

**Proceedings of the
27th International
Workshop
on Statistical Modelling**

July 16 – 20, 2012

Prague

**Arnošt Komárek, Stanislav Nagy
(editors)**

Proceedings of the 27th International Workshop on Statistical Modelling,
Prague, July 16–20, 2012,
Arnošt Komárek, Stanislav Nagy, editors,
Prague 2012.

Editors:

Arnošt Komárek, komarek@karlin.mff.cuni.cz
Stanislav Nagy, nagy@karlin.mff.cuni.cz

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University in Prague
Sokolovská 83, 186 75 Praha 8 – Karlín, Czech Republic

Scientific Programme Committee

- Arnošt Komárek (Chair)
Charles University in Prague, Czech Republic
- Jaromír Antoch
Charles University in Prague, Czech Republic
- Carmen Armero
Universitat de València, Spain
- Guadalupe Gómez Melis
Universitat Politècnica de Catalunya–BarcelonaTech, Barcelona, Spain
- Göran Kauermann
Ludwig-Maximilians-University, Munich, Germany
- Helmut Küchenhoff
Ludwig-Maximilians-University, Munich, Germany
- Philippe Lambert
University of Liège, Belgium
- Joseph B. Lang
University of Iowa, USA
- Emmanuel Lesaffre
*Erasmus Medical Centre Rotterdam, the Netherlands
and I-Biostat, Catholic University of Leuven, Belgium*
- Vito M. R. Muggeo
University of Palermo, Italy
- Jeffrey S. Simonoff
New York University, USA
- Mikis Stasinopoulos
London Metropolitan University, UK

Local Organizing Committee

- Arnošt Komárek (Chair)
Charles University in Prague, Czech Republic
- Jakub Fischer
University of Economics, Prague, Czech Republic
- Zdeněk Hlávka
Charles University in Prague, Czech Republic
- Lenka Komárková
University of Economics, Prague, Czech Republic
- Petr Mazouch
University of Economics, Prague, Czech Republic
- Stanislav Nagy
Charles University in Prague, Czech Republic
- Marek Omelka
Charles University in Prague, Czech Republic
- Ludmila Petkovová
University of Economics, Prague, Czech Republic
- Jaroslav Richter
Charles University in Prague, Czech Republic

Preface

This year, already the 27th edition of the INTERNATIONAL WORKSHOP ON STATISTICAL MODELLING (IWSM) will be held in Prague, Czech Republic. Many things have changed since 1986 when the inaugural workshop was held in Innsbruck, Austria. Originally a small meeting of the GLIM users turned into an annual conference attracting regularly 150–200 participants from around the world. Last but not the least, the Statistical Modelling Society was founded in 2003 and since then, it serves as an umbrella organization for the workshop and supports several related activities. For even longer period of time, the scholarly journal, *Statistical Modelling: An International Journal*, is associated with the workshops and is now already in its 12th year.

Nowadays almost unique feature of the IWSM among statistical conferences of similar size is the fact that it maintains only one plenary session during the whole week. Hence great thanks to all members of the Scientific Programme Committee who reviewed 103 contributed papers submitted for an oral presentation. They had a very tough task to identify only 58 of them that could fit into the programme of the plenary sessions of the workshop. Nevertheless, not only oral presentations, but also high-quality posters contribute to the scientific success of the workshop. It is our great pleasure to announce that more than sixty authors agreed to participate the poster session of the conference.

We are also pleased to have attracted M. Luz Calle, Paul H. C. Eilers, Thomas Kneib, Michal Kulich, and Robert B. O'Hara to deliver a one hour state of the art invited talk in their specific fields of expertise. Our special thanks are dedicated to Dimitris Rizopoulos for his one-day short course entitled 'An Introduction to Joint Models for Longitudinal and Survival Data with Applications in R' that directly preceded the workshop.

It is a good tradition of the IWSM to encourage students to not only attend but also to present, giving them the opportunity to speak to many of the worlds experts on statistical modelling. Traditionally, student participants compete for one of three IWSM student awards, and also during the 27th IWSM, we award three students for the best paper, the best oral presentation, and the best poster. Participation of students was even enhanced this year by two student travel grants provided kindly by the Statistical Modelling Society.

Finally, we thank to all authors who contributed to this proceedings volume, for participating in the workshop, and for carefully preparing their manuscripts and talks or posters.

Welcome to Prague. Enjoy the city, its admirable history and architecture, and have a great workshop!

Arnošt Komárek and Stanislav Nagy
Prague, July 2012

Contents

Part 1. Invited Papers

M. LUZ CALLE: Statistical challenges in modelling disease risk from multiple genotypes	3
PAUL H. C. EILERS: Composite link, the neglected model	11
THOMAS KNEIB: Beyond mean regression	23
MICHAL KULICH, ARNOŠT KOMÁREK, MAREK OMELKA: Estimating cross-sectional incidence from biological markers	33
ROBERT B. O'HARA: Bayesian variable selection and the (ab)use of priors	37

Part 2. Contributed Papers (Volume I)

ELENI-ROSALINA ANDRINOPOULOU, DIMITRIS RIZOPOULOS, JOHANNA J. M. TAKKENBERG, EMMANUEL LESAFFRE: Joint modeling of two longitudinal outcomes and competing risk data	45
FRANCESCO BARTOLUCCI, LEONARDO GRILLI, LUCA PIERONI: Latent class inverse probability weighting to estimate causal effects of sequential treatments under unobserved confounding	51
CARLO GIOVANNI CAMARDA, PAUL H. C. EILERS, JUTTA GAMPE: Additive decomposition of vital rates from grouped data ...	57
AMPARO Y. CASTRO-SÁNCHEZ, MARC AERTS, ZIV SHKEDY, PETER VICKERMAN, NIEL HENS: Inference for a dynamic model of HIV and HCV	63

JULIANA COBRE, FRANCISCO LOUZADA, MÁRIO DE CASTRO, GLE- ICI PERDONÁ, FERNANDA M. PERIA: Estimatives of lymph nodes metastasis rates and treatment effectiveness under two latent activation schemes	69
ENRICO A. COLOSIMO, GUSTAVO L. GILARDONI, MARISTELA D. OLIVEIRA: Heterogeneity identification of repairable systems	75
SUSANA CONDE, GILBERT MACKENZIE: Model selection in sparse contingency tables: LASSO penal- ties <i>vs</i> classical methods	81
IAIN D. CURRIE: Forecasting with the age-period-cohort model?	87
ANTOINE DE FALGUEROLLES: Cauchy, Prague and multiple regression	93
EMANUELE DEL FAVA, ZIV SHKEDY, MEHRETEAB FANTAHUN ARE- GAY, GEERT MOLENBERGHS: Modeling multivariate, overdispersed binomial data with ad- ditive and multiplicative random effects	99
FILIPPO DOMMA, SABRINA GIORDANO: A measure of household financial fragility	105
ACHMAD EFENDI, GEERT MOLENBERGHS, EDMUND NJAGI, PAUL DENDALE: A joint model with marginal interpretation for longitudinal continuous and time-to-event outcomes	111
JOCHEN EINBECK, BENJAMIN J. ISAAC, LUDGER EVERS, ALESSAN- DRO PARENTE: Penalized regression on principal manifolds with application to combustion modelling	117
CHELLAFE ENSOY, CHRISTEL FAES, MARC AERTS: A dynamic spatio-temporal model to investigate the effect of movements of animals on the spreading of Bluetongue BTV- 8 in Belgium	123

FRANCESCO FINAZZI:
A statistical model for spatial point aggregated data. The geostatistical potential model 129

GIANLUCA FRASSO, PAUL H. C. EILERS:
Smoothing parameter selection using the L-curve..... 135

AIMEE N. GOTT, IDRIS A. ECKLEY:
Detecting aliasing in locally stationary textured images..... 141

IL DO HA, NICHOLAS J. CHRISTIAN, JONG-HYEON JEONG, YOUNGJO LEE:
A general subhazard frailty model for multi-center competing risks data..... 147

FELIX HEINZL, GERHARD TUTZ:
Clustering in linear mixed models with Dirichlet process mixtures using EM algorithm..... 153

GILLIAN Z. HELLER, MAURIZIO MANUGUERRA:
Ordinal regression models for continuous scales..... 159

SAMUEL IDDI, GEERT MOLENBERGHS:
A joint marginalized multilevel model for longitudinal outcomes 165

JONATHAN JAEGER, PHILIPPE LAMBERT:
Bayesian ODE-penalized B-spline model with Gaussian mixture as error distribution 171

CHAITANYA JOSHI, DANIEL C. LAUGHLIN, PETER M. VAN BODEGOM, ZACHARY A. BASTOW, PETER Z. FULÉ:
Modeling trait based ecological community assembly..... 177

GÖRAN KAUERMANN, RENATE MEYER:
Flexible modeling of multivariate data by mixtures of Archimedean copulas..... 179

PHILIPPE LAMBERT:
Nonparametric estimation of conditional Archimedean copula..... 185

JOSEPH B. LANG: Upon closer inspection... Testing in comparative experiments	191
DAE-JIN LEE, MARÍA DURBÁN: Seasonal modulation smoothing mixed models for times series forecasting	197
NIRIAN MARTÍN, LEANDRO PARDO: Cook's distance in polytomous logistic regression	203
ANDREAS MAYR, TORSTEN HOTHORN, NORA FENSKE: Fitting prediction intervals for BMI patterns in childhood by boosting quantile regression	209
ANA MOREIRA, LUÍS MACHADO: Conditional estimation of the bivariate distribution under dependent right censoring	215
DAVID MORIÑA, PEDRO PUIG, JORDI VALERO: Autoregressive models with non-gaussian errors	221
VITO M. R. MUGGEO: Smoothed score confidence interval for the breakpoint in segmented regression	227
STANISLAV NAGY: Nonparametric classification of noisy functions	233
GERHARD NEUBAUER: Statistical models for deficient count data	239
PETR NOVÁK: Testing goodness-of-fit of the Accelerated Failure Time model with time-varying covariates	245
EKATERINA OGURTSOVA: Estimation of multi-state model parameters from panel data: A comparison of different methods	251
MARIA OLIVEIRA PEREZ, ROSA M. CRUJEIRAS, A. RODRÍGUEZ-CASAL: Nonparametric circular density estimation for temperature cycles	257

KONSTANTINOS PERRAKIS, DIMITRIS KARLIS, MARIO COOLS, DAVY JANSSENS, GEERT WETS:
Poisson mixture regression for Bayesian inference on large over-dispersed transportation origin-destination matrices... 263

IAIN PROCTOR, ROGNVALD I. SMITH, E. MARIAN SCOTT:
Fine-scale downscaling of environmental covariates in biodiversity modelling 269

LEENDERT PUNT, LINDA M. HAINES, CHRISTIEN THIART:
Modelling the movement of dusky kob in the Sundays River 275

REYHANEH RIKHTEHGARAN, IRAJ KAZEMI, GEERT VERBEKE, WIM DE KORT, EMMANUEL LESAFFRE:
Piecewise transition models with random effects for unequally-spaced measurements 279

DIMITRIS RIZOPOULOS:
A pseudo-adaptive Gaussian quadrature rule for fitting joint models for longitudinal and time-to-event data 285

VERONIKA ROČKOVÁ, EMMANUEL LESAFFRE:
Incorporating prior knowledge in Bayesian modeling of sparse networks..... 291

CARLES SERRAT, JAIME-ABEL HUERTAS, GUADALUPE GÓMEZ:
A semi-parametric joint model for two sequential times to events and one longitudinal covariate 297

MICHAEL G. SCHIMEK, MARCUS BLOICE:
Modelling the rank order of Web search engine results 303

SABINE K. SCHNABEL, FRED A. VAN EEUWIJK, PAUL H. C. EILERS:
Modeling latent curves for genotype by environment interaction 309

FABIAN SOBOTKA, RADOŠLAVA MIRKOV, BENJAMIN HOFNER, PAUL H. C. EILERS, THOMAS KNEIB:
Modeling flow in gas transmission networks using shape-constrained expectile regression..... 315

REGINA TÜCHLER, HELGA WAGNER: Analysing living conditions in Austria by a Bayesian mixed data model	321
KARIN AYUMI TAMURA, VIVIANA GIAMPAOLI: Comparison of prediction methods for mixed logistic regression	327
CÉLIA TOURAINE, PIERRE JOLY: Predictions from a Markov illness-death model: Application to dementia disease	333
GERHARD TUTZ, WOLFGANG PÖSSNECKER: Variable selection for the multinomial logit model	339
INSHA ULLAH, BEATRIX JONES: Hierarchical covariance selection models	345
ARDO VAN DEN HOUT, JEAN-PAUL FOX, GRACIELA MUNIZ: A longitudinal model for latent cognitive function	351
CRISTIANO VARIN, ANNAMARIA GUOLO: Marginal beta regression for time series analysis	357
ELISABETH WALDMANN, THOMAS KNEIB: Variational approximations in Bayesian geoadditve quantile regression	363
DEIRDRE WALL, CARL SCARROTT, JOHN NEWELL, HELEN INGOLDSBY, GRACE CALLAGY, MICHAEL J. KERIN: Identifying underlying structure in classification and regression trees using surrogate splits	369
JIXIAN WANG: Assessing surrogacy of progression free survival for overall survival: A multi-state model approach	375
SUSAN R. WILSON: Assessment of model stability for high- dimensional data with applications to complex genomic data	381
Author Index	387

Part 3. Contributed Papers (Volume II)

ARIEL ALONSO ABAD:
Misspecified random-effects distribution in non-linear mixed models with linear random effects 395

GIADA ADELFIGIO, GIOVANNI BOSCAINO, VINCENZA CAPURSI:
Regression quantiles to assess higher education performance 401

LEILA D. A. F. AMORIM, RAYDONAL OSPINA:
prLogistic: An R package for estimating prevalence ratios using logistic models 407

MANOOICHEHR BABANEZHAD:
Matrix correlation and matrix cross spectrum of mismeasured multivariate time series 413

IRANTZU BARRIO, INMACULADA AROSTEGUI, MARÍA XOSÉ RODRÍGUEZ-ÁLVAREZ, JOSE MARÍA QUINTANA:
Location of optimal cut-points to categorize continuous variables in clinical studies 419

ANA BORGES, INÊS SOUSA, LISANDRA ROCHA, RAQUEL MENEZES:
Fixed effects versus random effects in a longitudinal study: A simulation study 425

KEVIN BURKE, GILBERT MACKENZIE:
Multi-parameter regression survival models 431

NORZIHA CHE HIM, TREVOR C. BAILEY, DAVID B. STEPHENSON:
Climate variability and dengue incidence in Malaysia 435

D. COCCHI, MASSIMO VENTRUCCI, L. ALTIERI, E. MARIAN SCOTT:
Modelling urban sprawl patterns in binary raster maps 441

MARCO COSTA, MAGDA MONTEIRO, A. MANUELA GONÇALVES:
Kalman filtering approach in the calibration of radar rainfall data 447

JALILA DAOUDI:
Modeling operational risk losses 455

MARISTELA DIAS DE OLIVEIRA, ENRICO A. COLOSIMO, GUSTAVO L. GILARDONI: Bayesian inference for power law processes with applications in repairable systems	459
CHRYSOULA DIMITRIOU-FAKALOU: The $(\mathbb{Z}^d \times \mathbb{Z})$ spatial-temporal Auto-Linear-Auto-Regressive model	465
DENISE DUARTE, WECSLEY PRATES, ENRICO A. COLOSIMO: The sample signature of probabilistic context trees with an application to Linguistics	471
MANUELA ENDER, LU ZONG: Analysis of temperature-based weather derivatives in mainland China: Pricing and simulation	479
JUDE EZE, E. MARIAN SCOTT, KEVIN POLLOCK, RUTH STIDSON, CLAIRE MILLER, DUNCAN LEE: Modelling the association between bathing water quality and gastrointestinal illness in South West Scotland	485
VERONIKA FENSTERER, HELMUT KÜCHENHOFF, JOSEF CYRYS, SUSANNE BREITNER, ALEXANDRA SCHNEIDER, MIKE PITZ, JIANWEI GU, ANNETTE PETERS: Measurement error in the personal exposure to air pollution	491
ROSEMEIRE FIACCONE, ROBIN HENDERSON, LEILA D. A. F. AMORIM: Correlated frailty model for multivariate longitudinal count data	495
PATRÍCIA A. FILIPE, DULCE GOMES, CARLA NUNES, MARÍLIA SILVA, BRUNO DE SOUSA, TEODORO BRIZ: Delay in diagnosis of pulmonary tuberculosis in Portugal ...	501
ADELAIDE FREITAS, SARA ROQUE: A comparison of several background correction and normalization methods on microarray data	507
IRENE GARCÍA-GARRIDO, J. LINARES-PÉREZ, R. CABALLERO-ÁGUILA: Recursive linear estimation from multi-sensor observations with correlated uncertainties	513

KATHAKALI GHOSH MUKHERJEE, CLAIRE MILLER, ADRIAN W. BOWMAN, GREGOR THUT: A flexible regression framework for TMS-EEG signals	519
GUSTAVO L. GILARDONI, MARISTELA D. OLIVEIRA, ENRICO A. COLOSIMO: Bootstrap confidence intervals for the optimal maintenance time of a repairable system	525
A. MANUELA GONÇALVES, MARCO COSTA: Water monitoring sites discrimination using clustering water variables time series data and main latent factors identification	531
MARKUS HAINY, WERNER G. MÜLLER, HELGA WAGNER: Simulation-based D_B-optimal designs: Conception and implementation issues	537
RADEK HENDRYCH: Conditional correlation modelling: Simulation study	543
CHIU-HSIEH HSU, QI LONG, YISHENG LI, ELIZABETH JACOBS: A robust nearest neighbor-based multiple imputation approach for data with missing covariate values	549
ŠÁRKA HUDECOVÁ, MICHAL PEŠTA: Generalized estimating equations in claims reserving	555
NIKOLA KASPŘÍKOVÁ: Application of text mining for media analysis	561
ANDRÉ KLIMA, HELMUT KÜCHENHOFF, PAUL W. THURNER: Effects of different spatial modeling: Ruralisation of the NS-DAP in the Weimar Republic?	567
MARIA KRÁLOVÁ, ALENA KLAPALOVÁ, JURAJ ŠIŠKA: Corporate financial performance and its predictors	573
ANTONIO JESÚS LÓPEZ-MONTOYA, M. L. GÁMIZ-PÉREZ: Semi-parametric regression models with reliability data	577

MÓNICA LÓPEZ-RATÓN, CARMEN CADARSO-SUÁREZ, ELISA M. MOLANES-LÓPEZ, EMILIO LETÓN: Inference of the symmetry point with different costs for the specificity and sensitivity	583
ANNOUSCHKA LAENEN: Model uncertainty and multimodel inference in reliability estimation	589
ANTONY LAWSON, JOCHEN EINBECK: Generative linear mixture modelling	595
KENAN M. MATAWIE, SARGON HASSO: Refined information retrieval and frequency distribution ...	601
LARISSA A. MATOS, MARCOS O. PRATES, MING H. CHEN, VICTOR H. LACHOS: Likelihood based inference for linear and nonlinear mixed-effects models with censored response using the multivariate-<i>t</i> distribution	607
CHRIS R. MCLELLAN, BRUCE J. WORTON, WILLIAM DEASY, A. NICHOLAS E. BIRCH: Modelling tracks of cabbage root fly larvae in a novel study of crop protection	613
RADOSLAVA MIRKOV, HOLGER THOMAE, MICHAEL FEIST, THOMAS MAUL, GORDON GILLESPIE, BASTIAN LIE: Modeling and forecasting customer behavior for revolving credit facilities	621
ELISA M. MOLANES-LÓPEZ, EMILIO LETÓN: Multivariate copula models in ROC analysis	627
ELISA M. MOLANES-LÓPEZ, JUAN ROMO: Multiple testing based on depth	633
DANIEL ALBERTO MOLINARI, LUDGER EVERS, ADRIAN W. BOWMAN: Smoothing parameter selection for spatiotemporal models with application to the analysis of contaminants in ground-water	637

ANA MOREIRA, ARTUR AGOSTINHO ARAÚJO, LUÍS MACHADO:
Estimation of the bivariate distribution function: A comparative study 643

CHENJERAI KATHY MUTAMBANENGWE, CHRISTEL FAES, MARC AERTS:
Spatial regression of quantiles based on parametric distributions 649

AINHOA OGUIZA, INMACULADA GALLASTEGUI, VICENTE NÚÑEZ-ANTÓN:
Analysis of pseudo-panel data with dependent samples 655

SAMUEL D. OMAN:
Shrinkage estimation when calibrating in the presence of random effects 661

MAREK OMELKA:
ANOSIM test revisited 667

MARTIN OTAVA, ADETAYO KASIM, ZIV SHKEDY, DAN LIN, BERNET S. KATO:
Bayesian variable selection method for modeling dose-response microarray data under simple order restrictions... 673

ROBERT L. PAIGE, A. ALEXANDRE TRINDADE:
Saddlepoint-based bootstrap inference in parametric and semiparametric models..... 679

VALENTIN PATILEA, CÉSAR SÁNCHEZ-SELLERO, MATTHIEU SAUMARD:
Projection-based nonparametric checks of regressions with functional covariates..... 685

CHRISTIAN PFEIFER, ACHIM ZEILEIS:
Trend analysis of snow avalanche accidents in Tyrol within the years 1989–2010..... 691

CHARLES ROHDE:
Standard statistics as likelihood statements 697

ERLIS RULI, LAURA VENTURA: Bayesian approximation methods for pseudo-posterior distributions in the presence of nuisance parameters	705
CIBELE M. RUSSO, EMMANUEL LESAFFRE, GILBERTO A. PAULA: A penalized elliptical mixture partially nonlinear mixed effects model	711
MARC SAEZ, ANNIBALE BIGGERI, DOLORES CATELAN, MARIA ANTÒNIA BARCELÓ, LAURA GRISSOTO, ALBERTO ALLEPUZ: Methods to control for the concavity in spatial ecological regressions (IneqCities project)	717
BRUNO R. SANTOS, SILVIA N. ELIAN: Analysis of residuals in quantile regression: An application to income data in Brazil	723
GUNTHER SCHAUBERGER, GERHARD TUTZ: Effect stars for categorical response models	729
LINDA SCHULZE WALTRUP, GÖRAN KAUERMANN, FABIAN SOBOTKA, THOMAS KNEIB: Comparing the estimation of expectiles and quantiles towards efficiency	735
ALI SHEIKHI, DAVID RAMSEY: A likelihood ratio test for detection of single nucleotide polymorphisms (SNPs)	741
GIOVANA O. SILVA, EDWIN M. M. ORTEGA, GAUSS M. CORDEIRO: Log-Beta modified Weibull regression models in survival analysis	747
KATIA STEFANOVA: Factor analytic mixed models with inclusion of pedigrees in the analysis of plant breeding trials	753
EWA STRZALKOWSKA-KOMINIAK, ELISA M. MOLANES-LÓPEZ, EMILIO LETÓN: Estimation of the conditional distribution of two censored gap times based on a nonparametric approach	759

ANATOLI TOROKHTI, STANLEY MIKLAVCIC:
Robust estimation of large data sets by piecewise-linear interpolating estimator 765

JAN VAN DE KASSTEELE, JAN VAN EIJKEREN, JACCO WALLINGA:
Age and sex specific social contact patterns stratified by location and day of the week 771

ZIHUA YANG, JACK CUZICK:
Modelling interactions of multiple events with application to human papillomavirus virus infections 777

ZAMIRA ZAMZURI, ROSS SPARKS, GRAHAM WOOD, GILLIAN Z. HELLER:
Spatial model for multivariate traffic accident count data... 783

Part 1. Invited Papers

Statistical challenges in modelling disease risk from multiple genotypes

M. Luz Calle¹

¹ Dept. of Systems Biology, University of Vic, Spain

E-mail for correspondence: `malu.calle@uvic.cat`

Abstract: Common human diseases have a complex etiology and the study of their genetic component poses a number of statistical challenges: number of genotyped variants usually larger than the number of individuals, small marginal effects, unknown genetic architecture, multifactor epistatic effects and linkage disequilibrium. We describe the performance of three statistical strategies for gene finding and disease risk prediction: marginal selection, LASSO and AUC-RF. Some modifications of the standard implementation of these methods will be proposed in order to improve their performance.

Keywords: AUC-RF, disease risk prediction, genetic epidemiology, LASSO, variable selection.

1 Introduction

Most common human diseases, such as cancer, diabetes, coronary heart disease or psychiatric disorders, have an inherited genetic component. Its contribution to disease risk has been inferred in aggregation studies that quantify the increased disease risk in relatives of affected individuals and heritability, the proportion of the phenotypic variance in the population that is due to genotypic differences among individuals. Unlike Mendelian diseases that are mainly governed by a single genetic mutation, common human diseases have a complex etiology; they are caused by the combined effect of multiple genetic and environmental factors. The usual study design for exploring the genetic basis of human diseases is a case-control study where single nucleotide polymorphisms (SNPs) are genotyped and differences in genotype frequencies between cases and controls are analyzed. A SNP is a polymorphic single nucleotide locus in the genome DNA where different variants (alleles) are observed among the individuals in a population. This is the most simple and common form of genetic variation among individuals. Throughout the human genome there are about 10 million SNPs and the variation in these polymorphic loci would explain an important part of our individual susceptibility to disease or the different individual

responses to treatments. Most SNPs have two possible alleles; the most frequent one in the population is denoted by "a" and the less frequent one by "A". Since human genome is diploid, that is, the DNA is duplicated in each cell of an individual, this yields to three possible genotypes per SNP: "aa" for the common homozygous subjects, "Aa" for the heterozygous subjects and "AA" for the variant homozygous subjects. From a statistical point of view an SNP can be thought as a categorical variable with three different categories that can be recoded numerically as the number of minor alleles, that is, zero for "aa", one for "Aa" and two for "AA". The number of genotyped SNPs for each individual is of the order of hundreds in candidate gene studies to 1 million or more in genome-wide association studies (GWAS). Candidate gene studies genotype SNPs in a number of genes that are thought to have some relation with the disease. Instead GWAS are designed to cover most of the human genetic variation by genotyping SNPs across the whole genome without any prior hypothesis of causality. Indeed, GWAS are indirect association studies where the genotyped SNPs act as markers of its region nearby; it is assumed that an associated SNP will be either a causing disease variant or will be in linkage disequilibrium (LD) with an unmeasured causing variant. In the last few years a large number of association studies have been carried out and many genes and genetic variants associated with disease risk have been identified. However, these well established variants only explain a small proportion of the inferred genetic contribution of disease and discovering the rest of genetic variants remains a major challenge.

2 Statistical challenges when searching the genetic basis of human diseases

One important reason for the statistical and computational difficulties encountered when searching the genetic basis of human diseases is the dimensionality problem, also known as "small n - large p paradigm", where the number of covariates far exceeds the number of available individuals. Apart from this obvious difficulty there are additional statistical issues that stand in the way of progress towards the identification of the genetic component of complex diseases. The challenges we discuss in this work are: small size effects of individual variants, unknown genetic architecture of disease, non additive multifactor effects and linkage disequilibrium (LD) or correlation between the genetic variants.

2.1 Small effect size of individual variants

Most associated variants detected up to now in GWAS have small effect sizes with odds-ratios of disease for the associated risk allele typically smaller than 1.5 but many around 1.2 (Wray N.R. et al. 2008). This fact

has two important consequences: First, that very large sample sizes are required for identification of causal variants in this context; otherwise we will only be able to detect a small proportion of causal variants, they will only explain a small part of the total genetic variation and models of individual disease risk based on these identified variants will have scarce prediction capacity. Second, as we will discuss next, if the effects act additively, a large number of such small effect variants are required, hundreds or thousands, to explain the total genetic contribution inferred for most common diseases.

2.2 Genetic architecture of complex diseases

The genetic architecture of a disease refers to the way in which a number of genetic variants interact to affect risk of disease. For most complex diseases their specific genetic architecture will be difficult to elucidate and will probably remain unknown.

Gibson G. (2012) proposes four conceptual models for the genetic architecture of common diseases: the CDCV model, the infinitesimal model, the rare allele model and the broad sense heritability model.

The CDCV model assumes the "common-disease common-variant" hypothesis. In contrast to Mendelian diseases that are usually under strong selection and consequently causing variants are highly penetrant and rare, the CDCV model assumes that the variants underlying common diseases would not have this strong selection and would have larger frequencies. This is the basis for large-scale association studies like GWAS where a small number of moderate-effect loci were expected to produce very strong signals, each of which explaining part of the genetic variance. This model has been questioned after the evidence that in most GWAS the associated variants only explain a small proportion of the genetic variance.

The infinitesimal model assumes that a very large number of common variants of very low effect (explaining less than 1% of the risk) are the major source of genetic variance for disease susceptibility. In this setting most of the missing heritability in current GWAS studies would be hidden beneath the thresholds used to define significant associated variants.

The rare allele model assumes that most of the variance for certain complex diseases is due to rare variants (allele frequencies typically smaller than 1%) with moderate or high effects. Standard GWAS analysis will not detect these rare variants because only a very small proportion of people in the sample will carry these causing variants.

The broad sense heritability model considers that additive contributions of common variants and large effects of rare variants are insufficient to explain the missing heritability. Gene-gene and gene-environment interactions would be responsible for the missing genetic component.

2.3 Epistatic multifactor effects

The search of genetic factors that underlie common diseases should not ignore the existence of interactions between loci. Genes do not act isolately but instead their function may depend on many other genes in a network or pathway that interact in a complex way. From a statistical point of view epistasis or genetic interactions represent departures from a linear model that describes how two or more variants explain the phenotype. A growing interest in this topic has led to many statistical proposals for epistasis analysis: from exhaustive searches using regression models including interactions (Marchini, J. et al. 2005) to data-mining methods such as the Model-Based Multifactor Dimensionality Reduction method (Calle, M.L. et al. 2010) that is an extension of the popular MDR method (Ritchie, M.D. et al. 2003) but allows adjusting for marginal effects and confounders. An overview on this topic is given by van Steen K. (2012).

2.4 Linkage disequilibrium

Linkage disequilibrium (LD) is a population level concept corresponding to non-random association of alleles in two loci. LD appears when a particular allele at one locus is found together with a specific allele at a second locus more often than expected if the loci were segregating independently in the population. LD together with the CDCV hypothesis are the basis for the current genome-wide association designs: LD allows that not all genetic variants should be genotyped since common causing variants would be detected directly or indirectly through LD with an associated genotyped marker. The existence of SNPs in high LD corresponds to collinearity among predictors and this introduces an additional difficulty in the analysis. Specific strategies to deal with this problem will be necessary.

3 Statistical approaches for disease risk prediction

The goal of disease risk prediction is to discriminate between individuals who will develop the disease and those who will not. This requires the identification of a set of variables that maximizes the prediction accuracy of the model based on these variables.

3.1 Marginal variable selection

The usual strategy for building disease models in GWAS is marginal variable selection. Each locus is evaluated individually for its marginal association with disease by performing, for instance, a marginal chi-square test. Those genotypes with a p-value below a specified threshold ($5 \cdot 10^{-7}$ in GWAS) are included in the prediction model.

This is an attractive strategy mainly because of computational reasons, but it has important limitations: variants with small effects will be difficult to capture, features are selected based on association significance but this not necessarily corresponds with prediction capacity, redundant variants are selected because of linkage disequilibrium and, finally, this marginal approach cannot identify interacting loci.

3.2 Variable selection with LASSO

An alternative approach to marginal variable selection is LASSO (Least Absolute Shrinkage and Selection Operator) which performs simultaneously estimation and variable selection (Tibshirani, R. 1996).

The LASSO is a \mathcal{L}_1 penalized maximum likelihood method where the regression coefficients are constraint to $\sum_j^p |\beta_j| \leq t$ where p is the number of genotyped variants $X_j, j = 1, \dots, p$. For logistic regression, where Y is the indicator of disease, the LASSO procedure corresponds to maximization of:

$$\sum_{i=1}^n \left\{ y_i(\beta_0 + \sum_j^p \beta_j x_{ij}) - \log(1 + \exp(\beta_0 + \sum_j^p \beta_j x_{ij})) \right\} - \lambda \sum_j^p |\beta_j|$$

LASSO shrinks some coefficients and sets others to zero performing a variable selection that will select the subset of variables exhibiting the strong effect. The parameter λ controls the amount of shrinkage that is applied to the estimates.

3.3 Variable selection with Random Forest

Random Forest methodology (RF) is a classification algorithm developed by (Breiman, L. 2001) consisting in the aggregation of multiple classification trees. Classification trees are attractive in genetic association studies for its simplicity and because their structure may highlight complex relationship between genetic variants. The main disadvantage of classification trees is their instability: small changes in the initial dataset can derive in important changes in the predictions. Random Forest overcomes this important limitation by aggregating several classification trees, each one built in a different bootstrap sample of the original dataset. This process is called "bagging" from "bootstrap aggregating" (Breiman, L. 1994). The expected reduction in variance due to aggregation may be limited because the different classifications trees are not completely independent. In order to further reduce the correlation among the trees additional randomization is introduced in tree building by considering only a random subset of predictors at every split. The use of RF is increasingly common in genetic epidemiology, and its behavior in different scenarios LD has been extensively studied in the last few years.

Recently, a new algorithm for variable selection using random forest has been proposed (Calle et al. 2011). The AUC-RF algorithm performs an iterative backward elimination process. A first RF is built using all predictor variables that provides the ranking of the variables. In the subsequent steps a fraction of the less important variables according to the initial ranking is eliminated (by default 20%). RF is built with the remaining variables and the AUC of the reduced model is computed on the out-of-bag data (data that has not been used for model building). This is repeated until the number of remaining variables is less or equal than a specified value. The elimination process can be visualized with a curve describing the AUC value of the different RFs as a function of the number of predictor variables. The optimal set of predictive variables is considered the one giving rise to the RF with the highest AUC.

4 Performance of marginal, LASSO and AUCRF variable selection in genetic studies

In this talk I will describe the performance of marginal, LASSO and AUC-RF variable selection in genetic studies. I will discuss the advantage and limitations of each approach in presence of the statistical challenges mentioned before: small sample size, unknown genetic architecture, epistasis and linkage disequilibrium. Some modifications of the standard implementation of LASSO and Random Forest will be proposed in order to improve the performance of these approaches.

