0} & \mathbf{R_j} \end{pmatrix}\right],$$

Random and fixed effects account for local and global spatial field trends and for some extraneous variation due to agronomic or trial management practices.

Selection of the final FA model is based on the Restricted/Residual Maximum Likelihood (REML)log-likelihood, Akaike Information Criterion (AIC) and on the proportion of the genetic variation explained by each of the FA models fitted. The percentage variance explained for example for the first factor can be calculated as the ratio $\frac{tr(\lambda_1\lambda_1')}{tr(\mathbf{\Lambda\Lambda}'+\mathbf{\Psi})} \times 100$.

In regards to the prediction of the $V \times E$ effects, we obtain their best linear unbiased predictors (BLUPs) by following the approach of Cullis *et al* (2010). More precisely, we obtain the empirical BLUPs (E-BLUPs), since the unknown variance parameters have been replaced by their REML estimates in the mixed model equation.

3 Example

This example illustrates the use of the techniques described above in the analysis of 2011 MET lupin *Angustifolius* (further in the text referred as lupin) data from the Australian National Variety Testing Program. There are in total 33 trials, all the latest year of testing prior to release. There is a good connectivity (varieties in common) between the 2011 lupin trials, ranging between 14 and 38 varieties in common with exception of 4 trials having 5-9 varieties in common.

Sequence of the fitted models, as described in is presented and followed in Table 1. The modelling first commences with fitting the diagonal (DIAG) model, which assumes different genetic variances for the trials and no genetic correlation between the trials. The genetic variances obtained from the DIAG model are further used as initial values when next fitting the FA1 model. Table 1 also presents the % variance accounted for in modelling VxE for each model, from 2 to 7, it starts with 53.5% and reaches 94.5%. The comparison of the models is based on the REML log-likelihood and AIC. The FA6 model shows the lowest values for AIC and is significantly better than FA5 (the REML log-likelihood ratio test used for comparing FA5 and FA6 showed significant difference, p=0.024).

The estimated genetic correlation matrix for model FA6 is used to produce the dendrogram on Figure 1. The trials are ordered in accordance with the agglomerative hierarchical clustering of their genetic correlations. The dendrogram cut-off point of 0.4 (see the line at height 0.4 on Figure 1)

TABLE 1. Summary of the models fitted for the NVT lupin *Angustifolius* MET 2011 data.

Model	Name	REML Log-Likelihood	% Variance Accounted	No Var Parameters	AIC
1	DIAG	664.60		33	-1263.19
2	FA1	831.48	53.50	66	-1530.96
3	FA2	892.83	69.90	97	-1591.66
4	FA3	929.79	77.90	123	-1613.59
5	FA4	963.49	87.00	149	-1628.98
6	FA5	989.38	90.30	173	-1632.76
7	FA6	1019.47	94.50	195	-1648.94

defines 8 clusters of trials, 4 of which consist of a single trial. Apart from the dendrogram, a heat map (Figure 2) is a very helpful tool in selection of the clusters by visualizing the genetic correlations within clusters. The analyses have been performed in R environment (R Core Development Team, 2009), using ASREML-R (Butler *et al*, 2010) and the ***agnes*** R function for the hierarchical clustering.

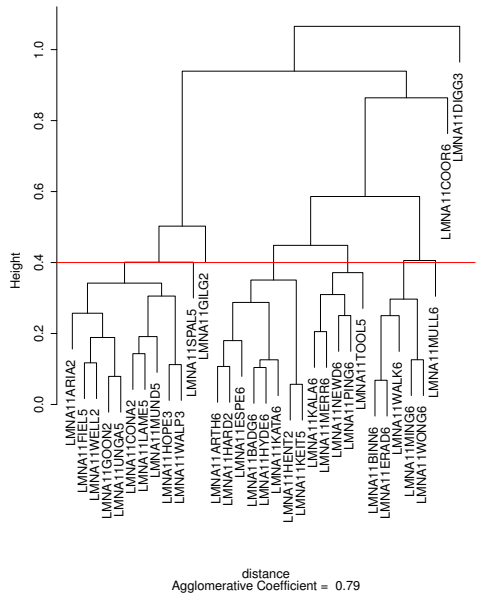


FIGURE 1. Cluster diagram of the trials based on the REML estimate of the genetic correlation for the FA6 model

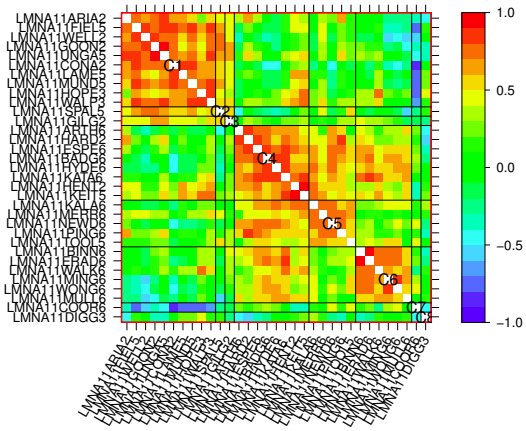


FIGURE 2. Heat map representation of the REML estimate of the genetic correlation for the FA6 model

Both, the dendrogram and the heat map, indicate that clusters 1, 4, 5 and 6 consists of relatively large number of trials, ranging between 5 and 10. The average genetic correlation between trials within these clusters is around 0.7. An interesting and very distinctive feature of the grouping is that cluster 1, the biggest cluster, consists of trials entirely of Eastern States and South Australia. Cluster 6 consists of trials only from Western Australia and clusters 4 and 5 predominantly, of trials from Western Australia. This reflects the typical environmental diversity in Australia.

Acknowledgments: The financial support of the Grains Research and Development Corporation of Australia is gratefully acknowleged.

References

Cullis, B.R., Smith, A.B., Beeck, C.P., and Cowling, W.A. (2010). Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. *Genome*, **53**, 1002 – 1016.

Smith, A.B., Cullis, B.R., and Gilmour, A. (2001). The analysis of crop variety evaluation data in Australia. *Australian and New Zealand Journal of Statistics* **43(2)**, 129-145.

Analyzing SET over time using multilevel multidimensional explanatory IRT models

Isabella Sulis, Vincenza Capursi

[1] Dipartimento di Scienze Sociali e delle Istituzioni, Cagliari, Italy
[2] Dipartimento di Scienze Economiche, Aziendali e Statistiche, Palermo, Italy

E-mail for correspondence: isulis@unica.it

Abstract: This contribution makes an attempt to analyze students' ratings of university teaching on a broad prospective, trying to adjust the final assessment from a wide range of factors which jointly may influence the process under evaluation: academic year peculiarities, course characteristics, students' characteristics and item dimensionality. From a methodological point of view, by setting complex Item Response models as special case of *Generalized Linear* or *Mixed Models* a large flexibility is introduced in the specification of ad hoc modelling approaches for the analysis of students' ratings.

Keywords: IRT; multidimensional; multilevel; longitudinal; Mokken Scale Analysis.

1 Introduction

The survey on university teaching quality aims to assess the quality of university course by indirect measurements provided by students' ratings. To make meaningful evaluations, the final measure of synthesis of students' ratings has to fulfill the following conditions: i) uni-dimensionality of the indicator items which define the latent variable; ii) absence of association with relevant units' characteristics (at level of student, course, class) which are external to the process under evaluation; iii) adjustment of the characteristics of the units under comparison (teachers, courses, etc) to remove the effect of confounding factors on the final assessment. Specifically, a good measurement scale should be characterized by items which contribute to build up a reliable measure of the locations of the units on the scale. This requires firstly the items measure the same aspect of the phenomenon under analysis. Whenever the uni-dimensionality condition is not satisfied the risk is to summarize in a single statement trends related to different aspects, balancing the positive and negative performances in a measure which is not adequate to highlight outstanding cases. For concepts that are multidimensional in their nature, firstly the construct needs to be structured in more sub-dimensions, secondly each sub-dimension has to be operationally

defined via a set of suitable indicator items. The detection of association between covariates external to the process under evaluation is a signal that factors which exist independently of the concept may potentially have influenced the intensity of the latent variable. In the analysis of students' ratings of teaching quality these external factors may be related to students, lecturers, courses or more generally, environment characteristics or annual/term disturbances (Sulis et al, 2011). Previous researches on the topic indicate students' previous interest towards the subject regardless the way the course is taught together with students self-state assessment on the sufficiency of their previous knowledge of the topic as the individual factors (among students'personal and curricula details) which have the greatest power in explain variation between ratings (Sulis and Capursi 2013). Thus, it is likely to attend that the same lecturer will receive, in average, different evaluations in classes where there are students with different levels of motivation in attending and studying the topic or whit different lack of basic knowledge that are supposed to be acquired in previous courses or in the secondary schools. As the same time, the omission of relevant covariates at course or teacher level, such as for instance the topic of the course, may contribute to leave unexplained the heterogeneity on the quality of university teaching not directly ascribable to lecturer's capabilities. For instance it is well known that in any formative path there are major and minor disciplines; a negative feeling towards the subject, together with inadequate skills to succeed, may have as a consequence low motivation and decreased levels of participation; all these status can affect the final assessment based on students' ratings.

2 A longitudinal analysis of SET using IRT models

Item Response Theory (IRT) (Fox, 2011) is considered the methodological approach that can provide the greatest information with respect to the aim of having a likely measure of an attribute measured on a set of categorical items . The classical IRT models are simple descriptive models since they investigate the characteristics of persons and items without considering the complexity of the relationships between external factors: the effect of observed subjective factors as well as the effect of group-level membership. In the last decades a number of IRT models have been developed in the statistics and psychometrics literature as extensions of the most basic descriptive models which allows us to easily deal with multidimensional latent trait, hierarchical data, analysis of the phenomena over time and the presence of significant covariates (Bacci and Caviezel 2011; Sulis and Capursi 2013). The most interesting part of these extensions concern the structural part of the model and the effect of the predictors, which can be either fixed or mixed and which can influence person or item parameters. Specifically, by considering the so called person parameters as random rather than fixed

effects, the basic descriptive IRT model (and its generalizations) is a simple level-2 multilevel model with measurement occasions nested within persons (descriptive two-level model) (Fox, 2011, Bacci, 2011). As well, a multilevel IRT model can be straightforwardly set up as a level-3 random-effect models where a further random term is added to take into account of the nesting of persons in groups (classes, schools, degree programs) (Bacci and Caviezel 2011; Sulis and Capursi 2013).

3 The data and the modelling approach

This study here presented concerns evaluation forms related to 42 lecturers gathered in a Faculty of an Italian university in three academic years. The items related to measure students' opinion on the quality of university course take the form of preposition on which the student has to declare the level of agreement on a four category scale:*decidedly no* (DN), *more no than yes* (MN), *more yes than no* (MY), *decidedly yes* (DY). To consider the role of peculiarities which can bias students' perception of the quality of the aspect related to teaching in a particular academic year the study considers the evaluations related to the same lecturer in three academic years (Sulis et al, 2011), even if they refer to different classes. The dimensionality of the items related to teaching activities has been analyzed using a non parametric IRT (NIRT) model, known as Mokken Scale Analysis (MSA). The main goodness of fit indexes associated to NIRT are based on the Loevinger' H coefficient (Sijtsma, 2000) which indicates how much the scale departs from the perfect Guttman scalogram. For a set of $j = 1, \ldots, J$ items $(Y_1, \ldots, Y_j)$, measured on K ordered $(k = 1, \ldots, K)$ categories H^S is defined as

$$H = \frac{\sum_{j=1}^{J} Cov(Y_j, R_{-j})}{\sum_{j=1}^{J} Cov(Y_j, R_{-j})^{max}}$$

with the sum score defined as $Y_+ = \sum_{j=1}^{J} Y_J$ and the rest score as $R_{-j} = Y_+ - Y_j$. On the basis of H value a set of item is defined weakly scalable if $0.3 \leq H < 0.4$, moderately scalable if $0.4 \leq H < 0.5$ and highly scalable if $H \geq 0.5$. If a Mokken Scale is weakly monotone it satisfies the minimum necessary conditions which are required by any parametric IRT models, specifically: uni-dimensionality, local independence, and latent monotonicity. The last assumption implies the probability to observe a responses not lower than a specific category (k) is a non-decreasing monotone function of the latent trait. In probabilistic terms, given two values of the latent trait θ_v and θ_w with $\theta_v \leq \theta_w$, $P(Y_j \geq k|\theta_v) \leq P(Y_j \geq k|\theta_w)\ \forall j, \forall Y$. Diagnostic tests to assess the weakly monotonicity assumption (WMA) classify respondents who show close values of the rest score in rest score groups $(s = 1, \ldots, S)$ of a minimum size and for any group s check if the condition $P(Y_j \geq k|R_{-j} \in s) \geq P(Y_j \geq k|R_{-j} \in r)\forall j, Y_j$ and $s > r$ holds.

Table 1 shows how many (#) times for each item the WMA is violeted in # comparisons, where the # of comparisons depends on the number of groups (van der Ark, 2007). On the basis of the MSA results, the items related to teaching have been structured in two dimensions: attitude to organize teaching activities -S^1- and ability in teaching -S^2-(Table 1).

TABLE 1. Mokken Scale Analysis.

Dimension 1: Attitude to organize teaching activities $H^{S_1} = 0.43$				
Item	Contains	H_j	# (co)	#(vi)
I_1	Clear exam rules	0.44	63	0
I_2	Clear indications on how to study the discipline	0.44	84	0
I_3	Presence at lecture	0.37	45	0
I_4	Clear course aims	0.49	84	0
I_5	Presence at office hours	0.43	108	0
I_6	Respect of lectures timetables	0.40	78	0
I_7	Suitability of the teaching materials	0.39	84	0
Dimension 2: Ability in teaching $H^{S_2} = 0.64$				
Item	Contains	H_j	# (co)	#(vi)
I_8	Ability to motivate the interest toward the topic	0.68	45	0
I_9	Ability to highlight the most important aspects	0.59	63	0
I_{10}	Availability to answer questions in class	0.56	58	0
I_{11}	Clarity explanations	0.67	41	0
I_{12}	Utility of teacher's lectures	0.66	63	0

From here on, given the strong negative skewness of the responses provided to the items of the two dimensions, we merged the responses DN, MN, MY in the same category and we focus the analysis on factors which influence the probability to be or not *definitely yes satisfied*. An explorative analysis has been carried out by considering a level-three variance component model (with logit link) and exploring the effect of the introduction of covariates on the variance of the random terms at level-two (students) and level-three (lecturers). Results show that at student-level, the variability is mainly explained by students' assessment on the sufficiency of their previous knowledge on the topic and on their interest towards the topic regardless the way the course has been carried out. The variability ascribable to lecturers is partially explained by introducing the following covariates at course-level: information on the subject of the course (the courses have been classified in macro-area, i.e. Statistics, Economics, etc.), the percentage of students in the class who declare to have sufficient knowledge on the topic and who are interested in the topic. None significant trend is detected at Faculty-level ascribable to the year the evaluation refer to. Indexing with j ($j = 1, \ldots, J$) the responses to single items of the scale, with i ($i = 1, \ldots, n$) the evaluation form/student and with s ($s = 1, \ldots, S$) the lecturers, we model the logit of the probability to be DY satisfied as follows

$$\text{logit}[P(Y_{jis} = 1)] = \alpha + X_j^{\mathrm{T}}\beta + D_{is}^{\mathrm{T}}\gamma + Z_s^{\mathrm{T}}\delta + \lambda^{\mathrm{T}}u_{js} + \eta^{\mathrm{T}}v_s \qquad (1)$$

where $u_{is} \sim N(0, \Omega_u)$ and $v_s \sim N(0, \Omega_v)$ are respectively a bivariate vector of random terms at student-level and v_s a trivariate vector of random

terms at lecturer-level. The last vector is introduced to inspect possible trajectories of performances at lecturer-level over the time. X_j is a vector of indicator variables which indicate on which item the response is related, D_{is} is a vector of students' or courses' covariates and Z_s is a vector of lecturer's covariates. λ is a binary vector which specifies on which dimension the item loads, whereas η is a binary vector which specifies to which academic year the evaluation belongs to. The model has been estimated with the runmlwin routine (Leckie and Charlton 2013) which adopts Markov Chain Monte Carlo bayesian estimation methods in order to approximate the likelihood. Comparisons across models have been made using the Bayesian Deviance Information Criterium (DIC) that summarize the fit and the complexity of the model.

References

Bacci S. (2012). Longitudinal data: different approaches in the context of item-response theory models. *Journal of Applied Statistics*, **29**, 2047 – 2065.

Bacci S. and Caviezel V. (2011). Multilevel IRT models for the university teaching evaluation. *Journal of Applied Statistics*, **28**, 2775 – 2791.

Fox J. (2012). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Leckie G. and Charlton C (2013). *runmlwin - A Program to Run the MLwiN Multilevel Modelling Software from within Stata*. Journal of Statistical Software, forthcoming.

Sijtsma K. and Hemker B. T (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, **25**, 391 – 415.

Sulis I. and Capursi V. (2012). Building up adjusted indicators of students evaluation of university courses using generalized item response models. *Journal of Applied Statistics*, **40** , 88 – 102. 2013.

Sulis I., Porcu M., and Tedesco N (2011). Evaluating lecturers capability over time. Some evidence from surveys on university course quality. In S. Ingrassia, R. Rocci, and M. Vichi, editors, *New Perspectives in Statistical Modeling and Data Analysis*. Heidelberg: Springer-Verlag.

van der Ark L. A. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software*, **7**, 1 – 19.

Handling missing data in Item Response Theory. Assessing the accuracy in estimation of two multiple imputation procedures

Isabella Sulis[1], Mariano Porcu[1]

[1] Università degli Studi, Cagliari

E-mail for correspondence: isulis@unica.it

Abstract: A critical feature in analyzing multi-item Likert-type scales using Item Response Theory models is the treatment of missing data. This work presents a simulation study aimed to validate two *ad hoc* stochastic procedures for categorical variables advanced by the authors to multiple impute missing observations in multi-item Likert-type scales

Keywords: Item Response Theory; Multiple Imputation; Categorical Data.

1 Introduction

Missing data is a problem in the analysis of multi-item Likert-type data collected to measure latent traits. *Item Response Theory* (IRT) is the modelling approach which provides the greatest insight to summarize individuals responses. In this work we discuss some imputation methods for Likert-type scales and we compare the performances of these methods with two *ad hoc* multiple imputation approaches for categorical variables: Multiple Imputation by Stochastic Regression (MISR) and Multiple Imputation by Latent Class Analysis (MILCA).

2 An overview (brief) of methods to handle missing data in the IRT framework

Methods that have been advanced in literature to cope with missing data can be classified as based on single or multiple imputation. The standard single imputation methods include the substitution of the missing value of an item with the mean of the item calculated over non missing units (Total Mean Substitution – TMS) or the intra-individual mean of the respondent for all non missing items (Valid Mean Substitution – VMS). A refinement of both methods especially designed for Likert-type scale is the Relative Mean Substitution (RMS). In the IRT framework, the imputation methods

that replace each missing observation with a single imputed value tend to underestimate the real variance (Raghunathan, 2004). *Multiple Imputation Analysis* (MIA) (Rubin, 1987) enable to overcome this drawback.

3 The MISR and MILCA procedures

The MISR (Sulis and Porcu 2010) procedure works in two steps. Consider a data matrix Y and set y_i as the vector which contains the responses of unit i to the J ($j = 1, \ldots, J$) categorical items measured on K categories. In the first step, the procedure starts building up for each unit i the distribution of the relative frequencies of ratings in each of the K response categories. Unobserved items for unit i are replaced by drawing values from a *Multinomial* distribution with parameters set equal to the relative frequencies of ratings observed for each category. In the second step a regression model is fitted: J regression equations are specified (one for each of the items in the data-set) where each of the J item is considered as a response variable whose values depend upon the set of the $(J-1)$ remaining predictors, and $(J-1)$ times as predictor (e.g. the value of y_1 is assumed to depend on $y_2 - y_6$ and so forth). Next, we fit a *proportional odds* logistic regression model to predict the probability to answer a category lower rather than greater than k: $\text{logit}[P(Y \leq k|x)] = \alpha_k + \beta^{\mathrm{T}} x$. The probability to provide a response in each category is

$$\pi_k = \Big[\frac{\exp(\alpha_k + \beta^{\mathrm{T}} x)}{1 + \exp(\alpha_k + \beta^{\mathrm{T}} x)} - \frac{\exp(\alpha_{k-1} + \beta^{\mathrm{T}} x)}{1 + \exp(\alpha_{k-1} + \beta^{\mathrm{T}} x)}\Big]. \tag{1}$$

The $\hat{\alpha}_k$s and $\hat{\beta}_k$s are estimated using the complete data-set generated in Step 1. Next, for each unobserved unit, a random draw is made from a *Multinomial* with the estimated parameters $[\hat{\pi}_1(x), \ldots, \hat{\pi}_K(x)]$.
The MILCA procedure relays on Latent Class Analysis (LCA). LCA assumes that any dependency across subjects' responses to manifest categorical indicators is explained by a single unobserved latent categorical variable z which has R categories $(z_1, \ldots, z_R)$. The idea to multiple impute variables using the results of the LCA has been introduced by Vermunt (2008). The main differences with MILCA are described in Sulis (2013). Denote with y_{ijk} the indicator variable which takes value 1 if observation i ($i = 1, \ldots, n$) selects category k ($k = 1, \ldots, K$) of item j, the joint probability density function of y_i is specified as $P(y_i|p, \pi) = \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K} (\pi_{rjk})^{y_{ijk}}$. π_{rjk} denotes the probability that an observation in latent class r provides the k outcome to item j; p_r is the probability to belong to each of the r classes. The EM algorithm is then used until convergence is reached: i) in the expectation step the posterior class membership $\hat{P}(r|y_i)$ of each unit i is calculated using Bayes' formula; (ii) in the maximization step new values of the key parameters are estimated; iii) the new values are set as new parameters in (i).

4 Assessing the accuracy in estimation of MISR and MILCA procedures

To assess the accuracy in estimation (AE) of the MISR and MILCA procedures we have compared their performances with those achieved with the RMS imputation and with the Multiple Imputation by Chained Equations (MICE) (Groothuis-Oudshoorn and van Buuren, 2011) procedure.

TABLE 1. AE of IRT model: item-threshold (β_k) and item-discrimination (λ_j)

		% of MAR obs.					
		5	10	15	20	25	30
		$MORAI_{\beta}$					
MICE		1.100	1.089	1.148	1.120	1.055	1.154
RMS		1.340	1.721	2.196	2.559	3.384	3.903
MILCA	3LC	1.037	1.009	1.374	1.250	1.025	1.065
	4LC	1.118	1.068	1.117	1.135	1.245	1.440
	5LC	1.164	1.095	1.254	1.078	1.079	1.190
	6LC	1.127	1.212	1.246	1.179	1.056	1.038
MISR		1.079	1.004	1.017	1.129	1.196	1.347
		$MORAI_{\lambda}$					
MICE		0.114	0.126	0.141	0.150	0.172	0.146
RMS		0.122	0.184	0.330	0.435	0.665	0.985
MILCA	3LC	0.114	0.134	0.155	0.162	0.159	0.209
	4LC	0.112	0.119	0.127	0.184	0.156	0.289
	5LC	0.108	0.117	0.132	0.143	0.195	0.229
	6LC	0.113	0.120	0.124	0.166	0.162	0.197
MISR		0.110	0.116	0.143	0.158	0.172	0.223
		$MORAI_{\beta,\lambda}$					
MICE		1.214	1.215	1.289	1.270	1.226	1.300
RMS		1.462	1.905	2.526	2.994	4.049	4.888
MILCA	3LC	1.152	1.143	1.529	1.412	1.184	1.274
	4LC	1.230	1.187	1.243	1.319	1.401	1.729
	5LC	1.272	1.212	1.387	1.220	1.273	1.419
	6LC	1.239	1.332	1.369	1.345	1.218	1.235
MISR		1.189	1.120	1.160	1.287	1.368	1.571

To this end a *Graded Response Models* (GRM) have been fitted for a complete dataset with 8 categorical variables measured on a 4 category Likert-type scale; estimates of the category-threshold (β_{jk}) parameters, of the item-discrimination parameters (λ_j), and of the person random parameter $\theta_i \sim \mathcal{N}(0, \sigma_\theta^2)$ have been calculated. y_{ij} is the response given on a k categories scale by individual i to variable j. Six data sets with different rates (5%, 10%, 15% ,20%, 25%, 30%) of missing units have been generated

deleting observations from the complete dataset according to two missing data generating processes: Missing Completely at Random (MCAR) and Missing at Random (MAR) (Rubin, 1987). In the following we will consider only the results obtained in the 6 datasets with observations simulated MAR. In the three multiple imputation procedures (MISR, MILCA and MICE) M has been set equal to 5. The imputed data sets have been used to estimate β_{jk} and λ_j parameters of the GRM. Two measures of efficiency have been considered: the Mean Square Error (MSE) and the Relative Accuracy Index (RAI). The latter has been defined by the authors as $RAI(\hat{\vartheta}_t) = (\frac{\hat{\vartheta}_t}{\vartheta_t} - 1)^2 + Var(\hat{\vartheta}_t) \quad \forall \hat{\vartheta}_t, \vartheta_t \neq 0$.
For the MILCA procedure, 4 models with different number of latent classes (ranging from 3 to 6) have been specified to evaluate the effect of overfitting and underfitting the LCA model on the AE (Vermunt et al. 2008, Sulis 2013). To make easier an assessment of the overall accuracy of the procedures in terms of MSE and RAI, both indexes have been summarized by taking the sum over the 24 item-threshold and 8 item-discrimination parameters of each of the 42 estimated models (6 rate of missingness $\times$ 7 missing data recovering methods). Table 4 summarizes the results. Specifically: the Model Overall Relative Accuracy Index of discrimination parameters - $MORAI_\beta = \sum_j \sum_k RAI(\beta_{jk})$; the Model Overall Relative Accuracy Index of discrimination parameters $--MORAI_\lambda = \sum_j RAI(\lambda_j)$; the sum of $MORAI_{\beta,\lambda} = MORAI_\beta + MORAI_\lambda$. The higher the values are, the worse is the overall estimation accuracy of the related imputation procedure.

References

Groothuis-Oudshoorn, K. and van Buuren, S. (2011) mice: Multivariate Imputation by Chained Equations in R. *J. of Stat. Soft.*. **45**, 3.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. NY: Wiley.

Sulis, I. and Porcu, M. (2010) A MI approach in a survey on university teaching evaluation. In: Palumbo F., Lauro, C. and Greenacre, M. (Eds.) *Data Anal. and Class.*. Berlin: Springer.

Sulis, I. (2013) A further proposal to MI missing categorical values using Latent Class Analysis. In: Giudici, P., Ingrassia, S. and Vichi, M. (Eds.) *Stat. Models for Data Anal.*. Berlin: Springer.

Vermunt, J.K. *et al.* (2008). Multiple imputation of categorical data using latent class analysis. *Sociological Methodology*. **33**, 269 – 297.

Nearest Neighbors Prediction Method for mixed logistic regression

Karin A. Tamura [1], Viviana Giampaoli[1], Alexandre Noma[2]

[1] Departamento de Estatística, Universidade de São Paulo, Brazil
[2] Centro de Matemática, Computação e Cognição, UFABC, Brazil

E-mail for correspondence: karinat@ime.usp.br

Abstract: This paper proposes a new approach to predict the random effects for new groups of a mixed logistic regression, by using nearest neighbors technique. The method, named as nearest neighbors prediction method (NNPM), computes the distances between a new group and the groups with known random effects, based on the covariates. Then, the random effects of the nearest neighbors can be considered in order to define the value of the random effects for the new group. The approach was compared with recent methods and presented superior performance in an application study.

Keywords: nearest neighbors; mixed logistic regression; outcome prediction.

1 Related Works

Mixed logistic model considers that, conditional on $\boldsymbol{\alpha}_i$, y_{ij}'s are independent Bernoulli's, in which i indexes the group, $i = 1, \ldots, q$, j indexes the observation within the i-th group, $j = 1, \ldots, n_i$. This model is given by

$$\text{logit}\{P(y_{ij} = 1|\boldsymbol{\alpha}_i)\} = \log\{\frac{p_{ij}}{1-p_{ij}}\} = \mathbf{x}_{ij}^t\boldsymbol{\beta} + \mathbf{z}_{ij}^t\boldsymbol{\alpha}_i, \quad (1)$$

in which $\boldsymbol{\beta}$ is an unknown vector of fixed effects $(p \times 1)$ and $\boldsymbol{\alpha}_i$ is an unknown vector of random effects $(k \times 1)$ of the i-th group. Vector $\mathbf{x}_{ij}^t$ of known covariates $(1 \times p)$ is associated with $\boldsymbol{\beta}$, defined by $\mathbf{x}_{ij}^t$=(1, x_{1ij}, x_{2ij},..., $x_{(p-1)ij}$). Vector $\mathbf{z}_{ij}^t$ of known covariates $(1 \times k)$ is associated with $\boldsymbol{\alpha}_i$, defined by $\mathbf{z}_{ij}^t$=(1, z_{1ij}, z_{2ij},..., $z_{(k-1)ij}$). This model assumes that $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q$ are i.i.d. with $\boldsymbol{\alpha}_i \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$, in which $\boldsymbol{\Sigma}$ is the unknown covariance matrix of the random effects.