Acknowledgments: This research was partially supported by grant MTM 2008-06747-C02-00 from the Ministerio de Ciencia e Innovacion (Spain) and grant 2009SGR-581 from Generalitat de Catalunya (Spain).

References

- Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**, 123-140.
- Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5-32.
- Calle, M. L., Urrea, V., Boulesteix, A.L., and Malats, N. (2011). AUC-RF: A New Strategy for Genomic Profiling with Random Forest. *Hum Hered*, **72**, 121-132.
- Calle, M. L., Urrea, V., Malats, N., and Steen, K. V. (2010). mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. *Bioinformatics*, **26**, 2198-2199.
- Gibson, G.(2011). Rare and common variants: twenty arguments. *Nat Rev Genet*, **13**, 135-145.

- Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, **37**, 413-417.
- Ritchie, M. D., Hahn, L. W., and Moore, J. H. (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet.Epidemiol.*, **24**, 150-157.
- Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso, *Journal Of The Royal Statistical Society Series B-Methodological*, **58**, 267-288.
- Van Steen, K.(2012). Travelling the world of gene-gene interactions. *Brief Bioinform*, **13**, 1-19.
- Wray, N. R., Goddard, M. E., and Visscher, P. M. (2008). Prediction of individual genetic risk of complex disease. *Current Opinion in Genetics & Development*, **18**, 257-263.
- Wray, N. R. and Goddard, M. E.(2010). Multi-locus models of genetic risk of disease. *Genome Med*, **2**, 10.
- Wray, N. R., Yang, J., Goddard, M. E., and Visscher, P. M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet*, **6**, e1000864.

Composite link, the neglected model

Paul H. C. Eilers¹

¹ Department of Biostatistics, Erasmus University Medical Centre, The Netherlands

E-mail for correspondence: p.eilers@erasmusmc.nl

Abstract: The composite link model (CLM) is an extension of the generalized linear model (GLM). It introduces an additional layer, generating a weighted sum of GLMs. It is a great model for grouped counts, digit preference and various similar types of observation schedules. Often a latent distribution is to be estimated and the model is ill-posed. Appropriate penalties solve this issue. The penalized CLM, parameter estimation and model diagnostics are presented. To motivate the model some existing applications are shortly discussed. Two new applications, inspired by real consulting questions, are described in more detail.

Keywords: Digit preference; GLM; Grouped data; Overdispersion.

1 Introduction

Forty years ago, the generalized linear model (GLM) was published by Nelder and Wedderburn (1972). It took the world by storm. Today it is a standard part of the toolbox of most statisticians, and it has found a place in any decent statistical software package.

The popularity of the GLM is easy to understand. Combining maximum likelihood estimation with the exponential family of distributions, it offers a unified way to build and estimate models for many types of observations, not only the over-employed normal ones, but also counts, binary observations, and many more. The appreciation of the statistical community is borne out by the number of citations to the GLM paper. When I checked the Web of Science in early May 2012, their number was almost 1685. Actually the all time yearly record was in 2011, with 87 citations.

It is a little over 30 year ago that the composite link model (CLM) was published by Thompson and Baker (1981). It extends the GLM in a very useful way. The easiest way to describe it is as a linear combination (often a sum) of GLMs. At first sight this might not look spectacular, but in fact there are many types of observations for which the CLM is a natural choice. Once you know the model well, you see many applications for which it is the perfect choice. I will try to convince you by showing a number of examples.

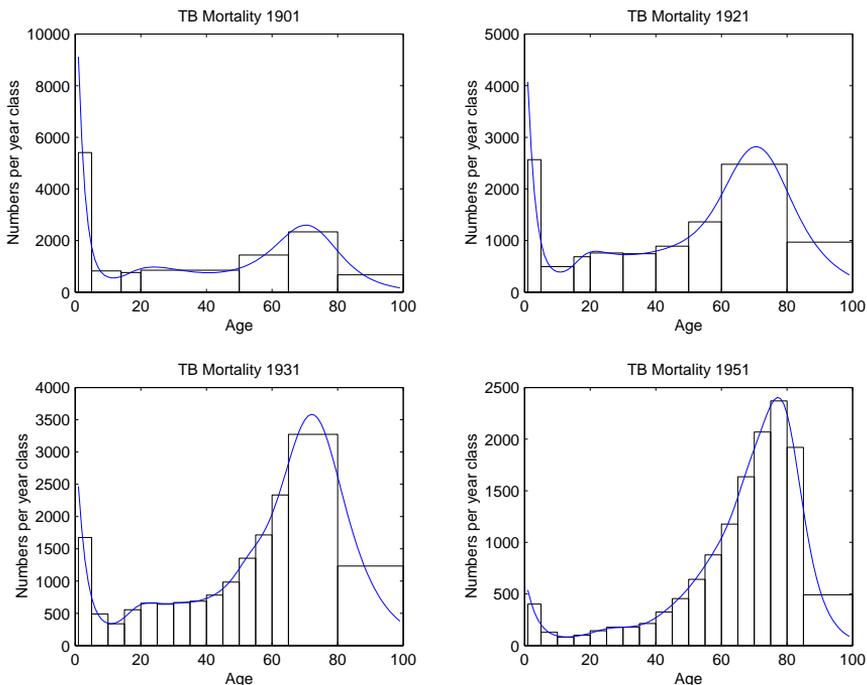


FIGURE 1. Deaths caused by Tuberculosis in The Netherlands, as reported in age groups that change over time. The smooth curves have been estimated with a penalized CLM.

Despite its usefulness, the CLM has been largely neglected by the statistical community. According to the Web of Science the CLM paper has been cited 82 times, of which 14 in the last decade. I would like to see this change.

My goal is not only to advertise the CLM itself, but also an extension that makes it even more useful. As said, the CLM proposes a linear combination of GLMs. The individual components cannot be observed directly, so they have a latent character. Some types of CLMs are also ill-conditioned, meaning that there is not enough information in the data to estimate their parameters reliably. This can lead to numerical problems and silly results. A powerful remedy is to introduce a penalty. When we are estimating a distribution or a time series, a natural choice is to penalize differences between neighbors, to enforce a smooth result. In other cases, with no natural order, a ridge-like penalty on the size of the parameter vector is effective.

In the next section I will describe the model in global terms, motivating why it is attractive, using three existing applications as examples. In Section 3 I will present some theory. To make this contribution more than a review, I present two new applications, based on my own consulting experience, in Section 4. As a close there will be a short Discussion.

2 Motivation of the model

In this section I will introduce the CLM in the context of counts and the Poisson distribution. Consider a GLM with design matrix X , coefficients β , and linear predictor $\eta = X\beta$. With the usual logarithmic link function we then have that $E(y) = \gamma = \exp(X\beta)$ and realizations from Poisson distributions with expectations γ . If we could observe them directly, we could stay in the GLM world and my story would end here. Instead, we observe grouped data with $E(y) = \mu = C\gamma$. The expected values, μ , are linear combinations of the latent expectations, γ , and the elements of C describe how they are combined. Many practical problems fit into this scheme.

A first example is grouping of age classes in a table of counts of disease cases or deaths. Then γ might represent the expected numbers in one-year age classes, while the data y are only available as totals of five-year classes. Say that γ covers the ages from 0 to 99, and so has 100 elements. The matrix C has 20 rows and 100 columns, while μ has 20 elements. Most elements of C are zero, but in row 1 we find a 1 in columns 1 to 5, in row 2 in columns 6 to 10, and so on, reflecting the sums over five-year intervals. This example is a special case of grouped counts. Histograms with coarse bins are very similar. The width of the bins does not have to be constant: variable widths can be handled by appropriate patterns of zeros and ones in C . In fact it is even possible for each individual observation to have a different interval.

An example of this type of data is shown in Figure 1, presenting counts of deaths due to Tuberculosis in The Netherlands. Historical records often used quite wide intervals, that became narrower over time. The goal is to estimate a smooth distribution of deaths, to combine it with population records (which are more precise) for mortality estimates, see Eilers and Borgdorff (2003).

A second example is digit preference. When people report a number from their memory, or estimate it, they have a tendency to round it, to even numbers, or to multiples of five or ten. This is called digit preference or heaping, and an example is shown in Figure 2, showing reported numbers of deaths in Spain in 1910 and 1960. There was a strong tendency to round to numbers ending with a zero and to a lesser degree to those ending with a five. Over the years the quality of birth records has improve a lot; in 1960 digit preference has decreased a lot and in present day statistics it has disappeared completely.

It is reasonable to assume that the actual distribution does not have the spikes at multiples of five. What we see is an artificial increase at these multiples, resulting from transfers of counts from the age classes at both sides, that are artificially decreased. If we know the probability, say p_1 , that an observation in age class j is reported in ages class $j - 1$, and similarly p_2 for reporting in age class $j + 1$, then we can build a C matrix that reflects

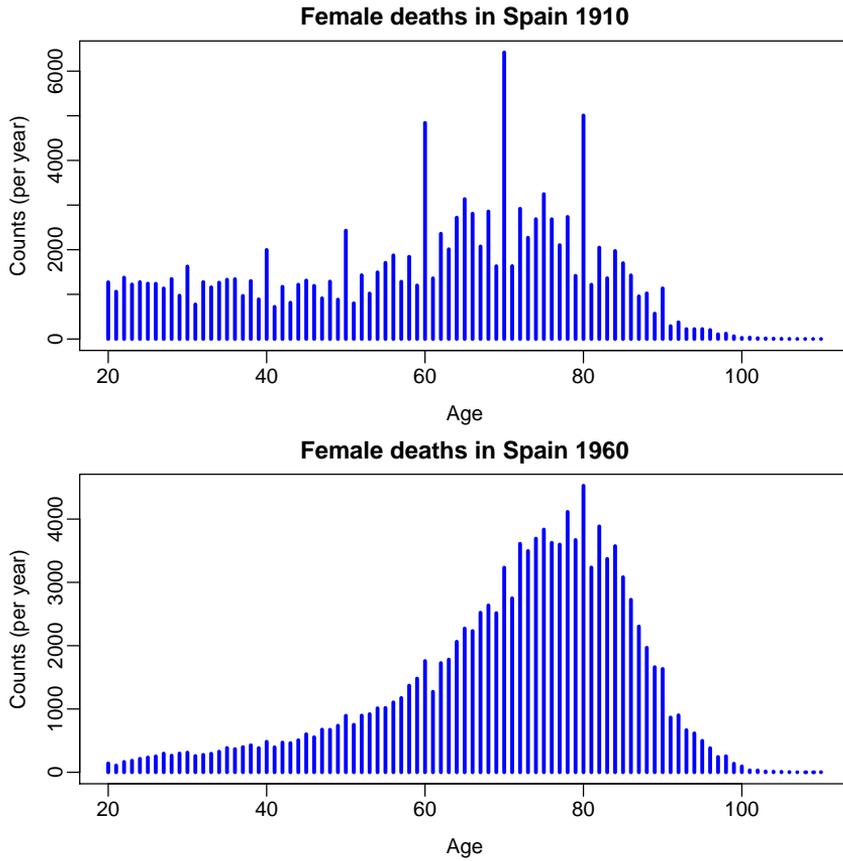


FIGURE 2. Counts of female deaths in Spain. In 1910 strong digit preference occurred. Over the years it has disappeared gradually.

these patterns: $c_{j-1,j} = p_1$, $c_{j+1,j} = p_2$ and $c_{j,j} = 1 - p_1 - p_2$. In this way the whole matrix can be filled with a diagonal band of probabilities. A more sophisticated model would involve transfers to second neighbors. Note that here μ has the same dimension as γ and C is a square matrix. This example assumes that we know the probabilities, have observed y and want to estimate γ . In practice we have to estimate the probabilities from the data as well. Details can be found in Camarda et al. (2009).

The final example comes from genetics. It is well known that normal human DNA is contained in pairs of chromosomes. The majority of the DNA is identical in all people, but in many places we find so-called single nucleotide polymorphisms (SNP), where individuals differ. For the purpose of this example it is sufficient to know that a SNP can be in one of two states, indicated by 0 and 1. So a set of neighboring SNPs on one chromosome can

be represented by a binary vector, which is called a haplotype. The same is true for the other chromosome, and the vectors generally are different. With L SNPs, 2^L different vectors are possible. It is not possible to observe individual haplotypes, only their sum, the genotype, which is a ternary vector (its elements can have the values 0, 1 or 2). We want to estimate in the probabilities of all possible haplotypes, when we have observed counts of genotypes.

The haplotype problem can be written as a CLM if β represents the logarithms of the haplotype probabilities, and $\gamma = \exp(X\beta)$ the probabilities of ordered pairs of haplotypes, called diplotypes, with a proper choice of X . The genotypes are unordered pairs and C connects them to the diplotypes. There are 2^L haplotypes, 4^L diplotypes and 3^L genotypes possible. See Uh and Eilers (2011) for details.

In the first two examples there is a natural order and it is reasonable to assume an underlying smooth distribution. In the next section it will be shown how to introduce a difference penalty to enforce it. In the haplotype example no natural order is present. There a ridge-like penalty will be added, that pushes estimated probabilities towards independence of the SNP states.

The list of references points to a number of applications with a similar flavor. Heisterkamp et al (1999) compute a time series of HIV infections, using the known distribution of incubation times and observed new AIDS cases. Eilers (2007) models a strongly over-dispersed discrete distribution of environmental complaints. Lambert and Eilers (2009) use the CLM for Bayesian analysis of grouped counts. Lambert (2011) extended this idea to the two-dimensional case. Yavuz and Lambert (2011) analyze interval censored survival data. Van den Hout (2010) et al analyzed randomized response models (without a penalty).

3 The penalized composite link model

This section presents a short technical overview of the (penalized) CLM. The model is given by

$$\mu = C\gamma = C \exp(\eta) = C \exp(X\beta); \quad y \sim \text{Pois}(\mu).$$

The design matrix X may be determined by the subject matter, as in the haplotype problem, or it may be a convenient basis, e.g. of B-splines, to represent a smooth series in a compact way. In the latter case, when γ is not too large, X may even be the identity matrix and $\gamma = e^\beta$. We observe a vector of counts y , assumed to be drawn from Poisson distributions with expected values μ .

Thompson and Baker (1981) showed how to estimate the parameter vector β by iteratively solving

$$(\check{X}'\check{W}\check{X})\beta = \check{X}'(y - \check{\mu} + \check{W}\check{X}\check{\beta}). \quad (1)$$

Here a tilde, as in $\tilde{\beta}$, indicates the current approximation to the solution, $W = \text{diag}(\mu)$ and $\check{X} = \tilde{W}^{-1}C\tilde{\Gamma}X$, with $\Gamma = \text{diag}(\gamma)$. Notice that the equation (1) has exactly the same structure as that for a GLM. The difference is that in a GLM we would have X while here \check{X} occurs in (1). Thus \check{X} is a “working” X matrix.

Suppose we subtract a penalty $\lambda\beta'P\beta/2$ from the log-likelihood, to constrain the size or the roughness of β . That would lead to the following modification of (1):

$$(\check{X}'\tilde{W}\check{X} + \lambda P)\beta = \check{X}'(y - \tilde{\mu} + \tilde{W}\check{X}\tilde{\beta}). \quad (2)$$

The penalty brings only a small change. However, we should be careful when we constrain the size of β . We have that $\gamma = \exp(X\beta)$, so if we push β in the direction of zero, we push γ towards 1, which is probably not what we want. In most applications, like in the haplotype problem, it will make more sense to push β in the direction of a given vector, say α , and the penalty becomes $\lambda\|\beta - \alpha\|^2/2$. The iterative estimating equation change to

$$(\check{X}'\tilde{W}\check{X} + \lambda I)\beta = \check{X}'(y - \tilde{\mu} + \tilde{W}\check{X}\tilde{\beta}) + \lambda\alpha. \quad (3)$$

It is desirable to have an automatic, data-driven, procedure to decide on a “good” value of the penalty parameter λ . One choice is AIC, combining deviance and effective model dimension: $\text{AIC} = \text{Dev} + 2 * \text{ED}$. Here

$$\text{Dev} = 2 \sum_i y_i \log(y_i/\hat{\mu}_i); \quad \text{ED} = \text{tr}[(\check{X}'\hat{W}\check{X} + \lambda P)^{-1}(\check{X}'\hat{W}\check{X})].$$

As usual, we have to be careful when using AIC (or any other information criterion). If the data are over-dispersed relative to the Poisson distribution, or if there is un-modeled serial correlation, AIC will indicate light smoothing. That is not a fault of AIC, but a consequence of incorrect modeling assumptions.

In many cases the algorithm of Thompson and Baker works without any changes and converges quickly. But in other cases it can diverge and one has to apply more or less sophisticated line search methods in every iteration. A simple approach is to evaluate the penalized log-likelihood at a proposed new value of the parameter vector. If it shows a decrease, accept the change. If not try a new position halfway the proposal and the old one. Repeat this halving of the interval as long as needed.

4 Two new applications

In Section 2 I mentioned a number of applications of the (penalized) CLM. In this section I will present two new ones, that were inspired by practical consulting problems. One is non-parametric density estimation after taking logarithms. This is not challenging when the data are given with high

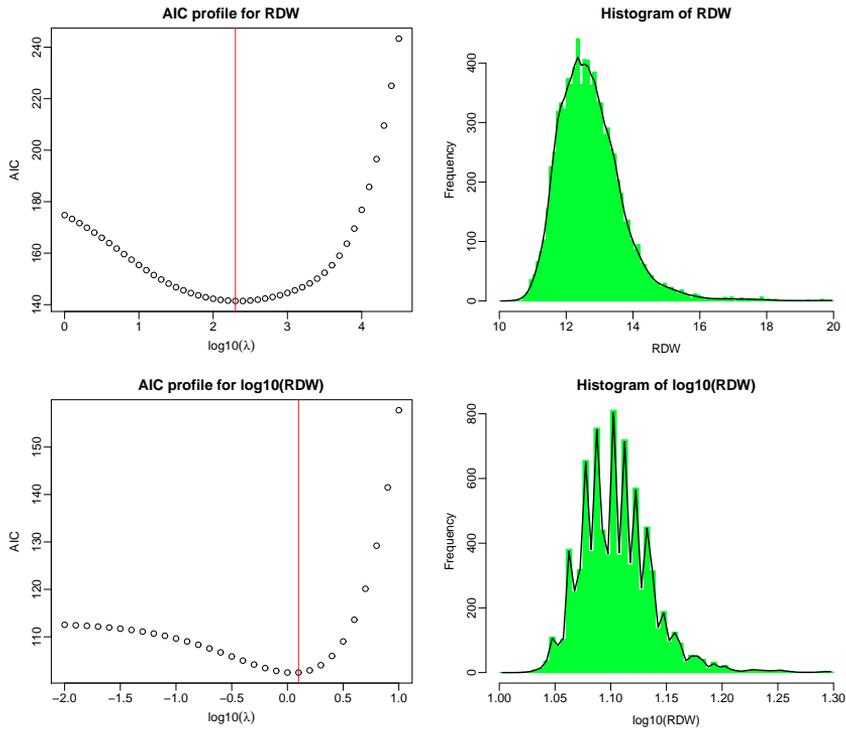


FIGURE 3. Histograms of RDW (red cell distribution width) on linear and logarithmic scales. Discrete smoothing was applied, minimizing AIC to set the penalty, resulting in the red lines. The disturbing pattern in the histogram of logarithms is caused because the original data were rounded to one decimal place.

enough precision. But when they have been rounded, automatic optimization of the amount of smoothing, using AIC, no longer works. The other application comes from X-Ray diffractometry (XRD). The X-Ray source delivers radiation at two slightly shifted wavelengths. This causes each peak in a diffraction profile to show a “shadow” with known relative size and at a known distance. The goal is to remove the shadows.

Figure 3 shows histograms and non-parametric smooths, based on a penalized GLM, which is equivalent to the PCLM with identity matrices for C and X . The amount of smoothing was decided by minimizing AIC; the profiles are shown in the same figure. The data come from an epidemiological database at the Erasmus University Medical Centre; they represent RDW (red cell distribution width), a property of the distribution of the size of red blood cells. On the linear scale everything looks quite fine: the histogram is OK and the smooth estimate is not too bad, although it might be a bit smoother. But we notice a skew distribution, so it might be a good

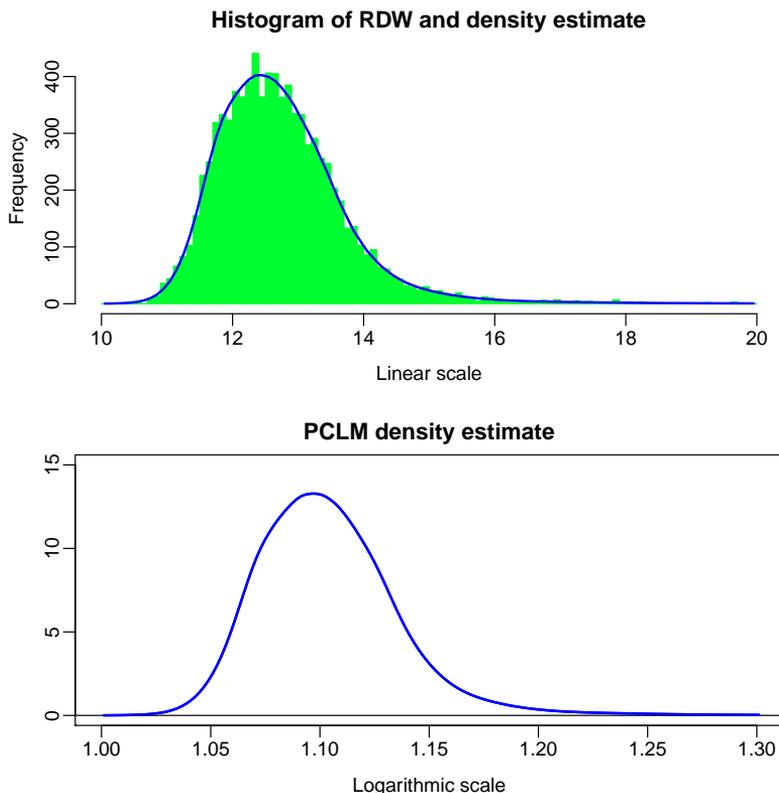


FIGURE 4. Histogram smoothing on the logarithmic scale, using the PCLM. The top panel shows the histogram on the original scales and the back-transformed density. The lower panel shows the estimated density on the logarithmic scale.

idea to take logarithms first. This was done and the result is shown in the lower part of the figure. It looks terrible, what happened? The original data were rounded (by the measuring equipment) to one decimal. After taking logarithms, some of the new bins take the count in one 0.1 wide bin on the linear scale, while other takes the counts of two of them. This explains the comb-like pattern. The large systematic variation is picked up by the smoother as being a real signal, leading to a very wild result.

Actually this phenomenon is quite common, because it is unavoidable when transforming rounded data. Even linear transformation are not free of it. Think of distances that have been reported as whole numbers of feet, and are summarized as counts in one-meter intervals. This no problem for parametric models, but it is disastrous for automatic smoothing.

The PCLM is ideally suited for this problem. The goal is to estimate a smooth density on the logarithmic scale, which is γ (after normalization).

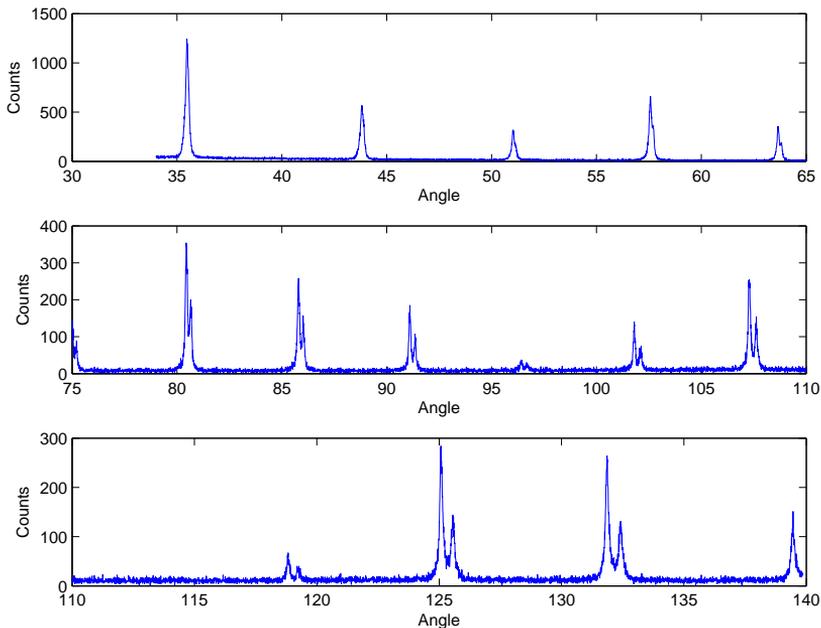


FIGURE 5. A complete X-Ray diffraction profile, distributed over three panels (with unequal vertical scales).

It is short enough that no matrix X is needed, so $\gamma = e^\beta$. The elements of C are computed as follows: c_{ij} is the fraction of logarithmic interval j that overlaps with linear interval i . Working this way and using AIC again to choose the penalty weight, we arrive at the results shown in Figure 4. The improvement is clear.

The data for the second application are shown in Figure 5. The peaks correspond to layers of a crystal, which is slowly rotated relative to a beam of X-Ray radiation; two times the rotation angle, indicated by ϕ , is on the horizontal axis. All peaks consist of two components, although this is hardly visible for the ones at low ϕ . The double peaks do not reflect a property of the crystal, but of the X-Ray source, which emits radiation at two close wavelengths, one being about 0.2% larger than the other. As can be seen from the data, the shift between the two components increases with ϕ . It can be computed exactly, using Braggs' diffraction law, as

$$\delta = 2 \arcsin(\rho \sin(\phi/2)) - \phi,$$

where ρ is the ratio between the longest and the shortest wavelength. The instrument counts photons, so we model the data as Poisson distributed with expectation $\mu = \check{\gamma} + \bar{\gamma}$, where $\check{\gamma}$ represent the main peaks

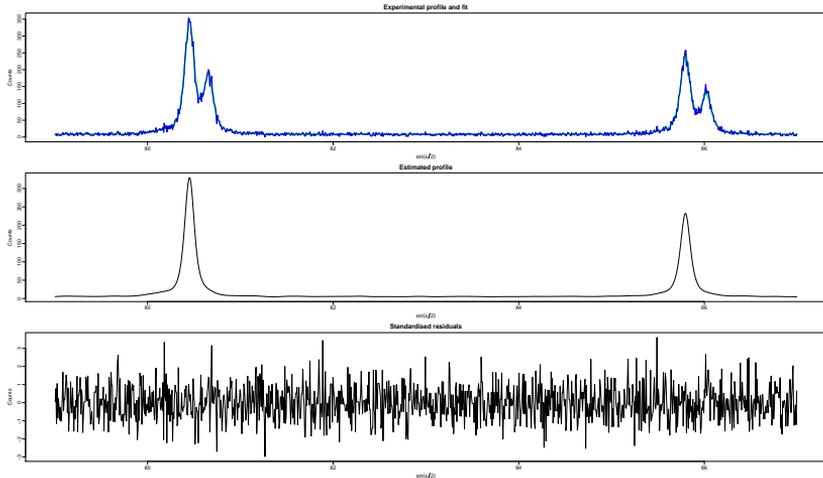


FIGURE 6. The penalized composite link model used to eliminate shadow peaks for a segment of the X-Ray diffraction profile. Top: data (blue) and fit (green). Middle: the estimated main profile $\check{\gamma}$. Bottom: standardized residuals, $(y - \hat{\mu}) / \sqrt{\hat{\mu}}$.

and $\bar{\gamma}$ the shifted ones. Let

$$\check{\gamma} = \exp(\check{B}\alpha); \quad \check{\gamma}_j = \exp\left(\sum_k \alpha_k B_k(\phi_j)\right)$$

be a B-spline representation of $\check{\gamma}$ and

$$\bar{\gamma} = \exp(\bar{B}\alpha); \quad \bar{\gamma}_j = \exp\left(\sum_k \alpha_k B_k(\phi_j + \delta_j)\right).$$

We use the same parameters (knots and spline degree) for both bases, but one is evaluated at ϕ and the other at $\phi + \delta$. The relative strength of the second radiation component is also known, it is $\tau = 0.47$. If we now form matrices and vectors as

$$B' = [\check{B}' \mid \bar{B}']; \quad \gamma' = [\check{\gamma}' \mid \bar{\gamma}']; \quad C = [I \mid \tau I],$$

with I an m -by- m identity matrix (m is the length of ϕ), we are in the PCLM world again. Results are shown in Figure 6. The shadow peaks have been eliminated completely. The standardized residuals show that the Poisson assumption is realistic: their RMS value is 0.97 and they show no structure.

Some computational details. The complete XRD profile is almost 9000 observations long. The peaks are quite sharp, so a large number of B-splines is needed. It is practical to work on shorter segments, with a small overlap. In any case all matrices allow a very efficient sparse representation. I can recommend the R package `spam` for that.

5 Discussion

The (penalized) composite link model deserves more attention from the statistical community. I have described three existing applications in some detail, mentioned many others, and presented two new ones.

Rabe-Hesketh and Skrondal (2007) suggest extensions that have not been explored yet. Between μ and $C\gamma$ a second link function could be inserted. They even consider the possibility that each element of C works as a function on γ instead as a weighted contribution to a sum.

Looking back, it is remarkable and gratifying to see that most recent publications on this subject started out as contributions to International Workshops on Statistical Modelling. It is my hope that future meetings will turn out to be a breeding ground for many more fruitful ideas.

References

- Camarda, C.G., Eilers, P.H.C., and Gampe, J. (2008) Modelling general patterns of digit preference. *Statistical Modelling*, **8**, 385–401.
- Camarda, C.G., Eilers, P.H.C. and Gampe, J. (2009) Modelling trends in digit preference patterns. *Proceedings of the 24th International Workshop on Statistical Modelling*.
- Eilers, P.H.C. (2007) III-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, **7**, 239–254.
- Eilers, P.H.C, and Borgdorff M. (2003) Analysis of TB mortality patterns in The Netherlands in the first half of the twentieth century. *Unpublished manuscript*.
- Eilers, P.H.C., and Borgdorff, M. (2004) Modeling and correction of digit preference in tuberculin surveys. *International Journal of Tuberculosis and Lung Disease*, **8**, 232–239.
- Lambert, P. (2011) Smooth semiparametric and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data, *Computational Statistics & Data Analysis*, **55**, 429–445.
- Lambert, P., and Eilers, P. H. C. (2009) Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis*, **53**, 1388–1399.
- Rabe-Hesketh, S., and Skrondal A. (2007) Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika*, **72**, 123–140.
- Skrondal, A., and Rabe-Hesketh, S. (2007) Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, **34**, 712–745.

- Uh, H.-W., and Eilers, P.H.C.(2011) Haplotype Estimation from Fuzzy Genotypes Using Penalized Likelihood. *PLOS ONE*, **6**.
- van den Hout, A., Gilchrist, R., and van der Heijden, P.G.M. (2010) The randomized response log linear model as a composite link model. *Statistical Modelling*, **10**, 57–67.
- Yavuz, A.C., and Lambert, P.(2011) Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines *Statistics in Medicine*, **30**, 75–90.

Beyond mean regression

Thomas Kneib¹

¹ Chair of Statistics, Department of Economics, Georg-August-University
Göttingen, Germany

E-mail for correspondence: tkneib@uni-goettingen.de

Abstract: Usual exponential family regression models focus on only one designated quantity of the response distribution, namely the mean. While this entails easy interpretation of the estimated regression effects, it may often lead to incomplete analyses when more complex relationships are indeed present and also bears the risk of false conclusions about the significance / importance of covariates. We will therefore give an overview on extended types of regression models that allow us to go beyond mean regression. More specifically, we will study generalized additive models for location, scale and shape as well as semiparametric quantile and expectile regression.

Keywords: expectile regression, GAMLSS, quantile regression, semiparametric regression

1 Introduction

Consider a regression situation with n observations (y_i, \mathbf{z}_i) , $i = 1, \dots, n$, on a continuous response variable y and covariates \mathbf{z} . Then a typical regression models for the mean takes the form

$$y_i = \eta_i + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2,$$

with regression predictor η_i specified in dependence of the covariate \mathbf{z} (we will discuss specific choices later). The two assumptions on the error term imply that, on the one hand, the predictor describes the expectation of the response since

$$E(y_i) = \eta_i + E(\varepsilon_i) = \eta_i$$

and, on the other hand, that ordinary least squares estimation can be used due to homoscedastic errors. Typically, the error term will additionally be assumed to follow a normal distribution such that ε_i i.i.d. $N(0, \sigma^2)$.

Such a mean regression model has the advantage of being easy to understand and estimate and to entail easy interpretation of the regression effects contained in the predictor. However, it is often also too restrictive due to the strong assumptions on the error term. For example, in case of

heteroscedasticity, also the variance (or the standard deviation) of the response may depend on covariates. This can easily be incorporated in the model formulation by modifying the regression equation to the location-scale model

$$y_i = \eta_{i1} + \exp(\eta_{i2})\varepsilon_i, \quad E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = 1, \quad (1)$$

with two predictor structures η_{i1} for the mean and η_{i2} for the standard deviation such that

$$E(y_i) = \eta_{i1} \quad \text{Var}(y_i) = \exp(\eta_{i2})^2.$$

When the errors additionally follow a normal distribution, this is the easiest example of a regression model for location, scale and shape (GAMLSS, in fact without an effect on the shape) as introduced by Rigby and Stasinopoulos (2005) as a comprehensive class of models where different parameters of the response distribution are related to regression predictors. We will introduce GAMLSS in more detail in Section 2.2.