Tamura and Giampaoli (2011) proposed to use the empirical best predictor (EBP) to predict the outcome probability of the i-th new group in the observation level of model (1) based on the conditional expectation of the random effects. In this case, numerical integration methods must be used to solve the k-dimensional integration.

Non-parametric prediction method (NPPM) models the dependence of the outcome (empirical random effects obtained by model (1)) in relation to

the covariates aggregated in the group level by considering the additive non-parametric model (Hastie and Tibshirani (1990)). For the new groups, the random effects are predicted based on estimate function of the non-parametric model. Then, the predicted random effects are inserted in the linear predictor of the mixed logistic regression, providing the outcome probability of a new group in the observation level. More details, see Tamura and Giampaoli (2012).

2 Nearest Neighbors Prediction Method (NNPM)

The proposed approach is based on the nearest neighbors (NN) technique, which is commonly used for supervised pattern classification. According to Cover and Hart (1967), based on feature vectors, the task is to assign a nominal outcome to a given element. The classification is based on the feature vectors (covariates) by using a distance (e.g. Euclidian, Mahalanobis, City Block, etc). Considering l nearest neighbors, the original method chooses the most voted nominal outcome. Recently, the technique has allowed continuous outcome, e.g. Nigsch et. al (2006).
In this paper, the goal is to assign random effects to new groups based on feature vectors, corresponding to covariates in the observation level or aggregated in the group level. Since the random effects are continuous outcomes, we applied the NN technique, in which the assignment is performed by considering some centrality measurement (e.g. mean, median, medoid, etc) of the known random effects of the l nearest neighbors.
Applied to mixed logistic regression, NNPM is described as follows. For each $i \in G = \{1, \ldots, q\}$, there is a feature vector g_i, and a known random effects vector $\hat{\boldsymbol{\alpha}}_i = (\hat{\alpha}_{1i}, \ldots, \hat{\alpha}_{ki})$ estimated by model (1). For each new group i', with $i' = 1, \ldots, q'$, there is only a feature vector $g_{i'}$, with $i' \notin G$. The aim is to predict the values of the random effects for the i'-th new group represented by $\boldsymbol{\alpha}_{i'}$. The NNPM algorithm is described as follows:

1:For i' in 1 to q' {
2: For i in 1 to q {
3: Compute the distance $d_{(i',i)}$ between $g_{i'}$ and g_i;
4: }
5: Sort the elements of $d_{(i',.)} = (d_{(i',1)}, d_{(i',2)}, \ldots, d_{(i',q)})$ in increasing order;
6:}
7:For l in 1 to q {
8: For i' in 1 to q' {
9: Compute a centrality measurement of the known random effects, $\boldsymbol{\alpha}_{i'} = (\bar{\alpha}_{1.}, \ldots, \bar{\alpha}_{k.})$, corresponding to the l first elements of the sorted $d_{(i',.)}$;
10: The random effects $\boldsymbol{\alpha}_{i'}$ are inserted in the linear predictor of the mixed logistic regression, providing the outcome probability of the i'-th new group in the observation level;
11: }
12:}

13: Select l which maximizes the performance prediction of the mixed logistic model.

In the algorithm, lines 1-6 compute the distances, which can be stored in a matrix with elements $d_{(i',i)}$. In lines 7-13, the approach computes l that maximizes the performance prediction, based on the predicted random effects for new groups.
In the following application, we considered NNPM as centrality measurement the mean of the known random effects, and the Euclidian distance.

3 Application

The nutritional information of 241 newborns were collected in 7 periods of observation after the birth. The outcome is the HAZ-score, a nutritional classification based on the height of the children, classified into two categories: 1 - heavy unnourished and 0 - otherwise. To model the HAZ-score, we considered the covariate z_{ij}, which is a function of the weight of the i-th child in the j-th period of observation. Mixed logistic regression with 2 random effects was fitted by

$$\text{logit}[P(y_{ij} = 1|(\alpha_{1i}, \alpha_{2i}))] = \beta_0 + \beta_1 z_{ij} + \alpha_{1i} + \alpha_{2i} z_{ij}. \quad (2)$$

We considered a random sample of 50% of the children (training data set) to fit model (2). The remaining children were considered in the validation data set to predict the outcome of new groups, based on the parameters estimate from the training data set. Table 1 presents the estimate parameters of model (2) by Laplace approximation.

TABLE 1. Estimates, standard error and p-value of the mixed logistic regression.

Parameters	Estimate	Standard Error	P-value
β_0	1.635	0.390	<0.001
β_1	-5.294	0.594	<0.001
$(\sigma_1; \sigma_2)$	(2.226; 3.212)		
ρ	-0.167		

To evaluate the performance prediction, we considered AUC (Area Under the Curve) and KS (Komolgorov-Smirnov) measurements. The values of AUC and KS vary between 0% and 100%, in which the higher the value, the better the performance of the model.
Table 2 presents the prediction performance of EBP, NPPM and NNPM, applied in the validation data set. NPPM and NNPM I considered $\bar{z}_{i.}$, the average of z_{ij}, as the covariate aggregated in the group level. NNPM II considered the covariates in the observation level, taking into account the tendency over the time. We noticed that NNPM I and NNPM II outperformed the other prediction methods. For [NNPM I; NNPM II], the maximum value of AUC=[85.1%; 85.6%] considered the average of $l = [25; 10]$

TABLE 2. Performance of the prediction methods.

Prediction Method	AUC	KS
EBP	82.9%	50.0%
NPPM	84.4%	53.4%
NNPM I	85.1%	57.2%
NNPM II	85.6%	59.0%

nearest neighbors, while the maximum value of KS= [57.2%; 59.0%] considered the average of $l = [16; 9]$ nearest neighbors. Comparing NNPM I with NNPM II, we observed that by considering the covariates in the observation level improved the prediction performance and required less nearest neighbors. Moreover, we noticed that if the empirical random intercept and the random slope do not follow a normal distribution, as seen in this application, NPPM and NNPM can be more appropriate than EBP, because these methodologies do not require this assumption.

4 Conclusions

The main advantages of the proposed NNPM are: prediction of the outcome without assuming any distribution of the random effects and a better performance by considering covariates in the observation level. For a future work, it is interesting to compare these methodologies (EBP, NPPM and NNPM) by simulation studies, and to evaluate the impact of the misspecification of the random effects distribution in the prediction methods.

Acknowledgments: The work received financial support from FAPESP and CNPq.

References

Cover, T.M., Hart, P.E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13 (1)**, 21 – 27.

Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall.

Nigsch, F., Bender, A., Van Buuren, B. , Tissen, J., Nigsch. E. and Mitchell, J.B. (2006). Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *Journal of Chemical Information and Modeling*, **46 (6)**, 2412 – 2422.

Tamura, K.A., Giampaoli, V. (2011). Prediction for an observation in a new cluster for Multilevel Logistic Regression considering k random coefficients. In: *26th International Workshop on Statistical Modelling*, Valencia,593-596.

Tamura, K.A., Giampaoli, V. (2012). Comparison of prediction methods for mixed logistic regression. In: *27th International Workshop on Statistical Modelling*, Prague,327-332.

Illness-death model for interval-censored and left-truncated data with random effects: Application to dementia

Célia Touraine[1,2], Pierre Joly[1,2]

[1] Univ. Bordeaux, ISPED, INSERM U897, Bordeaux, F-33000, France
[2] INSERM, ISPED, INSERM U897, Bordeaux, F-33000, France

E-mail for correspondence: celia.touraine@isped.u-bordeaux2.fr

Abstract: The illness-death model allows individuals to move from the "healthy" state to the "dead" state either directly or having been in the "diseased" state before. Regression models like proportional transition intensities models are used to assess the impact of individual factors on each transition. We propose to incorporate random effects for group-specific factors in them. The motivating data set on dementia lead us to an illness-death model and its estimation that allow for left-truncated and interval-censored data.

Keywords: Illness-death model; Shared frailty; Random effects; Interval censoring.

1 Introduction

Dementia research in epidemiology is often based on data from prospective cohort studies. The individuals are initially non demented and have follow-up visits at more or less regular intervals to determine their health status. Consequently, for demented individuals the onset of dementia is interval censored between the diagnostic visit and the previous one. Moreover, these cohorts are composed of elderly individuals who are likely to die during the follow-up period. For dead individuals who were non demented at their last visit, there is an uncertainty about their health status since they may have been demented between last visit and death. These special features of the data can be handled by an illness-death model that allows individuals to move from the "non-demented" state (state 0) to the "dead" state (state 2) either directly or having been in the "demented" state (state 1) before (see Figure 1).

2 Motivating data set

The motivating data set is from the french elderly Paquid cohort study (Letenneur *et al*, 1999). The individuals were randomly sampled from the general population of 75 communities of South-Western France. 3675 participants aged 65 years or older entered the study in 1988. The number of

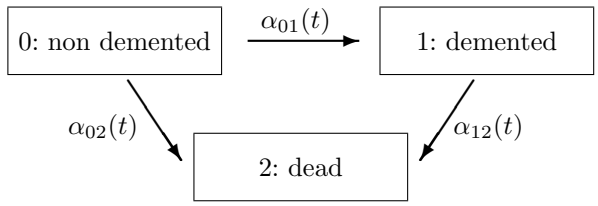


FIGURE 1. Illness-death model

participants by community varies from 15 to 724. Then, visits for dementia screening were planned at 1, 3, 5, 8, 10, 13, 15, 17, 20 years of follow-up. To enter the study, individuals had to be non demented. Note that we have to deal with left-truncated data because age is chosen as the basic time scale. Over a 20-year period, 2937 deaths occurred and 832 cases of dementia were observed (among them, 639 died afterwards). In some analyses of Paquid data set, an illness-death model for left-truncated and interval-censored data have been fitted with individual factors measured at baseline, like level of education (Commenges *et al*, 2004). However, it is likely that individuals belonging to the same community share certain factors (environmental factors – like climate, aluminium content in tap water, *etc.* –, social factors, *etc.*) which may impact risks of dementia and risk of death. Our aim is to improve an illness-death model model for left-truncated and interval-censored data by accounting for both individual factors and group-specific factors, where groups are communities.

3 Model

In an illness-death model, the functions of interest are the transition intensities $\alpha_{01}(.)$, $\alpha_{02}(.)$, $\alpha_{12}(.)$ that are instantaneous probabilities of direct transition between states 0 and 1, 0 and 2, 1 and 2. Regression models are used to assess the impact of individual factors on each transition. Usually, one assume proportional transition intensities models: for individual i and for $(k,l) \in \{(0,1),(0,2),(1,2)\}$,

$$\alpha_{kl,i}(t|\mathbf{z}_{kl,i}) = \alpha_{kl,0}(t)e^{\boldsymbol{\beta}_{kl}^{\mathrm{T}}\mathbf{z}_{kl,i}}$$

where $\alpha_{kl,0}(.)$ is the baseline transition intensity and $\mathbf{z}_{kl,i}$ the covariate vector for transition $k \to l$. In a similar way to shared frailty models in survival analysis (see Duchateau and Janssen, 2008 and Hougaard, 2000), we propose to take into account random effects of group specific factors in addition to fixed effect of individual factors using proportional intensities models with frailty terms on each transition. For individual j of group i:

$$\alpha_{01,ij}(t|\mathbf{z}_{01,ij}, u_{01,i}) = \alpha_{01,0}(t)e^{\boldsymbol{\beta}_{01}^{\mathrm{T}}\mathbf{z}_{01,ij}+u_{01,i}}$$

$$\alpha_{02,ij}(t|\mathbf{z}_{02,ij}, u_{02,i}) = \alpha_{02,0}(t)e^{\boldsymbol{\beta}_{02}^{\mathrm{T}}\mathbf{z}_{02,ij}+u_{02,i}}$$

$$\alpha_{12,ij}(t|\mathbf{z}_{12,ij}, u_{12,i}) = \alpha_{12,0}(t)e^{\boldsymbol{\beta}_{12}^{\mathrm{T}}\mathbf{z}_{12,ij}+u_{12,i}}$$

where $\boldsymbol{\beta}_{01}$, $\boldsymbol{\beta}_{02}$, $\boldsymbol{\beta}_{12}$ are the fixed effects vectors; $u_{01,i}$, $u_{02,i}$, $u_{12,i}$ are the random effects (frailties) of the ith group, that are the actual values of a sample from normal distributions:
$U_{01} \sim \mathcal{N}(0, \sigma_{01}^2)$, $U_{02} \sim \mathcal{N}(0, \sigma_{02}^2)$, $U_{12} \sim \mathcal{N}(0, \sigma_{12}^2)$.

The likelihood is:

$$\prod_{i=1}^{I} \frac{\int\int\int \prod_{j=1}^{n_i} \mathcal{L}_{ij}(x_{ij}|u_{01,i}, u_{02,i}, u_{12,i}) f_{U_{01}}(u_{01,i}) f_{U_{02}}(u_{02,i}) f_{U_{12}}(u_{12,i}) du_{01,i} du_{02,i} du_{12,i}}{\int\int \prod_{j=1}^{n_i} \mathbb{P}(X(t_{0,ij}) = 0|u_{01,i}, u_{02,i}, u_{12,i}) f_{U_{01}}(u_{01,i}) f_{U_{02}}(u_{02,i}) du_{01,i} du_{02,i}}$$

where: I is the number of groups; n_i is the number of individuals in group i; $f_{U_{01}}$, $f_{U_{02}}$, $f_{U_{12}}$ are the probability density functions for U_{01}, U_{02}, U_{12}; $t_{0,ij}$ is the age of individual j of group i at entry into the cohort; $X(t)$ is the state in which he is observed at age t, x_{ij} contains observations of individual j of group i.
The contribution of individual j of group i, $\mathcal{L}_{ij}(x_{ij}|u_{01,i}, u_{02,i}, u_{12,i})$, and the term due to left truncation, $\mathbb{P}(X(t_{0,ij}) = 0|u_{01,i}, u_{02,i}, u_{12,i})$, can be expressed in terms of transition intensities (see Joly *et al*, 2002 in a context without random effect). The parameters to estimate are the parameters of the baseline transition intensities, the regression parameters and the variance of the random effects. We use either a parametric approach with Weibull baseline intensities, or, a semi-parametric approach with M-spline approximation of the baseline intensities. We maximise the likelihood in the parametric approach or a penalized likelihood in the semi-parametric approach.
Note that in the above formula, we consider different random effects on each transition. The likelihood would be simplified if we consider for example the same random effect on mortality *i.e.* on transition $0 \to 2$ and $1 \to 2$. Note also that we can consider dependant random effects by adding covariance parameters.
Some statistical tests for the significance of the random effects can be done. For example, one can test for the significance of the random effect on transition $0 \to 1$ by testing the null hypothesis "$\sigma_{01}^2 = 0$". A corrected likelihood ratio test is used because the null hypothesis is on the boundary of the parameter space (Verbeke and Molenbergghs, 2000).

Acknowledgments: Special Thanks to the *Région Aquitaine* for financial support and to the Paquid team for making the Paquid data available to us.

References

Commenges, D., Joly, P., Letenneur, L. and Dartigues, J. F. (2004). Incidence and mortality of Alzheimer's disease or dementia using an illness-death model. *Statistics in Medicine*, **23**, 199 – 210.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Berlin: Springer, Statistics for Biology and health.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer, Statistics for Biology and health.

Joly, P., Commenges, D., Helmer, C. and Letenneur, L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*, **3**, 433 – 443.

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, J. M. and Dartigues, J. F. (1999). Are sex and educational level independent predictors of dementia and Alzheimer's disease? Incidence data from the PAQUID project. *Journal of Neurology, Neurosurgery & Psychiatry*, **66**, 177 – 183.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Modeling heterogeneity by fixed effects models

Gerhard Tutz[1], Margret-Ruth Oelker[1]

[1] Institute of Statistics, LMU Munich, Germany

E-mail for correspondence: margret.oelker@stat.uni-muenchen.de

Abstract: The classical approach to the modeling of heterogeneity in repeated measurements is to allow for random effects. Random effects are an efficient tool but inference may depend on the assumed mixing distribution. Moreover, potential correlation between random effects and predictors is ignored. Fixed effects models are an alternative but suffer from the large number of parameters. Regularized estimation methods allow to obtain estimates. We are in particular interested in the identification of clusters of units that share the same effect and propose corresponding regularized estimation procedures.

Keywords: Fixed effects, random effects, heterogeneity.

1 Introduction

When the observations are clustered, as in longitudinal studies, or as in subsamplings of the primary sampling units in cross-sectional studies, the heterogeneity amongst the units has to be considered. Often, random effects are employed to model the heterogeneity of effects (see Verbeke and Molenberghs, 2009; Tutz, 2012). *Fixed effects models* where the effects are considered as unknown but fixed, are an alternative. Fixed effects models are especially useful when one is interested in the performance of specific units. Moreover, they are not affected by the potential misspecification of the mixing distribution. The objective of the paper is to show that subject-specific approaches in combination with regularized estimates are an attractive alternative to existing approaches in cases where the units themselves are of interest. In particular, they allow to identify clusters with identical effects on the response.

2 Modeling heterogeneity

We briefly consider two approaches that are able to model heterogeneity – methods that are based on random effects and subject-specific models that are more flexible but call for regularized estimation procedures.

2.1 Random effects models

Let the observations be given in clusters with y_{it} denoting the observation t in cluster i, $i = 1, \ldots, n$, $t = 1, \ldots, T_i$. In addition, let $x_{it}^{\mathrm{T}} = (1, x_{it1}, \ldots, x_{itp})$ be a covariate vector associated with fixed effects and $z_{it}^{\mathrm{T}} = (z_{it1}, \ldots, z_{itq})$ be a covariate vector associated with random effects.
The structural assumption in a generalized linear mixed effects model (GLMM) for clustered data specifies that the conditional means $\mu_{it} = E(y_{it}|b_i, x_{it}, z_{it})$ have the form

$$g(\mu_{it}) = x_{it}^{\mathrm{T}}\beta + z_{it}^{\mathrm{T}}b_i = \eta_{it}^{\mathrm{par}} + \eta_{it}^{\mathrm{rand}}, \tag{1}$$

where g is a monotonic and continuously differentiable link function and $\eta_{it}^{\mathrm{par}} = x_{it}^{\mathrm{T}}\beta$ is a linear parametric term with the parameter vector $\beta^{\mathrm{T}} = (\beta_0, \beta_1, \ldots, \beta_p)$, which includes an intercept. The second term, $\eta_{it}^{\mathrm{rand}} = z_{it}^{\mathrm{T}}b_i$, contains the vector with the cluster-specific random effects b_i that model the heterogeneity of clusters. For the random effects, one assumes a distributional form, typically a normal distribution, $b_i \sim N(\mathbf{0}, Q)$, with covariance matrix Q.
In a GLMM, one assumes that $y_{it}|b_i, x_{it}, z_{it}$ follows a simple exponential family. Moreover, it is assumed that the observations y_{it} are conditionally independent with means $\mu_{it} = E(y_{it}|b_i, x_{it}, z_{it})$.
The focus of random effects models is on the fixed effects; the distribution of the random effects is used to account for the heterogeneity over clusters. However, assuming a distribution for the random effects prevents that units are clustered. Moreover, the random effects and the explanatory variables are assumed to be independent.

2.2 Fixed effects models

In the subject-specific model, one assumes

$$g(\mu_{it}) = x_{it}^{\mathrm{T}}\beta + z_{it}^{\mathrm{T}}\beta_i \tag{2}$$

for the link between explanatory variables and the mean $\mu_{it} = E(y_{it}|x_{it}, z_{it})$. In addition to the global parameter vector β, each unit has its own parameter $\beta_i = (\beta_{i0}, \ldots, \beta_{iq}), i = 1, \ldots, n$; whereby parameters $\beta_i = (\beta_{i0}, \ldots, \beta_{iq})$ are weights on the vector $z_{it}^{\mathrm{T}} = (1, z_{it1}, \ldots, z_{itp})$. The large number of parameters can render the estimates unstable and encourage overfitting. However, under the assumption that observations form clusters with respect to their effect on the response, the number of parameters reduces and estimates are available. In contrast to common approaches, and in order to avoid identifiability problems, we assume that z_{it} is not a subset of x_{it}.
The subject-specific term in (2) can also be seen as a varying-coefficient term. It represents the interaction between the variable z_{it} and the factor

that represents the clusters. Let us consider the simple intercept model $g(\mu_{it}) = x_{it}^{\mathrm{T}}\beta + \beta_{i0}$, where $z_{it} = 1$. By using the dummy variables $x_{C(1)}$, $\ldots$, $x_{C(n)}$ to code the individuals in $C = \{1, \ldots, n\}$, the model has the form

$$g(\mu_{it}) = x_{it}^{\mathrm{T}}\beta + x_{C(1)}\beta_{10} + \cdots + x_{C(n)}\beta_{n0}.$$

3 Regularized estimation for subject-specific models

The basic concept to enforce the clustering of units by their effect strengths, is penalized maximum likelihood estimation. Let all the parameters be collected in $\alpha^{\mathrm{T}} = (\beta^{\mathrm{T}}, \beta_1^{\mathrm{T}}, \ldots, \beta_n^{\mathrm{T}})$, with β_j, $j = 1, \ldots, n$, denoting the subject-specific parameters. Instead of maximizing the log-likelihood, one maximizes the penalized log-likelihood $l_p(\alpha) = l(\alpha) - \lambda J(\alpha)$, where $l(\alpha)$ denotes the familiar un-penalized log-likelihood, λ is a tuning parameter, and $J(\alpha)$ is a penalty term that enforces clustering. A specific penalty term with this property is

$$J(\alpha) = \sum_{s=1}^{q} \sum_{i>j} |\beta_{is} - \beta_{js}|. \tag{3}$$

A simple example is the penalty with $z_{it} = 1$ representing a random intercept only. With the model given by $g(\mu_{it}) = x_{it}^{\mathrm{T}}\beta + \beta_{i0}$, the penalty has the form $J(\alpha) = \sum_{i>j} |\beta_{i0} - \beta_{j0}|$. If $\lambda = 0$, one obtains the usual un-penalized estimates of α and each unit has its own intercept. If $\lambda \to \infty$ the penalty enforces that all parameters are set to equal; that is, just one cluster is identified. Thus, λ determines the number of clusters in the data.
Adaptive versions of penalties that include weights have been shown to have better properties in terms of selection consistency; see, Zou (2006) who considered adaptive versions of the Lasso (Tibshirani, 1996). With $w_{ijs} = |\tilde{\beta}_{is} - \tilde{\beta}_{js}|^{-1}$, where $\tilde{\beta}_{is}$ denotes an $\sqrt{n}$-consistent estimate like the maximum likelihood (ML) estimate, an adaptive version of penalty (3) is obtained.

4 Mortality after myocardial infarction

We consider a 22-center clinical trial of beta blockers for reducing mortality after myocardial infarction (see for example, Aitkin, 1999). In each center, the number of deceased/successfully treated patients in control/test groups was observed. Hence, the first-level units are the hospitals; the patients represent the second-level units. The binary response (deceased/not deceased) suggests a logit model. A classical model that has been used on this data set is the random intercept model, which has the form $\mathrm{logit}P(y_{it} = 1) = \beta_0 + b_i + \beta_T \cdot \mathrm{Treatment}_{it}$, where the random effects

b_i follow a normal distribution and where $\text{Treatment}_{it} \in \{0, 1\}$ codes the treatment in hospital i.
The subject-specific model with varying intercepts has the form

$$\text{logit}P(y_{it} = 1) = \beta_{0i} + \beta_T \cdot \text{Treatment}_{it},$$

$i = 1, \ldots, 22$ Centers, $t \in \{\text{control}, \text{test}\}$, where β_i are fixed unknown parameters. To identify clusters of hospitals, parameters β_i are penalized based on penalty (3) but with adaptive weights. In the resulting estimates, basically five clusters of hospitals are distinguished in terms of the basic risk captured by the intercepts while subtle distinctions between the centers are still present.
As alternative approach, we fitted the finite mixture model of Grün and Leisch (2008) with adjusted coding and predictor

$$g(\mu_{it}) = \beta_{0m} + \beta_T \cdot Treatment_{it},\ i = 1, \ldots, 44 \text{ Cases},$$

where $m \in \{1, \ldots, K\}$ refer to the partition of the 22 centers into K groups. The predictor contributes to a mixture likelihood that is maximized by an EM algorithm. We assume 3 and 5 clusters. The clusters obtained by regularization show the same structure as in the finite mixture models.
The fitted treatment effect has approximately the same size in all models. However, only the regularization approach combines data driven clustering with stable results.

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models *Biometrics*, **55**, 117 – 128.

Grün, B. and Leisch, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, **25**, 225 – 247.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267 – 288.

Tutz, G. (2012). *Regression for categorical data*. New York: Cambridge University Press.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. New York: Springer.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418 – 1429.

Weight smoothing models to estimate survey estimates from binary data

Yannick Vandendijck[1], Christel Faes[1], Hens Niel[1,2]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[2] Centre for Health Economic Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

E-mail for correspondence: yannick.vandendijck@uhasselt.be

Abstract: In surveys, when the number of respondents in a post-stratum is small relative to the population size in that post-stratum, post-stratification weights are inflated and modifications are required to obtain less variable estimates. Weight smoothing models, random-effects models that induce shrinkage across post-stratum means, are such modifying methods. We describe the empirical Bayes weight smoothing model approach to estimate the overall mean of a binary survey outcome. The generalized linear mixed model formulation of this model allows easy fitting. Two extensions of the model are presented. The estimation of the prevalence and incidence trend of influenza-like illness using the Great Influenza Survey in Flanders, Belgium, is considered as an application.

Keywords: Binary data; Empirical Bayes; Post-stratification; Random-effects.

1 Introduction

Stratification is the process of dividing the population into homogeneous mutually exclusive strata before sampling to improve the representativeness of the sample. In observational studies post-stratification can be used to correct for known differences between the obtained sample and the population. This is done by equating the distribution of a secondary variable (*e.g.*, age) measured in the sample with its distribution in the population, and adjusting estimates using weighting techniques. This can improve both the accuracy and precision of estimates (Little, 1991).
Let Y denote a binary survey outcome variable and X a discrete post-stratifying variable with H levels and known population distribution. Let N_h and n_h denote the population and sample size in post-stratum h, respectively. We assume that N_h is known. Define $N = \sum_{h=1}^{H} N_h$ and $n = \sum_{h=1}^{H} n_h$. We consider inference for the finite population mean $\bar{Y} = \sum_{h=1}^{H} P_h \bar{Y}_h$, where $\bar{Y}_h$ is the population mean in post-stratum h and $P_h = N_h/N$ is the population proportion in post-stratum h.

An estimate for the population mean is of the form $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} w_{i(h)} y_i$, where $w_{i(h)}$ is the weight of observation i belonging to post-stratum h. The unweighted sample mean, $\bar{y}_{unw}$, is obtained when $w_{i(h)} = 1$ $(\forall i)$, and can be written as $\bar{y}_{unw} = \sum_{h=1}^{H} p_h \bar{y}_h$, where $p_h = n_h/n$ is the sample proportion and $\bar{y}_h$ is the sample mean in post-stratum h. Whenever p_h deviates from its population proportion P_h, the unweighted mean is a biased estimate. The post-stratified mean estimate, $\bar{y}_{ps}$, is obtained when $w_{i(h)} = P_h/p_h$ $(\forall i)$. While $\bar{y}_{ps}$ is an unbiased estimate of $\bar{Y}$, it has greater variance than $\bar{y}_{unw}$. This increase in variance can overwhelm the reduction in bias, so that the post-stratified mean estimate actually increases the mean squared error. This happens especially when some weights are large.
A common approach to deal with this problem is weight trimming. This procedure uses the bias-variance trade-off by introducing some bias in the estimate, but effectively reducing the variance. An alternative model-based strategy is to model the stratum means directly by random-effects. These so-called weight smoothing models make a distributional assumption for the Y_i and use the model to predict the non-sampled values of Y. For a Gaussian survey outcome these models are well explained in literature (see *e.g.*, Elliott and Little, 2000). For a binary survey outcome only the full Bayesian approach has been discussed (Elliott, 2007) in the context of generalized linear regression estimators. We describe the empirical Bayes estimation approach of weight smoothing models for binary data and present two extensions of these models.