While GAMLSS retain the assumption of a parametric distribution for the responses (or equivalently the error terms) it may also be useful to completely drop this assumption and to formulate nonparametric models that still enable to describe more than the mean of the response. This may in particular be the case if interest is not on identifying covariate effects on specific parameters of the response distribution but on the relation of “extreme” observations in the tails of the distribution on covariates. This is enabled in quantile and expectile regression models where we go back to the initial model formulation

$$y_i = \eta_{i\tau} + \varepsilon_{i\tau},$$

but modify the assumptions on the error terms appropriately to model observations in the tails as denoted by the asymmetry parameter $\tau \in [0, 1]$ that specifies the desired “extremeness” and is therefore added to both the predictor and the error terms as a subscript. In quantile regression, we assume that the τ -quantile of the error term is zero (i.e. $F_{\varepsilon_{i\tau}}(0) = \tau$, where $F_{\varepsilon_{i\tau}}(\cdot)$ denotes the cumulative distribution function of the i -th error term). This assumption implies that the predictor $\eta_{i\tau}$ specifies the τ -quantile of y_i and, as a consequence, the regression effects can be interpreted on the quantiles of the response distribution. Estimating results for a dense set of quantiles then also allows to characterize the complete distribution of the responses in terms of covariates, see Section 2.3 for details. An alternative to quantile regression is expectile regression where basically the assumptions on quantiles are replaced with expectiles which provide an alternative way of describing the tails of distributions in terms of a generalization of the mean instead of a generalization of the median as in quantile regression. We will introduce expectiles in more detail in Section 2.4.

Regression models going beyond mean regression are of interest in several areas of application. Some examples, we have worked on so far include

- the Munich rental guide, where interest is on determining flexible interval estimates for ranges of usual rents of flats instead of only point predictions for rents.
- childhood malnutrition in developing countries, where the impact of covariates on extreme forms of malnutrition is of higher relevance than models for the average nutritional status.
- efficiency estimation in agricultural production, where we are particularly interested in covariates impacting above-average performance of farms.
- modelling gas flow networks, where the behavior of the network in high or low demand situations shall be studied.

In the remainder of this paper, we will first introduce the class of semiparametric predictor structures we are interested in (Section 2.1), followed by brief introductions to GAMLSS as well as quantile and expectile regression. We will then discuss different estimation principles and their suitability for each of the model classes (Section 3). Finally, we will summarize advantages and disadvantages of GAMLSS, quantile and expectile regression both from a methodological and an applied perspective (Section 4).

2 Model Types

2.1 Semiparametric Regression Models

Instead of restricting our attention to regression models with linear predictors $\eta_i = \mathbf{z}_i' \boldsymbol{\beta}$, we are interested in semiparametric regression models with predictors of the generic form

$$\eta_i = \beta_0 + \sum_{j=1}^p f_j(\mathbf{z}_i) \quad (2)$$

where β_0 is an intercept and the functions $f_j(\mathbf{z}_i)$ reflect different types of regression effects depending on subsets of the covariate vector \mathbf{z}_i . Associated with each function is a penalty term $\lambda_j \text{pen}(f_j)$ that enforces specific properties of the function such as smoothness or sparsity and $\lambda_j \geq 0$ are the corresponding smoothing parameters that govern the impact of the penalty. A broad and flexible class of function types is obtained with the following assumptions:

- The functions f_j are approximated in terms of basis function representations

$$f_j(\mathbf{z}) = \sum_{k=1}^K \beta_{jk} B_k(\mathbf{z})$$

where $B_k(\mathbf{z})$ are the basis functions and β_{jk} denote the corresponding basis coefficients.

- The penalty is quadratic in the vector of basis coefficients $\beta_j = (\beta_{j1}, \dots, \beta_{jK})'$, i.e.

$$\text{pen}(f_j) = \beta_j' \mathbf{K}_j \beta_j$$

with penalty matrix \mathbf{K}_j chosen such that the desired regularisation properties are achieved. In a Bayesian formulation, we would equivalently assume that the regression coefficients are assigned a normal prior $\beta_j \sim N(\mathbf{0}, \tau_j^2 \mathbf{K}_j^-)$, where the penalty matrix defines the precision of the normal distribution and \mathbf{K}_j^- denotes the generalized inverse.

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. Note that the semiparametric predictor (2) may either act on specific parameters of the response distribution in case of GAMLSS or on quantiles / expectiles of the response distribution as detailed in the following sections.

2.2 Generalized Additive Models for Location, Scale and Shape (GAMLSS)

GAMLSS provide a unified framework for estimating semiparametric regression models when assuming that the responses y_i follow distributions depending on up to four parameters $(\mu_i, \sigma_i, \nu_i, \xi_i)$, where usually μ_i and σ_i are a location and a scale parameter, respectively, while ν_i and ξ_i correspond to shape parameters such as skewness or kurtosis. The limitation to four parameters is only chosen for convenience since common distributions rarely have more than four distributional parameters and because interpretation becomes quite messy in more complex cases. Each of the distributional parameters is related to a predictor via a suitable link function, i.e.

$$\mu = g_1(\eta_1), \quad \sigma = g_2(\eta_2), \dots$$

The class of distributions covered by GAMLSS is very broad and comprises, at the moment, about 50 different distributions. For continuous responses, the most prominent examples are the normal distribution (with two parameters), the power exponential distribution (with three parameters), the gamma distribution (with up to three parameters), the t-distribution (with three parameters) or the Box-Cox power exponential distribution (with four parameters). Estimation in GAMLSS usually relies on likelihood principles and requires the (at least numerical) availability of first (and optimally

second) derivatives to facilitate optimization (but we will also discuss some alternatives in Section 3). The major advantages of GAMLSS is that the predictors act directly on interpretable response quantities and therefore facilitate the understanding of the estimated regression effects.

2.3 Quantile Regression

Quantile regression for the τ -quantile starts from the model

$$y_i = \eta_{i\tau} + \varepsilon_{i\tau}, \quad F_{\varepsilon_{i\tau}}(0) = \tau$$

where $F_{\varepsilon_{i\tau}}(\cdot)$ denotes the cumulative distribution function of $\varepsilon_{i\tau}$. This defines a specification that is similar to usual mean regression but replaces the assumption of zero means for the error terms with the assumption of zero τ -quantiles. As a consequence, the predictor $\eta_{i\tau}$ is the τ -quantile of the response y_i since

$$\tau = F_{\varepsilon_{i\tau}}(0) = P(\varepsilon_{i\tau} \leq 0) = P(\eta_{i\tau} + \varepsilon_{i\tau} \leq \eta_{i\tau}) = P(y_i \leq \eta_{i\tau}) = F_{y_i}(\eta_{i\tau}).$$

Note that no further assumptions are made on the error terms and therefore quantile regression is also applicable in situations with heteroscedastic error terms. In fact, quantile regression is only of interest in such situations, where not only the mean depends on covariates but also other properties of the response distribution.

Classical estimation in quantile regression relies on optimizing an asymmetrically weighted absolute error criterion. Recall that empirical quantiles \hat{q}_τ based on an i.i.d. sample of observations y_1, \dots, y_n can be estimated as

$$\hat{q}_\tau = \operatorname{argmin}_q \sum_{i=1}^n w_\tau(y_i, q) |y_i - q|$$

with asymmetric weights

$$w_\tau(y_i, q) = \begin{cases} 1 - \tau & y_i < q \\ 0 & y_i = q \\ \tau & y_i > q \end{cases}$$

that basically weight observations below and above the quantile of interest differently to shift the estimate to upper or lower parts of the sample. Regression quantiles can then in analogy be determined by replacing the common quantile q with the predictor $\eta_{i\tau}$ of the semiparametric regression model and augmenting the penalty terms, yielding

$$\sum_{i=1}^n w_\tau(y_i, \eta_{i\tau}) |y_i - \eta_{i\tau}| + \sum_{j=1}^p \lambda_j \operatorname{pen}(f_j). \quad (3)$$

We will give an overview on possible optimisation approaches in the next section.

Note that in fact any parametric regression model also implicitly defines a quantile regression model. For example, in case of a simple mean regression model with homoscedastic normal errors, the τ -quantile of response y_i is given by $\eta_i + \sigma z_\tau$, where z_τ denotes the τ -quantile of the standard normal distribution, and as a consequence the simple mean regression structure implies parallel quantile curves. This can be overcome in the location-scale model (1), where the quantiles are determined as $\eta_{i1} + \exp(\eta_{i2})z_\tau$, but flexibility is still limited as compared to “real” quantile regression models.

2.4 Expectile Regression

Expectile regression is obtained when replacing the asymmetric absolute deviations employed in quantile regression with asymmetric quadratic deviations, yielding the optimisation criterion

$$\sum_{i=1}^n w_\tau(y_i, \eta_{i\tau})(y_i - \eta_{i\tau})^2 + \sum_{j=1}^p \lambda_j \text{pen}(f_j).$$

This criterion exhibits a closer connection to ordinary least squares estimation and in fact usual mean regression appears as a special case with $\tau = 0.5$. We will compare quantile and expectile regression in more detail in the following sections but already note here that expectiles are an alternative way of characterising tail properties of a distribution. The theoretical τ -expectile e_τ for a random variable y is obtained as the solution of the equation

$$\tau = \frac{\int_{-\infty}^{e_\tau} |y - e_\tau| f_y(y) dy}{\int_{-\infty}^{\infty} |y - e_\tau| f_y(y) dy} = \frac{G_y(e_\tau) - e_\tau F_y(e_\tau)}{2(G_y(e_\tau) - e_\tau F_y(e_\tau)) + (e_\tau - \mu)}$$

where $f_y(\cdot)$ and $F_y(\cdot)$ denote the density and cumulative distribution function of y , $G_y(e) = \int_{-\infty}^e y f_y(y) dy$ is the partial moment function of y and $G_y(\infty) = \mu$ is the expectation of y . While expectiles do not enjoy the easy interpretation of quantiles, they have a number of computational advantages that we will detail in the next sections.

3 Inference

In this section, we will review several possible estimation principles for the three types of models introduced in the previous section. We will start with direct optimisation approaches, will then discuss Bayesian inference and finally consider functional gradient descent approaches from statistical learning theory.

3.1 Direct Optimisation

In this section, we will focus on approaches that directly aim at optimizing a suitable optimization criterion that may either be given by the likelihood or a criterion like (3). For quantile regression, linear programming is the standard approach in parametric approaches that allows for routine and fast optimisation of the asymmetrically weighted L_1 -loss function. This approach can still be used in combination with L_1 penalty terms arising for example from the LASSO or in total variation penalization for spline regression but can not easily be combined with the class of quadratic L_2 penalties that we are interested in. In addition, simultaneous estimation of smoothing parameters is still challenging and largely unsolved in this framework.

For expectile regression, estimation typically relies on iteratively weighted least squares estimates

$$\hat{\boldsymbol{\beta}}_j^{[t+1]} = (\mathbf{B}'_j \mathbf{W}_\tau^{[t]} \mathbf{B}_j + \lambda_j \mathbf{K}_j)^{-1} \mathbf{B}'_j \mathbf{W}_\tau^{[t]} \mathbf{y},$$

where \mathbf{B}_j is the design matrix associated with the j -th model term, \mathbf{y} is the vector of responses and $\mathbf{W}_\tau = \text{diag}(w_\tau(y_1, \eta_{1\tau}), \dots, w_\tau(y_n, \eta_{n\tau}))$ is a diagonal matrix containing the weights. Since the weights also depend on the current estimates, an iteration loop is required to obtain the final estimates. Still, the approach already shows that the asymmetrically weighted quadratic loss fits well with the class of quadratic penalties we are considering. Expectile regression also enables the incorporation of smoothing parameter selection, for example by making use of the mixed model representation of penalized regression. Iteratively weighted least squares estimation can also be used to determine quantile regression estimates when the weights are modified to

$$\tilde{w}_\tau(y_i, \eta_{i\tau}) = \frac{w_\tau(y_i, \eta_{i\tau})}{|y_i - \eta_{i\tau}|}.$$

For GAMLSS, the optimality criterion is given by the log-likelihood and, given that first (and optimally second) derivatives of the log-likelihood are available, Newton-type optimization can be performed to obtain the estimates. Such derivative-based optimization approaches also easily incorporate quadratic penalties since these can be differentiated with respect to the parameters and therefore induce only minor additions to the score vector and Hessian matrix.

3.2 Bayesian Inference

While Bayesian inference is obviously an alternative for estimation in GAMLSS, it seems to be more complicated to relate Bayesian approaches to the nonparametric formulations of quantile and expectile regression. However,

due to the formal equivalence between penalized estimates and posterior modes based on suitable auxiliary error distributions, such a connection is indeed possible. For example, in case of quantile regression, assuming the asymmetric Laplace distribution $\text{ALD}(0, \sigma^2, \tau)$ with density

$$p_{\varepsilon_{i\tau}}(\varepsilon_i) = \frac{\tau(1-\tau)}{\sigma^2} \exp\left(-w_\tau(\varepsilon_i, 0) \frac{|\varepsilon_i|}{\sigma^2}\right).$$

for the error terms induces the likelihood

$$\exp\left(-\sum_{i=1}^n w_\tau(y_i, \eta_{i\tau}) \frac{|y_i - \eta_{i\tau}|}{\sigma^2}\right)$$

and therefore maximizing the penalized likelihood is equivalent to minimizing (3). The main advantage of this connection is that – via a location-scale mixture representation of the asymmetric Laplace distribution – the Bayesian model formulation can be related to Gaussian regression models with an offset and weights. This enables Bayesian inference either based on Markov chain Monte Carlo simulations within a Gibbs sampler or approximate variational Bayes inference and, in particular, incorporates estimation of the smoothing parameters which was a major difficulty when using linear programming. Similarly, an asymmetric normal distribution can be considered in Bayesian expectile regression but this avenue has not yet been explored in detail (most probably since estimation via iteratively weighted least squares already allows for data-driven determination of smoothing parameters).

3.3 Boosting

A third way to perform inference in any of the three model classes considered in this paper is given by functional gradient descent boosting, an approach that originated from the machine learning community for optimization in complex problems. The basic idea is to fit simple base-learning procedures to iteratively updated gradients of an optimization problem to achieve a final estimate. When using a componentwise boosting approach, where separate base-learners are specified for each of the model components in the semiparametric predictor (2), boosting has the particular advantage that it combines model estimation (including data-driven determination of the appropriate amount of smoothness) with automatic variable selection and model choice.

Boosting algorithms have been developed for both quantile and expectile regression (where the gradients of the optimality criteria are easy to derive) as well as GAMLSS where the derivatives of the log-likelihood are required also in the Newton-type optimisation procedures.

4 Summary and Conclusions

In this section, we will summarize a number of advantages and disadvantages for each of the three approaches discussed in the previous sections. GAMLSS require a distributional assumption for the responses and restrict inference to specific response properties. This can be seen both as an advantage and a disadvantage since, on the one hand, it enables easy and direct interpretation of estimated effects but on the other hand may be too restrictive in certain situations. The assumption of one joint parametric model also has the advantage that derived results for quantiles or expectiles will always be consistent in the sense that crossing quantile or expectile curves can not occur. Finally, GAMLSS can be treated in both the frequentist likelihood-based and the Bayesian framework.

Quantile regression allows to work fully nonparametrically and still allows for an easy interpretation of the regression effects in terms of conditional quantiles for the response distribution. Standard linear programming tools seem less suited for complex semiparametric predictor specifications involving quadratic penalties due to the lack of automatic smoothing parameter selection. As a consequence, the Bayesian formulation of quantile regression seems to provide a promising alternative that enables great flexibility. One theoretical disadvantage of quantile regression is that it estimates a step function for the continuous cumulative distribution function (similar as the empirical cumulative distribution function is a step function). This may also be an issue when studying the efficiency of quantile regression relative to quantiles derived from expectile regression.

Finally, expectile regression has the great advantage that it avoids complex optimization approaches since it relies on iteratively weighted least squares fits. Yet it enables large flexibility in the model specification in combination with automatic smoothing parameter determination. In addition, the estimated expectile function for the response distributions is a smooth estimate and can also be used to estimate quantiles of interest. Therefore, the drawback of less intuitive interpretation can at least partly be compensated. In addition, simulation-based evidence seems to suggest that crossing expectile curves are less frequent than crossing quantile curves. We are also currently studying the efficiency of quantiles derived from expectiles relative to that of quantile regression and have some preliminary, promising results.

Acknowledgments: This paper summarizes joint work with Nora Fenske, Benjamin Hofner, Torsten Hothorn, Göran Kauermann, Stefan Lang, Andreas Mayr, Matthias Schmid, Linda Schulze Waltrup, Fabian Sobotka, Elisabeth Waldmann and Yu Yue. Financial support by the German Research Foundation (DFG), grant KN 922/4-1 is gratefully acknowledged.

References

- Fenske, N., Kneib, T. and Hothorn, T. (2011). Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression. *Journal of the American Statistical Association*, **106**, 494–510.
- Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.
- 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**, 403–427.
- Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. Cambridge: Cambridge University Press. *Applied Statistics*, **54**, 507-554.
- Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, **56**, 755-767.
- Sobotka, F., Kauermann, G., Schulze Waltrup, L. and Kneib, T. (2012). On Confidence Intervals for Geoadditive Expectile Regression. *Statistics and Computing*, to appear.
- Waldmann, E., Kneib, T., Lang, S. and Yue, Y. (2012). Bayesian Semiparametric Additive Quantile Regression. Technical Report.

Estimating cross-sectional incidence from biological markers

Michal Kulich¹, Arnošt Komárek¹, Marek Omelka¹

¹ Dept. of Probability and Statistics, School of Mathematics and Physics, Charles University, Prague, Czech Republic

E-mail for correspondence: kulich@karlin.mff.cuni.cz

Abstract: Incidence of an infectious disease such as HIV is traditionally estimated by following a cohort of susceptible individuals and registering new cases that occur during the follow-up period. However, in certain situations cohort follow-up may be unfeasible but a cross-sectional study can be done easily. Biological markers correlated with time since infection have been proposed for estimating HIV incidence in cross-sectional studies. The incidence is estimated from the number of “recent infections”, i.e. the samples whose biomarker is below some prespecified threshold. However, it has been observed that the cross-sectional incidence estimates obtained in this way can be severely biased.

We investigate the theoretical sources of the bias in cross-sectional incidence estimation and identify conditions on the biomarker that make the bias tolerable. We show how to evaluate a collection of several biomarkers measured in conjunction on a set of infected blood samples with “known” times since infection and explain how to use such validation data to develop cross-sectional incidence estimates with optimized performance across a range of different populations.

Keywords: HIV; sensitivity; specificity; bias

Background

Disease incidence is traditionally measured by following a cohort of disease-free individuals and recording new cases over a period of time. A consistent incidence estimator is obtained by dividing the number of new cases with the total duration of the follow-up. For infectious diseases such as HIV, the ascertainment of the new cases requires repeated testing of the enrolled individuals. This method is expensive, time-consuming, and bias-prone. Selection bias is an issue because it is difficult to enroll and maintain cohorts that are representative of the target population. With HIV, future incidence could be affected by counselling, which is an integral part of the testing procedure, and the knowledge of the disease status.

To overcome these limitations, methods have been developed for measuring HIV incidence by a cross-sectional study. With this approach, a single

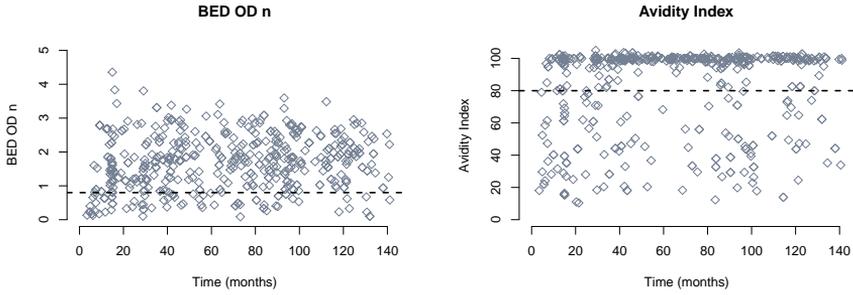


FIGURE 1. Scatterplots of BED assay (left panel) and avidity index (right panel) against estimated time since infection in samples infected by HIV subtype D.

blood sample is obtained from each participant and tested for HIV infection. Samples that are found to be HIV-positive undergo further testing in order to measure one or more biological markers associated with the duration of the infection. HIV-positive samples with biomarker level below a prespecified threshold are considered “recent” infections. The incidence estimator is obtained by dividing the number of recent infections by the number of HIV-negative samples and by so called “time window”, which represents the duration of the follow-up for a single person.

We consider two previously proposed biomarkers: the BED assay (Parekh et al., 2002) and the avidity index [AI] (Suligoi et al., 2003), in conjunction with CD4 cell levels. The BED assay measures the concentration of anti-HIV immunoglobulin G in the infected serum. The AI measures the strength of anti-HIV antibody binding. Both markers generally increase with the time since infection but the association is weak and non-linear (Figure 1). No matter how the thresholds on these markers are set, a certain proportion of samples that have been infected for a long time will be classified as “recent”. It follows that incidence estimates obtained from such markers are generally biased, and the bias can be quite severe. The bias cannot be adjusted for by considering the misclassification rates or by a clever choice of the time window, because these quantities vary from population to population.

Objectives

We study the bias in the cross-sectional incidence estimators based on biomarkers by considering the probability limit of the estimator. It turns out that the bias depends on quantities that vary with the investigated population and cannot be measured. We identify conditions under which the bias is small and give a formula for the time window when these conditions are met. Next, we turn our attention to the problem of selecting

optimal threshold values for procedures that combine several biomarkers, such as BED, AI, and CD4. We make use of a validation dataset consisting of thousands of samples with approximately known times since infection and available biomarker results. We consider a grid of threshold values for the biomarkers and evaluate each combination of thresholds in a simulation study with infected samples randomly picked up from the validation data according to several prespecified sampling mechanisms. We explain how to conduct such simulation studies, how to assess the performance of the incidence estimators, and to evaluate the results in order to choose a proper combination of thresholds.

References

- Parekh, B.S., Kennedy, M.S., Dobbs, T., Pau, C.P., Byers, R. *et al.* (2002). Quantitative Detection of Increasing HIV Type 1 Antibodies after Seroconversion: A Simple Assay for Detecting Recent HIV Infection and Estimating Incidence. *AIDS Research and Human Retroviruses*, **18**, 295–307.
- Suligoi, B., Massi, M., Galli, C., Sciandra, M., Di Sora, F. *et al.* (2003). Identifying Recent HIV Infections Using the Avidity Index and an Automated Enzyme Immunoassay. *Journal of Acquired Immune Deficiency Syndromes*, **32**, 424–428.

Bayesian variable selection and the (ab)use of priors

Robert B. O'Hara¹

¹ BiK-F, Biodiversity and Climate Change Research Centre, Germany

E-mail for correspondence: bohara@senckenberg.de

Keywords: Variable Selection; Bayes; Shrinkage; Priors.

1 Introduction

There are many aspects to data analysis. Here I will be concerned with just one: variable selection in generalized linear models and their extensions. These methods are useful when we have "large- p , small- N " problems, where there are many potential covariates and a limited amount of data, e.g. in identifying genetic loci that affects traits (i.e. QTL mapping and association studies) when there are many genetic markers, only a few of which actually affect a trait. It is common with this sort of data to have $p > N$. Even for less extreme problems, variable selection can be helpful for us, as there is often several a sub-set of variables that are affecting our response but which are not themselves of direct interest. Removing the variables that have little or no effect can improve the model and the estimates of the other parameters.

The Bayesian revolution over the last 20 or so years has affected many areas of statistics, and variables selection has not been immune. Writing a model as a hierarchical probability model has stimulated the development of several methods for variable selection. Much of this development has been motivated by computational concerns, i.e. how to make sure your MCMC sampler doesn't take near-infinite time to converge. One great advantage of being a Bayesian nowadays is that graphical models and MCMC techniques have freed us up to develop complex models that accurately capture what we are trying to model.

An apparent problem of variable selection is that shifting between models of different size changes the dimension of the model, which makes the model fitting more complex. However, these models can be re-constructed as fixed dimension models by using indicator variables to define which variables are "in" the model. This avoids the practical problems of fitting variable dimension models, at the computational cost of always fitting the maximal

model. But the relative ease with which these models can be developed and fitted means that they can easily be adapted to new situations and ideas. In practice almost all Bayesian model fitting is done using MCMC, which this creates some practical challenges for variable selection. The development of Bayesian variable selection can be viewed as a struggle against the limits of MCMC, which has resulted in a plethora of ingenious solutions that will be discarded once better methods for numerical integration have been developed.

My aim is to argue that, when they are not being abused, Bayesian variable selection techniques are a valuable tool in the toolbox. And their flexibility there suggests there is a scope for extending their use into new and stranger directions.

2 Approaches to Bayesian Variable Selection

Bayesian variable selection methods are all just different ways of creating a "slab and spike" prior (see O'Hara and Sillanpää, 2009, for a review). This prior has a spike at zero (or some other interesting number), representing the prior probability that the variable is not 'in' the model, or equivalently that the effect of the variable is insignificant. The slab then represents the prior density for when the variable has an appreciable effect, i.e. that it is 'in' the model. The different methods for doing this lead to different models with different properties, and it is difficult to claim that one is intrinsically superior to any other.

We can assume a regression model of a response, y_i being (potentially) explained by a vector of covariates X_{ij} :

$$E(y_i) = \sum_{j=1}^P \beta_j X_{ij} \quad (1)$$

Variable selection is a problem of deciding whether β_j can be safely ignored, i.e. removed, set to zero or to a value so small that it barely affects $E(y_i)$. For notational convenience we will sometimes write $\beta_j = I_j \theta_j$, where $I_j \in \{0, 1\}$ is an indicator function for whether the variable is in the model, and θ_j is an auxiliary variable.

2.1 Mixture distribution with a point mass

An obvious way of creating a slab and spike is to make the spike a point mass at zero. The simplest way of setting this up is to make I_j a random variable:

$$E(y_i) = \sum_{j=1}^P I_j \theta_j X_{ij} \quad (2)$$

$$I_j \sim \text{Bern}(p) \quad (3)$$

Although simple this works poorly in practice. Problems arise in the MCMC sampler because when $I_j = 0$ because θ_j makes no contribution to the likelihood, so it is sampled from its prior. If the prior is very wide compared to the posterior for β_j , most of these values will have such a low posterior probability that the MCMC sampler prefers to stick with a value of zero. This problem can be circumvented by having different priors for $\theta_j|I_j = 1$ and $\theta_j|I_j = 0$. The latter prior does not affect the likelihood, and so can be tuned to improve mixing.

A conceptually different way of producing the same model is to use a variable dimension model, i.e. only include the X_{ij} 's in eq. 1 that are "in" the model and introduce a step into the MCMC which moves between different model spaces. This can be more efficient if eq. 1 has many fewer terms. For our purposes, it is the same as the other methods in this section, as multiplying by zero obviously has the same effect as leaving the covariate out.

2.2 Mixture Distribution of two densities

Rather than making the spike a point mass at zero, we can make it a very tight peak. A simple way of doing this, called SSVS (Stochastic Search Variable Selection), is to make it a normal distribution with a small prior, and the slab is also a normal distribution but with a wider prior:

$$\beta_j \sim N(0, \sigma^2(I_j)) \quad (4)$$

with $\sigma^2(1) = c\sigma^2(0)$ and c set to some suitable large value (e.g. 500). The conditional likelihood of β_j is thus always normal, so the MCMC sampler merely has to change its prior variance. In practice this can make SSVS run much faster than having a spike at exactly zero. However, the performance depends on the width of the prior spike, so tuning for performance also means tuning the prior distribution, which will affect the posterior.

2.3 Adaptive Shrinkage

An alternative approach to creating a spike is to use a single density that happens to be spiked, i.e. one which has a large kurtosis. This can be done by making the prior for the a mixture distribution. For example the Bayesian LASSO assumes that the variance follow an exponential distribution:

$$\beta_j \sim N(0, \sigma_j^2) \quad (5)$$

$$\sigma_j^2 \sim \text{Exp}(\lambda) \quad (6)$$

which then makes the prior for β a double exponential, and perhaps not spiked enough. An alternative is to use a Jeffrey's prior for the variance:

$$\beta_j \sim N(0, \sigma_j^2) \tag{7}$$

$$\log(\sigma_j^2) \sim U(-\infty, \infty) \tag{8}$$

In practice ∞ can be replaced with a number that is very large. This creates a distribution that is more spike, and also does not have any tuning parameters.

3 Adapting the Approach

Although the models outlined above were developed for regression and generalized linear models, there is a clear potential to extend their use to a wider range of problems. Here I outline a couple of adaptations of these models, which should illustrate their potential.

3.1 Shrinking Factors

The methods used above can be adapted for factors, by using the same prior variance on each level. For fixed effects the variances can obviously be set to be either a small (possibly zero) or large value. A slab and spike prior can also be devised for the variance of a random effect, by placing the spike at zero. All of the approaches outlined above can be used, by placing the spike exactly at zero, using a mixture of two densities (e.g. gamma distributions), or using adaptive shrinkage (for example using the Jeffrey's prior in eq. 8).

3.2 Smoothing

Polynomials are a form of smoothed curve. One obvious problem is to find the order of the polynomial. This can be done by making the order a random variable, and using the variable selection methods above to estimate the parameters given the order (obviously this is not possible using adaptive shrinkage).

High order polynomials often look too wiggly to be effective smooths. But a Bayesian fit will marginalise over the curves, and smooth out the wiggles, improving the predicted curve.

4 Are these methods of any use?

My continued interest in Bayesian variable selection has mainly been motivated by their intrinsic interest, rather than because they are important. There will be times when they are useful, principally when the number of potentially important covariates is large compared to the amount of data so reducing the size of the model is important.

The automation of variable selection implies that we view the variables are being equivalent. Significance is statistical, not practical. Variable selection generally ignores practical significance. However, within a Bayesian paradigm we can use prior distributions to introduce practical significance by tuning them to align statistical and practical significance. In practice, this is probably impractical for many analyses, as if we are bothering with variable selection, we are not too interested in the variables being selected. The main exception is in analyses like QTL mapping, where we expect there to be several genes to have an effect, but we do not know which. Variable selection methods are still being developed, but there is one question that remains unanswered (and possibly un-answerable): can we use these methods to focus our attention on interpretable models? We can certainly use them to focus on simpler models, but this does not guarantee they can easily be interpreted.

5 Conclusions

Lots of fun, but handle with care.

Acknowledgments: Thanks to Mikko Sillanpää and Crispin Mutshinda-Mwanza for discussions. This study was funded by the research funding program Landes-Offensive zur Entwicklung Wissenschaftlich-konomischer Exzellenz (LOEWE) of Hesses Ministry of Higher Education, Research, and the Arts, Germany.

References

- OHara, R.B., and Sillanpää, M.J. (2009). A Review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis*, **4**, 85–118.

Part 2. Contributed Papers
(Volume I)

Joint modeling of two longitudinal outcomes and competing risk data

Eleni-Rosalina Andrinopoulou¹, Dimitris Rizopoulos¹, Johanna J. M. Takkenberg², Emmanuel Lesaffre^{1,3}

¹ Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

² Department of Cardiothoracic Surgery, Erasmus Medical Center, Rotterdam, The Netherlands

³ L-Biostat Catholic University, Leuven, Belgium

E-mail for correspondence: e.andrinopoulou@erasmusmc.nl

Abstract: When multiple longitudinal outcomes are collected together with survival outcomes it may be of interest to analyze them together. However, the majority of the work has focused on the joint models of a single longitudinal outcome with a single time-to-event. Joint modeling is a relatively new approach in the statistical literature of the analysis of longitudinal and survival data with promising results. We propose a joint model consisting of two longitudinal outcomes and time-to two events using a dataset including patients who received a human tissue valve in the aortic position. The Bayesian approach is followed for this analysis.

Keywords: Mixed-effects model; Competing risks; Joint Modeling; Bayesian approach.

1 Introduction

The joint modeling of longitudinal and time-to-event data is an active area of biostatistics and statistics research that has received a lot of attention in the recent years. This interest comes from the fact that a variety of problems can arise when (1) investigating the relationship between longitudinal and survival processes, (2) investigating the evolution of biomarkers with informative dropout, and (3) the interest lies in the hazard of an event. Numerous papers appeared in the literature that have proposed several extensions of the standard joint model introduced by Faucett and Thomas (1996) and Wulfsohn and Tsiatis (1997). These extensions include among others the consideration of a flexible specification of the subject-specific profiles (Brown and Ibrahim, 2003), nonparametric modeling of the random effects distribution (Song et al., 2002), the consideration of multiple longitudinal outcomes (Rizopoulos and Ghosh, 2011; Brown et al., 2005),

competing risks problems (Huang et al., 2011; Elashoff et al., 2008), and calculating dynamic predictions and accuracy measures (Rizopoulos, 2011; Proust-Lima and Taylor, 2009). Nice overviews of some early work in this field are given by Tsiatis and Davidian (2004) and Yu et al. (2004).

Even though joint models are becoming increasingly popular and a great deal of work has been done in this framework, there has been relatively little work published about the joint analysis of multivariate longitudinal and competing risk outcomes. In this work we consider a joint model including multivariate longitudinal outcomes and competing risk data. The proposed model extends existing methods to handle multiple longitudinal data with possible missing data not at random. Specifically, our model enables one to make inference for the longitudinal outcomes and the failure times adjusting for (1) the correlation between the multiple repeated markers, (2) the potential correlations between the events, and (3) the missing data caused by the failure times.

The motivation of this research comes from a study which includes all patients who received a human tissue valve in the aortic position in Erasmus University Medical Center (Department of Cardio-Thoracic Surgery). These patients are followed prospectively over time by annual telephone interviews and biennial standardized echocardiographic assessment of valve function in patients 16 years and older, Bekkers (2011). Particularly, echo examinations were scheduled at 6 months and 1 year postoperatively and biennially thereafter. From 1987 until 2008, 275 patients who survived aortic valve or root replacement with an allograft valve were followed until 08-Jul-2010. The total number of follow-up years is 3,292 and the completeness of follow-up is 98%. During follow-up 61 (20.3%) patients died and 78 patients required a reoperation (26.2%) on the allograft. In the 275 patients, a total of 1,228 echocardiographic measures of aortic gradient and aortic regurgitation were performed for each patient at different time points (median number of measurements 4, range 1-11; median echocardiographic follow-up 6.7 years, range 0-19.5 years). Aortic gradient (mmHg) was collected as a continuous variable while aortic regurgitation was collected using an ordinal scale (grade: 0 (none), 0.5 (trace), 1+, 2+, 3+, and 4+). High values of aortic gradient indicate aortic stenosis and high values of aortic regurgitation indicate aortic regurgitation.

It is clear that aortic gradient and aortic regurgitation are measurements of the valve function, thus it is expected that they are biologically inter-related. Furthermore, both death and reoperation could result in missing data not at random due to the fact that they are highly related to the disease condition of the patient.