2 Weight Smoothing Models for Binary Data

The general form of the weight smoothing models for a binary survey outcome is

$$Y_{i(h)}|p_h \sim \text{Binom}(1, p_h) \quad \text{and} \quad \boldsymbol{\delta}^* \sim \mathcal{N}_H(\boldsymbol{\delta}, \mathbf{D}), \tag{1}$$

where $g(E[Y_{i(h)}|p_h]) = \delta_h^*$, $\boldsymbol{\delta}^* = (\delta_1^*, ..., \delta_H^*)^{\mathrm{T}}$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_H)^{\mathrm{T}}$ are unknown vectors, $\mathbf{D}$ is an unknown $H \times H$ covariance matrix and $g(\cdot)$ is the logit-link function. Under model (1) the weight smoothed estimate of $\bar{Y}$ is

$$\bar{y}_{ws} = E[\bar{Y}|\mathbf{y}] = \frac{1}{N}\sum_{h=1}^{H}\{n_h\bar{y}_h + (N_h - n_h)\hat{\mu}_h\}, \tag{2}$$

where $\hat{\mu}_h = E[\bar{Y}_h|\mathbf{y}] = E[\mu_h|\mathbf{y}] = g^{-1}(\delta_h^*)$. The unweighted and post-stratified mean are obtained as estimates of (2) if $\mathbf{D} \to 0$ and $\mathbf{D} \to \infty$, respectively. We consider four other cases of model (1) (Little, 1991; Lazzeroni and Little, 1998; Elliott and Little, 2000):
(a) Exchangeable random effects (XRE): $\delta_h = \beta$ for all $h, \mathbf{D} = \sigma_D^2 \mathbf{I}_H$.
(b) First order autoregressive (AR1): $\delta_h = \beta$ for all $h, (\mathbf{D})_{ij} = \sigma_D^2 \rho^{|i-j|}$ for $i, j \in \{1, \ldots, H\}$.

(c) Linear (LIN): $\delta_h = \beta_0 + \beta_1 X_h$ for all $h, \mathbf{D} = \sigma_D^2 \mathbf{I}_H$.
(d) Nonparametric (NPAR): $\delta_h = f(X_h)$ for all $h, \mathbf{D} = \sigma_D^2 \mathbf{I}_H$, where f is a nonparametric spline function. We use the approximating thin plate spline family.
All these models can be cast in the generalized linear mixed model (GLMM) framework. The model is fit by pseudo-restricted maximum likelihood estimation based on linearisation. At convergence, estimates of $\hat{\mu}_h$ are obtained and $\bar{y}_{ws}$ can be calculated. Calculation of the variance for $\bar{y}_{ws}$ can be either done analytically or by a bootstrap method.

Extension 1: Assume a binary survey outcome is measured at different time points, and interest is in the estimation of the time trend of the overall mean, namely $\bar{Y}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} Y_{it}$, for $t = 1, ..., T$. At each time point, the unweighted or post-stratified mean can be calculated. However, one can use a smoothed weight approach which exploits the time-trend. The general form of this model is

$$Y_{i(h),t} | p_{ht} \sim \text{Binom}(1, p_{ht}), \ \forall \ t, \quad \text{and} \quad \boldsymbol{\delta}_t^* \sim \mathcal{N}_H(\boldsymbol{\delta}_t, \mathbf{D}). \tag{3}$$

The unknown parameters $\boldsymbol{\delta}_t = (\delta_{h1}, ..., \delta_{hT})^{\mathrm{T}}$ are additively decomposed into $\delta_{ht} = \delta_h + \delta_t$. For δ_h and $\mathbf{D}$ we assume models (a)-(d). For the time trend a non-parametric trend, namely $\delta_t = f_t(t)$, is assumed.

Extension 2: Misspecification in (1) leads to biased estimates for $\hat{\mu}_h$ in (2) and consequently a biased estimate of $\bar{y}_{ws}$. We propose the use of a doubly robust weight smoothed estimate of the form

$$\bar{y}_{ws,dr} = \frac{1}{N} \sum_{h=1}^{H} \left\{ \frac{n_h}{\hat{\pi}_h} \bar{y}_h + (N_h - \frac{n_h}{\hat{\pi}_h}) \hat{\mu}_h \right\}, \tag{4}$$

in analogy with doubly robust estimates in the missing data context. The $\hat{\pi}_h$ represent inclusion probabilities and are calculated using a method that resembles a trimming weights approach.

3 Application

The Great Influenza Survey (GIS) is an observational survey based on the voluntary participation of individuals via the internet aiming at the monitoring of influenza-like illness (ILI). We use data from the Flemish GIS from the 2010-2011 influenza season ($n = 4551$). Interest is in the estimation of the overall prevalence and the incidence trend of ILI. The age distribution of the GIS population is very dissimilar to the overall Flemish population age distribution (Figure 1(a)). Post-stratification weights range from 0.46 to 35.70 (18 age groups of length 5 years as post-strata). The unweighted mean estimate of the prevalence is 5.12% (95% CI: 4.52-5.80%), whereas the post-stratified mean estimate is 7.10% (95% CI: 5.31-9.45%).

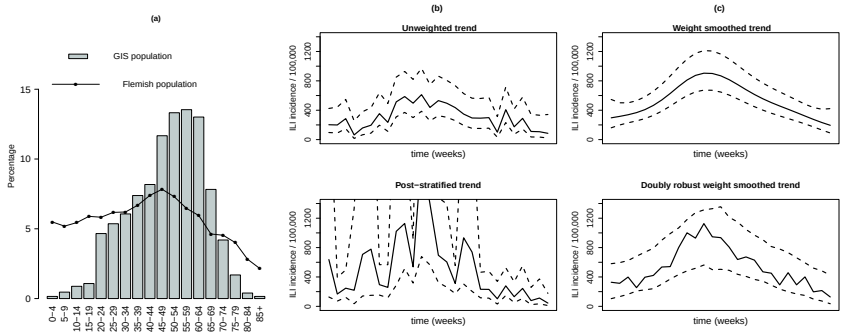


FIGURE 1. (a) Age distribution of the GIS and Flemish population. (b) Estimated unweighted and post-stratified trend with confidence intervals (CIs). (c) Estimated weight smoothed and doubly robust weight smoothed trend with CIs.

The weight smoothed estimate using the NPAR model yields 6.88% (95% CI: 5.69-8.30%). The doubly robust approach for the prevalence yields similar results, namely 6.82% (95% CI: 5.61-8.28%). The results of the incidence trend estimation (using the NPAR model) is shown in Figure 1(b) and Figure 1(c). It is seen that the weight smoothing results are a compromise between the unweighted and post-stratified trends.

4 Discussion

Weight smoothing models offer a good solution for inference of a binary survey outcome when some post-stratification weights are large. These models can be cast into the GLMM framework which allows for implementation in standard statistical software. In the real-life data application it was shown that the different approaches yield substantially different results. It is therefore important to use weight smoothing models in this specific data context.

References

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, **33**, 23 – 34.

Elliott, M.R. and Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, **16**, 191 – 209.

Lazzeroni, L.C. and Little, R.J.A. (1998). Random-effects models for smoothing poststratification weights. *Journal of Official Statistics*, **14**, 61 – 78.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, **7**, 405 – 424.

Symmetric and log-symmetric regression models: a semiparametric approach

Luis Hernando Vanegas[1,2] Gilberto A. Paula[1]

[1] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil
[2] Departamento de Estadística, Universidad Nacional de Colombia, Colombia

E-mail for correspondence: hvanegasp@gmail.com

Abstract: In this paper we propose a semiparametric regression model suitable for data set analysis in which the distribution of the response is strictly positive and asymmetric. In this model we describe the median of the response variable using a nonlinear function and skewness using a nonparametric function (natural cubic spline). We show that the proposed model allows the description of the response using distributions with tails heavier than those of log-normal (i.e., log-Student-t and log-power-exponential), providing the ability to reduce and control the influence of extreme observations in the parameter estimates and inference. A data set previously analyzed under parametric models is reanalyzed under log-symmetric regression models. Diagnostic methods are applied to select an appropriate model.

Keywords: Natural cubic spline, Skewness, Semiparametric regression models.

1 Introduction

Nonlinear regression models are commonly applied in areas such as Biology, Chemistry, Medicine, Economics and Engineering. The analysis based on models under normal errors and constant variance is the most popular when the variable of interest is continuous, due to desirable statistical properties and a comprehensive developed theory. However, the application of such models may be inadequate in some scenarios commonly found in practice. For example, if the response variable distribution is asymmetric, modeling the median instead the mean can provide a better description of the behavior of the response variable. On the other hand, if the response variance is not constant the application of a model under the heteroscedasticity assumption may introduce loss of efficiency in estimation and loss of power in the hypothesis testing. Furthermore, it is well known that modeling under the assumption of normally distributed errors can be highly influenced by extreme observations in the response. With these motivations, in this paper we propose a semiparametric modeling methodology that allows simultaneously dealing with asymmetry and extreme observations in the response variable (or with heteroscedasticity, symmetry and extreme observations),

even when skewness (or dispersion) depends on the explanatory variables in an unknown form. Recent results under homocedastic semiparametric symmetric models are given in Ibacache-Pulgar et al. (2013).

Initially in this paper we present the log-symmetric distribution class as a generalization of the log-normal distribution. We describe some of the properties of this class and list some distributions that are part of it, among them we can mention the log-normal, log-Student-t and log-power-exponential distributions. In the second part of this paper, we propose a semiparametric regression model suitable for data set analysis in which the distribution of the response is strictly positive and asymmetric. In this model, we describe the median of the response variable using a nonlinear function and skewness using a nonparametric function (natural cubic spline as described, for instance, in Green and Silverman, 1994). Furthermore, we show that the proposed model allows the description of the response using distributions with tails heavier than those of log-normal, providing the ability to reduce and control the influence of extreme observations in the parameter estimates and inference. The proposed model can be applied by fitting the transformed response variable (i.e., $Y = \log(T)$, where T is the response variable in the original scale) to a symmetric nonlinear model, where the variance of the random error is not constant and is modeled using a natural cubic spline. In the third part of the paper, we develop some diagnostic measures such as deviance for location (median or mean) as well as for skewness (or dispersion), residual analysis and local influence analysis under various perturbation schemes (see Cook, 1986). Finally, the proposed methodology is illustrated by applying it to a data set previously analyzed under parametric models.

1.1 Ultrasonic calibration

These data, analyzed previously by Lin *et. al* (2009) and Lachos *et. al* (2011), consists of 214 observations generated in an ultrasonic calibration study, in which the response variable, T, is the ultrasonic response and the explanatory variable, X, is the distance to the metal. In the *boxplots* of T versus the X values (omitted here) we observed that the ultrasonic response decreases with the distance from the metal and the skewness of the response distribution depends (in an unknown manner) on the distance to the metal, which motivates the data set to be described using a nonlinear function for the median and a nonparametric function for the skewness.

2 Model formulation

We assume that $T_1, \ldots, T_n$ is a set of n independent random variables representing the measure T performed on n individuals or experimental units. It is assumed that T_k can be written in the following form:

$$T_k = \eta_k \, \xi_k^{\sqrt{\phi_k}}, \qquad k = 1, \ldots, n, \tag{1}$$

where

η_k : median of the T_k distribution,
ξ_k : multiplicative random error affecting η_k,
ϕ_k : skewness (positive) of the T_k distribution,

and $\xi_1,\ldots,\xi_n$ is a set of independent random variables following distribution $\xi_k \sim \mathrm{LS}(1, 1, g(\cdot))$, which implies that $\xi_k^{\sqrt{\phi_k}}$ has unit median and that the probability density function of ξ_k is given by

$$f(\xi, g(\cdot)) \propto \frac{1}{\xi} g[\log^2(\xi)], \qquad \xi > 0, \tag{2}$$

with $g(\cdot)$ being a density generator function such as $g(u) > 0$ for $u > 0$ and $\int_0^\infty u^{-1/2} g(u) \partial u < \infty$. From (1) and (2) we have the T distribution belonging to the log-symmetric class denoted by $\mathrm{LS}(\eta, \phi, g(\cdot))$. For example, using $g(u) = \exp(-u/2)$, $g(u) = (1 + u/\nu)^{-\frac{\nu+1}{2}}$ and $g(u) = \exp[-u^{1/(1+\kappa)}/2]$, we have that the random error ξ follows log-normal, log-Student-t (with ν degrees of freedom) and log-power-exponential (with shape parameter $-1 < \kappa < 1$) distributions, respectively. It is further assumed that η_k and ϕ_k can be written as

$$\begin{cases} \log(\eta_k) = \mu_k = \boldsymbol{\mu}(\mathbf{x}_k; \boldsymbol{\beta}), & k = 1, \ldots, n, \\ \log(\phi_k) = \sum_{r=1}^{R} \gamma_r(b_{kr}), \end{cases}$$

where $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^{\mathrm{T}}$ and $\mathbf{b}_k = (b_{k1}, \ldots, b_{kR})^{\mathrm{T}}$ are vectors of explanatory variable values for η_k and ϕ_k, respectively, with μ_k being a continuous and twice differentiable function of $\boldsymbol{\beta}$ and $\gamma_r(\cdot)$, $r = 1, \ldots, R$, a nonparametric function (natural cubic spline).

Using the properties of log-symmetric distributions we can write the model (1) in the following way:

$$\underbrace{\log[T_k]}_{Y_k} = \underbrace{\log[\eta_k]}_{\mu_k} + \sqrt{\phi_k} \underbrace{\log[\xi_k]}_{e_k},$$

where

μ_k : mean of the Y_k distribution,
e_k : additive random error affecting μ_k,
ϕ_k : dispersion parameter of the Y_k distribution,

with $-\infty < \mu_k < \infty$ and $\phi_k > 0$ being the location and dispersion parameters of Y_k, respectively, and $e_1, \ldots, e_n$ being a set of independent random variables having a probability density function given by $f(e) \propto g(e^2)$ for $-\infty < e < \infty$. The Y distribution belonging to the symmetric class will be denoted denoted by $\mathrm{S}(\mu, \phi, g(\cdot))$.

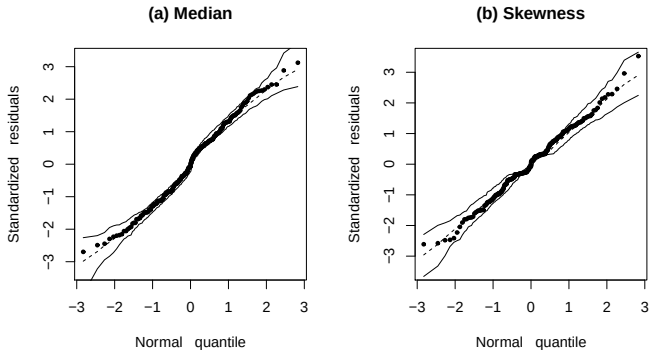


FIGURE 1. Ultrasonic calibration data: normal probability plots with simulated envelopes for the fitted model under log-power-exponential ($\kappa = 0.55$) errors.