2 Submodels and notation

Our joint model consists of three submodels: (1) a linear mixed-effects model for the longitudinal outcome, (2) a continuation ratio model for the

ordinal longitudinal outcome, and (3) a cause-specific hazard model for the competing risk failure time data. If we have n subjects under study, each with n_{1i} and n_{2i} observations for aortic gradient and aortic regurgitation, $i = 1, \dots, n$, let Y_{1i} and Y_{2i} denote the follow-up measurements for aortic gradient and aortic regurgitation respectively for patient i , where Y_{1i} takes values in $\{0, \dots, 144\}$ and Y_{2i} in $\{0 \text{ (none)}, 0.5 \text{ (trace)}, 1+, 2+, 3+, 4+\}$. Additionally, X_1 and X_2 are the design matrices for the fixed effects and Z_1 and Z_2 are the design matrices for the random effects.

$$\begin{aligned} Y_{1i} &= X_{1i}\beta_1 + Z_{1i}b_{1i} + \epsilon_i, \\ \pi_j &= P(Y_{2i} = j \mid Y_{2i} \leq j, X) \\ &= \frac{\exp[-(\theta_j + X_{2i}\beta_2 + Z_{2i}b_{2i})]}{1 + \exp[-(\theta_j + X_{2i}\beta_2 + Z_{2i}b_{2i})]}, \end{aligned}$$

where β_1 and β_2 are the fixed-effects for X_1 and X_2 respectively, $b_{1i} \sim N(0, \Sigma_{b1})$, $b_{2i} \sim N(0, \Sigma_{b2})$ are the random effects, $\epsilon_i \sim N(0, \sigma^2 I)$ is the measurement error and θ_j , with $j = 1, \dots, J$ denotes the intercept for each category of the aortic regurgitation.

For the survival part let T_i denote the observed failure time for patient i . $T_i = \min(T_{1i}^*, \dots, T_{Ki}^*, C_i)$ with $T_{\kappa i}^*$ denoting the failure time of each event and C_i the censored time. Moreover, let $\delta_i = 0, 1, \dots, K$ be the event indicator ($0 = \text{no event}$) and $\kappa = 1, \dots, K$ the type of failure. To model the risks of each of the competing events we postulate the proportional hazard models:

$$h_{i\kappa}(t, \theta_s) = h_{o\kappa}(t) \exp[\omega_i^T \gamma_\kappa + \alpha_{1\kappa}^T(\tilde{\beta}_1 + b_{1i}) + \alpha_{2\kappa}^T(\tilde{\beta}_2 + b_{2i})],$$

where θ_s is the parameter vector for the survival outcomes, ω_i is the design matrix, γ is a vector including the coefficients and $\alpha_{1\kappa}^T$, $\alpha_{2\kappa}^T$ are the coefficients that link the longitudinal and survival part. Particularly, they denote the strength of the relationship between aortic gradient and aortic regurgitation with death or reoperation. We denote $\tilde{\beta}_1$ and $\tilde{\beta}_2$ the coefficients for aortic gradient and aortic regurgitation from the fixed effects that correspond to the coefficients from the random effects. Moreover, b_{1i} and b_{2i} denote the random coefficients for aortic gradient and for aortic regurgitation. More details on the interpretation of α_1 and α_2 are given in Section 2.1. Furthermore, a piecewise constant baseline hazard function is assumed $h_{o\kappa}(t) = \sum_{q=1}^m h_{o\kappa q} I(t_{q-1} < t \leq t_q)$, where we consider $m-1$ intervals with t_{q-1} and t_q denoting the cutpoints and $I(\cdot)$ denoting the indicator function.

2.1 Interpretation of α

To clearly define the connection between the survival and longitudinal parts in our model, we explain the interpretation of the $\alpha_{1\kappa}$ with a simple example

where we assume a model for the aortic gradient with linear time effect and a random intercept and a random slope, i.e.,

$$Y_i = \beta_0 + \beta_1 t_i + b_{0i} + b_{1i} t_i.$$

The model for the two competing events takes the form

$$\begin{aligned} h_{i\kappa}(t, \theta_s) &= h_{o\kappa}(t) \exp[\omega^T \gamma_\kappa + \alpha_\kappa^T (\tilde{\beta} + b_i)] \\ &= h_{o\kappa}(t) \exp[\omega^T \gamma_\kappa + \alpha_{\kappa 1}(\beta_0 + b_{0i}) + \alpha_{\kappa 2}(\beta_1 + b_{1i})]. \end{aligned}$$

Hence, for one unit increase in the intercept of aortic gradient for patient i the hazard ratio of the κ -th event will be $\exp(\alpha_{\kappa 1})$, while for one unit increase in the slope of aortic gradient for patient i the hazard ratio of the κ -th event will be $\exp(\alpha_{\kappa 2})$.

3 Bayesian estimation

We adopt a Bayesian formulation for the proposed joint model, and derive posterior inferences using a Markov chain Monte Carlo (MCMC) algorithm. The posterior distribution is

$$\begin{aligned} P(\theta \mid Y_{1i}, Y_{2i}, T_i, \delta_i) &\propto P(Y_{1i} \mid b_{1i}, \theta_{Y_1}) P(Y_{2i} \mid b_{2i}, \theta_{Y_2}) P(T_i, \delta_i \mid b_{1i}, b_{2i}, \theta_s) \\ &\quad P(b_{1i} \mid \theta_{Y_1}) P(b_{2i} \mid \theta_{Y_2}) P(\theta_{Y_1}) P(\theta_{Y_2}) \Pr(\theta_s), \end{aligned}$$

where $\theta = (\theta_{Y_1}^T, \theta_{Y_2}^T, \theta_s^T)^T$ denotes the parameter vector for the longitudinal and survival outcomes. Respectively, θ_{Y_1} is the parameter vector for the aortic gradient, θ_{Y_2} for the aortic regurgitation and θ_s for the competing risk. We describe below the **likelihood** and **prior** functions.

The likelihood contribution for aortic gradient for the i -th subject terms is given by

$$\begin{aligned} P(Y_{1i} \mid b_{1i}; \theta_{Y_1}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \\ &\quad \times \exp\left[-\frac{1}{2\sigma^2}(Y_{1i} - X_{1i}\beta_1 - Z_{1i}b_{1i})^T(Y_{1i} - X_{1i}\beta_1 - Z_{1i}b_{1i})\right], \end{aligned}$$

The likelihood for the continuation ratio model (representing the aortic regurgitation) is the product of conditionally independent binomial terms, which is given by

$$P(Y_{2i} \mid b_{2i}; \theta_{Y_2}) = \prod_{j=2}^J \pi_j^{y_{2ij}} (1 - \pi_j)^{1 - \sum_{l=j}^J y_{2il}}$$

The likelihood contribution of the risk of death and reoperation can be written as

$$P(T_i, \delta_i \mid b_{1i}, b_{2i}, \theta_s) = \prod_{\kappa=1}^K [h_{o\kappa}(t) \exp(\omega^T \gamma_\kappa + \alpha_{1\kappa}^T b_{1i} + \alpha_{2\kappa}^T b_{2i})]^{I(\delta_i=\kappa)} \\ \exp \left[- \sum_{\kappa=1}^K \exp(\omega^T \gamma_\kappa + \alpha_{1\kappa}^T b_{1i} + \alpha_{2\kappa}^T b_{2i}) \sum_{q=1}^m h_{o\kappa q} T_{iq} \right],$$

where $T_{iq} = \min(T_i, t_q) - \min(T_i, t_{q-1})$.

We use standard prior distributions for the parameters. Particularly, for the regression coefficients β_1 , β_2 , b_1 , b_2 and the association coefficients α_1 , α_2 we take normal prior distributions. For the variance–covariance matrices for the random effects we take inverse wishart priors, while for the inverse variance for the continuous longitudinal outcome we take gamma prior.

We derive the full conditional distributions of each model parameter in order to implement the MCMC sampling.

The analysis is in preparation.

References

- Bekkers, J.A., Klieverik, L.M., Raap, G.B., Takkenberg, J.J.M., and Bogers, A.J.J.C. (2011). Re-operations for Aortic Allograft Root Failure: Experience from a 21-year Single-Center Prospective Follow-up Study. *Eur J Cardiothorac Surg*, **40**, 35–42.
- Brown, E.R. and Ibrahim, J.G. (2011). Bayesian Approaches to Joint Cure-Rate and Longitudinal Models with Applications to Cancer Vaccine Trials. *Biometrics*, **59**, 686–693.
- Brown, E.R., Ibrahim, J.G., and DeGruttola, V. (2005). A Flexible B-spline Model for Multiple Longitudinal Biomarkers and Survival. *Biometrics*, **61**, 64–73.
- Elashoff, R.M., Li, G., and Li, N. (2008). A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types. *Biometrics*, **64**, 762–771.
- Faucett, C.L. and Thomas, D.C. (1996). Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: a Gibbs Sampling Approach. *Statistics in Medicine*, **15**, 1663–1685.
- Huang, X., Li, G., Elashoff R.M., and Pan, J. (2011). A General Joint Model for Longitudinal Measurements and Competing Risks Survival Data with Heterogeneous Random Effects. *Lifetime Data Anal.*, **17**, 80–100.

- Proust-Lima, C. and Taylor, J.M. (2009). Development and Validation of a Dynamic Prognostic Tool for Prostate Cancer Recurrence Using Repeated Measures of Posttreatment PSA: a Joint Modeling Approach. *Biostatistics*, **10**, 535–549.
- Rizopoulos, D. (2011). Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics*, **67**, 819–829.
- Rizopoulos, D. and Ghosh, P. (2011). A Bayesian Semiparametric Multivariate Joint Model for Multiple Longitudinal Outcomes and a Time-to-Event. *Statistics in Medicine*, **30**, 1366–1380.
- Song, X., Davidian, M., and Tsiatis, A.A. (2002). A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-to-Event Data. *Biometrics*, **58**, 742–753.
- Tsiatis, A.A. and Davidian, M. (2004). Joint Modeling of Longitudinal and Time-to-Event Data: An Overview. *Statistica Sinica*, **14**, 809–834.
- Wulfsohn, M.S. and Tsiatis, A.A. (1997). A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, **53**, 330–339.
- Yu, M., Law, N. , Taylor, J., and Sandler, H. (2004). Joint Longitudinal-Survival-Cure Models and Their Application to Prostate Cancer. *Statistica Sinica*, **14**, 835–862.

Latent class inverse probability weighting to estimate causal effects of sequential treatments under unobserved confounding

Francesco Bartolucci¹, Leonardo Grilli², Luca Pieroni¹

¹ Department of Economics, Finance, and Statistics - University of Perugia, Italy

² Department of Statistics - University of Firenze, Italy

E-mail for correspondence: grilli@ds.unifi.it

Abstract: In many fields, it is of interest to analyze longitudinal studies designed to assess the causal effect of a sequential treatment on an outcome measured at the end of the period. We consider the common case of a binary treatment assigned repeatedly over time, where the treatment assignment at a given occasion depends on the sequence of previous assignments, as well as on time-varying confounders. A popular modeling strategy is represented by Marginal Structural Models; within this approach, the causal effect of the treatment is estimated by the Inverse Probability Weighted (IPW) estimator. This estimator is consistent provided that all the confounders are observed (sequential ignorability). To alleviate this serious limitation, we propose an extension of the IPW estimator to account for unobserved pre-treatment confounders. The proposed approach is based on the assumption that the unobserved confounders are summarized by a discrete latent variable, thus we estimate the probabilities of treatment using a latent class model. The new estimator, called Latent Class Inverse Probability Weighted (LC-IPW), is based on two steps. Its properties are assessed by a simulation study: the LC-IPW estimator outperforms the IPW estimator for all combinations of sample size and number of occasions considered in the simulation study, even when there is no unobserved confounding. The proposed approach is applied to the estimation of causal effects of wage subsidies on employment, using a dataset of Finnish firms observed for eight years.

Keywords: Causal inference; Longitudinal design; Mixture model; Potential outcomes; Sequential treatment.

1 Introduction

Consider a random sample of n subjects, or more generally units, with the chance of receiving treatment at T occasions (time points or intervals). To simplify the notation, we will usually omit the index i for the subject. We adopt the following notation: Y is the outcome of interest (measured after the last occasion), S_t is the binary indicator of treatment at occasion t , with $t = 1, \dots, T$, \mathbf{V} is a column vector of pre-treatment covariates (measured before the first occasion), and \mathbf{X}_t is a column vector of time-varying

covariates (possibly including previous measurements of the outcome variable). We use the subscript $1 : t$ to denote a column vector obtained by stacking vectors of variables measured from occasion 1 until occasion t , so that $\mathbf{S}_{1:t} = (S_1, \dots, S_t)'$ and $\mathbf{X}_{1:t} = (\mathbf{X}'_1, \dots, \mathbf{X}'_t)'$. Lowercase letters denote realizations of these variables.

The covariates \mathbf{V} and \mathbf{X}_t are confounders, namely they simultaneously affect S_t and Y . Following the potential outcome approach, the outcome Y has a potential version for each sequence of treatments, denoted with $Y^{(\mathbf{s}_{1:T})}$. The vector $\mathbf{Y}^{(all)}$ contains all the potential outcomes (2^T in the case of T binary treatments).

The *history* of a variable is the set of variables determined before it, thus potentially affecting it. We solve the simultaneity issue by postulating that S_t is determined before \mathbf{X}_t . Therefore, the history of S_t includes \mathbf{V} , $\mathbf{X}_{1:t-1}$, and $\mathbf{S}_{1:t-1}$, whereas the history of \mathbf{X}_t includes \mathbf{V} , $\mathbf{X}_{1:t-1}$, and $\mathbf{S}_{1:t}$. When $t = 1$ both $\mathbf{X}_{1:t-1}$ and $\mathbf{S}_{1:t-1}$ vanish.

In the spirit of Robins et al. (2000), we postulate a Marginal Structural Model (MSM) for the potential outcomes:

$$E(Y^{(\mathbf{s}_{1:T})}) = \beta_0 + \mathbf{g}(\mathbf{s}_{1:T})' \boldsymbol{\beta}_1, \quad (1)$$

where $\mathbf{g}(\mathbf{s}_{1:T})$ is a function summarizing the treatment sequence, for example the scalar $\mathbf{g}(\mathbf{s}_{1:T}) = s_+ = \sum_{t=1}^T s_t$ or the vector $\mathbf{g}(\mathbf{s}_{1:T}) = (s_+, I(s_T = 1))'$. Under the identification assumptions discussed below, the parameters $\boldsymbol{\beta}_1$ have a causal interpretation, allowing us to make inference on average treatment effects on the whole population.

Two standard assumptions for the identification of causal effects are: (1) *Stable Unit Treatment Value Assumption* (SUTVA), which implies that the potential outcomes of a given subject only depend on the treatment sequence of that subject, thus excluding any interference; (2) *positivity* or *random assignment*, which states that the conditional probability of being assigned to treatment is neither zero nor one. In this paper we focus on the third standard assumption, called *Sequential Ignorability Assumption* (SIA), which rules out unobserved confounders:

$$S_t \perp \mathbf{Y}^{(all)} \mid \mathbf{S}_{1:t-1}, \mathbf{X}_{1:t-1}, \mathbf{V} \quad t = 1, \dots, T. \quad (2)$$

The SIA states that, conditionally on the (observed) history up to occasion $t-1$, the treatment assignment at occasion t is independent of the potential outcomes; therefore, for each sub-population defined by the history up the previous occasion, the treatment is assigned as if it were randomized.

Under SIA, the causal parameters of the MSM based on assumption (1) can be consistently estimated by the IPW method, where each subject is weighted by the inverse of the probability of its observed treatment sequence. Higher efficiency is achieved by using the so called *stabilized weights*:

$$sw_i = \frac{\prod_{t=1}^T Pr(S_{it} = s_{it} \mid \mathbf{S}_{i,1:t-1})}{\prod_{t=1}^T Pr(S_{it} = s_{it} \mid \mathbf{S}_{i,1:t-1}, \mathbf{X}_{i,1:t-1}, \mathbf{V}_i)}. \quad (3)$$

The probabilities at the denominator are usually estimated through a pooled logistic regression, namely a standard logistic regression applied to the subject-occasion dataset (any subject contributes with one record for each occasion where it is observed). The probabilities at the numerator are estimated by the same approach used for those at the denominator, except that the set of regressors is restricted to the treatment indicators.

2 The proposed approach: latent class inverse probability weighting

We aim at extending the IPW method in order to get a consistent estimator of causal effects in the presence of a pre-treatment unobserved confounder. We assume that the confounder may be represented by a discrete latent variable U with values $c = 1, \dots, k$ corresponding to latent classes. Therefore, the population is divided into a finite number of latent classes having different parameters for the distribution of the observed variables. It is worth noting that the number of latent classes, k , and the corresponding probabilities, $\pi_c = Pr(U = c)$, are parameters to be estimated, thus the approach is flexible enough to satisfactorily approximate also continuous unobserved confounders.

We relax the ignorability assumption (SIA) defined in (2) by requiring that the independence holds within the latent classes induced by the unobserved confounder U . We call it the *Latent Class Sequential Ignorability Assumption* (LC-SIA):

$$S_t \perp \mathbf{Y}^{(all)} \mid \mathbf{S}_{1:t-1}, \mathbf{X}_{1:t-1}, \mathbf{V}, U \quad t = 1, \dots, T.$$

Clearly, LC-SIA is weaker than SIA because the independence statement is conditional on U , thus in general it does not hold marginally (namely, if U is not considered). Therefore, under LC-SIA the standard IPW estimator may be biased, but it is possible to devise a suitable modification to correct it. A route consists in computing the weights using probabilities conditional on U . Since the latent class c_i of subject i is unknown, it has to be predicted on the basis of the available data. For this end, we fit an auxiliary latent class model for the treatment indicators and the observed covariates. Therefore, we propose a two-step estimation procedure: (i) fit an auxiliary latent class model to assign subjects to latent classes; (ii) fit the MSM (1) using weights computed with the latent-class-specific probabilities.

2.1 First step: auxiliary latent class model

In order to assign subjects to latent classes, we fit a latent class model for the treatment indicators and the observed covariates. For this end, the joint distribution of the observed variables is written as a finite mixture

over the latent classes $c = 1, \dots, k$, and each component of the mixture is recursively factorized:

$$f(\mathbf{V}|c) \prod_{t=1}^T f(s_t | \mathbf{S}_{1:t-1}, \mathbf{X}_{1:t-1}, \mathbf{V}, c) f(\mathbf{X}_t | \mathbf{S}_{1:t}, \mathbf{X}_{1:t-1}, \mathbf{V}, c). \quad (4)$$

For every probability or density function in (4), we have to choose a suitable model. For the component $f(s_t | \mathbf{S}_{1:t-1}, \mathbf{X}_{1:t-1}, \mathbf{V}, c)$ we adopt a Bernoulli distribution parametrized as in a logistic regression model, namely $p_{t|c}^{s_t} (1 - p_{t|c})^{(1-s_t)}$ with

$$p_{t|c} = \text{expit}(\eta_{t|c}^{(1)} + \mathbf{S}'_{1:t-1} \boldsymbol{\eta}_{t|c}^{(S)} + \mathbf{X}'_{1:t-1} \boldsymbol{\eta}_{t|c}^{(X)} + \mathbf{V}' \boldsymbol{\eta}_{t|c}^{(V)}), \quad (5)$$

where $\text{expit}(x) = 1/(1 + \exp(-x))$. Note there are specific parameters for every combination of occasion t and latent class c .

The other components of the distribution in (4), namely $f(\mathbf{V}|c)$ and $f(\mathbf{X}_t | \mathbf{S}_{1:t}, \mathbf{X}_{1:t-1}, \mathbf{V}, c)$, should be modeled according to the nature of the variables: for example, for continuous variables the simplest choice is a multivariate normal distribution parametrized as in a linear regression model.

The parameters of the auxiliary latent class model are estimated with maximum likelihood using an EM algorithm. The number of support points k is chosen by the Normalized Entropy Criterion (NEC) of Celeux and Soromenho (1996). Once the parameters have been estimated, every subject is assigned to the latent class with the highest estimated posterior probability.

2.2 Second step: weighted estimation of the causal model

The second step of the proposed LC-IPW method entails fitting the MSM (1) with a modified IPW procedure where the weight for each sample unit is computed conditionally on the assigned latent class. Specifically, the stabilized weights (3) become

$$sw_{i, \hat{c}_i} = \frac{\prod_{t=1}^T Pr(S_{it} = s_{it} | \mathbf{S}_{i,1:t-1})}{\prod_{t=1}^T Pr(S_{it} = s_{it} | \mathbf{S}_{i,1:t-1}, \mathbf{X}_{i,1:t-1}, \mathbf{V}_i, U_i = \hat{c}_i)}.$$

The probabilities at the denominator are estimated using the logistic model (5) after assigning the latent class and replacing parameters with maximum likelihood estimates. Such estimates are available from the model fitted at the first step. The standard errors of the parameters of the MSM (1) are estimated via non-parametric bootstrap.

3 Simulation study

We studied the performance of the proposed LC-IPW estimator through a simulation study based on a model with a continuous outcome Y , a sequen-

tial binary treatment S_t , a continuous time-varying covariate X_t , a continuous pre-treatment covariate V , and a discrete unobserved pre-treatment covariate U . The unobserved covariate U is a potential confounder since it can influence both the outcome Y and the treatment indicators S_t . However, the effect on the treatment indicators depends on a parameter ϕ such that for $\phi = 0$ the effect is null and thus U is not a confounder.

Table 1 reports some of the results. The estimators under comparison are the standard regression estimator, the IPW regression estimator, the proposed LC-IPW with a number of latent classes k chosen by NEC, and an unfeasible version of the IPW estimator based on the true weights, which are known in the simulation. The results confirm that, in case of unobserved confounding, the LC-IPW estimator actually corrects most of the bias of the IPW estimator. Interestingly, the proposed estimator is slightly better than the standard estimator even when the unobserved covariate is not a confounder ($\phi = 0$). This finding may appear as surprising, but it can be explained by results on over-adjustment in inverse probability weighting (Rotnitzky et al. 2010). Moreover, it is in line with the simulation results of Lefebvre et al. (2008), who show that adding pure predictors of treatment is deleterious, whereas adding pure predictors of outcome may be beneficial.

4 Application

The proposed approach has been applied to the estimation of causal effects of wage subsidies on the number of employees. The dataset, extracted from the registers compiled by the Finnish Tax Authority from 1995 to 2002, comprises 1640 firms from 20 to 200 employees that applied at least once for wage subsidies. The dataset includes several variables measured at each of the 8 observation years: wage subsidy, number of employees, fixed capital, wage rate, sales, net profit. The wage subsidy is a sequential binary treatment variable, whereas the number of employees at the end of the period is the outcome. Previous values of the outcome and all the other variables are considered as time-varying confounders.

The analysis was based on marginal structural models with different specifications fitted by the standard IPW method and the proposed LC-IPW method. The main finding is as follows: both IPW and LC-IPW methods yield positive estimates of the causal parameters and indicate a significant causal effect for these subsidies; however, using the proposed method, the magnitude of this effect results much smaller. In particular, the LC-IPW estimate is around one half of the IPW estimate and this result is stable using a number of latent classes, k , from 2 to 5. For these data, the number of classes selected by NEC is equal to 4.

TABLE 1. Simulation results (number of time occasions $T = 8$, sample size $n = 2000$, unobserved covariate U with three latent classes, 1000 Monte Carlo replicates, true model $E(Y^{(s_{1:T})}) = \beta_0 + \beta_1 s_+$ with $\beta_0 = -2$ and $\beta_1 = 1$).

	$\phi = -0.5$		$\phi = 0$		$\phi = 0.5$	
	β_0	β_1	β_0	β_1	β_0	β_1
<i>Standard regression</i>						
Mean	-2.1280	1.0655	-3.9310	1.9452	-3.9933	1.8496
Bias	-0.1280	0.0655	-1.9310	0.9452	-1.9933	0.8496
St.dev.	0.0912	0.0386	0.0794	0.0312	0.0621	0.0191
RMSE	0.1571	0.0760	1.9326	0.9457	1.9943	0.8499
<i>IPW regression</i>						
Mean	-1.1463	0.5800	-2.0614	1.0367	-3.0942	1.4985
Bias	0.8537	-0.4200	-0.0614	0.0367	-1.0942	0.4985
St.dev.	0.1296	0.0571	0.3054	0.1497	0.5027	0.2053
RMSE	0.8634	0.4239	0.3114	0.1541	1.2041	0.5391
<i>LC-IPW regression</i>						
Mean	-2.0194	1.0114	-2.0694	1.0391	-2.5496	1.2703
Bias	-0.0194	0.0114	-0.0694	0.0391	-0.5496	0.2703
St.dev.	0.1370	0.0636	0.2949	0.1462	0.7001	0.2669
RMSE	0.1383	0.0646	0.3028	0.1513	0.8898	0.3798
<i>IPW regression (true weights)</i>						
Mean	-2.0004	1.0029	-2.0601	1.0356	-2.5166	1.2492
Bias	-0.0004	0.0029	-0.0601	0.0356	-0.5166	0.2492
St.dev.	0.1757	0.0796	0.3195	0.1553	0.7147	0.2787
RMSE	0.1756	0.0796	0.3249	0.1592	0.8816	0.3738

References

- Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195–212.
- Lechner, M. and Miquel, R. (2010). Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics*, **39**, 111–137.
- Lefebvre, G., Delaney, J.A.C., and Platt, R.W. (2008). Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine*, **27**, 3629–3642.
- Robins, J.M., Hernán, M.A., and Brumback, B. (2000). Marginal structural models and causal inference. *Epidemiology*, **11**, 550–560.
- Rotnitzky, A., Li, L., and Li, X. (2010). A note on overadjustment in inverse probability weighted estimation. *Biometrika*, **97**, 997–1001.

Additive decomposition of vital rates from grouped data

Carlo Giovanni Camarda¹, Paul H. C. Eilers², Jutta Gampe¹

¹ Max Planck Institute for Demographic Research, Rostock, Germany.
camarda@demogr.mpg.de, gampe@demogr.mpg.de

² Department of Biostatistics, Erasmus Medical Centre, Rotterdam,
The Netherlands. p.eilers@erasmusmc.nl

Abstract: Age specific rates are often given only for age groups that can be rather coarse. Furthermore, they tend to show complex patterns, which often can be explained by contributions of several subgroups, such as infants or the elderly. A penalized composite link model is used to decompose the complex rate trajectory into smooth additive components from grouped data. Monotonicity or shape constraints can be incorporated by special penalty matrices.

Keywords: Additive components; Composite Link Model; Grouped data; Penalized likelihood; Shape constraints; Vital rates.

1 Introduction

Demographic events, such as marriage, migration or death, have characteristic age specific patterns of occurrence. Finding so called model schedules to summarize the age pattern has a long tradition, however, parametric models are predominantly used (Siler, 1983; Rogers and Little, 1994). One feature of many demographic rates is that their overall shape is rather complex, but the pattern can be attributed to different distinct components. As an example, migration rates are high at very young ages (infants migrating with their parents), in the young adult ages (labour motivated migration), and again after retirement. Models that can separate these components are particularly welcome.

While some of the components can be described well by a parametric model, such as adult mortality by the Gompertz hazard, many others cannot. An additional complication arises if data are provided only in age groups, which is still the case in many official statistics, and is standard if one goes back in time. Such problems are not limited to demographic applications, but they can also be found in epidemiology, such as age specific disease incidence.

In the following we propose a general model that allows to specify a (demographic) rate across a wide range of ages as the sum of several components, which are modelled on the log scale and are assumed to be smooth, but do not have to follow a particular parametric form. If several nonparametric

components are additively combined, shape constraints are necessary to identify the components correctly. The data can be given in grouped form, and the age groups can be of variable lengths. The model will be demonstrated by two applications, the age specific migration rates from Germany to Spain, and age specific tuberculosis mortality in the Netherlands.

2 Additive components from grouped data

Camarda et al. (2010) proposed a model for estimating a complex rate trajectory by decomposing it into several components, which are modelled on the log scale and are then additively combined. This approach will be extended to deal with grouped data and to meet shape constraints for some of the components. The essentials of the model are as follows.

For a sequence of m different ages the vector $\mathbf{e} \in \mathbb{R}^m$ denotes the exposures at these ages, while $\mathbf{y} \in \mathbb{R}^m$ is the corresponding number of observed events. The elements of \mathbf{y} are realizations of a Poisson distribution, $\mathbf{y} \sim \mathcal{P}(\boldsymbol{\mu})$. The expected values $\boldsymbol{\mu}$ are the product of exposures \mathbf{e} and the actual vital rates at the respective ages, which we denote by $\boldsymbol{\theta}$.

The rates are assumed to be the sum of K components $\boldsymbol{\gamma}^k$, $k = 1, \dots, K$, each of length m , so that $\boldsymbol{\theta} = \boldsymbol{\Gamma} \cdot \mathbf{1}$, and where $\boldsymbol{\Gamma}$ is a matrix with K columns and m rows; column k of $\boldsymbol{\Gamma}$ is $\boldsymbol{\gamma}^k$.

If $\boldsymbol{\gamma} = \text{vec}(\boldsymbol{\Gamma})$ is the vector of length mK , then we can write $\boldsymbol{\mu} = \mathbf{C}_0 \boldsymbol{\gamma}$ with

$$\mathbf{C}_0 = \mathbf{1}_{1,K} \otimes \text{diag}(\mathbf{e}). \quad (1)$$

The matrix $\mathbf{C}_0 \in \mathbb{R}^{m \times mK}$ additively combines the $\boldsymbol{\gamma}^k$ and simultaneously matches the exposures. This is a composite link model (Thompson and Baker, 1981). For each component we assume that the discrete sequence $\boldsymbol{\gamma}^k$ can be written as $\boldsymbol{\gamma}^k = \exp(\mathbf{X}^k \boldsymbol{\beta}^k)$, $k = 1, \dots, K$. The design matrices $\mathbf{X}^k \in \mathbb{R}^{m \times p_k}$ can represent parametric or non-parametric structures, as needed. In this way, the composed mean $\boldsymbol{\mu}$ can be viewed as sum of K exponential components, which generally are smooth.

If the observed counts \mathbf{y} are given only for age groups but the exposures are available by single years of age, which is not uncommon, then the composition matrix can be adapted to represent also the additional grouping. For n age classes, which can be of different widths, the new composition matrix $\mathbf{C} \in \mathbb{R}^{n \times mK}$ has the following form:

$$\mathbf{C} = \mathbf{E} \cdot \mathbf{C}_0. \quad (2)$$

The elements η_{ji} of $\mathbf{E} \in \mathbb{R}^{n \times m}$ are equal to 1, if age i is contained in age group j , and zero otherwise. Again $\boldsymbol{\mu} = \mathbf{C} \boldsymbol{\gamma}$ and a composite link model results.

If the exposures are also grouped, then the methodology presented in Lambert and Eilers (2009) is used in a first step to ungroup the exposure numbers. The single age exposures obtained by this approach are then used in equation (2).

3 A Penalized Composite Link Model with shape constraints

In composite link models, which are overparametrized like in our case, the K coefficient vectors β^k can be estimated by a penalized iteratively re-weighted least squares (IRWLS) algorithm (Eilers, 2007):

$$(\check{\mathbf{X}}' \tilde{\mathbf{W}} \check{\mathbf{X}} + \mathbf{P}) \tilde{\beta} = \check{\mathbf{X}}' \tilde{\mathbf{W}} \tilde{\mathbf{z}}, \quad (3)$$

where $\tilde{\mathbf{W}} = \text{diag}(\tilde{\mu})$ and $\tilde{\mathbf{z}} = \tilde{\mathbf{W}}^{-1}(\mathbf{y} - \tilde{\mu}) + \check{\mathbf{X}} \tilde{\beta}$. The vector β concatenates the coefficients β^k of the K components and \mathbf{P} is the penalty matrix.

The overall design matrix $\check{\mathbf{X}}$ is put together from the component specific matrices $\check{\mathbf{X}}^k$

$$\check{\mathbf{X}} = \left(\check{\mathbf{X}}^1 : \dots : \check{\mathbf{X}}^k : \dots : \check{\mathbf{X}}^K \right). \quad (4)$$

The $\check{\mathbf{X}}^k$ are of dimension $n \times p_k$, where p_k is the number of parameters used to model the k -th component. Their elements are

$$\check{x}_{ij}^k = \sum_{l=1}^m c_{il} x_{ij}^k \gamma_l^k / \mu_i,$$

where c_{il} are elements of the composition matrix \mathbf{C} in equation (2).

The matrix \mathbf{P} combines the penalty matrices \mathbf{P}^k for the K components, i.e.,

$$\mathbf{P} = \text{diag}(\mathbf{P}^1, \dots, \mathbf{P}^k, \dots, \mathbf{P}^K).$$

The specific form of the \mathbf{P}^k depends on the assumptions we make on the different components γ^k (via the β^k). If smoothness is the only constraint, then a simple difference penalty on neighbouring coefficients in β^k suffices:

$$\mathbf{P}^k = \lambda^k \mathbf{D}_d^{k'} \mathbf{D}_d^k.$$

Here \mathbf{D}_d^k is a matrix that computes d -th order differences of the coefficients β^k and λ^k is the smoothing parameter controlling the roughness of the vector γ^k .

If several nonparametric components γ^k are to be estimated, then additional constraints are necessary to make the decomposition feasible. In practice, such assumptions often come naturally. For example, infant mortality is supposed to quickly drop from high values to zero, while the ‘accident hump’ at young adult ages, as the name suggests, is supposed to have a unimodal, log-concave shape. Old-age mortality is a strictly increasing function of age.

We follow the proposal of Bollaerts et al. (2006) to incorporate shape constraints. If a component is to be monotone, we add a term $\mathbf{P}_m^k = \kappa \mathbf{D}_1^{k'} \mathbf{W}^k \mathbf{D}_1^k$ to \mathbf{P}^k , where \mathbf{W}^k is a diagonal matrix with $w_j^k = 1$ if

$\beta_j^k \leq \beta_{j-1}^k$ and $w_j^k = 0$ otherwise. Alternatively, a log-concave shape is forced by adding a penalty term $\mathbf{P}_c^k = \kappa \mathbf{D}_2^{k'} \mathbf{W}^k \mathbf{D}_2^k$, where the diagonal matrix \mathbf{W}^k operates on the second differences of the coefficients β^k . These penalties both exert influence only if the shape constraint is violated and the weights in \mathbf{W}^k are computed iteratively. The size of κ regulates how strictly the constraint is enforced and, in the following examples, κ will be equal to 10^3 .