3 Ultrasonic calibration

We fitted the model (1) to the ultrasonic calibration data set for an appropriate nonlinear function $\boldsymbol{\mu}(\mathbf{x}_k; \boldsymbol{\beta})$ under log-normal, log-Student-t and log-power-exponential errors. Based on residual analysis and sensitivity studies we noticed the the log-power-exponential model with $\kappa = 0.55$ seems to present the better fit. Figures 1a-1b describe the normal probability plots with generated envelope for standardized residuals corresponding to the fitted mean and fitted skewness, respectively. We observe no unusual features neither outlying observations in both graphs.

Acknowledgments: The authors are grateful to CNPq and CAPES, Brazil, for the financial support.

References

Cook, R.D. (1986). Assessment local influence (with discussion). *Journal of the Royal Statistical Society B*, **48**, 133 – 169.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall: Boca Raton.

Ibacache-Pulgar, G., Paula, G.A., and Cysneiros, F.J.A. (2013). Semiparametric additive models under symmetric distributions. *Test* (to appear).

Lachos, V.H., Bandyopadhyay, D. and Garay, A.M. (2011). Heteroscedastic nonlinear regression models based on scale mixtures of skew-normal distributions. *Statistics & Probability Letters*, **81**, 1208 – 1217.

Lin, J.G., Xie, F.C., and Wei, B.C. (2009). Statistical diagnostics for skew-t-normal nonlinear models. *Communications in Statistics. Simulation and Computation*, **38**, 2096 – 2110.

A Score Test for Zero-adjusted Effect in Claim Severity Modeling

Alicja Wolny-Dominiak[1]

[1] University of Economics in Katowice, Poland

E-mail for correspondence: alicja.wolny-dominiak@ue.katowice.pl

Abstract: Classification ratemaking in non-life insurance is focus on the risk segmentation via rating variables and outlining criteria to consider when using a certain risk characteristics as a rating variables. At present, the typical statistical technique used in such ratemaking is the generalized linear model GLM. In this paper we consider multiplicative model for estimating claim severity. As the insurance data are usually non-negative and skewed to the right, we assume gamma distribution. Under this assumption consequently in estimation no claims policies are omitted. However, in classification ratemaking the valuable information is "what is going on in the group of non-claims policies", e.g. if policyholders do not no report of small claims, what is not cost-effective because of the bonus-malus system or they really do not cause claims. That is why it is reasonably to assume modified gamma distribution with zero-adjusted effect proposed by (Rigdby and Stasinopoulos (2007)). The goal of this paper is to obtains the score test for testing the significance of this effect.

Keywords: classification ratemaking; zero-adjusted effect; score test.

1 Introduction

Important part of data analysis in insurance business is the construction of a fair tariff structure called classification ratemaking. The goal of this classification is partition all policies in particular portfolio into homogeneous classes. Within every class, all policyholders pay the same premium. To design classification rating plans, actuaries use the generalized linear models (GLM) technique. In GLM model, the dependent variable Y_g is usually the claim severity or the claim frequency. In the paper we focus on the claim severity defined as the total claim amount divided by the number of claims. As Y_g variable is positive and skewed to the right, we assume gamma distribution. The rating variables $X_1, \ldots, X_p$ are usually categorical with few categories like e.g. gender or a large number of categories like e.g. spatial variables. Consider the independent observations $y_1, \ldots, y_n$ of

claim severity with gamma density of the form

$$f\left(y_i \mid \mu_i, \nu\right) = \left(\frac{\nu}{\mu_i}\right)^{\nu} \frac{y_i^{\nu-1} e^{-\frac{\nu y_i}{\mu_i}}}{\Gamma(\nu)}, \quad y_i > 0,$$

where $E(Y_g) = \mu_i$ and $V(Y_g) = \frac{{\mu_i}^2}{\nu}$. Assume $\eta_i = \sum_i \beta_i X_i$ be the linear predictor of rating variables connected with μ_i via log-link function. Then the structure of GLM model, according to McCullagh and Nelder (1989) notation, is as follows: link function – $\eta_i = \ln(\mu_i)$, variance function – $Var(y_i) = \frac{\mu_i^2}{t_i}$, t_i – prior weigh (exposure) and a dispersion parameter $\phi = \frac{1}{\nu}$. However this GLM model for claim severity is only for positive claim size, so non-claim policies in portfolio are omitted. Taking the exposure as the duration of a policy measured in year, we assume $t_i = 1$, $i = 1, \ldots, n$. However this GLM model for claim severity is only for positive claim size, so non-claim policies in portfolio are omitted. The possibility to take those policies into account is modified the gamma distribution as in (Heller, Stasinopoulos et al., 2007). In this modified distribution zero-adjusted effect is added. Suppose ϖ_i is the probability of zero claim y_i. Than the density is defined by

$$f_{ZA}\left(y_i \mid \mu_i, \varpi_i, \nu\right) = \begin{cases} \varpi_i, & \text{if } \; y_i = 0 \\ \left(1 - \varpi_i\right) f\left(y_i \mid \mu_i, \nu\right), & \text{if } \; y_i > 0 \end{cases} \tag{1}$$

Taking into account that the parameter ϖ_i is for $y_i = 0$ and does not affect the expected value, we have $E(Y) = (1 - \varpi_i)\mu_i$. From the same reason the general form of variance is $Var(Y) = (1 - \varpi_i) = (1 - \varpi_i)E({Y_g}^2) - E(Y)^2$. As $E({Y_g}^2) = \mu_i(\frac{1}{\nu} + 1)$, finally we obtain $V(Y) = (1 - \varpi_i){\mu_i}^2(\frac{1}{\nu} + 1) - (1 - {\varpi_i}^2){\mu_i}^2) = (1 - \varpi_i){\mu_i}^2(\varpi_i + \frac{1}{\nu})$. We assume log-link and logit-link functions for covariates

$$\eta_i = \ln(\mu_i), \quad \eta_i^{\varpi} = \ln\left(\frac{\varpi_i}{1 - \varpi_i}\right), \tag{2}$$

where $\eta_i^{\varpi} = \sum_{i=1}^{r} \gamma_i Z_i$ is the linear predictor of variables $Z_1, \ldots, Z_r$ influence the occurrence of claim. The set of rating variables affect the mean of claim severity and the occurrence of claim may be the same or different. The zero-adjusted effect is similar to the zero-inflation effect, which appears in claim frequency modeling. The difference between those two effects, sometimes used in literature incorrectly interchangeable, is that in the first case there is a possibility to draw a zero value from Poisson distribution, so there are two kinds of zeros in data structural zeros and sampling zeros. But in case of gamma distribution, zeros are simply added to non-zero data. Thus testing the zero-inflation effect can be interpreted as testing if there are sampling zeros in data. In case of zero-adjusted effect, the test shows if there are significant fraction of zeros in data. If this fraction is insignificant, it is reasonable to skip no claim policies in classification ratemaking.

2 The score test for zero-adjusted effect

In further calculations we apply gamma density (1) in exponential form

$$f\left(y_i|\mu_i,\nu\right)=e^{\left(\frac{-y_i}{\mu_i}-\log\mu_i\right)\nu+(\nu-1)\log(y_i)+\nu\log(\nu)-\log(\Gamma(\nu))}. \tag{3}$$

Only the case with no covariates for ϖ is under consideration, so this parameter is treated as constants value, $0<\varpi<1$. The log-likelihood function is as follows

$$\begin{aligned} l(\mu_i,\gamma,\nu;y_i) &= \sum_{i=1}^{n}\ln\left[\varpi+(1-\varpi)f(y_i\mid\mu_i,\nu)\right]= \\ &= \sum_{i=1}^{n}\left[\ln(1-\varpi)+\frac{\varpi+(1-\varpi)}{(1-\varpi)}f(y_i\mid\mu_i,\nu)\right]= \\ &= \sum_{i=1}^{n}\left[-\ln(1+\gamma)+\ln\left(\gamma+f(y_i\mid\mu_i,\nu)\right)\right], \end{aligned} \tag{4}$$

where $\gamma=\frac{\varpi}{1-\varpi}$. We test the null hypothesis $H_0:\varpi=0$ by using the score test (Cox and Hinkley (1974)). The score statistics $S\left(\beta,\varpi,\nu\right)$ is define by

$$S\left(\beta,\varpi,\nu\right)=U^{\mathrm{T}}\left(\beta,\varpi,\nu\right)I^{-1}\left(\beta,\varpi,\nu\right)U\left(\beta,\varpi,\nu\right), \tag{5}$$

where $U_{(p+2)}$ and $I_{(p+2)\times(p+2)}$ denote $(p+2)$ respectively the score function and Fisher information matrix. Differentiating the log likelihood (5) with respect to $\beta_1,\ldots,\beta_p,\gamma$ and ν gives:

$$\frac{\partial l}{\partial\beta_r} = \sum_{i=1}^{n}\frac{f(y_i|\mu_i,\nu)}{\gamma+f(y_i|\mu_i,\nu)}\left[\frac{y_i}{\mu_i}-1\right]x_{ir}\nu,\quad r=1,\ldots,p, \tag{6}$$

$$\frac{\partial l}{\partial\gamma} = \sum_{i=1}^{n}\frac{1-f(y_i|\mu_i,\nu)}{(1+\gamma)\left(\gamma+f(y_i|\mu_i,\nu)\right)}, \tag{7}$$

$$\frac{\partial l}{\partial\nu} = \sum_{i=1}^{n}\frac{f(y_i|\mu_i,\nu)}{\gamma+f(y_i|\mu_i,\nu)}\left[-\frac{y_i}{\mu_i}+\log\left(\frac{y_i\nu}{\mu_i}\right)+1-\left(\log\left(\Gamma(\nu)\right)\right)'\right] \tag{8}$$

Taking MLE estimates $\widehat{\beta}$ and $\widehat{\mu}_i$, under the null hypothesis $\varpi=0$ derivatives (6), (7), (8) simplify and the score function partitioned on two vectors:

$$U\left(\widehat{\beta},0,\widehat{\nu}\right)=\left[\widehat{U}_1,\widehat{U}_2\right] \tag{9}$$

has block vectors: $\widehat{U}_1=[0]_{p\times1}$ and
$\widehat{U}_2=\left[\sum_{i=1}^{n}\left[\frac{\hat{\mu}_i\Gamma(\hat{\nu})}{\hat{\nu}^{\hat{\nu}}e^{-\hat{\nu}}}-1\right]\quad\sum_{i=1}^{n}\left[\log\left(\hat{\nu}\right)+1-\left(\log\left(\Gamma\left(\hat{\nu}\right)\right)\right)'\right]\right]^{\mathrm{T}}$.

Now, partition Fisher information matrix $\widehat{I}\left(\widehat{\beta},0,\widehat{\nu}\right)$, we have

$$\widehat{I}\left(\widehat{\beta},0,\widehat{\nu}\right)=\begin{bmatrix}\left[\hat{I}_{11}\right]_{p\times p} & \left[\hat{I}_{12}\right]_{p\times 2}\\ \left[\hat{I}_{12}^{\mathrm{T}}\right]_{2\times p} & \left[\hat{I}_{22}\right]_{2\times 2}\end{bmatrix}_{(p+2)\times(p+2)}.$$

To inverse the matrix $\widehat{I}\left(\widehat{\beta},0,\widehat{\nu}\right)$, we use the Schur complement yield:

$$\widehat{I}^{-1}\left(\widehat{\beta},0,\widehat{\nu}\right)=\begin{bmatrix}E^{-1} & -E^{-1}\hat{I}_{12}\hat{I}_{22}^{-1}\\ \hat{I}_{22}^{-1}\hat{I}_{12}^{\mathrm{T}}E^{-1} & \hat{I}_{22}^{-1}+\hat{I}_{22}^{-1}\hat{I}_{12}^{\mathrm{T}}E^{-1}\hat{I}_{12}\hat{I}_{22}^{-1}\end{bmatrix}, \tag{10}$$

where $[E]_{p\times p}=\hat{I}_{11}-\hat{I}_{12}\hat{I}_{22}^{-1}\hat{I}_{12}^{\mathrm{T}}$. Substituting (9) and (10) to (5) the score statistics has a form:

$$S\left(\widehat{\beta},0,\widehat{\nu}\right)=\widehat{U}_2^{\mathrm{T}}\left(\hat{I}_{22}^{-1}+\hat{I}_{22}^{-1}\hat{I}_{12}^{\mathrm{T}}E^{-1}\hat{I}_{12}\hat{I}_{22}^{-1}\right)\widehat{U}_2\,. \tag{11}$$

Suitable Hessian matrix and Fisher information matrix is calculated in the appendix not included in this paper (contact author). In ongoing studies we accomplish a simulation study to check if the score statistics (11) corresponds to the chi-square approximation.

Acknowledgments: This research is supported by the grant of Polish Ministry of Science and Higher Education (nr NN 111461540).

References

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

de Jong, P. and Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models. Second Edition*. London: Chapman & Hall.

Rigby, R.A. and Stasinopoulos, D.M. (2007). *Generalized Additive Models for Location Scale and Shape (GAMLSS) in R*. Journal of Statistical Software, Volume 23, Issue 7.

Hidden Markov modelling of diffusion with an application in entomology

Bruce J. Worton[1], Chris R. McLellan[1]

[1] School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, Edinburgh, UK

E-mail for correspondence: Bruce.Worton@ed.ac.uk

Abstract: We investigate the use of Hidden Markov modelling of two-dimensional point data with an attraction towards a known location. The hidden states of the models are based on features of the data of primary interest as well as other characteristics that arise from limitations of the equipment used in the experimental study. Posterior distributions of the parameters corresponding to the features of primary interest are used to compare data collected under different conditions. An application to an extensive data set from an entomological study illustrates the advantages of using complex Hidden Markov models.

Keywords: Diffusion processes; Hidden Markov modelling; Model comparison.

1 Introduction

Consider an observed bivariate diffusion process, in which the initial location, $\mathbf{x}_0$, is known and the subsequent path, $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, is determined by various component processes. To be specific, we have a global attraction to a point $\mathbf{a}$ together with more localized features and these may be represented by processes

(P1) $\mathbf{X}_{t+1}|\mathbf{X}_t \sim \mathrm{N}\{\mathbf{a} + \boldsymbol{\Gamma}(\mathbf{X}_t - \mathbf{a}), \boldsymbol{\Phi}\}$,

(P2) $\mathbf{X}_{t+1}|\mathbf{X}_t \sim \mathrm{N}(\mathbf{X}_t, \boldsymbol{\Sigma})$.

Although we could use mixture models to switch between states, Hidden Markov modelling (HMM) offers a more appropriate and flexible approach (Frühwirth-Schnatter, 2006). In the application in Section 3 the global process (P1) is attraction towards or repulsion from a particular chemical of interest, and the localized process (P2) is associated with small movements related to the way the position of the experimental subject was observed. Of course, we may include further processes, (P3), (P4), ..., to account for other features.

2 Hidden Markov modelling

We initially considered a single state model for the data introduced in Section 3 based on (P1) alone. In this case it is possible, for a particular prior distribution, to obtain an explicit posterior distribution for the parameters of the model (Tiao and Zellner, 1964). However, we found that important features of the data could not be adequately accounted for using this approach as the model was oversimplistic. We therefore adopted models based on (P1) and (P2) above, with the inclusion of a variety of other processes. The following (independent) prior distributions were used,

$$\begin{aligned}(i,j) \text{ element of } \boldsymbol{\Gamma} &\sim \text{Normal}(1, 10^2), \quad i,j = 1,2,\\ \boldsymbol{\Phi} &\sim \text{Inverse-Wishart}\left(10^{-5}\mathbf{I}, 2\right),\\ \boldsymbol{\Sigma} &\sim \text{Inverse-Wishart}\left(10^{-5}\mathbf{I}, 2\right), \end{aligned} \tag{1}$$

and represent vague prior information (Leonard and Hsu, 1999). The prior distributions of π_{ij}, the elements of a probability transition matrix $\boldsymbol{P}$ of moving between states for the HMMs, were

$$(\pi_{i1}, \ldots, \pi_{im}) \sim \text{Dirichlet}\left(\frac{1}{2}, \ldots, \frac{1}{2}\right), \quad i = 1, \ldots, m, \tag{2}$$

where m is the number of states. Again these represent vague prior information. Posterior distributions were estimated with a Markov chain Monte Carlo approach using WinBUGS/OpenBUGS (Lunn et al., 2000).

3 Application

In this section we consider the application of Hidden Markov modelling to observed diffusion path data collected in a research project concerned with developing approaches to pest management. In this study, it was of interest to investigate chemicals to which the larvae of a pest are attracted or repelled, and assess the level of attraction or repulsion. The locations of a larva were sampled at a rate of 5 per second for 30 minutes. The point $\mathbf{a}$ in (P1) is the position of the chemical being studied, and is at a given distance in the y direction from the initial point $\mathbf{x}_0$.

As the off-diagonal elements of $\boldsymbol{\Gamma}$ are approximately zero, the (2,2) element of $\boldsymbol{\Gamma}$, i.e. parameter γ_{22}, of fitted Hidden Markov models may be used to quantify the features of interest. Furthermore, posterior densities provide useful information on how the attraction and repulsion differ under different experimental conditions. Table 1 presents the summary statistics of estimated posterior distributions for a three-state HMM. The third state considered here represents no change in location (due to intense sampling),

(P3) $\mathbf{X}_{t+1}$ is the same as $\mathbf{X}_t$,

and we have a 3×3 probability transition matrix

$$\boldsymbol{P} = \begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \\ \pi_{31} & \pi_{32} & \pi_{33} \end{pmatrix}.$$

The estimated posterior densities of the parameter γ_{22} are shown in Figure 1 for two experiments under different conditions, one with an attractant (Experiment I) and one with a repellent (Experiment II). Clearly, the posterior probability of being above unity is negligible for the former while being substantial for the latter.

4 Discussion

Model comparison using BIC indicates that the three-state model is far superior to a single state model based on (P1), or two-state HMM or mixture

TABLE 1. Summary statistics of estimated posterior distributions for the elements of the parameters $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{P}$ of a three-state HMM for Experiment I (test attractant chemical) and Experiment II (test repellent chemical). Only results for diagonal elements of the matrices are shown. Priors (1) and (2) used.

		Posterior summary statistics				
Experiment	**Parameter**	**Mean**	**Median**	**SD**	**2.5%**	**97.5%**
I	γ_{11}	0.9986	0.9986	0.0012	0.9964	1.0010
	γ_{22}	0.9998	0.9998	0.0003	0.9992	1.0000
	$\phi^{\dagger}_{11}$	3.4421	3.4360	0.1492	3.1760	3.7561
	$\phi^{\dagger}_{22}$	0.5811	0.5804	0.0249	0.5334	0.6318
	$\sigma^{\ddagger}_{11}$	0.1168	0.1167	0.0056	0.1063	0.1280
	$\sigma^{\dagger}_{22}$	3.4310	3.4240	0.1687	3.1280	3.7900
	π_{11}	0.2850	0.2847	0.0136	0.2584	0.3122
	π_{22}	0.2715	0.2713	0.0152	0.2427	0.3016
	π_{33}	0.8330	0.8330	0.0043	0.8244	0.8413
II	γ_{11}	1.0000	1.0000	0.0001	0.9998	1.0000
	γ_{22}	1.0021	1.0020	0.0031	0.9948	1.0080
	$\phi^{\ddagger}_{11}$	0.6232	0.6182	0.0711	0.5004	0.7806
	$\phi^{\dagger}_{22}$	2.6028	2.5890	0.2917	2.0394	3.2302
	$\sigma^{\dagger}_{11}$	3.1366	3.1025	0.3898	2.4370	3.9681
	$\sigma^{\ddagger}_{22}$	0.8225	0.8167	0.1035	0.6434	1.0411
	π_{11}	0.3295	0.3289	0.0364	0.2592	0.4017
	π_{22}	0.3626	0.3616	0.0483	0.2800	0.4450
	π_{33}	0.8107	0.8110	0.0130	0.7848	0.8347

† Values multiplied by 10^4.
‡ Values multiplied by 10^7.

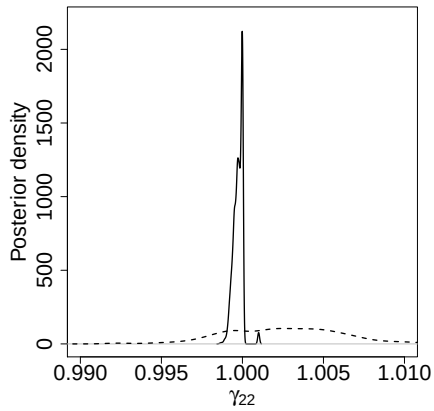


FIGURE 1. Estimated posterior densities of γ_{22}, the (2,2) element of $\mathbf{\Gamma}$, for Experiment I (solid line) and Experiment II (dashed line). Priors (1) and (2) used.

models based on (P1) and (P2). Inclusion of further states is possible, as is increasing the order of dependence for the state (P1), but similar conclusions are reached concerning the parameter of primary interest; posterior densities obtained from these models are similar to those shown in Figure 1.

Acknowledgments: Chris McLellan is supported by an EPSRC studentship. We are particularly grateful to William Deasy and Dr. A.N.E. Birch of The James Hutton institute, Invergowrie, Dundee, for kindly providing the experimental data.

References

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

Leonard, T. and Hsu, J.S.J. (1999). *Bayesian Methods.* Cambridge: Cambridge University Press.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS — A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

Tiao, G.C. and Zellner, A. (1964). On the Bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society, Series B*, **26**, 277–285.

Modelling Children's Journeys in Care: A Multistate Modelling Approach

Emily Yeend[1], Deborah Costain[1], Karen Broadhurst[2]

[1] Dept of Mathematics and Statistics, Lancaster University, UK
[2] Dept of Applied Social Science, Lancaster Univeristy, UK

E-mail for correspondence: e.yeend@lancaster.ac.uk

Abstract: The "journeys" of children in public care has been an issue of concern for some time. Whilst a number of studies have explored specific features of children's journeys, none have incorporated the sequence of placements and legal statuses experienced, the time spent in each, and the reasons for children subsequently exiting care. This study uses these features *together* to develop understanding of children's routes through care. A model was fitted to the first two "events" in children's journeys, yielding insight into factors associated with particular care pathways and into future model development.

Keywords: Children in Care; Administrative Data; Multistate Models.

1 Introduction

A recent Child Protection system review in England argued that the current system is not sufficiently child centered, and that performance management's target-focus fails to understand children's time in care holistically (Munro,2010, 2011a & 2011b). Research can be similarly critiqued; while a number of studies have explored individual features of children's journeys (eg Khoo et al, 2012; Akin, 2011; McSherry et al, 2010; Ward, 2009; Sinclair et al, 2007; Wulczyn et al, 2003; Webster et al, 2000; Rowe et al, 1989), few have explored movement through care explicitly; modelling transitions from one situation to another with regards their *nature* and *duration*. Existing longitudinal studies have utilised simplistic tables of placement orderings (Usher, 1999), flow diagrams, (Schofield et al, 2007), and qualitatively identified patterns in journeys (James et al, 2004). One study utilised competing risks modelling but was limited to contrasting restoration home and a subsequent placement (Fernandez, 1999). Moreover, legal status was typically not a focal point, yet a recent court judgment highlighted that legal status does have welfare consequences (A and S v Lancashire County Council, 2012). This study responds to increased interest in the journeys of children in care: seeking to explore and model placements, legal statuses and exits experienced over time, and incorporating potentially influential factors.

2 Data

The data came from records held by one England LA. Information was provided on all 1081 children who entered care for the first time between the 1st April 2006 and the 31st March 2009, with follow-up information capped at four years. The cohort was chosen to yield a representative sample of adequate size and ensure a reasonable period of follow-up. Data capture included information on the nature of placements and legal statuses over time and, where relevant, the circumstances of children's exits from care. Demographic information was extracted in addition to the reason for care being required and the district team responsible for the child's case.

Profile plots were used to visualise children's placement or legal journeys and motivate the modelling strategy. An example is shown in Figure 1: the graph shows the legal journeys of all children who entered care under a voluntary (S20) agreement; superimosed dots highlight the time-point children either exited care or were censored due to being in care at the end of the study. A S20 agreement occurs when parents agree to their child being accommodated by the LA, requires no court involvement, and can be terminated at any time by the parents (Children Act, 1989; Allen, 2005; Brammer, 2010). By contrast, all other legal routes require court involvement and effect *parental responsibility*. Whilst around 60% of chidren enter care under S20 only around 30% of the care population are in care under S20 (Department for Education, 2011). One explanation is that those starting on S20 exit quicker than those entering compulsorily. However, Figure 1 indicates a number of children actually switch from S20 to compulsory care. While there was variation in the legal status they progressed to and the timing of their transitions, those who transitioned tended to spend slightly longer in care than those who did not.

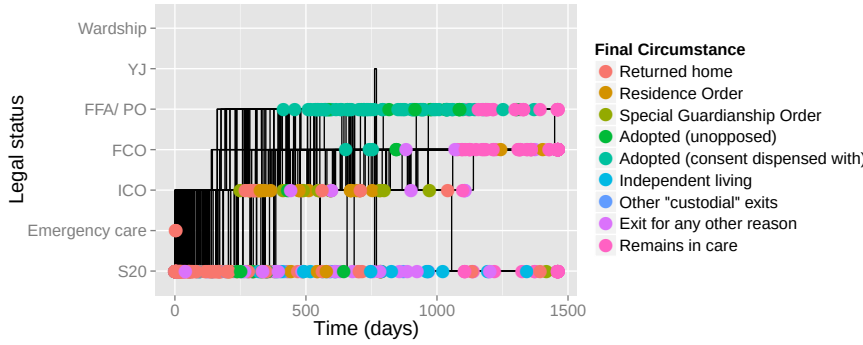


FIGURE 1. *Profile plot visualising the legal journeys of all children who entered care under a voluntary (S20) agreement.*

3 Modelling Approach

Since interest lies in modelling the *times* to *specific events* a Multistate approach was taken. As an extention of standard Survival Analyis, this technique can handle censoring, which features for children who have no events or who remain in care at the end of folllow-up, and can incorporate potentially influential factors (Putter et al, 2007; Hougaard, 1999).
The technique models the *instantaneous hazard* of transitioning from the current state, $X_{t^-} = m$, to the destination state, $D = k$, at time t. This hazard is given as

$$h_{mk}(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t, D = k | T \geq t, X_{t^-} = m, F_{t^-})}{\delta t}$$

where the numerator is the probability of transitioning from state m to k between time point t and $t + \delta t$, for δt tending to zero, given that the transition has not already occurred pior to time point t, and given the child's care history, F_{t^-} (Putter et al, 2007; Hougaard, 1999). The Cox-Proportional Hazard model in the multistate senario:

$$h_{imk}(t) = e^{\boldsymbol{\beta}_{mk}^{\mathrm{T}} x_i} h_{omk}(t)$$

assumes the hazard of individual i transitioning from state m to k at time t, $h_{imk}(t)$, is proportional to the *baseline hazard*, $h_{0mk}(t)$, of the same transition. $e^{\boldsymbol{\beta}_{mk}^{\mathrm{T}} x_i}$, termed the *hazard ratio* (HR), represents the quantity by which the hazard of transitioning from state m to k at time t for an individual i, with certain characteristics x_i, is increased (or reduced) beyond that of the baseline hazard (Putter et al, 2007).

4 Preliminary Results

A multistate model of time to the first two events was fitted using the software R. Possible events were change in placement, change in legal status, change in both, and, exit from care via, for example, Reunification with Parents or Adoption. Explanatory variables were gender, financial year of entry and starting placement and legal status, which were included based on forward selection using the Likelihood Ratio Test. Due to restrictions on space only a selection of the results have been reproduced here (Table 1), focusing on transitions *out of* the care system through reunification with parents and transitions *within* in terms of placement change.
Children whose starting legal status was *emergency measures* were more likely to be reunified than children who entered under S20. This indicates that despite serious legal involvement, reunification was not ony possible but more likely than when parents agreed to their child entering care. Children placed with kin on entry were less likely to experience reunification,

and, less likely have a placement move, than those placed with foster carers inside the LA, suggesting kin care is being used as a successful alternative permanence option to reunification.
Interestingly, factors associated with reunification or placement change *following entry* differ to those associated with these events *following a placement change*. In particular, the effect of entering care due to emergency measures on returning home did not remain after the child had experienced a placement change. Intriguingly, initially being placed with parents increased a child's chances of experiencing a second placement compared with that of foster care inside the LA. A gender bias was also identified for experieincing a second placement.