Finally, if a parametric model is chosen for component k , then \mathbf{P}^k is a matrix of zeros.

4 Applications

We now illustrate the performance of this model by two examples.

4.1 Migration schedule

In the first example we look at the age specific migration rates from Germany to Spain for 2004–2008, see Figure 1. The data were taken from the Eurostat web-site. The migration flows are given in five year age intervals, the population counts are available by single year of age.

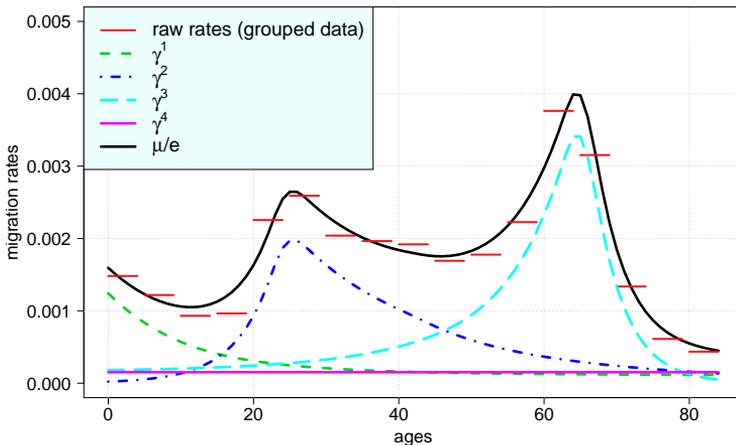


FIGURE 1. Migration rates from Germany to Spain for the years 2004–2008 and ages under 85. The step function shows the raw rates for 5 year age groups. Four (smooth) components are additively combined to give the overall trajectory.

The overall rate was decomposed into four (non-)parametric components. The first one was assumed to be monotone decreasing and relates to migration of children, which is mirrored in the relatively high rates of their

parents. Two other components were penalized to be log-concave and pertain to job and retirement related migration. A fourth component describes an age independent migration propensity and can be modelled either as γ^k strongly penalized with $d = 1$, or with an explicit modification of the \mathbf{C} matrix. Second order differences were taken to penalize γ^1 , and $d = 3$ for both γ^2 and γ^3 . To ensure sufficient flexibility, knots were placed every other year, and smoothing parameters λ^k were subjectively chosen.

4.2 Tuberculosis mortality pattern

The second example shows mortality from tuberculosis in the Netherlands in 1943, see Figure 2. The death counts are aggregated in age classes of varying widths, while the population counts are available by single years of age from Statistics Netherlands (CBS).

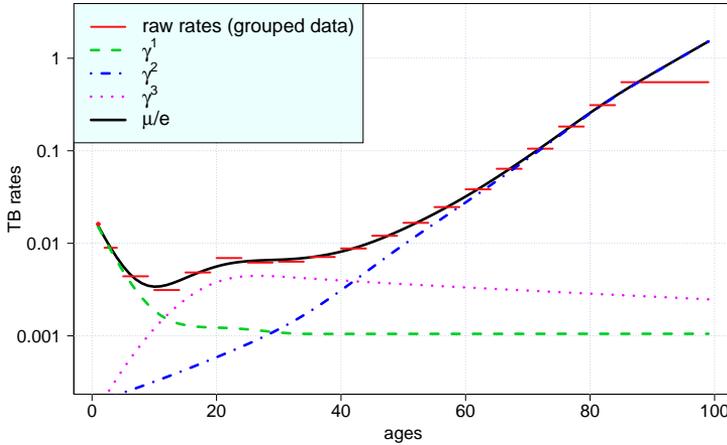


FIGURE 2. Mortality from tuberculosis in the Netherlands in 1943 for ages $[1, 100)$, both sexes combined. The step function shows the raw rates for age groups. Three smooth components are additively combined to give the overall trajectory.

Three components were modelled with penalized B -splines. Child mortality and ageing related mortality were forced to be monotonically decreasing and increasing, respectively. A third component for the accident hump is penalized to be log-concave. Here knots were placed at every fifth years of age. While a third order difference was taken for γ^3 , $d = 2$ was selected for infant and ageing component. Smoothing parameters λ^k were again subjectively chosen.

5 Conclusions

The proposed method allows us to model vital rates from grouped data as sum of several components. Instead of searching for parametric forms, we additively combine several nonparametric components.

Additional shape constraints can be accommodated within the estimation procedure to ensure convergence, but also to incorporate prior knowledge about features of the components. Moreover, we can cope with vital rates given in age groups of different length.

The trajectories of migration rates and mortality were successfully modelled over a wide range of ages, with good fit and interpretable components. In human mortality, the first year of life poses particular challenges. During the first months infant mortality plummets from comparatively high levels, and if the first year is contained in the first age group, this sharp drop is covered up. As a remedy, one can envision adapting the C matrix with a column that is specifically added for this age.

If one wants to study rate trajectories also over time, a generalization to two dimensions is needed. In case the length of the age groups varies over time, too, the composition matrix needs to be adapted accordingly. We plan to extend our approach along this line.

References

- Bollaerts, K., Eilers, P. H. C. and van Mechelen, I. (1996). Simple and multiple P -splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, **59**, 451–469.
- Camarda, C. G., Eilers, P. H. C. and Gampe, J. (2010). Sums of smooth exponentials. In: *Proceedings of the 25th International Workshop on Statistical Modelling*, Glasgow, UK, pp. 111–118.
- Eilers, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, **7**, 239–254.
- Lambert, P. and Eilers, P. H. C. (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis*, **53**, 1388–1399.
- Rogers, A. and Little, J.S.(1994). Parametrizing age patterns of demographic rates with the multiexponential model schedule. *Mathematical Population Studies*, **4(3)**, 175–195.
- Siler W. (1983). Parameters of mortality in human populations with widely varying life spans. *Statistics in Medicine*, **2**, 373–380.
- Thompson, R. and Baker, R. J. (1981). Composite Link Functions in Generalized Linear Models. *Applied Statistics*, **30**, 125–131.

Inference for a dynamic model of HIV and HCV

Amparo Yovanna Castro-Sánchez¹, Marc Aerts¹, Ziv Shkedy¹,
Peter Vickerman^{2,3}, Niel Hens^{1,4}

¹ Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Hasselt, Belgium

² London School of Medicine of Hygiene & Tropical Medicine, University of London, London, United Kingdom

³ School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom

⁴ Centre for Health Economics Research and Modeling Infectious Diseases, Centre for the Evaluation of Vaccination, Vaccine and Infectious Disease Institute University of Antwerp, Antwerp, Belgium

E-mail for correspondence: amparo.castrosanchez@uhasselt.be

Abstract: In infectious diseases, one of the main interests is to model the spread of a certain disease and to assess the impact of intervention policies. Therefore, mathematical transmission models provide a useful tool. However, calibrating such a model to real data is far from trivial. In this paper we focus on a joint transmission model for HCV and HIV among injecting drug users. The model accounts for co-infection with the two viruses and some biological complexities of the transmission process. Statistical concepts and methods are used to calibrate the model to real data and to perform sensitivity analyses.

Keywords: mathematical models; HCV/HIV; Injecting Drug Users; statistical models; sensitivity analysis

1 Introduction

In infectious diseases, mathematical models are used to characterize the transmission of certain diseases and to assess the impact of intervention policies. Nevertheless, fitting such a model to data is far from trivial, partially due to uncertainty about the parameters in the model. Furthermore, the number of parameters complicates the identification of a unique solution.

Hepatitis C and HIV have a considerable impact in terms of morbidity and mortality worldwide. One third of HIV infected individuals in Europe and the US are co-infected with HCV. Additionally, HIV accelerates the development of liver disease progress related with HCV (Rockstroh and

Spengler 2004). The majority of co-infected people are IDUs, so focusing on this population could give us insights about the co-infection.

In this sense, several proposals have been made to estimate the force of infection, specifically for HCV/HIV co-infection (among them: Del Fava et al. 2011; Sutton et al. 2008). The inputs are cross-sectional surveys containing information about the serostatus of each individual and the self-reported duration of injection (exposure time). These models are very flexible and allow to account for subject heterogeneity. However, they only provide a description of the data without focusing on the transmission process of the viruses (Garnett et al. 2011).

In the context of Injecting Drug Users (IDUs), we propose a deterministic compartmental model that accounts for Human Immunodeficiency Virus (HIV) and Hepatitis C virus co-infection attributed to sharing syringes and other paraphernalia. To select a parameter set that best fits the data, we adapt the multinomial likelihood function. Moreover, we propose several sensitivity analyses to determine important parameters. In order to quantify the variability we use a nonparametric bootstrap of the data.

Our data example corresponds to the baseline measure of a cohort study of young adult heroin users in Spain (Itinere project, collected between 2001 and 2003. De La Fuente et al. 2006). The main goals were to monitor the health impact of drug use and to identify related factors. The presence of antibodies for HIV and HCV was tested using dried blood samples.

2 The transmission model

The joint mathematical model considers the following stages for Hepatitis C: susceptible (S_{HCV}), acute infected (I_{HCV}), chronic carrier (CC_{HCV}), and susceptibles who were infected before but spontaneously cleared the virus (S_{HCV}^+). Conversely, for HIV we consider two stages: infected with HIV (I_{HIV}) and AIDS (A_{HIV}). The transmission process can be seen in detail in Figure 1; moreover, it is expressed using a set of differential equations.

Sharing injecting equipment is a known risk factor for the transmission of HCV and HIV among IDUs. Additionally, their behaviour is very heterogeneous, with some who hardly ever share injecting equipment and others who share more frequently. To account for that heterogeneity we consider several risk groups; assuming that an IDU will belong to a specific risk group during his/her entire injecting carrier.

For the i -th risk group, the force of infection for HCV can be defined in terms of the rate of contacts (κ_i), the transmission probability at each infection stage (b_{HCV1}, b_{HCV2}), the proportion of syringes shared with members of other risk groups (m_{ij} : mixing proportions among the risk groups), and the proportion of infected individuals in each risk group for every disease stage ($\text{Prev}_{HCV1_j}, \text{Prev}_{HCV2_j}$):

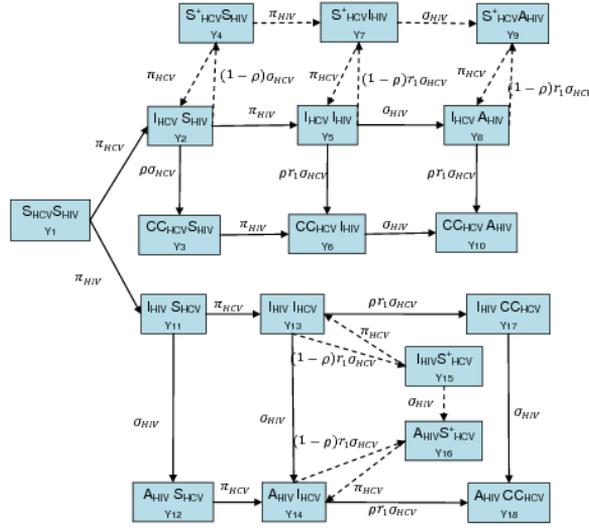


FIGURE 1. Flow diagram of the joint mathematical model for HCV/HIV. S: susceptible, I: Infected (Acute Infected in case of HCV and HIV positive), CC: Chronic Carrier of HCV, A: AIDS and S_{HCV}^+ : Susceptible for HCV after spontaneously clearing the virus.

$$\pi_{HCV_i} = \kappa_i \sum_{j=1}^R m_{ij} (b_{HCV1} \text{Prev}_{HCV1_j} + b_{HCV2} \text{Prev}_{HCV2_j}). \quad (1)$$

Analogously, we define the force of infection for HIV.

Estimation and inference process

For every parameter in the model we define a plausible range of values, taking into account the literature and auxiliary information coming from the data example. Using Latin hypercube sampling we select 500.000 parameter sets (Stein 1987).

The best parameter sets should reflect as much as possible the trends of the observed data (serostatus for HCV and HIV as well as the self-reported duration of injection). An individual at a certain exposure time (duration of injection) can be either positive for both viruses, negative for both, or positive for only one of them. Thus he/she can be classified in one of the four categories, and the classification of the subjects is assumed to be independent. Thus, we can consider a multinomial distribution with four different outcomes and four probabilities, one for each possible outcome, and the number of individuals at each exposure time constitutes the number

of independent trials. For infection acquired through injecting drug use only, let p_{00d} , p_{01d} , p_{10d} and p_{11d} denote the probabilities of IDUs with an injecting career length of d that are: uninfected for both, infected by HIV but not HCV, infected by HCV but not HIV, and both HIV and HCV, respectively. Given a specific parameter set we use integration methods to solve the set of ordinary differential equations that represent the dynamic model. Then, we compare the fitted joint probabilities for both viruses with the observed prevalences. The multinomial likelihood considering the four possible outcomes at each exposure time is given by:

$$L(p) = \prod_{d=1}^D (p_{00d})^{y_{00d}} (p_{01d})^{y_{01d}} (p_{10d})^{y_{10d}} (p_{11d})^{y_{11d}}, \quad (2)$$

where $y_{ij d}$ are the observed number of individuals at each of the possible outcomes with an exposure time of d .

For further analyses, we focus on the model that assumes two risk groups. To identify the more influential parameters we rely on several statistical methods. Furthermore, we assess variability using a nonparametric bootstrap. In total 500 bootstrap samples were selected and the estimation process was performed for each bootstrap sample.

3 Results

As a result of the estimation and inference process, we select the parameter sets that best represent the data using the Akaike Information Criteria (AIC) and the multinomial likelihood (the AIC was calculated taking into account the number of parameters in the model and the log likelihood values). The likelihood values and the AIC for one, two and three risk groups are shown in Table 1. The best model fit assumes two risk groups in the population (low and high). Figure 2 compares the observed joint probabilities with the model based probabilities.

TABLE 1. Description “best” model fits

# Risk groups	Binomial $p_{i.d}$ HCV prevalence	Binomial $p_{.jd}$ HIV prevalence	Multinomial $p_{ij d}$ joint prevalences	AIC using multinomial $p_{ij d}$
One (R=1)	-395.118	-381.017	-689.296	1404.59
Two (R=2)	-354.570	-320.239	-675.980	1387.96
Three (R=3)	-348.780	-337.548	-692.014	1434.03

The sharing rates in the low risk group (κ_1) and the transmission probability of HIV at infection stage (b_{HIV1}) seem to be the most influential

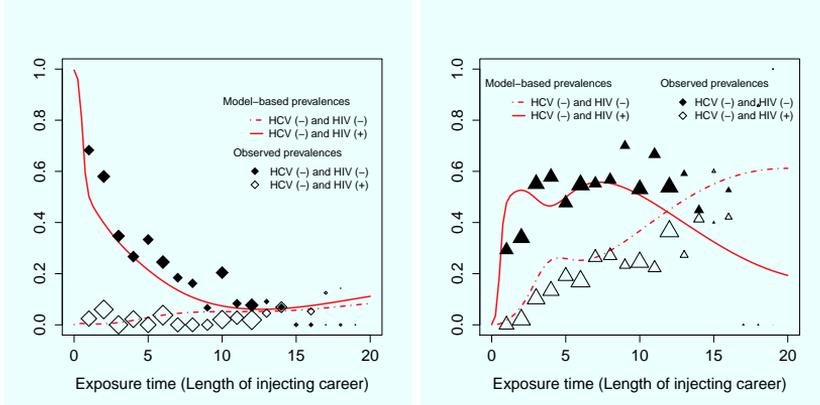


FIGURE 2. Comparison observed proportions (diamonds and triangles) vs model based proportions (lines) with two risk groups.

parameters. With respect to the variability analyses, for every bootstrap sample we determine how many times each of the parameter sets leads to the highest likelihood (Figure 3). In total 37 parameter sets lead to the best likelihood value in at least one bootstrap sample. However, only 6 of them were more frequently selected as the best.

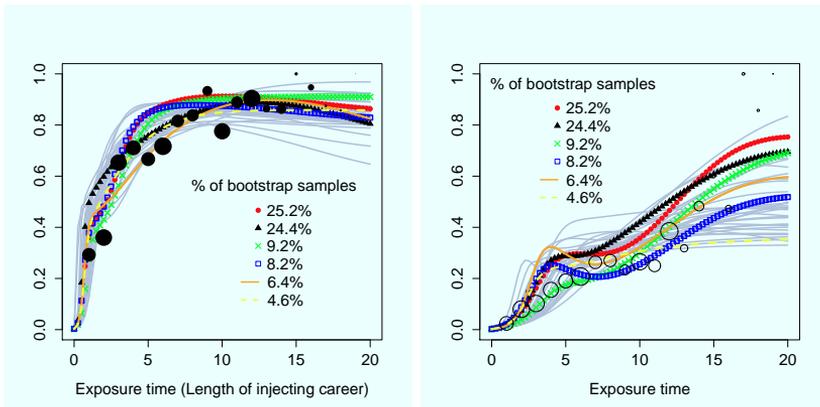


FIGURE 3. HCV Prevalence (left). HIV Prevalence (right). Observed prevalences (circles) and model based prevalences. The gray lines indicate parameters sets that lead to the highest likelihood value for at least one bootstrap sample. The other colors indicate the six parameter sets that produce the highest likelihood values more frequently.

4 Discussion

The model proposed here takes into account some of the biological complexities in the transmission of HCV and HIV in injecting drug users. We rely on an application of a likelihood function to assess the goodness of fit of the model to the data. According to the uncertainty analyses, the transmission probabilities and the contact rates seem to be the most important parameters. Some confirmatory analyses should be performed in order to verify these findings. Furthermore, it will be very interesting to determine the regions where the likelihood values are higher.

Some studies have pointed to temporal differences in HIV risk related behaviors among IDUs. In contrast, our model assumes that the force of infection is invariant with respect to calendar time. A possible extension would be to take calendar time into account in the model.

References

- De La Fuente, L., Bravo, M.J., Toro, C., Brugal, M.T., Barrio, G., Soriano, V., Vallejo, F., and Ballesta, R. (2006). Injecting and HIV prevalence among young heroin users in three Spanish cities and their association with the delayed implementation of harm reduction programmes. *Journal of epidemiology and community health*, **60**, 537–542.
- Del Fava, E., Shkedy, Z., Hens, N., Aerts, M., Suligoi, B., Camoni, L., Vallejo, F., Wiessing, L., and Kretzschmar, M. (2011). Joint modeling of HCV and HIV co-infection among injecting drug users in Italy and Spain using individual cross-sectional data. *Statistical Communications in Infectious Diseases*, **3**.
- Garnett, G.P., Cousens, S., Hallett, T.B., Steketee, R., and Walker, N. (2011) Mathematical models in the evaluation of health programmes. *The Lancet*, **378**, 515–525.
- Rockstroh, J. K. and Spengler, U. (2004) HIV and hepatitis C virus co-infection. *Lancet Infectious Diseases*, **4**, 537–544.
- Stein, M., (1987). Large sample properties of simulations using latin hypercube sampling. *Technometrics*, **29**, 143–151.
- Sutton, A.J., Hope, V.D., Mathei, C., Mravcik, V., Sebakova, H., Vallejo, F., Suligoi, B., Brugal, M.T., Ncube, F., Wiessing, L., and others (2008). A comparison between the force of infection estimates for blood-borne viruses in injecting drug user populations across the European Union: a modelling study. *Journal of Viral Hepatitis*, **15**, 809–816.

Estimatives of lymph nodes metastasis rates and treatment effectiveness under two latent activation schemes

Juliana Cobre ¹, Francisco Louzada ¹, Mário de Castro ¹,
Gleici Perdoná ², Fernanda M. Peria ²

¹ ICMC - Universidade de São Paulo, Brazil

² FMRP/USP - Universidade de São Paulo, Brazil

E-mail for correspondence: jucobre@icmc.usp.br

Abstract: In oncology studies, it is important to know the spread of the disease and the effectiveness of the applied treatment. The proposed model accommodates this scenario, assuming the lymph nodes to be latent causes which compete for the contamination of another lymph node, carrying metastasis. We propose that the number of latent competing causes follows a generalized negative binomial distribution and each lifetime related to each competing cause has a Weibull distribution. We also considered two activation schemes: the first activation scheme and the last activation scheme. The usefulness of our model is illustrated with computational tests carried out in real data on breast cancer.

Keywords: generalized negative binomial distribution; latent activation schemes; metastasis; cure fraction; Bayesian analysis.

1 Introduction

Metastasis is the spread of a disease from an organ to another caused by the lymph nodes. It happens when a contaminated lymph node infects a clean one, making other tumor to appear. To avoid such proliferation, there exist a number of treatments. The number of contaminated lymph nodes in patients with cancer is one of the factors observed for the prognostic of the disease (Kirkwood *et al.*, 2000). The proposed model accommodates this scenario. We assume the lymph nodes to be latent causes which compete for the contamination of another lymph node, in a process which might proliferate the disease. Our main interest is the estimation of the effectiveness of the applied treatment and the lymph nodes contamination rate. In this paper, we extend the cure rate survival models proposed by Yakovlev & Tsodikov (1996), using a generalized negative binomial (GNB) distribution (Hanin, 2001), the long-term survival function formulated by Rodrigues *et al.* (2009) and also two activation schemes: the first activation scheme and the last activation scheme. The importance of using a

GNB distribution comes from the fact that the proposed estimations can not be done if the number of causes competing follows Poisson or negative binomial distributions, as it is usual in this kind of studies. The proposed model also incorporates various important practical informations into the analysis, such as the initial number of competing causes and some characteristics of the treatment under consideration, such as the number of doses, the time interval between doses and the efficiency of each dose.

2 Model formulation

Considering that lymph nodes are competing causes of the metastasis process, we aim to estimate the contaminated lymph node proliferation rate, λ , and the treatment failure probability, i.e., the probability of the treatment not destroying the cancer cells of the contaminated lymph nodes, η . Let M be a random variable denoting the number of latent competing causes related to the contamination occurrence of another lymph node. Given M equals m , let $Z_l, l = 1, \dots, m$ be the lifetime due to the l th competing cause. Z_l are independent and equally distributed random variables with distribution function given by $F(z|\gamma) = 1 - S(z|\gamma)$, which does not depend on M and where $S(z|\gamma)$ denotes the survival function.

The lifetime for each subject at first-activation (FA) scheme and last-activation (LA) scheme are defined, respectively, as $Y = \min\{Z_1, \dots, Z_M\}$ and $Y = \max\{Z_1, \dots, Z_M\}$, with $P(Y = \infty|M = 0) = 1$, that is, there is a proportion of subjects not susceptible to the occurrence of the event, given by p_0 . FA scheme means that the contamination of a lymph node occurs when it receives the contamination from one contaminated lymph node. And LA scheme means that a lymph node will be contaminated if all the competing contaminated lymph node pass the disease.

We consider that M follows a generalized negative binomial (GNB) distribution, with a probability distribution, $p_m = P(M = m)$, $m = 0, 1, 2, \dots$, defined as in Hanin (2001) and expressed by

$$p_m = \left(\frac{a}{c}\right)^i \left(\frac{b}{a}\right)^m Q_m\left(\frac{ad-bc}{bc}\right), \quad m \geq 0, \quad (1)$$

where $Q_0(x) = 1$, $Q_m(x) = \sum_{r=1}^m \binom{m-1}{m-r} \binom{i+r-1}{r} x^r$, $m \geq 1$, and i denotes the initial number of contaminated lymph nodes. The treatment is given in k doses, the time interval between doses is τ , whereas a, b, c and d are related to the parameters of interest λ and η as proposed in Hanin (2001) and can be written as $a = \omega(\mu^k - 1)$, $b = \omega\mu^k - 1$, $c = \mu^k - \omega$ and $d = \mu^k - 1$, where, for $\alpha \neq 1$, $\alpha = e^{-\lambda\tau}$, $\omega = \frac{\lambda - \eta\lambda}{\lambda(1-\alpha)}$, $\mu = \frac{\eta}{\alpha} = \eta e^{\lambda\tau}$, and for $\alpha = 1$, by continuity, $\omega = (1 - \eta + \lambda\tau)/(\lambda\tau)$. The corresponding probability generating function is given by $A(s) = \left(\frac{a-bs}{c-ds}\right)^i$, $0 \leq s \leq 1$.

Consider $S_{\text{pop}}(y) = P(Y > y)$ as the improper survival function of the random variable Y and $S(\cdot|\gamma)$ as a proposed survival function associated with $Z_l, l = 1, \dots, M$. We have that

$$S_{\text{pop}}(y) = P(M = 0) + P(z_1 > y, Z_2 > y, \dots, Z_M > y, M \geq 1).$$

Following Rodrigues *et al.* (2009) we have that $S_{\text{pop}}(y) = A(S(y))$. Then, the survival function in this competing causes scenario is given by

$$S_{\text{pop}}(y) = \left[\frac{a - bS(y|\gamma)}{c - dS(y|\gamma)} \right]^i.$$

The cure fraction is given by

$$p_0 = P(M = 0) = \lim_{y \rightarrow \infty} S_{\text{pop}}(y) = \lim_{y \rightarrow \infty} \left[\frac{a - bS(y|\gamma)}{c - dS(y|\gamma)} \right]^i = \left(\frac{a}{c} \right)^i,$$

since $S(y|\gamma)$ is a proper survival function. For the GNB distribution the cure fraction may be different for each subject for it depends on i and k .

Given $M = m$, we assume that the lifetimes, $Z_l, l = 1, \dots, m$, are independent and identically distributed following a Weibull distribution with parameters $\gamma = (\gamma_1, \gamma_2)$, where γ_1 denotes the shape parameter and $\exp(\gamma_2)$ denotes the scale parameter.

Thus, we have a model for which the number of latent competing causes follows a GNB distribution and each lifetime related to each competing cause has a Weibull distribution, hereafter the Weibull generalized negative binomial model, or simply the WGNB model. As mentioned earlier, this model allows the estimation of the contaminated lymph node proliferation rate and the probability of the treatment not destroying the cancer cells of the lymph node. It also accommodates decreasing, increasing and monotone failure rate.

3 Inference

Let T_j and C_j denote, respectively, the observable lifetime and the cure time for the j th subject, such that $T_j = \min\{Y_j, C_j\}$ and $\delta_j = 1$ if $Y_j \leq C_j$ and $\delta_j = 0$ otherwise. Moreover, i_j is the initial number of tumors for the j th subject. Thus, considering that n subjects were observed, the data set is composed by vectors $\mathbf{t} = (t_1, \dots, t_n)^T$, $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)^T$ and $\mathbf{i} = (i_1, \dots, i_n)^T$. We assume that $M_j, j = 1, \dots, n$, are independent generalized negative binomial variables with probability function given by (1), with $\lambda > 0$ and $0 \leq \eta \leq 1$, and, given M_j equals m_j , the promotion times, Z_{j1}, \dots, Z_{jm_j} , are independent with Weibull distribution. We consider that the covariates are directly related to γ_2 . Thus, for each subject we have $\gamma_{2j} = \boldsymbol{\beta}^T \mathbf{z}_j, j = 1, \dots, n$, where \mathbf{z}_j is the vector of covariates for

each subject and β is the vector of unknown parameters without the intercept term. The corresponding likelihood function under non-informative censoring is given by

$$L(\lambda, \eta, \beta | \mathbf{t}, \delta, \mathbf{i}) \propto \prod_{j=1}^n \left[i_j f(t_j | \gamma_j) \frac{ad - bc}{[c - dS(t_j | \gamma_j)]^2} \right]^{\delta_j} \left[\frac{a - bS(t_j | \gamma_j)}{c - dS(t_j | \gamma_j)} \right]^{i_j - \delta_j}, \quad (2)$$

where, a, b, c and d are given above, $\gamma_j = (\gamma_1, \beta^T \mathbf{z}_j)^T$, $j = 1, \dots, n$, $S(\cdot) = 1 - F(\cdot)$, $f(\cdot)$ and $F(\cdot)$ are the probability density and cumulative distribution functions of Weibull distribution, respectively.

We consider proper prior distribution $\pi(\lambda, \eta, \beta)$ to assure that the joint posteriori is proper (Ibrahim *et al.*, 2001). Although it is not necessary, we assume that the parameters are *a priori* independent for the sake of simplicity. The parameters have prior distribution according to their parametric space.

Deviance information criterion (DIC) and Log pseudo marginal likelihood (LPML) were considered to select the best fit model.

4 Application

The data set were available from Gozzo (2008) and present the survival times (in months) for 40 women diagnosed with locally advanced invasive ductal carcinoma. They were treated at the Ribeirão Preto School of Medicine Clinic Hospital and were submitted to neo-adjuvant (pre-operative) and adjuvant (post-operative) chemotherapy between 2003 and 2006. The chemotherapy scheme consisted of two drugs, docetaxel and epirubicin, with cycle ranging from $k = 4$ to $k = 6$, monthly, $\tau = 1$. We define the disease-free time as being the interval between surgery and relapse, termed as disease-free survival.

We choose independent and vague prior distributions: $\lambda \sim \Gamma(1, 0.001)$, $\eta \sim \mathcal{B}(1, 1)$, $\gamma_1 \sim \Gamma(1, 0.001)$ and β 's $\sim \mathcal{N}(0, 1000)$. The computational code was implemented in OpenBUGS software version 3.2.1. Two parallel chains of the Metropolis-Hastings sampler with size 55,000 for each parameter were generated. The first 15,000 were discarded as the burn-in period and the remaining was thinned 20 to 20, resulting in a sample of size 4,000.

The covariates docetaxel, epirubicin, age, body surface (in mm^2), number of white cells (in $\times 10^4/mm^2$) were included into the analysis. We fitted to the data both models. The obtained DIC statistics values for the two fitted models, WGNB-FA and WGNB-LA, were 109.0 and 111.2, respectively, providing evidence in favor of WGNB-FA model. This evidence is supported by LPML scores that were -60.039 and -61.381, respectively.

The second analysis included only the significative covariates for the best model. Table 1 presents the posterior mean, standard deviation and 95% credible interval for the parameters estimated by the WGNB-FA model.

TABLE 1. Posterior mean, standard deviation (SD) and 95% credible intervals (95% CI).

Parameter	Posterior mean	SD	95% CI
λ	1.033	0.052	(1.001,1.153)
η	0.578	0.182	(0.247,0.933)
γ_1	3.689	0.943	(2.062,5.721)
β_{txt}	-0.376	0.127	(-0.664,-0.165)
β_{epi}	0.326	0.150	(0.084,0.667)

The cure probabilities estimated for $k = 4$, and 1, 3, 4, 5, 7, 9 and 10, contaminated lymph nodes are equal to 60,7%, 28,1%, 24,4%, 14,9%, 8,5%, 5,3% and 4,0%, respectively.

The results presented in Table 1 provide some interesting information to a practitioner. In particular, λ denotes the contaminated lymph node proliferation index, which is equal to approximately 1 lymph node per month. The parameter η denotes the probability of a contaminated lymph node remaining contaminated after each treatment dose, which is approximately equal to 58%. Therefore, for instance, assuming independent doses, the surviving probability of a contaminated lymph node after 4 or 6 doses is almost equal to 3,0% and 0,5%, respectively.

5 Conclusions

The proposed WGNB model accommodates survival data in the presence of latent competing causes assuming that their number follows a GNB distribution and their lifetimes follow a Weibull distribution. An advantage of the proposed model is its ability to estimate the treatment effectiveness and the lymph nodes proliferation rates. In the Bayesian analysis we chose non-informative prior distributions. Nevertheless, knowledge of a practitioner should be incorporated into the analysis through informative priors.

Acknowledgments: The research was partially supported by the Brazilian Organizations CAPES, CNPq and FAEPA.

References

- Gozzo, T.O. (2008). *Toxicidade ao tratamento quimioterápico em mulheres com câncer de mama*. Doctorate Thesis - Escola de Enfermagem de Ribeirão Preto, USP, So Paulo, Brazil.
- Hanin, L.G. (2001). Iterated birth and death process as a model of radiation cell survival. *Mathematical Biosciences*, **169**, 89–107.

- Ibrahim, J.G, Chen, M.H, and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer.
- Kirkwood, J.M., Ibrahim, J.G., Sondok, V.K., Richards, J., Flaherty, L.E., Ernstoff, M.S., Smith, T.J., Rao, U., Steele, M. and Blum, R.H. (2000). High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of intergroup trial E1690/S91111/C9190. *Journal of Clinical Oncology*, **18**, 2444–2458.
- Rodrigues, J., Cancho, V.G.m de Castro, M. and Louzada-Neto, F. (2009). On the unification of the long-term survival models. *Statistics & Probability Letters*, **79**, 753–759.
- Yakolev, A.V and Tsodikov, A.D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Singapore, World Scientific.

Heterogeneity identification of repairable systems

Enrico A. Colosimo¹, Gustavo L. Gilardoni², Maristela D. Oliveira³

¹ Federal University of Minas Gerais, Brazil

² University of Brasília, Brazil,

³ Federal University of Bahia, Brazil

E-mail for correspondence: `enricoc@est.ufmg.br`

Abstract: A repairable system, under minimal repair, is usually modeled according to a NonHomogeneous Poisson Process (NHPP) assuming a power Law intensity function. A traditional approach considers iid NHPPs in order to conduct a statistical analysis based on a sample of systems. However, systems might be heterogeneous due to unmeasured variables such as age, suppliers and so on. In order to verify this assumption a classical and a Bayesian approaches are proposed in this paper. Some possible model scenarios considering different systems heterogeneity are compared using likelihood ratio tests and hierarchical Bayesian model. A real data set illustrates the proposed methodology.

Keywords: BIC, Hierarchical Bayesian, NonHomogeneous Poisson Process; Power Law.

1 Introduction

Statistical models for recurrent events are of great interest in reliability and maintenance of repairable systems, i.e., systems that are allowed to experience more than one failure throughout its lifetime (Ascher and Feingold (1984), Bain and Engelhardt (1991), Rigdon and Basu (2000)). Typically, after each failure, an action is necessary to put the system back into operation. Since the earlier work of Barlow and Hunter (1960), a distinction has been made between two different types of interventions: (i) minimal repair (MR) or *as bad as old*, which typically consists in the replacement only of the damaged part of the equipment and returns it to the same situation it was immediately before the failure and (ii) Perfect Maintenance (PM) or *as good as new*, which is typically identified by a major overhaul of the entire system. Some of the developments stemming from this situation can be found in Gerstack (1997), Bolck et al. (1990), Park et al. (2000) and Gilardoni and Colosimo (2007).

Let $N(t)$ be the number of failures from the beginning of the follow-up until time t . When the system is subject only to MR actions, $N(t)$ is usually modeled according to a Non-Homogeneous Poisson Process (NHPP). More precisely, this means that the process $N(t)$ has independent increments, satisfies that $N(0) = 0$ and $N(t)$ follows a Poisson distribution with mean $\Lambda(t) = \int_0^t \lambda(u) du$. In this situation, $\Lambda(t)$ and $\lambda(t) = \lim_{\Delta t \rightarrow 0} P[N(t + \Delta t) - N(t) = 1] / \Delta t$ are called respectively the *mean* and *intensity* functions of the process. Following Crow (1974), it is usual in the repairable systems literature to assume the parametric form $\lambda(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta}\right)^{\beta-1}$ for the intensity function, or equivalently $\Lambda(t) = (t/\theta)^\beta$, where both β and θ are positive, in which case it is said that $N(t)$ follows a *Power Law Process* (PLP).

1.1 Real Data Motivation

This paper was motivated by a data set concerning the failure histories of 11 electrical power transformers operated by CEMIG, the Electrical Power Company of the State of Minas Gerais, Brazil. A traditional approach would consider independent PLPs with the same parameters β and θ . However, systems might be heterogeneous due to unmeasured variables related for instance to operating conditions, in which case the assumption that all 11 transformers share the same β and θ might be inadequate. From a classical viewpoint, one may entertain several scenarios ranging from the most complex that all eleven transformers have different β 's and θ 's to the simplest one that they share the same set of parameters. Although the former scenario is more general, it has many (22) parameters and hence their estimation may lack precision. Thus, it is crucial to use model selection tools to choose among the several possible scenarios.

1.2 Paper goal and outline

The goal of this work is to identify the best or most adequate model to fit PLP models to data from several systems under MR. The rest of the paper is organized as follows. In Section 2, six possible scenarios are considered taking into account different systems heterogeneities. Classical and Bayesian approaches are used in Section 3 in order to treat this problem. Finally, Section 4 illustrates the methodology using the set of eleven transformers mentioned in Section 1.1.

2 Scenarios of Interest

Some possible model scenarios considering different systems heterogeneities are presented in this section. These scenarios are described next considering a sample of K systems.

1. All systems are different. They are essentially K separate systems. Therefore, it is necessary to estimate $2K$ parameters.
2. λ functions have different β 's but they have the same θ . In this case, it is necessary to estimate $K + 1$ parameters.
3. λ functions have different θ but they have the same β . Again, it is necessary to estimate $K + 1$ parameters.
4. All PLPs are identical. It means, that there is only one θ and β to be estimated.
5. Each $N_i(\cdot)$ represents a different PPH ($\beta = 1$). Therefore, it is necessary to estimate K θ 's scale parameters.
6. All systems are identical, following a PPH. It remains just one scale parameter to be estimated in the modeling structure.

3 Classical and Bayesian Approaches

In this section, likelihood functions are established for each model associated with the corresponding scenario described in Section 2. In reality, only the likelihood function for scenario 1 is derived next. The others ones are easily obtained from it by algebraic manipulation since they are special cases of scenario 1. Analytic expressions of the Maximum Likelihood Estimators (MLE) for each one of the scenarios are also presented in this section. It is considered for each scenario that there are K processes with time truncations, respectively, at T_1, T_2, \dots, T_K .

It was observed n_i failure times for the i -th system, indexed by failures at times $t_{ij}, i = 1, \dots, K, j = 1, \dots, n_i$. The likelihood function is given by: (Rigdon and Basu, 2000):

$$L(\lambda) = \prod_{i=1}^K \prod_{j=1}^{n_i} \lambda(t_{ij}) \times \exp \left[- \sum_{i=1}^K \int_0^{T_i} \lambda(x) dx \right]. \quad (1)$$

Likelihood function for the model associated with Scenario 1, is one that allows for maximum heterogeneity among systems. It can be obtained by plugging the PLP intensity function in (1).

3.1 Likelihood Ratio Tests

MLEs are obtained, by solving the system of equations obtained by the partial derivatives of the logarithm of the expression (1) with respect to each of the parameters. That is, for the scenario 1, is given by

$$\hat{\beta}_i = \frac{n_i}{\sum_{j=1}^{n_i} \log \frac{T_i}{t_{ij}}} \quad \text{and} \quad \hat{\theta}_i = \frac{T_i}{n_i^{1/\hat{\beta}_i}}, \quad i = 1, \dots, K.$$

Scenarios restrictions as defined in Section 2 can be used in (1) in a way to obtain the likelihood functions for models associated to the remaining scenarios.

Likelihood ratio test (LRT) is used in order to compare the scenarios of interest. This test can be used because the scenarios are nested within each other. It can be observed that Scenarios 1 and 2 require at least one failure per system for the existence of MLEs.

3.2 Hierarchical Modeling for PLPs

Inference for hierarchical modeling is fundamentally Bayesian, in terms that population unknown quantities have a probabilistic specification as hyperparameters. The aim is to find a general model, sufficiently flexible to comport several scenarios, but very simple to data analysis and result interpretation. Bayesian hierarchical modeling might work with more parameters than data.

A hierarchical model structure with three levels describes the Scenario 1. In the second level, $\eta_i = (\theta_i, \beta_i)$; $i = 1, \dots, K$, is treated as a sample from a common population distribution, indexed by ϕ in the third level.

In the hierarchical modeling, not only $\eta = (\eta_1, \dots, \eta_K)$ is unknown, but so is ϕ , with probability distribution $\pi(\phi)$. The goal is get a joint probability distribution for the vector (η, ϕ) . Thus, prior distribution for the unknown quantities is given by:

$$\pi(\eta, \phi) = \pi(\phi)\pi(\eta|\phi),$$

and the joint posterior distribution is:

$$\begin{aligned} p(\eta, \phi|N_1(t), \dots, N_K(t)) &\propto \pi(\eta, \phi)L(N_1(t), \dots, N_K(t)|\eta, \phi) \\ &= \pi(\eta, \phi)L(N_1(t), \dots, N_K(t)|\eta), \end{aligned}$$

where $L(N_1(t), N_2(t), \dots, N_K(t)|\eta)$ is the likelihood function (1). In order to satisfy the exchangeability principle, consider $\xi_i = (1/\theta_i)^{\beta_i}$ as an alternative parametrization. ξ_i has an interpretation as the expected number of events in one unit of time. Let's consider gamma prior distributions for ξ_i and β_i .

Let $\phi = (a_\beta, a_\xi, b_\beta, b_\xi)$ be the hyperparameters vector of the third level. As established above, for $i = 1, \dots, K$,

$$\pi(\xi_i|\phi) = \frac{b_\xi^{a_\xi} \xi_i^{a_\xi-1} e^{-b_\xi \xi_i}}{\Gamma(a_\xi)} \qquad \pi(\beta_i|\phi) = \frac{b_\beta^{a_\beta} \beta_i^{a_\beta-1} e^{-b_\beta \beta_i}}{\Gamma(a_\beta)}, \quad (2)$$

where $a_\beta, a_\xi, b_\beta, b_\xi > 0$.

Using the jacobian for transformation of variables, it follows that

$$\pi(\theta_i|\beta_i, \phi) = \frac{\beta_i b_\xi^{a_\xi} \exp(-b_\xi/\theta_i^{\beta_i})}{\Gamma(a_\xi)\theta_i^{1+\beta_i a_\xi}}, \quad (3)$$

and, therefore, $\pi(\beta, \theta|\phi) = \pi(\theta|\beta, \phi)\pi(\beta|\phi)$.

In the third level, consider also a Gamma hyperprior distribution for each of the four components of ϕ . That is,

$$a_\beta \sim \gamma(a_\beta|a_{a_\beta}, b_{a_\beta}); \quad a_\xi \sim \gamma(a_\xi|a_{a_\xi}, b_{a_\xi}); \quad (4)$$

$$b_\beta \sim \gamma(b_\beta|a_{b_\beta}, b_{b_\beta}); \quad b_\xi \sim \gamma(b_\xi|a_{b_\xi}, b_{b_\xi}),$$

where $\gamma(x | , a , b) = b^a x^{a-1} e^{-bx} / \Gamma(a)$ ($x, a, b > 0$) is the density of the Gamma distribution with shape and scale parameters equal to a and b , respectively.

It follows then that the prior distributions are specified by (2) and (3) for the second level and (4) for the third level. This developments are the ones for the most complex scenario. In order words, in order to obtain prior specifications for the other scenarios is just a matter of exclude some pieces of prior specifications.

Under prior and hyperprior specifications (2), (3) and (4), respectively, conditional posterior densities for Model 1 are obtained.

Adaptive Metropolis Rejection Sampling - ARMS (Gelman et al., 2003) algorithm is used in order to get numeric results. This algorithm is implemented in Ox software. 300,000 samples from the posterior density of each parameter and hyperparameter were generated by this algorithm. After considering a burn-in of 50,000 and lag of 10, the final sample is 25,000.

4 Numerical Example

Data set of this work represents the failure histories of eleven 500 kvolt transformers belonging to a state electrical power company in Brazil, between 1999 and 2009. Electric power transformers are complex and most of their repairs involve the replacement of only a small fraction of their parts. Therefore, its reasonable to suppose that the system's reliability after a repair is essentially the same as before the failure. This fact characterizes MR and, hence, justifies the statistical analysis by using a NHPP for these data.

LRT in Table 1 indicates the most simple is the best fitted model. Model 6 consists of only one parameter.

In hierarchical Bayesian approach non-informative prior distributions were considered in this situation. DIC criterion agrees with AIC and BIC criteria for this case. They are in Table 2.

TABLE 1. LRT Transformer data.

Scenario	Model 2	Model 3	Model 4	Model 5	Model 6
Model 1	0.416 (12)	0.102 (12)	0.278 (24)	0.135 (13)	0.322 (25)
Model 2	–	Not applicable	0.228 (12)	Not applicable	0.284 (13)
Model 3		–	0.698 (12)	0.620 (1)	0.757 (13)
Model 4			–	Not applicable	0.647 (1)
Model 5				–	0.701 (12)

Degrees of freedom for chi square distribution in parenthesis.

TABLE 2. Decision criteria for Transformer data.

Scenario	AIC	BIC	DIC
Model 1	411.013	419.767	200.940
Model 2	409.414	414.189	200.434
Model 3	406.937	411.712	198.623
Model 4	394.226	395.021	197.113
Model 5	405.182	409.559	197.998
Model 6	392.436	392.833	196.170

References

- Ascher, H. and Feingold, H.W. (1984). *Repairable Systems Reliability: Modeling, Inference, Misconception and their Causes*. Marcel Dekker: New York.
- Bain, L.J. and Engelhardt, M. (1991). *Statistical Analysis of Reliability and Life-Testing Models, Theory and Methods*. Marcel Dekker: New York.
- Barlow, R.E. and Hunter, L.C. (1960). Optimum preventive maintenance policies. *Operations Research*, **8**, 90–100.
- Block, H.W., Borges, W.S., and Savits, T. H. (1990). A General age replacement with minimal repair. *Naval Research Logistics*, **35**, 365–372.
- Crow, L.R. (1974). Reliability analysis for complex systems. *Reliability and Biometry*, F. Proschan and J. Serfling (Eds), 379–410.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*. Chapman and Hall: New York.
- Gerstack, I.B. (1977). *Models of Preventive Maintenance*. North Holland: Amsterdam.
- Gilardoni, G.L. and Colosimo, E.A. (2007). Optimal maintenance time for repairable systems. *Journal of Quality Technology*, **39**, 48–53.
- Park, D.H., Jung, G.M., and Yum, J.K. (2000). Cost Minimization for Periodic Maintenance Policy of a System Subject to Slow Degradation. *Reliability Engineering & System Safety*, **68**, 105–112.
- Rigdon, S.E. and Basu, A.P. (2000). *Statistical Methods for the Reliability of Repairable Systems*. John Wiley: New York.

Model selection in sparse contingency tables: LASSO penalties *vs* classical methods

Susana Conde^{1,2,3}, Gilbert MacKenzie^{2,4}

¹ Behaviour and Health Research Unit, University of Cambridge, Institute of Public Health, Robinson Way, Cambridge CB2 0SR

² Centre of Biostatistics, Department of Mathematics and Statistics, The University of Limerick, Ireland

³ Department of Statistics, School of Mathematical Sciences, Western Gateway Building, Western Road, University College Cork, Ireland

⁴ CREST, ENSAI, Rennes, France

E-mail for correspondence: sc778@medsch1.cam.ac.uk

Abstract: We compare improved classical backward elimination and forward selection methods of model selection in sparse contingency tables with methods based on a regularisation approach involving the least absolute shrinkage and selection operator (LASSO) and the Smooth LASSO. The results show that the modified classical methods outperform the regularisation methods, by producing sparser models which are always hierarchical. Curiously, models selected by the regularisation methods often include effects which are known to be inestimable in the classical paradigm. Our findings support the use of classical methodology.

Keywords: Contingency tables; Model selection; Regularisation, Smooth LASSO; Sparseness.

1 Introduction

Penalized likelihood (Eilers and Marx, 1996) has received a lot of attention recently as a method for achieving smoothness, sparsity, etc. In contingency table analysis, Dahinden *et al* (2007), Park and Hastie (2008), and Conde and MacKenzie (2011) each propose different penalized likelihood approaches. The first of these papers proposes an optimization algorithm using a least absolute shrinkage and selection operator (LASSO) and other penalties while the second develops an L_2 -norm penalty in a logistic model. The third paper develops the Smooth LASSO and other LASSO-related penalties. All three approaches are intended to be used in sparse contingency tables that can arise from genetic data or multivariate comorbidity data. Such data sets are typically high-dimensional and accordingly pose a major challenge to model selection.

In this paper, we compare model selection methods in sparse contingency tables. Specifically we compare penalized likelihood approaches with our

classical stepwise algorithms. The penalized likelihood approaches involve the LASSO with the implementation appeared in Dahinden *et al* (2007), the LASSO using the Bayes Information Criterion (BIC), which is novel in this context, and the smooth parametric approximation to the LASSO which appeared in Conde and MacKenzie (2011); the classical algorithms involve modified backwards elimination (MacKenzie-Conde Backwards Elimination (MCBE); Conde, 2011, pp. 137-138, BE2) and forward selection (FS).

2 Methods

Let assume the same notation and model of the expected frequencies as in Conde and MacKenzie (2011) i.e. consider a p -dimensional contingency table with $q = 2^p$ cells, a hierarchical log-linear regression model $\ln(\mu_i) = \sum_{j=1}^k a_{ij} \theta_j$ with Yates' constraints where θ is the vector of unknown parameters measuring the influence of constant, main effects and interactions, and all the other quantities as defined in the mentioned paper. The penalised negative log-likelihood is

$$-\ell^{\mathcal{P}}(\theta, \lambda) = -\ell_{\text{mult}}(\theta) + \text{pen}_{\lambda}$$

where $\text{pen}_{\lambda} = \lambda \sum_{j=2}^k |\theta_j|$ (LASSO) or $\text{pen}_{\lambda} = \lambda \sum_{j=l}^k \omega \ln[\cosh(\theta_j/\omega)]$ (Smooth LASSO) with ω a certain parameter. We estimate λ using 5-fold cross-validation (CV) and again by BIC. For the smooth (LASSO) we set $\omega = 1$ and use the 95% confidence interval around 0 to determine when a parameter is zero. Note that the LASSO penalty in binary variables coincides with the group- L_1 norm, which is invariant to the choice of design matrix. We have that $BIC = -2\hat{\ell} + k \ln n$ where $\hat{\ell}$ is the maximized log-likelihood, k is the number of parameters in the model, and n is the sample size. The algorithms MCBE, BE2, and FS are likelihood-ratio based and work in a stepwise fashion (Conde, 2011, pp. 66-78, 81-85); MCBE starts with the *sparse saturated model* (SSM), that is, the fullest model that can be fitted after eliminating effects with non-existent maximum likelihood estimates that are detected by MacKenzie's theorem (Conde, 2011, pp. 37-38); BE2 starts with a fitting model with up to and including a certain order of interactions, and FS, with a null (or main effects) model. They remove or add one effect at a time until no other effect can be removed or until they find a model that fits.

When considering 5-fold cross-validation, we selected 20 random samples (i.e. sets of 5 training tables and 5 testing tables). For the LASSO with CV, we used the `logilasso` package in R. For the BIC, we used the same path following algorithm in this package in order to have the estimates of the parameters. In all the penalized likelihood approaches we rescaled the original $\lambda \in [0, \infty)$, into $\lambda^* \in [0, 1]$ using the bijective mapping

$$\lambda = 1/\alpha [\ln \{(1 + \lambda^*)/(1 - \lambda^*)\}]$$

with $\alpha = 0.03$.

3 Results

We present the results of a small simulation study and then illustrate the methods by analysing some real data from a COPD study of comorbidities (GSK COPD, 2006).

TABLE 1. Percentages of final models found; $p = 2$, in 100 simulated tables. CV: 5-fold cross-validation. BIC: BIC criterion with a LASSO penalty. The Smooth LASSO approximation is used with $\omega = 1$ and 5-fold cross-validation.

p	n	$model$	%					
			MCBE	BE2	FS	LASSO		Smooth LASSO*
CV	BIC							
2	50	null	4	8	5	0	1	23
		{c1}	11	8	13	0	3	12
		{c2}	6	5	6	3	2	9
		{c1, c2}	20	20	20	15	11	10
		sat.	55	55	52	78	79	40
		**	4					
		Total	100	96	96	96	96	94
2	10	null	18	27	25	7	13	69
		{c1}	15	8	15	4	4	4
		{c2}	15	13	16	5	5	4
		{c1, c2}	13	13	13	14	11	0
		sat.	22	22	14	53	50	4
		**	17					
		Total	100	83	83	83	83	81

* We removed tables when `nlm` did not converge; (2, 2 respectively in each scenario).

** SSM does not fit.

3.1 Simulation study

We simulated 100 2×2 random contingency tables (Conde, 2011, pp. 86-90, 188) and used each of the methods given above to find a final best fitting model.

The sample sizes were $n = 50$ and $n = 10$ leading to sparse tables albeit of low dimension - 21% and 79% of the tables have some zero respectively. According to MacKenzie's theorem (Conde, 2011, pp. 37-39), there are 4 tables in the first scenario and 22 tables in the second scenario with at least one inestimable effect; these lead to 4 and 17, respectively, tables where the sparse saturated model (SSM) does not fit, as indicated by MCBE, which is the only algorithm of the above that can detect this. We removed the tables whose SSM does not fit from these analyses.

Table 1 presents the results of the simulation study. In all the scenarios studied, the classical stepwise algorithms find sparser, i.e., more parsimonious models, and furthermore being in the case of MCBE, free of inestimable effects models (for example, in the first scenario, all the other algorithms found the saturated model in the four tables whose SSM, which is smaller than the saturated, does not fit. Moreover, none of the penalised likelihood approaches take into account the hierarchical rules for model building.

We note in passing that these tables are simulated at random and we do not know the underlying true models. However, for these scenarios with 2×2 tables, the final models found from any of the classical algorithms can be very reliably taken as the true models (Conde, 2011, pp. 99-100). Furthermore, we note that the results are more homogeneous within methods (classical, penalized likelihood) than between methods.

3.2 Real data analysis

As a first step we constructed a three-dimensional contingency table from our comorbidity data, composed by the binary variables: mild liver disease, diabetes, and lung cancer. In Fortran standard order (and the variables in the mentioned order), the table is $\mathcal{Y} = (45426, 20, 2568, 0, 136, 0, 8, 0)$. We note that according to MacKenzie's theorem, the maximum likelihood estimates (MLEs) of the effects $c1c2$, $c1c3$, and $c1c2c3$ are nonexistent. Figure 1(a) displays the values of BIC along the path of λ^* ; BIC is minimum in the MLEs.

Table 2 displays the final models found in this table: the three classical algorithms and the Smooth LASSO found the main effects model, i.e. conclude that the three comorbidities are statistically independent. Having lung cancer is not affected by mild liver disease and diabetes, and *vice versa* with all the combinations of the three comorbidities. The LASSO, in contrast, found either the all 2-ways model (CV) or the saturated model (BIC) so these approaches would conclude that there is a heavy load of interaction pattern between the comorbidities. Furthermore, the CV and BIC LASSO methods include effects which are formally inestimable in the classical paradigm in their final best fitting models. The Smooth LASSO

TABLE 2. Final models found with the comorbidity table. Variables mean c1: mild liver disease; c2: diabetes; c3: lung cancer.

		LASSO		
MCBE, BE2, FS		CV	BIC	Smooth LASSO*
Comorb.	[c1, c2, c3]	[c1c2, c1c3,	[c1c2c3]	[c1, c2, c3]
data		c2c3]		

* We removed $\lambda^* = 0$ from the path as `nlm` did not converge.

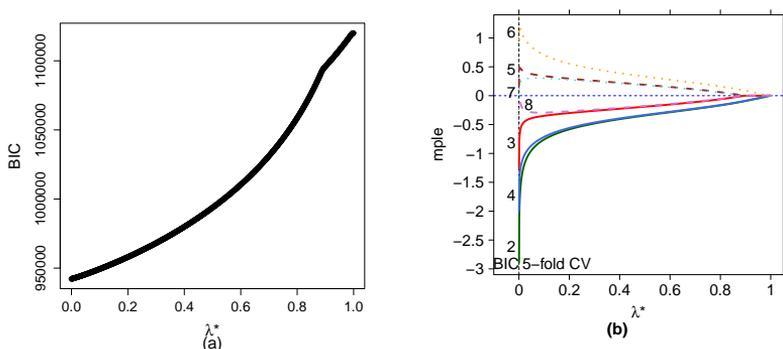


FIGURE 1. Graphs with comorbidity data and the LASSO penalty. (a) BIC along the path of λ^* . (b) Values of the MPLEs along the path of λ^* s. The estimates of λ^* from 5-fold cross-validation and BIC are indistinguishable (and ≈ 0). For both estimates we used the path following algorithm in `logilasso` to maximise the penalised likelihood. The numbers of each line mean 2: c1; 3: c2, ..., 5: c1c2, 6: c1c3, ..., 8: c1c2c3.

is more successful, a result which is in agreement with previous findings, perhaps as a consequence of the larger sample size.

Figure 1(b) displays the values of the estimates along the path of λ^* . The estimates of λ^* are very close to 0; in the case of the BIC, none of the parameters is zero and in CV, the three-way interaction is 0 (the path of this effect is not monotonic in this case, it is zero for the first λ^* s, then different from 0 until λ^* is close to 0.88).

4 Conclusions and discussion

The LASSO penalty is viewed as a method for finding sparse final models. The findings in this paper contradict this overview, whilst comparing

LASSO approaches with classical stepwise algorithms in contingency tables. While the methods in the `logilasso` package succeed in some applications (Dahinden and Bühlmann, 2009), it is not the case here.

The classical methods outperform all of the penalized likelihood approaches by finding the most parsimonious models which are always hierarchical, and in the case of MCBE, free of inestimable effects that are detected by MacKenzie's theorem. The results are based on a small simulation but confirm the work carried out in the PhD thesis of the first author.

We have used the Smooth LASSO approximation considering $\omega = 1$ and the use of a normal-based 95% confidence interval to detect a zero parameter. According to the results, it can be considered to be a contender.

Finally, we have alluded to, but not fully discussed, the issue of the nonexistence of maximum likelihood estimates in sparse tables. Accordingly, we hope to discuss these points in more detail at the workshop when we will present the results of a more comprehensive simulation study.

References

- Conde, S. (2011). *Interactions: Log-Linear Models in Sparse Contingency Tables*. PhD Thesis. The University of Limerick, Ireland.
- Conde, S. and MacKenzie, G. (2008). Search Algorithms for Log-Linear Models in Contingency Tables: Comorbidity Data. In: *Proceedings of the 23rd International Workshop on Statistical Modelling*, Utrecht, 184-187. Ed.: Eilers, P.H.C.
- Conde, S. and MacKenzie, G. (2011). LASSO Penalised Likelihood in High-Dimensional Contingency Tables. In: *Proceedings of the 26th International Workshop on Statistical Modelling*, Valencia, 127-132. Ed.: Conesa, D. and Forte, A. and López-Quílez, A. and Muñoz, F.
- Dahinden, C., Parmigiani, G., Emerick, M.C. and Bühlmann, P. (2007). Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics*, **8**:476.
- Dahinden, and Bühlmann, P. (2009). Decomposition and Model Selection for Large Contingency Tables. *arXiv:0904.1510v2* [stat.ME].
- Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science*, **11**(2) 89-121.
- Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**(1), 30-50.

Forecasting with the age-period-cohort model?

Iain D. Currie¹

¹ Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, UK

E-mail for correspondence: I.D.Currie@hw.ac.uk

Abstract: Future pension liabilities depend critically on the life expectancy of the members of a pension scheme. This remark applies equally to public and private schemes, and to defined-benefit and defined-contribution schemes. Some pension providers use the age-period-cohort model to forecast mortality. We examine some of the pitfalls in this approach. Our methods are based on the theory of constrained generalized linear models and are illustrated with mortality data from the UK Office of National Statistics.

Keywords: Age-period-cohort model, constraints, forecasting, generalized linear model, identifiability.

1 Introduction

The course of future mortality has a major impact on society, since improving longevity implies greater demands in many areas, such as the funding of public and private pensions, the care of the elderly, and the provision of health services. The forecasting of mortality is thus of fundamental importance to providers (whether public or private) of these services. In this paper we concentrate on the forecasting of mortality in the pensions area. Actuaries use forecasts of mortality to determine the level of reserves required to meet future liabilities. The age-period-cohort model, which we refer to as the APC model, is just one of many models that have been used here. For example, in an influential paper, Cairns *et al.* (2011) estimated the parameters in the APC model and then forecast these parameters to produce a forecast of mortality. The Continuous Mortality Investigation also uses the APC model in this way to produce a forecast of mortality (Continuous Mortality Investigation, 2010). There are some problems with this approach since the APC model is not identifiable and so the estimated coefficients depend on the particular constraints that are used to force a unique solution. In this paper we examine critically the forecasting of mortality with the APC model. We illustrate our methods with data from the UK Office of National Statistics (ONS).

The plan of the paper is as follows. In section 2 we define the APC model and give a general formula for estimation in a constrained generalized linear model or GLM. In section 3 we use canonical correlations to examine the level of association among the estimates of the age, period and cohort parameters. In section 4 we use random constraints to emphasise the dependence of the estimated coefficients on the constraints. The paper closes with some general remarks on forecasting with the APC model.

2 The age-period-cohort model

We suppose that we have mortality data, deaths and exposures to the risk of death, arranged in two matrices, $\mathbf{D} = (d_{i,j})$ and $\mathbf{E} = (e_{i,j})$, each $n_a \times n_y$, whose rows and columns are classified by age at death and year of death respectively. The assumption in the APC model is that the force of mortality or hazard rate, λ_{ij} , at age i and in year j is given by

$$\log \lambda_{ij} = \alpha_i + \kappa_j + \gamma_{j-i}, \quad i = 1, \dots, n_a, \quad j = 1, \dots, n_y. \quad (1)$$

Let $n_c = n_a + n_y - 1$ be the number of distinct cohorts, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{n_a})'$, $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_{n_y})'$, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{n_c})'$ and $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\kappa}', \boldsymbol{\gamma}')'$. A tentative interpretation is that $\boldsymbol{\alpha}$ describes the age effect, $\boldsymbol{\kappa}$ the year effect and $\boldsymbol{\gamma}$ the year of birth or cohort effect. The APC model has $n_a + n_y + n_c$ parameters, but the model matrix, \mathbf{X} , has column rank $n_a + n_y + n_c - 3$ so model (1) is not identifiable; any interpretation of the parameters must respect the particular constraints employed to achieve identifiability.

We assume $d_{i,j} \sim \mathcal{P}(e_{i,j} \lambda_{i,j})$ and then (1) defines a GLM with data $\mathbf{d} = \text{vec } \mathbf{D}$, log link and offset $\log \mathbf{e}$ where $\mathbf{e} = \text{vec } \mathbf{E}$; here vec is the operator which stacks the columns of a matrix in column order into a vector. Suppose we number the cohorts from 1 (oldest) to n_c (youngest), and let w_c be the number of times that cohort c appears in the data. We use the following three constraints: $\sum \kappa_j = 0$, $\sum w_c \gamma_c = 0$ and $\sum c w_c \gamma_c = 0$ which we express compactly as $\mathbf{H}\boldsymbol{\theta} = \mathbf{0}$ where \mathbf{H} has dimension $3 \times (n_a + n_y + n_c)$. The condition that these constraints define a unique solution is that the augmented model matrix $\mathbf{X}_{aug} = [\mathbf{X}' : \mathbf{H}']'$, $(n_a n_y + 3) \times (n_a + n_y + n_c)$ has full column rank. It is easily checked that this condition holds for this \mathbf{H} . Currie (2012) shows that the maximum likelihood estimate of $\boldsymbol{\theta}$ subject to the constraint $\mathbf{H}\boldsymbol{\theta} = \mathbf{0}$ is the unique solution of the iterative least squares equation

$$\begin{pmatrix} \mathbf{X}' \tilde{\mathbf{W}} \mathbf{X} & : & \mathbf{H}' \\ \mathbf{H} & : & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\omega}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}' \tilde{\mathbf{W}} \tilde{\mathbf{z}} \\ \mathbf{0} \end{pmatrix}. \quad (2)$$

Here, $\hat{\boldsymbol{\omega}}$ is a vector of Lagrange multipliers, the tilde represents an approximate solution, as in $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\theta}$ is an improved estimate, and $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{z}}$ are the usual *diagonal matrix of weights* and *working variable* respectively in the standard GLM algorithm. Figure 1 shows the estimated parameter values

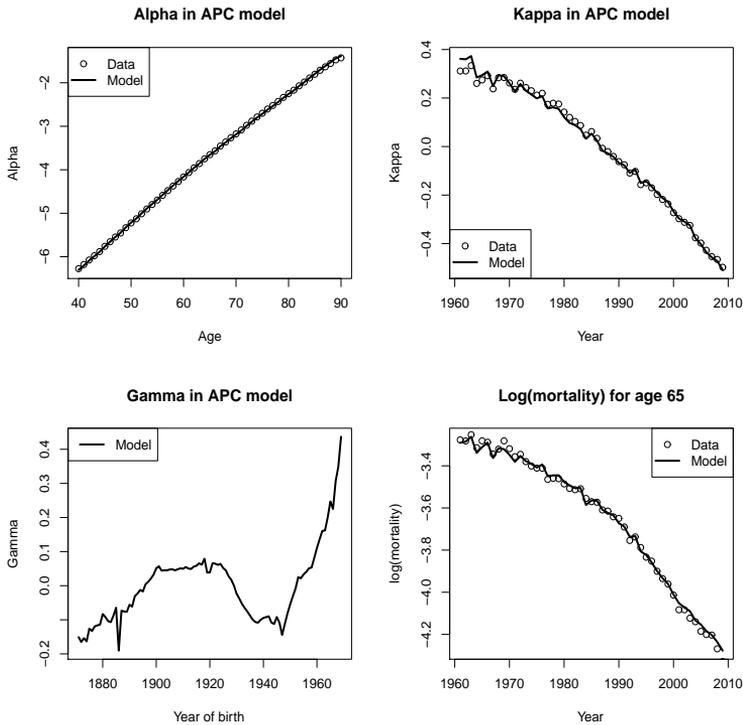


FIGURE 1. Parameter estimates $\hat{\alpha}$, $\hat{\kappa}$ and $\hat{\gamma}$ in the APC model under the constraints \mathbf{H} defined above. Observed and fitted $\log(\text{mortality})$ for age 65 under the APC model for any constraint. Data plots in the upper panels are the row and column averages of $\log(\mathbf{D}/\mathbf{E})$ (column averages have been centred).

with male ONS data for England and Wales for ages 40 to 90, years 1961 to 2009 and the above constraints. Of course, the fitted log mortality is an invariant and does not depend on the particular choice of constraints. With these constraints, Figure 1 suggests that it is reasonable to interpret $\hat{\alpha}$ and $\hat{\kappa}$ as age and period effects respectively, at least for these data. Figure 1 also suggests that it should be possible to forecast the $\hat{\kappa}$ values. On the other hand, the prospects of producing a sensible forecast of the $\hat{\gamma}$ values look bleak indeed.

3 Canonical correlations

Let $\Delta = \mathbf{X}'\hat{\mathbf{W}}\mathbf{X} + \mathbf{H}'\mathbf{H}$ then Currie (2012) shows that

$$\text{var}(\hat{\theta}) = \Delta^{-1} - \Delta^{-1}\mathbf{H}'(\mathbf{H}\Delta^{-1}\mathbf{H}')^{-1}\mathbf{H}\Delta^{-1}. \quad (3)$$

We denote the first canonical correlation between $\hat{\alpha}$ and $\hat{\kappa}$ by $r(\hat{\alpha}, \hat{\kappa})$. Mardia et al. (1979) explain how to calculate $r(\hat{\alpha}, \hat{\kappa})$ from $\text{var}(\hat{\theta})$. We find $r(\hat{\alpha}, \hat{\kappa}) = 0.442$, and in a similar fashion $r(\hat{\alpha}, \hat{\gamma}) = 0.555$ and $r(\hat{\kappa}, \hat{\gamma}) = 0.659$. In their discussion of forecasting with the APC model, Cairns *et al.* (2011) state: “we follow earlier studies (e.g., Renshaw and Haberman (2006)) and assume that the cohort effect γ_{j-i} has dynamics that are independent of the period effect, κ_j ”. Our view is that the dependence of the $\hat{\gamma}$ values on the $\hat{\kappa}$ values is so strong that any forecast of $\hat{\gamma}$ cannot ignore the information in $\hat{\kappa}$. The same remark applies equally to the forecasting of $\hat{\kappa}$.