TABLE 1. *Results for time-to reunification and change of placement from the multistate model of the first and second events to occur after entry to care.*

Event Predictor	**HR (95% CI)**
Reunification as the first event after entry to care	
Start legal status: Emergency Measures v Voluntary (S20)	6.25 (4.23, 9.23)
Start legal status: Youth Justice (YJ) v Voluntary (S20)	3.93 (2.35, 6.55)
Start legal status: Interim Care Order (ICO) v Voluntary (S20)	0.16 (0.10, 0.26)
Start placement: Kin care v Foster care (inside LA)	0.50 (0.27, 0.89)
Start placement: Other v Foster care (inside LA)	0.23 (0.08, 0.62)
Remain in care but experience a change in placement	
Start placement: Kin care v Foster care (inside LA)	0.27 (0.17, 0.43)
Start placement: Foster care (outside LA) v Foster care (inside LA)	0.36 (0.20, 0.65)
Reunification after experiencing a change in placement	
Start legal status: Youth Justice (YJ) v Voluntary (S20)	5.50 (2.22, 13.61)
Start legal status: Interim Care Order (ICO) v Voluntary (S20)	0.41 (0.17, 1.00)
Remain in care after experiencing a second change in placement	
Gender: Male v Female	1.51 (1.08, 2.13)
Start legal status: Interim Care Order (ICO) v Voluntary (S20)	1.63 (1.02, 2.61)
Start legal status: Emergency measures v Voluntary (S20)	0.13 (0.02, 0.94)
Start placement: Home or hostel v Foster care (inside LA)	3.43 (2.02, 5.83)
Start placement: Parents v Foster care (inside LA)	1.77 (1.11, 2.83)

5 Discussion

This study makes use of administrative data, routinely collected on children in care in England for performance monitoring, but rarely analysed extensively using sophisticated modelling techniques accounting for its longitudinal nature. This study has began to explore children's journeys in care with regards to placement and legal status and has started to yield insight into care pathways. Future work will incorporate additional explanatory factors and extend modelling to subsequent events.

References

Hougaard, P. (1999). Multi-state Models: A review. *Lifetime Data Analysis*, **5**, 239 – 264.

Putter, H., Fiocco, M. and Geskus, R.B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389 – 2430.

Index

Abbruzzo, Antonino, 59
Abrams, Steven, 475
Adelfio, Giada, 65, 479
Adler, Robert J., 71
Aerts, Marc, 191, 307
Amorim, Leila, 115
Andrade, Mercedes, 595
Ankinakatte, Smitha, 77
Araújo, Mariana C., 547
Arima, Serena, 485
Armero, Carmen, 83
Azorit, Concepción, 659

Bühlmann, Peter, 15
Bai, Jiawei, 3, 489
Bailey, Marc A., 497
Bani-Mustafa, Ahmed S., 493
Barriga, Glagys D.C., 669
Bartz, Kevin, 71
Baxter, Paul D., 497
Bazzi, Marco, 501
Bellera, Carine A., 357
Bellio, Ruggero, 507
Bergersen, Linn Cecilie, 405
Bertolotto, Patricia, 611
Blas Achic, Betsabé G., 511
Bolfarine, Heleno, 369
Bollaerts, Kaatje, 773
Boscaino, Giovanni, 479
Bowman, Adrian W., 289, 295
Bremhorst, Vincent, 87
Broadhurst, Karen, 827
Burke, Kevin, 93

Caballero-Águila, Raquel, 515, 521
Cadarso-Suárez, Carmen, 663
Camarda, Carlo G., 97, 103
Campos, José A.G., 525
Cancho, Vicente G., 669
Capursi, Vincenza, 139, 789
Cascone, Marcos H., 525
Cattelan, Manuela, 529
Cavenague, Hayala, 735
Cederbaum, Jona, 533
Chiodi, Marcello, 65
Christou, Vasiliki, 539
Cobre, Juliana, 543
Cocchi, Daniela, 435
Colombi, Roberto, 109
Colosimo, Enrico A., 115, 615
Conde-Sánchez, Antonio, 761
Costa, André G.F.C., 115
Costa, Marco, 625
Costain, Deborah A., 765, 827
Crainiceanu, Ciprian M., 3, 157, 399, 489
Crujeiras, Rosa María, 313
Cuevas, Francisco, 121
Cysneiros, Audrey H.M.A., 547
Cysneiros, Francisco José A., 567

Díaz, Margarita, 555
Díaz, Maria del Pilar, 555
Danesi, Ivan Luciano, 551
Darwent, David, 177
Davies, Vinny, 559
Dawson, Drew, 177
De Castro, Mário , 543
De Sousa, Bruno, 621
De Toledo, Maria Luiza G., 615
Deiana, Luca, 681
Dey, Dipak K., 669

Diggle, Peter J., 83
Donald, Margaret R., 563
Donaldson, Ana Nora A., 595
Dondelinger, Frank, 673
Duller, Christine, 731
Durbán, María, 233, 343
Dvorzak, Michaela, 127

Edwards, David, 77
Eilers, Paul H.C., 97, 103, 133, 151, 343, 381, 459, 757
El Ghouch, Anouar, 727
Enea, Marco, 139
Estrada, Francisco, 685
Evers, Ludger, 289

Fabio, Lizandra C., 567
Faes, Christel, 307, 811
Fang, Zhou, 573
Faria, Susana, 577, 581
Faschingbauer, Florian, 269
Fassò, Alessandro, 585
Fernández-Alcalá, Rosa M., 713
Ferrara, Giancarlo, 591
Fierens, Sébastien, 773
Filipe, Patrícia A., 621
Finazzi, Francesco, 145
Firth, David, 215
Flórez, Álvaro, 595
Fokianos, Konstantinos, 325
Forte, Anabel, 83
Frühwirth-Schnatter, Sylvia, 599
Francart, Julie, 773
Frasso, Gianluca, 151
Freitas, Marta A., 615
Fussl, Agnes, 599

Gallacher, Kelly, 603
Gampe, Jutta, 97, 757
García-Garrido, Irene, 515, 521, 659
Gellar, Jonathan, 157
George, Edward I., 165
Geraerts, David, 773
Gertheiss, Jan, 607
Giampaoli, Viviana, 611, 799
Gilardoni, Gustavo L., 615
Giordano, Sabrina, 109
Glad, Ingrid K., 405
Glass, Thomas A., 489
Gomes, Dulce, 621
Gomes, Rui, 577
Gonçalves, Arminda Manuela, 577, 625
Gonzalez, Javier, 465
Gosselin, Pol, 773
Gottard, Anna, 171
Greven, Sonja, 363, 533
Grilli, Leonardo, 693
Guédon Yann, 331
Guerrero, Victor, 685
Gutiérrez, Ramón, 659

Höller, Peter, 743
Ha, Il Do, 257
Haberman, Steven, 551
He, Bing, 489
Hendrych, Radek, 631
Hens, Niel, 103, 475
Hochreiter, Ronald, 697
Hofhuis, Agnetha, 423
Holmström, Lasse, 635
Hothorn, Torsten, 15
Hudson, Irene L., 177
Husmeier, Dirk, 559, 673

Iacus, Stefano M., 27
Iannario, Maria, 183
Iddi, Samuel, 641

Jaspers, Stijn, 191
Jiménez-Gamero, M. Dolores, 727
Joly, Pierre, 803
Jowaheer, Vandna, 197

Küchenhoff, Helmut, 203
Kiers, Henk A.L., 607
Kincl, Tomáš, 655
King, Gary, 27
Klein, Nadja, 645
Klima, André, 203
Klingenberg, Bernhard, 651
Klinten Grand, Mia, 45
Kneib, Thomas, 15, 209, 263, 343, 363, 393, 447, 645, 753
Komárek, Arnošt, 655
Komárková, Lenka, 655
Konstantinos, Konstantinos, 539
Kosmidis, Ioannis, 215
Kou, Samuel C., 71

López-Montoya, Antonio J., 659
López-Ratón, Mónica, 663
Lachos, Victor H., 689
Lambert, Philippe, 87, 221
Lang, Stefan, 645
Launonen, Ilkka, 635
Leday, Gwenaël G.R., 227
Lee, Dae-Jin, 233, 343
Lee, Xing Ju, 739
Leemaqz, Shalem Y., 177
Leemaqz, Sharon X., 33
Lesaffre, Emmanuel, 165, 301, 337, 459
Letón, Emilio, 663
Linares-Pérez, Josefa, 515, 521
Longford, Nicholas T., 239
Louw, Nelmarie, 245
Louzada, Francisco, 669, 735
Lovison, Gianfranco, 387

Macdonald, Benn, 673
Machado, Luís, 701
Machado, Robson J.M., 251
MacKenzie, Gilbert, 93, 257
Malá, Ivana , 677
Mameli, Valentina, 681
Manitz, Juliane, 263
Martínez-Gómez, Elizabeth, 685
Martínez-Rodríguez, Ana Maria, 761
Matawie, Kenan M., 493
Matos, Larissa A., 525, 689
Maul, Thomas, 697
Mayr, Andreas, 269, 393
McKeone, James P., 275
McLachlan, Geoffrey J., 33
McLellan, Chris R., 823
Mesue, Nicholas, 281
Metelli, Silvia, 693
Miller, Claire, 145, 295, 603
Millossovich, Pietro, 551
Mineo, Angelo M., 59
Mirkov, Radoslava, 697
Molanes-López, Elisa M., 663
Molenberghs, Geert, 641
Molinari, Daniel A., 289
Monod, Anthea, 71
Moreira, Ana , 701
Moro, Javier, 659
Morris, Darcy Steeg, 707
Mukherjee, Kathakali Ghosh, 295

Muniz, Graciela, 429
Murawska, Magdalena, 301
Musio, Monica, 681

Navarro-Moreno, Jesús, 713
Neupane, Rajendra, 197
Nicolussi, Federica, 717
Niel, Hens, 811
Noma, Alexandre, 799
Nummi, Tapio, 281
Nunes, Carla, 621
Nysen, Ruth, 307

Oelker, Margret-Ruth, 807
Oliveira, María, 313
Olmo-Jiménez, María José, 761
Oya, Antonia, 713

Pace, Luigi, 723
Papiez, Monika, 779
Pardo-Fernández, Juan Carlos, 727
Pauger, Daniela, 731
Paula, Gilberto A., 567, 815
Pedeli, Xanthi, 325
Perdona, Gleici Castro, 735
Pereira, Marcos Antonio A., 511
Perpiñán, Hèctor, 83
Pettitt, Anthony N., 275, 739
Peyhardi, Jean, 331
Pfeifer, Christian, 743
Piccolo, Domenico, 183
Pimanda, John E., 563
Pinheiro, Aluísio, 749
Plaia, Antonella, 139
Poffijn, André, 773
Porcu, Emilio, 121
Porcu, Mariano, 795
Pourahmadi, Mohsen, 325
Proust-Lima, Cécile, 357
Putter, Hein, 45
Pößnecker, Wolfgang, 319

Racugno, Walter, 351
Rampichini, Carla, 693
Reich, Daniel S., 399
Reulen, Holger, 753
Rizopoulos, Dimitris, 301
Rizzi, Silvia, 757
Ročková, Veronika, 165, 337
Roach, Greg, 177
Rodríguez - Álvarez, María Xosé, 343
Rodríguez–Casal, Alberto, 313
Rodríguez-Avi, José, 761
Ruiz-Molina, Juan Carlos, 713
Ruli, Erlis, 351
Russo, Cibele M., 251, 769

Séne, Mbéry, 357
Sáez-Castillo, Antonio J., 761
Saefken, Benjamin, 363
Salvan, Alessandra, 723
Santos, Bruno, 369
Sartori, Nicola, 723
Schauberger, Gunther, 375
Schmid, Matthias, 269
Schnabel, Sabine K., 381
Sciandra, Mariangela, 387
Scott, D. Julian A., 497
Scott, Marian, 145, 435, 603
Sen, Pranab Kumar, 749
Sharples, Stuart J., 765
Sherlock, Chris, 765
Shinohara, Russel T., 399
Silva, Danilo A., 769

Simons, Koen, 773
Smiech, Slawomir, 779
Sobotka, Fabian, 209, 393
Sofronov, Georgy, 453
Sonck, Michel, 773
Soromenho, Gilda, 581
Steegers, Regine, 459
Stefanova, Katia, 783
Stein, Anke, 447
Sulis, Isabella, 789, 795
Sutradhar, Brajendra, 197
Sweeney, Elizabeth M., 399

Tüchler, Regina, 441
Tamura, Karin A., 799
Teixeira, Lara, 625
Tellaroli, Paola, 501
Teunis, Peter, 423
Tharmaratnam, Kukatharmini, 405
Thurner, Paul W., 203
Thut, Gregor, 295
Torres, Javier, 595
Touloumis, Anestis, 411
Touraine, Célia , 803
Trevisani, Matilde, 417
Trottier, Catherine, 331
Tutz, Gerhard, 319, 375, 807
Tuzzi, Arjuna, 417

Unnikrishnan, Ashwin, 563
Usuga, Olga, 611

Vallejos, Ronny, 121
Van Bladel, Lodewijk, 773
Van de Kassteele, Jan, 423
Van de Wiel, Mark A., 227
Van den Hout, Ardo, 429
Van der Vaart, Aad W., 227
Van Eeuwijk, Fred A., 381
Van Nieuwenhuyse, An, 773
Van Oyen, Herman, 773
Van Pelt, Wilfrid, 423
Vandendijck, Yannick, 811
Vanegas, Luis Hernando, 815
Vargas, José M., 555
Varin, Cristiano, 529
Ventrucci, Massimo, 435
Ventura, Laura, 351
Verbeke, Geert, 191
Vidoli, Francesco, 591
Vidoni, Paolo, 507
Vogelstein, Joshua T., 399
Vujacic, Ivan, 59, 465

Wagner, Helga, 127, 441, 731
Waldmann, Elisabeth, 209, 447
Weerasinghe Jayawardana Rathambalage, Madawa P., 453
Willemsen, Sten, 459
Wilson, Nairn, 595
Wilson, Susan R., 563
Wit, Ernst C., 59, 465
Wolny-Dominiak, Alicja, 819
Worton, Bruce J., 823

Yeen, Emily, 827

Zanardo, Cleyton, 735
Zeileis, Achim, 743
Zipunnikov, Vadim, 3
Zulehner, Christine, 599

28th IWSM 2013 Sponsors

We are very grateful to the following organisations for sponsoring 28th IWSM 2013.

- University of Palermo
- Municipality of Palermo
- Toyota Motor Corporation
- Leonard N. Stern School of Business
- Springer Publisher
- Italian Statistical Society