4 Random constraints

Forecasting with the APC model is done by forecasting the estimated coefficients $\hat{\kappa}$ and $\hat{\gamma}$. However, the estimated coefficients are not invariants of the model and depend on the particular constraints used. We can demonstrate this dependence by using random constraints. This idea may seem odd at first but it serves to emphasise the critical influence of the constraints on the resulting parameter estimates. We define three random constraints by setting $\mathbf{r}'_1 \boldsymbol{\alpha} = 0$, $\mathbf{r}'_2 \boldsymbol{\kappa} = 0$ and $\mathbf{r}'_3 \boldsymbol{\gamma} = 0$ where the elements in \mathbf{r}_1 , \mathbf{r}_2 and \mathbf{r}_3 are independent random draws from the Poisson distribution with mean 1.

Figure 2 shows the results of five simulations of this exercise. The estimates of the parameters under the constraints \mathbf{H} in section 2 are shown by the thick line, **—**. The estimates are quite unstable under these random constraints. Nevertheless, all constraints lead to the same table of fitted $\log(\text{mortality})$, as illustrated in the lower right panel of Figure 2. In terms of fitting the model any set of constraints is as good as any other but it seems unlikely that one would attempt a forecast with any of the sets of random estimates shown in Figure 2. What is it about the particular constraints in section 2 that allow forecasting to take place? We do not have an answer to this question.

5 Conclusion

Many forecasting models make the assumption of independence between estimates of sets of parameters. In this paper we have shown that this assumption is not valid for the APC model as applied to data in England and Wales. We are not the first to raise concerns about forecasting with the APC model. Clayton and Schifflers (1987) wrote: “In recent years, there have been several attempts to use an APC model fitted to past data to forecast rates. It should come as no surprise to a reader of this paper that we should doubt the wisdom of this course”. The purpose of our paper is to quantify some of the fears expressed in the Clayton and Schifflers paper.

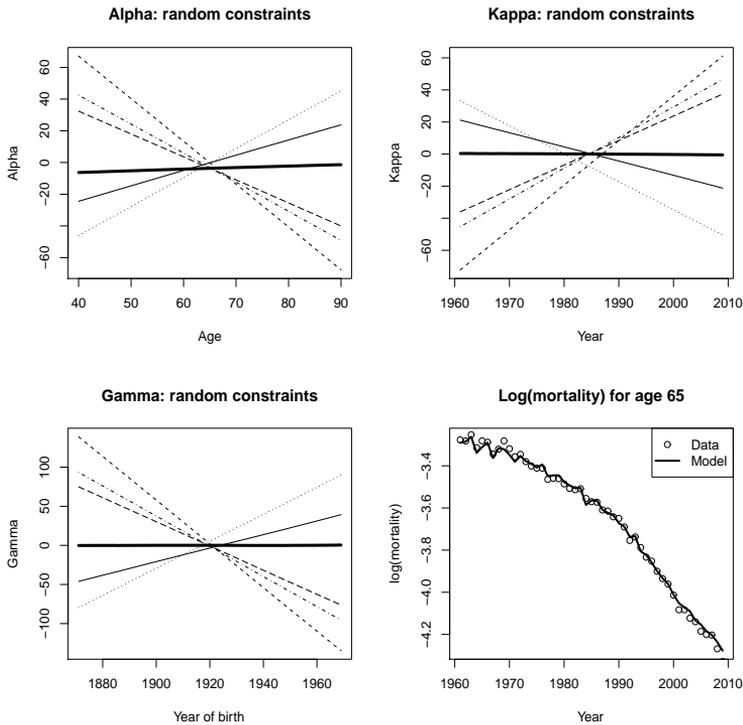


FIGURE 2. Parameter estimates $\hat{\alpha}$, $\hat{\kappa}$ and $\hat{\gamma}$ in the APC model under random constraints. Observed and fitted log(mortality) for age 65 under the APC model for any constraint.

We have done this in two ways, first, by calculating canonical correlations between the estimates, and second, by demonstrating the critical role of the constraints in the estimation of the parameters. Our methods are quite general and can be applied to the Lee-Carter model (Lee and Carter, 1992), for example; see Currie (2012). Forecasts of mortality play a central role in the pricing and reserving of pensions so forecasting methods must be soundly based. We hope that we have contributed to the discussion of such methods.

Acknowledgments: I am grateful to Maria Durban and the Spanish Ministry of Science and Innovation (project MTM2011-28285-C02-02), and to Longevitas Ltd for financial support.

References

- Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M. (2011). Mortality density forecasts: an analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, **48**, 355–367.
- Clayton, D. and Schifflers, E. (1987). Models for temporal variation in cancer rates. II: age-period-cohort models. *Statistics in Medicine*, **6**, 469–481.
- Continuous Mortality Investigation (2010). The CMI Mortality Projections Model, CMI 2010. *Working Paper 49*, 8.
- Currie, I.D. (2012). Efficient estimation in constrained generalized linear models with applications to models of mortality. In preparation.
- Lee, R.D., and Carter, L. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, **87**, 659–675.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*. London: Academic Press.
- Renshaw, A.R. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38**, 556–570.

Cauchy, Prague and multiple regression

Antoine de Falguerolles¹

¹ Université Paul Sabatier - Toulouse III (Retired)

E-mail for correspondence: antoine@falguerolles.net

Abstract: The early contributions to parameter estimation in linear models by French authors, Pierre-Simon de **Laplace** (1749-1829) and Adrien-Marie **Legendre** (1752-1833), are well known. Also well documented are those of the non-French pioneers Tobias **Mayer** (1723-1762) and Carl Friedrich **Gauss** (1777-1855) to name the most famous. Less known is that of Augustin Louis **Cauchy** (1789-1857) who, during his exile to Prague (c. 1835), proposed a simple method for multiple linear regression. Cauchy's method is now forgotten but several publications show that his method was used in parallel with least squares in French speaking circles, even in the early 20th century. This is exemplified by two publications, one by Vilfredo **Pareto** (1848-1923) and one by Lucien **March** (1859-1933). Both comprise seed ideas to the fitting of generalized linear models.

Keywords: Regression, Cauchy, weighted least squares, iteratively weighted least squares.

1 Introduction

Charles X (1757 – 1836), King of France from the House of Bourbon, was over-throned in the 1830 July Revolution. Louis-Philippe I (1773 – 1850), a member of the Orléans branch of the House of Bourbon, proclaimed himself King of the French. (His reign did end with the Revolution of 1848!) Henri d'Artois (1820 - 1883), Charles X's grandson, became the apparent heir to the French throne for the Legitimists. (The succession is more complicated but this is another story.) Former King Charles X went first to Holyrood Palace (Scotland), then to Prague (1832-1836) in the Old Royal Palace and finally to České Budějovice (South Bohemia) and Görz (Slovenia nowadays) where he died in 1836. The young Henri d'Artois followed the whereabouts of his grandfather and finally settled in Frohsdorf (Austria) in 1851.

A corollary to the French dynastic change was the voluntary exile (1830-1838) of Augustin Cauchy, a legitimist and a devout catholic who will refuse to swear the compulsory oaths of allegiance to any of the forthcoming French regime. Cauchy (1789-1857) went first to Fribourg (Switzerland), then to Turino (Italy), and finally to Prague (1833) where he became the science tutor to Charles X's grandson. Cauchy eventually came back to Paris in 1838.

As expected, Cauchy was productive during his exile. In particular, he introduced circa 1835 a new method for stepwise linear fitting. Although this can be seen now as a minor work, Cauchy's proposal offered a decent competitor to the established method of least squares which greatly reduced its computational burden. Cauchy's motivation was the fitting of data to series expansions of functions in physics. But during the 19th century the fitting of data in other fields than physics, geodetic and cosmology emerged. This is exemplified by two articles where Cauchy regression is used in parallel with least squares and where some seed ideas for the use of generalized linear models can be found. One problem is the estimation of the population rate of growth (Pareto, 1897). Another is the fitting of theoretical distribution curve to an observed distribution of wages (March, 1898). Other memories of the scientific work of Cauchy while in Prague can be found. See Cauchy (1836) and Richlík (1957). Proofs of the personal encounters between Cauchy and Bernard Bolzano (1781 - 1848), are given in Richlík (1962).

2 Cauchy's heuristic for regression

Cauchy considered the regression situation in which the response is a non-linear function of an explanatory variable which can be approximated by a linear combination of the leading terms of a series of simpler functions; special attention was given to the estimation of the order of the approximation to be carried out (Cauchy, 1837, p. 195). With this situation in mind, Cauchy introduces a sequential procedure based on the repeated use of simple regressions. Although the least squares estimator could have been used in these simple regressions, Cauchy proposed a fast estimator for the slope.

2.1 Cauchy's linear estimators for simple regression

Cauchy keeps the intuitive form of the least squares estimator for the intercept: $\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}$. His proposal for the slope is:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \text{sign}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n |x_i - \bar{x}|}.$$

In the formula above, the explanatory variable can be centered with respect to any other location parameter. If a median rather than the mean value is taken, the estimator for the slope then coincides with a particular form of the estimator proposed by Mayer (Falguerolles, 2009).

Clearly, Cauchy's approach avoided the $2 \times n$ multiplications implied by the least squares formula, namely $\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$. This simplification was central to Cauchy's approach to multiple regression at a time when computations were hand-made.

2.2 Cauchy's multiple regression

In his publication on least squares (1805), Legendre did show that the first order conditions formed a system of linear equations (now called the normal equations) which fully summarizes the information needed for estimation. For systems larger than order 2, triangularization was likely used to obtain the estimated regression coefficients. Cauchy's approach to multiple regression combines the ideas of triangularization and step by step conditioning of the variables.

First, all variables are centered to their means. Then iteratively at each step, a not-yet-considered currently revised explanatory variable is introduced; the currently revised response variable and all other not-yet-considered currently revised variables are regressed onto this variable according to Cauchy's formula; the slope coefficients thus obtained are stored; the residuals from all these regressions are computed and constitute the updated revised variables. The inspection of the values of the revised response variable may tell if the iterative process is to be stopped; that of the revised explanatory variables may help in finding the 'best' explanatory variable to introduce next. When the process is stopped, back calculation gives the estimated regression coefficients.

The main drawback of Cauchy's heuristic method for multiple regression is that the estimated values for the regression coefficients may depend on the order of introduction of the explanatory variables. However, if Legendre's formula for the slope is used in place of Cauchy's formula, the unique solution given by multiple regression based on least squares is obtained.

2.3 The fate of Cauchy's regression

Cauchy's approach was severely criticized for its lack of theoretical foundation by the *aficionados* of least squares (Bienaymé, 1853). It nevertheless received some persistent attention in France due to Cauchy's fame. Cauchy and least squares were often simultaneously applied as exemplified by two articles published in the *Journal de la Société de statistique de Paris*, namely those authored by Pareto (1897) and March (1898). Cauchy's method was also taught in French textbooks as a possible procedure among others. Examples are Carvallo (1912) and March (1930).

3 (Iteratively) weighted least squares

Pareto's motivating example consists in the estimation of the coefficients of an evolution curve over time for the population size in England and Wales (Pareto, 1897, pp. 371-372). In modern notation, the model for the mean is $\mu_i = \exp\{\eta_i\}$ with $\eta_i = \beta_0 + \beta_1 x_i$ where x_i denotes the time of observation (a log **link function**) and constant variance (a **constant variance function**).

Firstly Pareto notes that regression of the $\log y_i$ on the x_i by ordinary least squares or by Cauchy method might be improper since the minimisation of $\sum_{i=1}^n (\log y_i - \eta_i)^2$ and the minimisation of $\sum_{i=1}^n (y_i - \mu_i)^2$ are not equivalent.

Pareto recalls that the gradient $\nabla_S(\beta_0, \beta_1)$ for the later is:

$$- \left[\begin{array}{c} \sum_{i=1}^n (y_i - \mu_i) \mu_i \\ \sum_{i=1}^n (y_i - \mu_i) \mu_i x_i \end{array} \right]$$

while the Hessian $H_S(\beta_0, \beta_1)$ is:

$$- \left[\begin{array}{cc} \sum_{i=1}^n ((y_i - \mu_i) \mu_i - \mu_i^2) & \sum_{i=1}^n ((y_i - \mu_i) \mu_i x_i - \mu_i^2 x_i) \\ \sum_{i=1}^n ((y_i - \mu_i) \mu_i x_i - \mu_i^2 x_i) & \sum_{i=1}^n ((y_i - \mu_i) \mu_i x_i^2 - \mu_i^2 x_i^2) \end{array} \right].$$

At a stationary point $(\beta_0^*, \beta_1^*)'$, the Hessian simplifies to:

$$H_S(\beta_0^*, \beta_1^*) = \sum_{i=1}^n \mu_i^{*2} \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} - \sum_{i=1}^n \mu_i^* x_i^2 \begin{bmatrix} 0 & 0 \\ 0 & y_i - \mu_i^* \end{bmatrix}$$

Pareto proposes to search for a stationary point by using a Newton-Raphson algorithm in which a simplified version of the Hessian is used:

$$\begin{bmatrix} \beta_0^{(k+1)} \\ \beta_1^{(k+1)} \end{bmatrix} = \begin{bmatrix} \beta_0^{(k)} \\ \beta_1^{(k)} \end{bmatrix} - \left(\sum_{i=1}^n \mu_i^{(k)2} \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} \right)^{-1} \nabla_S(\beta_0^{(k)}, \beta_1^{(k)}).$$

Has Pareto realized that $E[\mu_i x_i^2 (Y_i - \mu_i)] = 0$ and that he uses what is called now Fisher's scoring method?

Secondly, Pareto notes that this iterative method is computationally too burdensome to be used in practice and states that, by properly weighting the regression of the log-transformed response, a fair approximation to the coefficients can be easily calculated. Pareto's proposal is to consider the minimisation of

$$\sum_{i=1}^n y_i^2 (\log(y_i) - (\beta_0 + \beta_1 x_i))^2 .$$

In a later paper, Pareto (1899) justifies his choice as follows. Posing $\omega_i = \log y_i - \eta_i$, the residuals in the improper regression, he notes that $y_i - \mu_i = y_i \omega_i T_i$ where $T_i = \frac{1 - \exp(-\omega_i)}{\omega_i} = 1 - \frac{\omega_i}{2!} + \frac{\omega_i^2}{3!} - \dots$. Assuming that $T_i \approx 1$, $y_i - \mu_i \approx y_i \omega_i$.

It turns out that this weighting scheme simply relates to the well known iteratively weighted least squares procedure. Assuming independent observations with constant variance and log link ($\log(\mu_i) = \eta_i = \beta_0 + \beta_1 x_i$), the loss function is $\sum_{i=1}^n (y_i - \exp\{\beta_0 + \beta_1 x_i\})^2$. At iteration k , the loss function is

$$\sum_{i=1}^n (\exp\{\eta_i^{(k-1)}\})^2 (z_i^{(k)} - (\beta_0 + \beta_1 x_i))^2$$

where $z_i^{(k)} = \frac{y_i - \exp\{\eta_i^{(k-1)}\}}{\exp\{\eta_i^{(k-1)}\}} + \eta_i^{(k-1)}$. Taking the convenient starting values $\eta_i^{(0)} = \log(y_i)$, the loss function at iteration 1 is exactly what Pareto has in mind.

4 Fitting a Gamma density?

March, a former graduate from the *École Polytechnique*, introduces a theoretical model for the density of wages distribution of the form: $f(x) = \alpha x^\beta \exp\{-\gamma x\}$ (March, 1898). Curiously, March ascribes the theoretical curve to the German Otto Ammon. (However, Kleiber and Kotz (2003) claim that they did not trace any such thing in Ammon's publications!) In any case, it is an early attempt of Gamma modelling.

Daringly, March proceeds to the estimation of the unknown coefficients by using a regression. Noting the linear form of the log transformed density, $\log f(x) = \log \alpha + \beta \log x - \gamma x$, and using grouped data, March considers the regression of the log transformed empirical density onto the centres of classes and their logarithms. As expected in France, March uses the two strategies for regression, namely least squares and Cauchy's heuristic, and mentions the possible introduction of the weights suggested by Pareto in these. Nowadays March's approach is highly questionable. First, the fact that α is a function of β and γ is not recognised. Second, empty classes cannot easily be taken into account. The introduction of null weights for these, which bias the estimations, is a possibility but there is no clear justification for weighting by the squared empirical density the non-empty classes. A constant model for the weighted mid-class values in the generalized linear model settings for the Gamma distribution would be a starting point nowadays.

5 An historical *addendum*

Another famous science tutor of Henri d'Artois is Joachim Barrande (1799-1883), a top graduate from the *École Polytechnique* (1819) like Cauchy (1805). Barrande followed the Legitimist heir to the French Crown in Holyrood Palace and then in Prague. Discharged in 1833, Barrande worked as a railway engineer in Bohemia and became an acclaimed paleontologist, specialist of the Bohemian fossils and the trilobites in particular. Barrande, executor of Henri d'Artois's will, died in Froshdorf a few months after his former Royal pupil.

Barrande is well remembered in the Czech republic. A district in south-west Prague, above the Vltava river, is named after him. In this district, in connection with the entrepreneurial Havel family, stands the Czech Hollywood, the Barrandov studios. Famous films were shot there: *Amadeus*, *Casino Royal* ... Not surprisingly the annual film awards are in the form of a Golden Trilobite!

The Barrande program is also a current bilateral exchange program between the Czech and the French republics.

Acknowledgments: Special Thanks to Jaromír Antoch who tipped me on the Barrandov district in Prague.

References

- Bienaymé, J. (1853). Remarques sur les différences qui distinguent l'interpolation de M. Cauchy de la méthode des moindres carrés, et qui assurent la supériorité de cette méthode. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences*, pp. 5-13.
- Carvalho, E. (1912): *Le calcul des probabilités et ses applications*. Gauthier - Villars : Paris.
- Cauchy, A. (1837). Mémoire sur l'interpolation. *Journal de mathématiques pures et appliquées*, **Série 1, vol. 2**, pp. 193-205.
- Cauchy, A. (1836). *Mémoire sur la dispersion de la lumière*. Prague: publié par la Société Royale des Sciences.
- Falguerolles, A. de (2009). Quelques remarques sur la méthode d'ajustement de Mayer : lien avec les méthodes de classifications. *Mathématiques et sciences humaines / Mathematics and Social Sciences*, **189(3)**, pp. 43-58.
- Kleiber, Chr. & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Series in probability and statistics. Wiley.
- March, L. (1898): Quelques exemples de distribution des salaires, contribution à l'étude comparative des méthodes d'ajustement, *Journal de la Société de Statistique de Paris*, **39 (June)**, pp.193-206.
- March, L. (1930): *Les principes de la méthode statistique avec quelques applications aux sciences naturelles et à la science des affaires*. Félix Alcan: Paris.
- Pareto, V. (1897): Quelques exemples d'application des méthodes d'interpolation à la statistique. *Journal de la Société de statistique de Paris*, **58 (novembre)**, pp. 367-379.
- Pareto, V. (1899): Tables pour faciliter l'application de la méthode des moindres carrés. *Zeitschrift Schweizerische Statistik*, pp. 121-150.
- Richlík, K. (1957): Un manuscrit de Cauchy aux Archives de l'Académie tchécoslovaque des Sciences. *Revue d'histoire des sciences et de leurs applications*, **10(3)**, pp. 259-261.
- Richlík, K. (1962) Sur les contacts personnels de Cauchy et de Bolzano. *Revue d'histoire des sciences et de leurs applications*, **15(2)**, pp. 163-164.

Modeling multivariate, overdispersed binomial data with additive and multiplicative random effects

Emanuele Del Fava ¹, Ziv Shkedy ¹, Mehreteab Fantahun Aregay ², Geert Molenberghs ^{1,2}

¹ I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium.

² I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium.

E-mail for correspondence: emanuele.delfava@uhasselt.be

Abstract: Often, when modeling longitudinal binomial data we need to take into account both clustering and overdispersion. When we are interested in accommodating both phenomena, we can use separate sets of random effects that capture the extra variability due to overdispersion and within-cluster association. We propose a series of hierarchical Bayesian random-effects models that deal simultaneously with both phenomena, and we apply them to a sample of multivariate data about HCV and HIV infection prevalence in injecting drug users in Italy from 1998 to 2007.

Keywords: HCV and HIV infection; clustering; overdispersion; MCMC.

1 Introduction

There are two main issues that should be taken into consideration when modeling non-Gaussian longitudinal data, namely, (1) the clustering in data, when we have repeated measurements over time or multivariate data, and (2) the occurrence of overdispersion, meaning that there is more variability in the data than the one prescribed by the mean-variance relation of the distribution. Clustering can be accommodated using subject-specific random effects, usually assumed to be normally distributed, which take into account the association between the observations within the cluster. Overdispersion can be taken into account through a variety of overdispersion models, e.g., the beta-binomial model for grouped data. However, if interest lies in simultaneously combining these two phenomena, the implementation of the numerical solution is not usually straightforward within the frequentist approach. Molenberghs *et al.* (2010) proposed a class of generalized linear models that account for clustering and overdispersion with two separate sets of random effects, using maximum likelihood (ML) estimation.

In this paper we extend their work with a series of generalized linear mixed models (GLMMs) that accommodate overdispersion through different sets of random effects, either additively or multiplicatively included in the model, while dealing with clustering. In order to avoid the difficulties encountered with the ML estimation, we opt for a Bayesian approach, using MCMC methods.

We apply this methodology to a sample of prevalence data about hepatitis C virus (HCV) and human immunodeficiency virus (HIV) infection status of injecting drug users (IDUs) in treatment from the twenty Italian regions from 1998 to 2007.

2 Methodology

2.1 The Basic Joint Model

The data consist of aggregate repeated measurements, $(y_{i1k}, \dots, y_{i10k})$, which represent the number of reported cases with infection k in the i th region in year j . The data have been discussed and analyzed in Del Fava *et al.* (2011) using joint hierarchical GLMMs for HCV and HIV infection prevalence. We refer to Figure 1 for a graphical representation of the data. In the first stage of the hierarchical model, we assume that the distribution of y_{ijk} is binomial with index n_{ijk} :

$$y_{ijk} \sim \text{Bin}(\pi_{ijk}, n_{ijk}) \quad i = 1, \dots, 20 \quad j = 1, \dots, 10 \quad k = 1, 2.$$

Then we define for π_{ijk} a logistic model, which adjusts for clustering in data, but not for overdispersion:

$$\begin{cases} \text{logit}(\pi_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1}, \\ \text{logit}(\pi_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2}. \end{cases} \quad (1)$$

For infection k , the model includes time-specific fixed effects, β_{jk} and region-specific random intercepts, γ_{ik} . The best model for the data is a correlated random-effects model where γ_{ik} follows a bivariate normal distribution with unstructured covariance matrix D :

$$\begin{pmatrix} \gamma_{i1} \\ \gamma_{i2} \end{pmatrix} \sim \text{MVN} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, D = \begin{pmatrix} \sigma_{\gamma_1}^2 & \rho_{\gamma_1\gamma_2} \sigma_{\gamma_1} \sigma_{\gamma_2} \\ \rho_{\gamma_1\gamma_2} \sigma_{\gamma_1} \sigma_{\gamma_2} & \sigma_{\gamma_2}^2 \end{pmatrix} \right]. \quad (2)$$

The infection-specific variances $\sigma_{\gamma_1}^2$ and $\sigma_{\gamma_2}^2$ account for the within-region association, whereas the coefficient $\rho_{\gamma_1\gamma_2}$ measures the correlation between the two infections at the level of the linear predictor, averaged all over the years.

In the second stage of the hierarchical model, we specify the prior distributions for the unknown parameters. We use diffuse priors in order to include as much uncertainty as possible in the estimation process. For the fixed

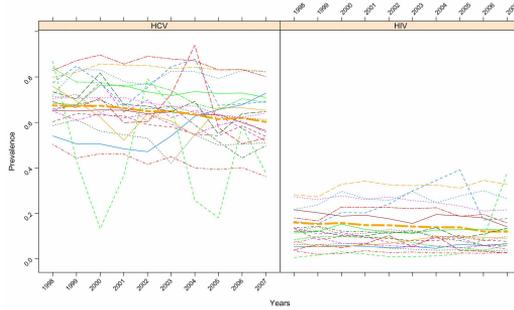


FIGURE 1. Observed regional profiles for HCV (left panel) and HIV (right panel) infection from 1998 to 2007.

effects, α_k and β_{jk} , we use normal priors with mean zero and large variance. For the random intercepts, we specify a bivariate normal prior with a mean zero and precision matrix $S = D^{-1}$, for which, in turn, we specify a Wishart prior. For further details about the prior distributions, see Del Fava *et al.* (2011).

2.2 Joint GLMM with additive overdispersion parameters

In this section we present an extension to the GLMM formulated in (1) to account also for overdispersion, due to the extra binomial variability within the years. We thus update model (1) with additive overdispersion parameters, θ_{ijk} (McLachlan, 1997):

$$\begin{cases} \text{logit}(\pi_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1} + \theta_{ij1}, \\ \text{logit}(\pi_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2} + \theta_{ij2}. \end{cases} \quad (3)$$

The parameters θ_{ijk} are random effects for which we wish to test three different configurations. In the first case, they are shared between the two infections, with $\theta_{ij1} = \theta_{ij} \sim N(0, \sigma_\theta^2)$ and $\theta_{ij2} = \delta\theta_{ij} \sim N(0, \delta^2\sigma_\theta^2)$. For the hyperparameters δ , which relaxes the hypothesis of common variance between HCV and HIV infection, and σ_θ^2 , flat priors are chosen.

In the second case, θ_{ijk} is bivariate normally distributed with the following alternative covariance matrices:

$$D_I = \begin{pmatrix} \sigma_{\theta_1}^2 & 0 \\ 0 & \sigma_{\theta_2}^2 \end{pmatrix} \text{ vs. } D_U = \begin{pmatrix} \sigma_{\theta_1}^2 & \rho_{\theta_1\theta_2}\sigma_{\theta_1}\sigma_{\theta_2} \\ \rho_{\theta_1\theta_2}\sigma_{\theta_1}\sigma_{\theta_2} & \sigma_{\theta_2}^2 \end{pmatrix}. \quad (4)$$

Selecting the independent model with D_I implies that there is overdispersion within the years for each infection, but no further correlation between the infections within the years. Instead, selecting the correlated model with D_U implies that it exists additional correlation, $\rho_{\theta_1\theta_2}$, between infections

at the level of the yearly measurements. For these two matrices, we specify Wishart prior distributions similar to the one for the random intercepts γ_{ik} .

2.3 Joint GLMM with multiplicative overdispersion parameters

A second modeling approach considers a setting in which we account for overdispersion using multiplicative effects. Hence we assume that

$$\begin{cases} y_{ijk} \sim \text{Bin}(\pi_{ijk} = \theta_{ijk} \cdot \kappa_{ijk}, n_{ijk}), \\ \text{logit}(\kappa_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1}, \\ \text{logit}(\kappa_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2}. \end{cases} \quad (5)$$

Note that, since we include the overdispersion parameter at the level of prevalence, it follows that $0 \leq \theta_{ijk} \leq 1$ to ensure that $0 \leq \theta_{ijk} \kappa_{ijk} \leq 1$. To fit the model in a Bayesian framework, we specify a Beta prior distribution for θ_{ijk} , with uniformly distributed hyperparameters taking positive values within a large range:

$$\theta_{ij} \sim \text{Be}(a, b), \quad \text{with } a, b \sim U(1, 100).$$

We even test a simpler specification for the Beta parameters, that is, $a = b = 1$, which implies a $U(0, 1)$ distribution.

3 Results

All the models are fitted to data using MCMC methods through JAGS software and the best models are selected using the penalized expected deviance (PED, Plummer *et al.* 2008), which tends to penalize more the complex models. We refer to Table 1 for a summary of our main results. The best additive model has shared overdispersion parameters, while the best multiplicative model has hyperparameters $a, b \neq 1$. We see that the additive model outperforms both the basic and the multiplicative models. In the additive model, the correlation arising from clustering, $\hat{\rho}_{\gamma_1 \gamma_2}$, is larger than 0.7, while in the multiplicative model it decreases to 0.56 and shows a larger variability. We refer to Figures 2–4 for a graphical representation of the results. Note that the basic model fits poorly because unable to capture the extra variability in each year, while this can be accomplished by both overdispersion models.

4 Conclusions

In this paper we used joint GLMMs for HCV and HIV infection prevalence data, repeatedly measured from 1998 to 2007, in order to account for clustering and overdispersion. As concerns the latter, we used either additive or

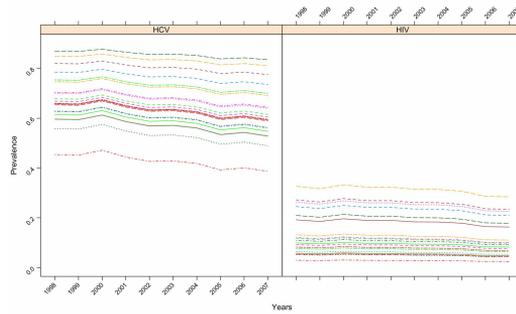


FIGURE 2. Fitted regional profiles from the basic GLMM for HCV (left panel) and HIV (right panel) infections from 1998 to 2007.

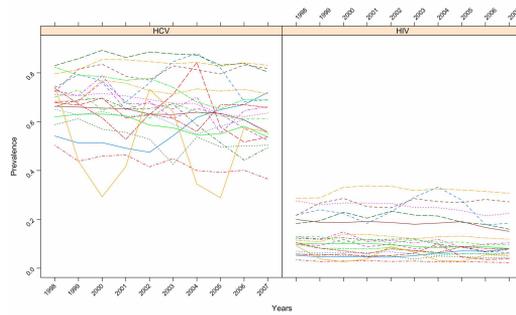


FIGURE 3. Fitted regional profiles from the GLMM with additive overdispersion parameters for HCV (left panel) and HIV (right panel) infections from 1998 to 2007.

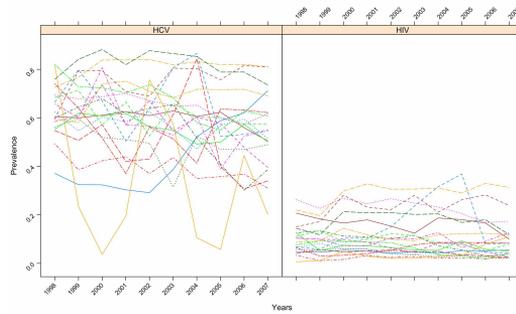


FIGURE 4. Fitted regional profiles from the GLMM with multiplicative overdispersion parameters for HCV (left panel) and HIV (right panel) infections from 1998 to 2007.

TABLE 1. Comparison of the estimated models by PED.

Type	Model	PED
Basic	No θ_{ijk}	10351
Additive	Shared θ_{ijk}	6176
Additive	Independent θ_{ijk}	7469
Additive	Correlated θ_{ijk}	7515
Multiplicative	$Be(1, 1)$	8270
Multiplicative	$Be(a, b)$	7224

multiplicative random effects to accommodate the extra variability along the years. Dealing only with clustering ignoring overdispersion (and vice versa, possibly) can result in a poor fit, as we see from Figure 2. Further research will investigate the time patterns of correlation, by allowing the overdispersion parameters correlation to differ along the years.

References

- Del Fava, E., Kasim, A., Usman, M., Shkedy, Z., Hens, N., Aerts, M., Bollaerts, K., Scalia Tomba, G., Vickerman, P., Sutton, A.J., Wiessing, L., and Kretzschmar, M. (2011). Joint Modeling of HCV and HIV Infections among Injecting Drug Users in Italy Using Repeated Cross-Sectional Prevalence Data. *Statistical Communications in Infectious Diseases*, **3**, 1–24.
- McLachlan, G.J. (1997). On the EM algorithm for overdispersed count data. *Statistical Methods in Medical Research*, **6**, 76–98.
- Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A.M.C. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.
- Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523–539.

A measure of household financial fragility

Filippo Domma¹, Sabrina Giordano¹

¹ Department of Economics and Statistics, University of Calabria, Italy

E-mail for correspondence: sabrina.giordano@unical.it

Abstract: The paper is inspired by the stress-strength models in the reliability literature, in which given the strength (Y) and the stress (X) of a component, its probability of failure is measured by $P(X > Y)$. In this literature, X and Y are typically modeled as independent. Since in many applications such an assumption might not be realistic, we propose a copula approach in order to take into account the dependence between X and Y . We then apply a copula-based approach to the measurement of household financial fragility. Specifically, we define as financially fragile those households whose yearly consumption (X) is higher than income (Y), so that $P(X > Y)$ is the measure of interest and X and Y are clearly not independent. Modeling income and consumption as Dagum distributed variables and their dependence by a Frank copula, we show that the proposed method improves the estimation of household fragility on data from the Bank of Italy.

Keywords: Stress-strength models; Reliability; Copula; Dagum distribution.

1 Introduction

The stress-strength term comes from a reliability problem: it describes the life of a component which has a random strength Y and is subject to a random stress X . If the stress exceeds the strength ($X > Y$) the component will fail, while the component works whenever $X < Y$. Thus, $\mathfrak{F} = P(X > Y)$ measures the likelihood of failure and $\mathfrak{R} = P(X < Y)$ is a measure of component reliability. There is a well-developed theory for the case in which \mathfrak{R} , \mathfrak{F} are calculated under the assumption that X and Y are independent variables, see e.g. Kotz et al. (2003). Yet little attention has been paid to the more realistic problem in which X and Y are dependent. In this paper, we calculate the failure measure \mathfrak{F} taking the stress-strength relationship into account through a copula-based approach.

2 Household financial fragility

The recent crisis has highlighted that understanding the ability of households to offset adverse changes in their financial circumstances is of great relevance for policy makers. Attention to household financial stress has

been rising in recent years. Several authors explore measures of financial fragility to identify which households are potentially more vulnerable to unfavorable changes in the economic environment. Despite the increasing contributions, a highly debated, but still untangled point in the literature concerns exactly the definition of household financial fragility. Recently, Lusardi et al. (2011) measure financial fragility by examining the household ability to come up with an unexpected expense of 2000 dollars in one month regardless of the source of funds (i.e. savings, borrow from family/friends, traditional access to credit, work more etc). In most studies, the financial fragility measures are compared across countries suggesting that the interest in this issue is not confined to the Italian experience. Our idea is to identify a stress-strength problem in the household budget management: the component is the household having at its disposal a random strength Y , income, which is subject to a stress X , consumption. If the stress exceeds the strength the component fails, so when expenses outpace the disposable income the household has a financial problem and needs to borrow or to decumulate its wealth and becomes vulnerable. To put it bluntly, when consumption exceeds income, households face financial stress, so $\mathfrak{F} = P(X > Y)$ is directly interpretable as a measure of household financial fragility. It is worth noting that it may be the case, however, that households are reliant on their savings to support their consumption without difficulties in managing financially. But, our definition of fragility implicitly assumes that the household is also susceptible to financial stress when it needs to access its cumulated savings, as income does not suffice, to cope with expected or unexpected consumption within the year. Income and consumption of Italian households are drawn from the 2008 wave of the Bank of Italy's Survey on Household Income and Wealth (SHIW08). The variable Y indicates the net disposable household income obtained in 2008 as sum of income coming from payroll employment, pension and net transfers, self-employment and properties of all the income-earners in the household; X denotes the year's durable and non-durable consumption.

3 \mathfrak{F} for dependent variables

Dependence between variables is completely described by the joint distribution function which can be specified by modelling the univariate margins and the dependence structure separately. A flexible tool to model different kinds of dependence is the copula function (Joe, 1997). A two-dimensional copula is merely a bivariate distribution function with Uniform (0,1) margins. According to Sklar's theorem, any bivariate distribution $H(x, y) = P(X \leq x, Y \leq y)$ with continuous margins $F(x)$ and $G(y)$ can be written as $H(x, y) = C(F(x), G(y))$, where C is a unique copula. The joint density function is denoted by $h(x, y) = c(F(x), G(y)) f(x)g(y)$ where $c(F(x), G(y)) = \frac{\partial^2 C(F(x), G(y))}{\partial F(x) \partial G(y)}$ is the copula density, and $f(x), g(y)$

indicate the marginal density functions. Consequently, the measure \mathfrak{F} allowing for the dependence between X and Y , with positive values, turns out to be

$$\mathfrak{F} = \int_0^{+\infty} \int_0^x h(t, y) dy dt = \int_0^{+\infty} \int_0^x c(F(t), G(y)) f(t)g(y) dy dt. \quad (1)$$

Copulas and marginal distributions depend on one or more parameters, hereafter considered implicitly. In the rest, we concentrate on two steps: 1. to find appropriate marginal parametric distributions and a copula which serve the need to suitably model the joint distribution of income-consumption data; 2. to estimate the measure \mathfrak{F} of Italian household fragility accounting dependent income and consumption variables.

Choice of marginal distributions. The Dagum distribution has been appreciated by economists and is often preferred to its parametric competitors to model income data as highlighted in several empirical applications. For this reason, we assume that the marginal densities of household income and consumption belong to the three-parameter Dagum family with cumulative distribution function $F(x; \gamma) = (1 + \lambda x^{-\delta})^{-\beta}$ and density function $f(x; \gamma) = \beta \lambda \delta x^{-\delta-1} (1 + \lambda x^{-\delta})^{-\beta-1}$ where $x \in \mathbb{R}^+$, and $\gamma = (\beta, \lambda, \delta)$, positive. Dagum distributions for consumption and income are henceforth denoted by $F(x; \gamma_c)$ and $G(y; \gamma_i)$, respectively. We fit the Dagum distri-

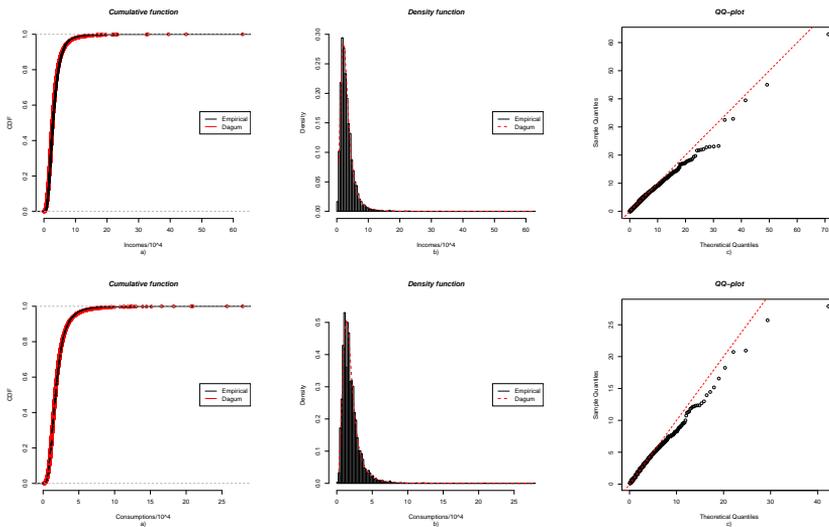


FIGURE 1. Graphical goodness-of-fit on income-consumption data: a) empirical and fitted cumulative distribution functions; b) empirical and fitted density functions; c) QQ-plots

bution to the SHIW08 data. The maximum likelihood estimates (MLEs)

TABLE 1. MLE, standard errors (SE), 95% asymptotic confidence intervals, max log-likelihood values (l), the Anderson-Darling (AD) statistic test (p-value)

	γ	MLE	SE	95% CI	l	AD (<i>p-value</i>)
Consumption	β_c	1.08	0.050	0.98–1.18	-11118.5	1.07 (<i>0.32</i>)
	δ_c	2.99	0.054	2.89–3.10		
	λ_c	4.33	0.417	3.51–5.15		
Income	β_i	0.86	0.035	0.79–0.93	-15405.1	1.34 (<i>0.22</i>)
	δ_i	2.97	0.054	2.87–3.08		
	λ_i	23.8	2.795	18.3–29.3		

$\hat{\gamma}_c = (\hat{\beta}_c, \hat{\lambda}_c, \hat{\delta}_c)$ and $\hat{\gamma}_i = (\hat{\beta}_i, \hat{\lambda}_i, \hat{\delta}_i)$ are obtained by numerical methods as the solution of the maximum likelihood equations is not in closed form. Table 1 reports the summary of the estimation. The Anderson-Darling goodness-of-fit statistic values confirm that the Dagum distribution is suitable for these data. Moreover, the probability plots in Figure 1 show an excellent goodness-of-fit of the Dagum model.

Choice of copula. The next practical step is the choice of an appropriate copula function to model the association of income-consumption data. The coverage of an appropriate dependence range, a graphical approach and a goodness-of-fit test are the three criteria that oriented our choice. The copula functions depend on one or more association parameters, θ . Common association measures, such as Kendall's τ and Spearman's ρ , are usually expressed as function of θ . It is worth noting that τ and ρ values corresponding to the θ domain do not necessarily cover the whole interval $[-1, 1]$ and the range of dependence that can be really achieved varies for different copulas. So, since the empirical values of τ and ρ on SHIW08 data are $\tau^E = 0.494$ and $\rho^E = 0.677$ showing a medium-high degree of positive association, a critical step involves choosing one copula whose association parameter lies within a range which allows τ, ρ to cover at least that empirical value. We restrict our search to a copula which possesses this property. Moreover, a graphical tool provides us with a rough copula selection strategy: the comparison between the plot of the pairs of fitted margins (\hat{F}_j, \hat{G}_j) , $j = 1, \dots, n$, where $\hat{F}_j = F(x_j; \hat{\gamma}_c)$, $\hat{G}_j = G(y_j; \hat{\gamma}_i)$, and the scatterplot of simulated data (U_j, V_j) , $j = 1, \dots, n$, (Joe, 1997, page 146) from different copulas may orient the choice towards certain copula families whose scatterplot looks like that of the fitted marginal values on the unit square. The third but key criterion for the choice of a copula consists in performing a goodness-of-fit test. A natural goodness-of-fit test involves quantifying a distance between the empirical copula C_n , which is a non-parametric estimator of the true unknown copula, and the estimated copula $C(\hat{\theta})$. We considered the statistic S_n which performed well in Genest et al. (2009). Large values of this statistic lead to the rejection of the hypothesized

copula family. Approximated p-values are obtained by a bootstrap-based procedure. The Frank copula satisfies all the three criteria, other common Archimedean copulas, e.g. Clayton and Gumbel, do not perform all the same. The Frank family allows a range of both negative and positive dependence, shows a good fitting to the data ($S_n = 0.019, p\text{-value} = 0.2$) and is also in accordance with the graphical check. In this respect, shown in Figure 2, the scatterplot of data simulated from a Frank copula with $\hat{\theta} = 5.489$ (i.e. the estimated value for θ as shown below) and that of the pairs of Dagum margins seem very similar.

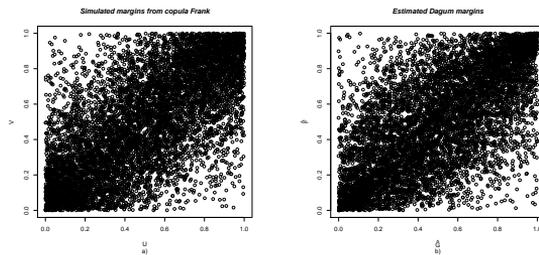


FIGURE 2. Scatterplots: a) U_j, V_j are simulated from a Frank copula with $\hat{\theta} = 5.489$; b) $\hat{F}_j = F(x_j; \hat{\gamma}_c)$, $\hat{G}_j = G(y_j; \hat{\gamma}_i)$ are the estimated Dagum cumulative functions on income-consumption data

Next step is the estimation of θ . The Frank copula density, with $\theta \in \mathbb{R} \setminus \{0\}$, is

$$c(F, G; \theta) = \frac{\theta(1 - e^{-\theta})e^{-\theta(F+G)}}{[1 - e^{-\theta} - (1 - e^{-\theta F})(1 - e^{-\theta G})]^2}.$$

The estimation procedure employed is the IFM, e.g. Joe (1997). It consists of two steps: the MLEs $\hat{\gamma}_c$ and $\hat{\gamma}_i$ of the consumption and income Dagum distributions are provided in the first step. They are then plugged into the log-likelihood function $l(\theta) = \sum_{j=1}^n \log[c(F(x_j; \hat{\gamma}_c), G(y_j; \hat{\gamma}_i); \theta)]$ which is maximized with respect to θ , $c(\cdot)$ being the Frank density copula and $F(x_j; \hat{\gamma}_c), G(y_j; \hat{\gamma}_i)$ are the estimated Dagum cumulative functions.

The starting value $\theta_0 = 5.62$ used in the maximization procedure is obtained by the inversion of Kendall's τ , a method-of-moment estimate for one-parameter copulas; the resulting MLE on income-consumption data is $\hat{\theta} = 5.489$, $SE = 0.08$, the estimated Kendall's τ , $\tau(\hat{\theta}) = 0.486$, and Spearman's ρ , $\rho(\hat{\theta}) = 0.678$, well approximate the empirical values $\tau^E = 0.494$ and $\rho^E = 0.677$.

Estimate of household fragility. We shall proceed to estimate the measure of Italian household fragility. Substituting the Dagum and Frank copula densities in the expression (1), \mathfrak{F} turns out to be

$$\mathfrak{F} = a \int_0^{+\infty} \int_0^x \frac{e^{-\theta(F+G)} (1 + \lambda_c t^{-\delta_c})^{-\beta_c - 1} (1 + \lambda_i y^{-\delta_i})^{-\beta_i - 1}}{t^{\delta_c + 1} y^{\delta_i + 1} [1 - e^{-\theta} - (1 - e^{-\theta F})(1 - e^{-\theta G})]^2} dy dt \quad (2)$$

where $a = \theta(1 - e^{-\theta})\beta_i\lambda_i\delta_i\beta_c\lambda_c\delta_c$. From the invariance property of MLEs, computing $\hat{\mathfrak{F}}$ as function of the estimated parameters $\mathfrak{F}(\hat{\gamma}_c, \hat{\gamma}_i, \hat{\theta})$ provides the MLE of \mathfrak{F} . We solve (2) numerically as it does not admit an explicit solution. The estimated value is $\hat{\mathfrak{F}} = 0.172$, whereas the empirical value is $\mathfrak{F}^E = \frac{\text{n. of households with (consumption - income)} > 0}{\text{n. of surveyed households}} = 0.156$ evaluated for the surveyed households in 2008, i.e. the yearly consumption of 15 households out of 100 is beyond their current income. The estimated $\hat{\mathfrak{F}}$ accounting for the existing dependence in the data seems to well approximate the empirical value. Now we calculate \mathfrak{F} in case of the consumption and income variables are assumed independent and compare it with $\hat{\mathfrak{F}}$ and \mathfrak{F}^E . For $\theta \rightarrow 0$ in the

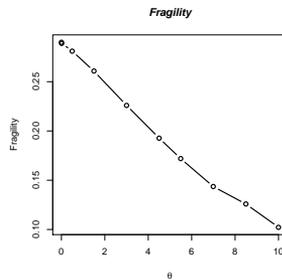


FIGURE 3. \mathfrak{F} for $\theta = 0.001, 0.01, 0.5, 1.5, 3, 4.5, 5.489, 7, 8.5, 10$

Frank copula, the marginal variables are independent. So, values of θ less than $\hat{\theta}$ approaching to zero reveal departure from the actual presence of dependence in the data towards the independence assumption. Decreasing θ leads to increasing values of \mathfrak{F} , Figure 3 shows this behavior for some values $\theta \in (0, 10]$. Hence, neglecting the existing dependence between income and consumption actually overestimates household financial fragility. The contribution of the dependence on \mathfrak{F} is remarkable and cannot be omitted.

References

- Genest, C., Remillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: a review and a power study. *Insurance: Mathematics and Economics*, **44**, 199–213.
- Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Monographs in Statistics and Probability 73. New York: Chapman & Hall.
- Kotz, S., Lumelskii, Y., and Pensky, M. (2003). *The Stress-Strength Model and Its Generalizations. Theory and Applications*. World Scientific.
- Lusardi, A., Schneider, D., and Tufano, P. (2011). *Financially fragile households: evidence and implications*. CeRP Working Papers 116.

A joint model with marginal interpretation for longitudinal continuous and time-to-event outcomes

Achmad Efendi¹, Geert Molenberghs^{2,1}, Edmund Njagi¹, Paul Dendale³

¹ I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium

² I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium

³ Virga Jesse Hospital, Heart Center Hasselt, Hasselt, Belgium

E-mail for correspondence: achmad.efendi@student.kuleuven.be

Abstract: This paper proposes a marginalized joint model for longitudinal continuous and repeated time-to-event outcomes, extending work of Njagi et al. (2012), as well as a marginalized joint model for bivariate repeated time-to-event outcomes.

Keywords: Joint Modeling; Combined Model; Marginalization.

1 Introduction

Ever more commonly does one jointly collect longitudinal and time-to-event outcomes, the latter possibly censored. While an extensive amount of literature is available for Gaussian and other longitudinal outcomes, and literature on the joint modeling of a longitudinal outcome and a single time to event is rapidly growing, methods for the more general setting where at least two longitudinal sequences of perhaps different data types are jointly recorded has received less attention so far, especially when one or more of the sequences consist of times-to-event. Nevertheless, such designs are not uncommon in practice, as our two case studies, introduced in the next section, underscore. Recently, Njagi et al. (2012) formulated joint models for pairs of jointly measured outcomes where for each type of outcome, two sets of random effects are considered, the conjugate and the normal random effects, extending the so-called combined model introduced by Molenberghs et al. (2010). However, the joint model is formulated conditionally upon the random effects, with then the random-effects distribution specified, the parameters have a subject-specific interpretation. This poses difficulties when scientific research is geared towards marginal, population-averaged effects. To allow for such interpretation nevertheless, we supplement the work of Njagi et al. (2012) by a model with marginal interpretation. Focus

is on the case where a repeated continuous and a repeated time-to-event outcome are measured simultaneously (the base model referred to as JCS; the marginalized version JCS-M), as well as on the situation of a bivariate repeated time-to-event outcome (BSS and BSS-M). The marginalization is done following ideas of the so-called marginalized multilevel model (MMM) proposed by Heagerty (1999).

2 Motivating case study

The first set data are from a study with the objective to check whether the follow-up of chronic heart failure (CHF) patients, by means of a tele-monitoring program, reduced mortality and re-hospitalization rates. Heart rate was longitudinally collected from 80 patients, recorded each day for a period of between 182 to 186 days. In addition, the following variables were also recorded: patient's gender, age, and heart rhythm at baseline. Our analysis of these data will be focusing on testing for a joint effect of heart rhythm on repeated time-to-hospitalization (as patients might experience multiple hospitalization) as well as on the longitudinal heart rate. The second set is a so-called comet assay. The data were collected in four groups of six male rats that received a daily oral dose of a compound in three dose levels (low, medium, high) or vehicle control. A cell suspension was prepared for each animal, from each of which three replicate samples were prepared for scoring. There were 50 randomly selected non-overlapping cells per sample, scored for DNA damage using a semi-automated scoring system. A total of 150 liver cells per animal was scored. DNA damage was assessed through the software system by measuring percentage of tail intensity and tail moment, these two responses has heavy tailed distribution and more or less similar to Weibull's. The data take the form of a multi-level structure where a cell suspension or slide, containing three replicate samples, is nested within an animal. In this paper, we target one clustering level, i.e., the slide. We also target two dose levels, low and medium.

3 Method and Estimation

3.1 Ingredients

There are three components to be involved to propose the joint models: linear mixed model (for longitudinal continuous outcomes), the combined model (for repeated, overdispersed time-to-event data), and the marginalization approach. We refer to Verbeke and Molenberghs (2000) to review the linear mixed model (LMM),

$$Y_{ij} = x'_{ij}\boldsymbol{\xi} + z'_{ij}\mathbf{b}_i + \varepsilon_{ij}. \quad (1)$$

where Y_{ij} denotes the response of interest, for the i th subject, measured at time τ_{ij} , $i = 1, 2, \dots, N$, $j = 1, 2, \dots, n_i$. The \mathbf{x}_{ij} and \mathbf{z}_{ij} are p - and q -vectors of known covariates, with $\boldsymbol{\xi}$ a p -dimensional vector containing the fixed effects. The \mathbf{b}_i and ε_i are assumed to be independent and distributed $\mathbf{b}_i \sim N(0, D)$ and $\varepsilon_i \sim N(0, \Sigma_i)$, respectively. This assumption is also applied to the rest of the paper. Meanwhile, the combined model (Molenberghs et al, 2012) for time-to-event outcomes can be the Weibull-gamma-normal model, specified as

$$Y_{ij} | \mathbf{b}_i, \theta_{ij} \sim \text{Weibull}(\rho, k_{ij}), \tag{2}$$

$$k_{ij} = \lambda \theta_{ij} e^{\tilde{\mathbf{x}}'_{ij} \boldsymbol{\xi} + \tilde{\mathbf{z}}'_{ij} \mathbf{b}_i}, \tag{3}$$

$$\theta_{ij} \sim \text{Gamma}(\alpha, \beta), \tag{4}$$

with Y_{ij} the time-to-event outcome of individual i at occasion j . The design vectors $\tilde{\mathbf{x}}_{ij}$ and $\tilde{\mathbf{z}}_{ij}$ play a role similar to their counterparts in the linear mixed model. Further, κ_{ij} is the mean function, ρ is the shape parameter, and the parametrization of the linear predictor is chosen in analogy with (1). Furthermore, regarding the marginalization, we adopt the idea of Heagerty (1999). A fully general MMM formulation is:

$$g(\mu_{ij}^m) = \tilde{\mathbf{x}}'_{ij} \boldsymbol{\xi}^m, \tag{5}$$

$$g(\mu_{ij}^c) = \Delta_{ij} + \tilde{\mathbf{z}}'_{ij} \mathbf{a}_i, \tag{6}$$

$$\mathbf{a}_i \sim F_a(0, D), \tag{7}$$

$$Y_{ij}^c = Y_{ij} | \mathbf{a}_i \sim F_{Y^c}(\mu_{ij}^c, v). \tag{8}$$

Retaining notational conventions used so far, (5) and (6) can be seen as specifying the marginal and conditional means, respectively, thereby linking them through so-called connector function Δ_{ij} . Each outcome Y_{ij} follows an exponential family model with distribution F_{Y^c} , as specified in (8). The $g(\cdot)$ is a link function applied to both means. The function Δ_{ij} depends on the covariates, marginal parameters, and random-effects specification. It connects the marginal and conditional means and can be obtained from solving the integral equation: $g^{-1}(\mathbf{x}'_{ij} \boldsymbol{\xi}^m) = \mu_{ij}^m = \int_a g^{-1}(\Delta_{ij} + \tilde{\mathbf{z}}'_{ij} \mathbf{a}_i) dF_a$. The MMM idea applies without difficulty to the combined model, with the integral equation now becomes:

$$g^{-1}(\tilde{\mathbf{x}}'_{ij} \boldsymbol{\xi}^m) = \mu_{ij}^m = \int_a \int_{\theta} g^{-1}(\Delta_{ij} + \tilde{\mathbf{z}}'_{ij} \mathbf{a}_i) d\Theta_{\theta} dF_a.$$

For the Weibull-gamma-normal MMM, with gamma distributed overdispersion effects as in (4), the connector becomes:

$$\Delta_{ij} = -\log(\alpha\beta) + \tilde{\mathbf{x}}'_{ij} \boldsymbol{\xi}^m - \tilde{\mathbf{z}}'_{ij} D \tilde{\mathbf{z}}_{ij} / 2. \tag{9}$$

3.2 The Proposed Joint Models

We introduce the following notation for a general Weibull model with both conjugate and normal random effects:

$$\omega(t, \lambda, \rho, \theta, \mu, b) = \lambda \rho t^{\rho-1} \theta e^{\mu+b} e^{-\lambda^\rho \theta e^{\mu+b}}.$$

Then, the joint distribution for the continuous and time-to-event outcomes, conditional upon the random effects is:

$$\begin{aligned} f(\mathbf{t}_i, \mathbf{y}_i | \mathbf{b}_i, \boldsymbol{\theta}_i) &= \prod_k \omega(t_{ik}, \lambda, \rho, \theta_{ik}, \mu_{ik} = \tilde{\mathbf{x}}'_{ik} \boldsymbol{\xi}, \tilde{\mathbf{z}}'_{ik} \mathbf{b}_i) \\ &\times \frac{1}{(2\pi)^{\frac{n_i}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}_i - X_i \boldsymbol{\xi} - Z_i \mathbf{b}_i)' \Sigma_i^{-1} (\mathbf{y}_i - X_i \boldsymbol{\xi} - Z_i \mathbf{b}_i)} \end{aligned} \quad (10)$$

Here, $\Sigma_i = \sigma^2 I_{n_i}$, with I_n denoting the identity matrix of dimension n . Then, implementing the MMM requires marginalization over the Weibull model only, given that the linear mixed model contribution trivially marginalizes. This implies that the connector function (9) applies without any problem. Moreover, in the same spirit, one can consider a joint model for two repeated time-to-event sequences. The association is induced by shared normal random effects:

$$\begin{aligned} f(\mathbf{t}_{1i}, \mathbf{t}_{2i} | \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i}, \mathbf{b}_i) &= \prod_j \omega(t_{1ij}, \lambda_1, \rho_1, \theta_{1ij}, \mu_{1ij}, \mathbf{b}_i) \\ &\cdot \prod_k \omega(t_{2ik}, \lambda_2, \rho_2, \theta_{2ik}, \mu_{2ik}, \gamma \mathbf{b}_i). \end{aligned} \quad (11)$$

The $\boldsymbol{\theta}_{1i}$ and $\boldsymbol{\theta}_{2i}$ are assumed to be independent. This process is closely related to the marginalization of a single sequence of repeated time-to-event outcomes, presented above. Also here, connector function (9) is used. Finally, regarding estimation, the fitting method of Molenberghs et al. (2010) is employed. It consists of analytically integrating the marginal form of (10) and (11) over the gamma and numerically over the normal random effects. This result that a standard software, such as the SAS procedure NLMIXED can be used to fit the model.

4 Application

With $\psi(t, \lambda, \rho, \mu, b, \alpha, \beta) = \frac{\lambda \rho t^{\rho-1} e^{\mu+b} \alpha \beta}{(\lambda t^\rho \beta e^{\mu+b} + 1)^{\alpha+1}}$ and $\xi(C, \lambda, \rho, \mu, b, \alpha) = \frac{1}{\left(\frac{\lambda C^\rho e^{\mu+b}}{\alpha} + 1\right)^\alpha}$, respectively the marginal conditional density of the Weibull combined model and its form with allowing right censoring, we fit the model of

$$\begin{aligned} f(y_{ij}, t_{ik} | b_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} [y_{ij} - (\beta_0 + \beta_1 x_i + \beta_2 \tau_{ij} + \beta_3 x_i \tau_{ij} + b_i)]^2} \\ &\cdot \xi(C_{ik}, \lambda, \rho, \mu_{ik}, \gamma b_i, \alpha). \end{aligned}$$

to analyze the chronic heart failure data, and the model of

$$f(t_{1ij}, t_{2ij} | b_i) = \psi(t_{1ij}, \lambda_1, \rho_1, \mu_{1ij}, b_i, \alpha_1) \cdot \psi(t_{2ij}, \lambda_2, \rho_2, \mu_{2ij}, \gamma b_i, \alpha_2).$$

employed for the comet assay data. The analysis results can be seen in Table 1. In the comet analysis, we observe similar point estimates and precision for both the BSS and the BSS-M. There is statistically significant evidence that the dose level has an effect on the hazard of the tail intensity and of the tail moment in both models (conditional and marginal). Direct marginal interpretation is possible. The shared parameter's presence is statistically significant, indicating that the two survival processes are correlated. Also here, the likelihood ratios are similar. Turning attention to the heart failure analysis, a few observations are in place. First, estimates of the two models are similar. Second, there is no statistically significant evidence that heart rhythm has an effect on the evolution of heart rate, both in the joint model and its marginalized one. In contrast, however, there is no significant effect of heart rhythm on the hazard of time-to-hospitalization in the marginalized joint model whereas this is not true in the joint model. This is important and requires careful qualification. Third, the shared estimate is statistically significant, pointing to non-negligible correlation between the continuous and survival processes. Overall, such a result should not be treated as problematic, but rather as resulting from genuine differences in parameter interpretation between the marginal and conditional formulations.

5 Concluding Remarks

Our work builds upon and extends work of Molenberghs et al.(2010), Molenberghs et al. (2012), Njagi et al. (2012), and Heagerty (1999), bringing in additional features e.g. the model can be marginalized in the sense of carrying marginal parametric regression functions that have a population-averaged interpretation; and the time-to-event outcomes are allowed to be right censored. Furthermore, even though the model is relatively complex in the sense that it extends and amends a conventional generalized linear mixed model in various ways, the marginalization using a so-called connector function on the one hand and the numerical technique of partial marginalization, renders the model relatively easy to fit, through standard statistical GLMM software, with minimal additional programming. While focus has been placed on bivariate longitudinal sequences, the methodology could be extended without trouble to more than two outcomes. Additionally, left-censoring and even interval censoring could be considered as well. For conciseness, this has not been made explicit here.

TABLE 1. The Chronic Heart Failure Data (With censoring) and The Comet Data. 'JCS' refers to joint continuous survival model; 'BSS' refers to the bivariate survival model; 'M' and 'Cens' means marginalized and with censoring, respectively.

Par.	JCS-Cens	JCS-Cens-M	Par.	BSS	BSS-M
	Est.(s.e.)	Est.(s.e.)		Est.(s.e.)	Est.(s.e.)
<i>Longitudinal process</i>			<i>The first survival process</i>		
β_0	3.4683(0.2393)	3.6728(0.0852)	ξ_1	-3.3509(0.1109)	-3.3521(0.1109)
β_1	-0.1853(0.3543)	-0.1487(0.1140)	λ_1	2.7610(0.2343)	2.8741(0.2468)
β_2	-0.0003(0.0001)	-0.0004(0.0001)	α_1	9.7570(2.3903)	9.7713(2.4071)
β_3	-0.0003(0.0002)	-0.0002(0.0002)	σ_1^2	0.0773(0.0236)	0.0770(0.0234)
σ^2	0.1530(0.0021)	0.1531(0.0021)			
<i>Survival process</i>			<i>The second survival process</i>		
ξ	-0.1812(0.0438)	-0.3724(0.3233)	ξ_2	-2.4161(0.0911)	-2.4167(0.0910)
λ	0.0018(0.0024)	0.0047(0.0012)	λ_2	0.2351(0.0158)	0.2411(0.0164)
α	10.241(3.5079)	5.1688(6.7443)	α_2	52.870(40.013)	52.871(39.968)
σ_b^2	3.9922(9.1030)	0.1821(0.1413)	σ_2^2	0.0391(0.0142)	0.0393(0.0146)
γ	0.3775(0.1740)	1.1670(0.4422)	γ	0.7958(0.1161)	0.7958(0.1161)
-2LL	11745	11688		19624	19624

References

- Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688–698.
- Molenberghs, G., Verbeke, G., Demetrio C.G.B., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.
- Molenberghs, G., Verbeke, G., Efendi, A., Braekers, R., and Demétrio, C.G.B. (2012). A combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event data. *Submitted for publication*.
- Njagi, E. N., Molenberghs, G., Verbeke, G., and Kenward, M. G. (2012). A flexible joint-modeling framework for longitudinal and time-to-event data with overdispersion. *Submitted for publication*.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Penalized regression on principal manifolds with application to combustion modelling

Jochen Einbeck¹, Benjamin J. Isaac^{2,3}, Ludger Evers⁴,
Alessandro Parente²

¹ Durham University, Department of Mathematical Sciences, England

² Université Libre de Bruxelles, Service d'Aéro-Thermo-Mécanique, Brussels, Belgium

³ University of Utah, Department of Chemical Engineering, United States

⁴ University of Glasgow, School of Mathematics and Statistics, Scotland

E-mail for correspondence: `jochen.einbeck@durham.ac.uk`

Abstract: For multivariate regression problems featuring strong and non-linear dependency patterns between the involved predictors, it is attractive to reduce the dimension of the estimation problem by approximating the predictor space through a principal surface (or manifold). In this work, a new approach for non-parametric regression onto the fitted manifold is provided. The proposed penalized regression technique is applied onto data from a simulated combustion system, and is shown, in this application, to compare well with competing regression routines.

Keywords: Smoothing; principal component analysis; local principal manifolds; combustion model; numerical simulation.

1 Chemical Background

Combustion systems constitute a particular challenge for numerical modelling due to their high-dimensional and non-linear character. Typically, such systems involve a set of variables $\Phi = [T, Z_1, \dots, Z_{n_s-1}]$ where T is the temperature, and $Z_j, j = 1, \dots, n_s - 1$ are the chemical species mass fractions of n_s chemical species. For instance, for simple fuels such as methane, the transport equations form a system of more than 50 highly coupled PDEs, of type

$$\rho \frac{D\Phi}{Dt} = -\nabla \cdot (j_\Phi) + s_\Phi \quad (1)$$

where $\frac{D}{Dt}$ is the material-derivative operator, j_Φ is the mass-diffusive flux of Φ , and s_Φ is the “source term”, that is the volumetric rate of production of Φ . Increasingly complex fuels lead to an increase in the number of chemical species and reactions, and, hence, in the number of coupled PDEs as well as the computational costs. Moreover, large chemical mechanisms are usually

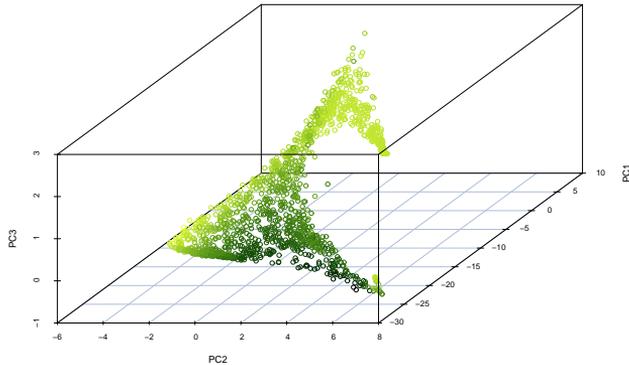


FIGURE 1. PC scores for a combustion system. Dark (green) color corresponds to small values of the first PC source term.

stiff, i.e. a broad range of chemical time-scales exist, thus complicating the numerical simulations including detailed chemistry. Recognizing that the thermodynamic state of a reacting system relaxes onto a low-dimensional, strongly attracting manifold, Sutherland and Parente (2009) suggested the substitution of Φ in (1) by a subset of its principal components, say $\boldsymbol{\eta}$, leaving a more tractable system of 2 or 3 transport equations,

$$\rho \frac{D\boldsymbol{\eta}}{Dt} = -\nabla \cdot (\boldsymbol{j}_{\boldsymbol{\eta}}) + s_{\boldsymbol{\eta}}. \quad (2)$$

However, now the PC source terms $s_{\boldsymbol{\eta}}$ are unknown, and have to be found by regression onto the principal component scores. This tends to lead to unsatisfactory results, due to the nonlinear shape of the manifold. The ability to obtain precise regressions of the source terms is crucial to correctly solve the convection diffusion equation (2) that would describe the variation of the principal component during a numerical simulation. This paper addresses this problem by modelling the state space structure explicitly through local principal manifolds. A novel approach for penalized regression on the manifold surface is provided, and is shown to compare favourably with competing multivariate regression techniques. Of course, the applicability of the proposed method is not restricted to the chemical context considered in here.

2 Data and initial analysis

The data that we had available for this study was a “high-fidelity” data set (i.e., with tabulated source terms) provided by the University of Utah. The data were obtained through an ODT (one-dimensional turbulence) model (Punati et al., 2011) and comprised a total of 11 variables, that is, 10 species mass fractions plus temperature.

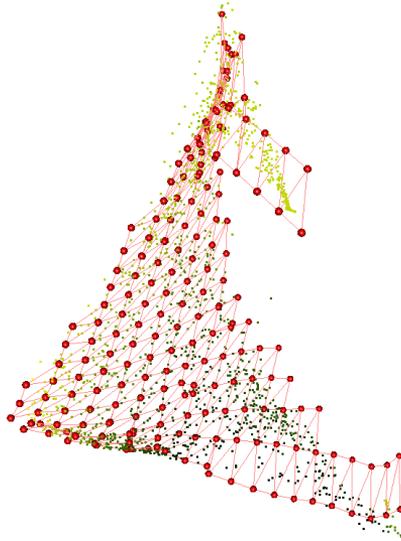


FIGURE 2. Principal surface (red vertices), with (training) data colored according to fitted response (see Sec. 4).

Initially, PCA was applied onto (a scaled version of) a training data set of size $n = 4000$. Fig. 1 provides a plot of the first three principal component scores ($\boldsymbol{\eta}$). Dark (green) colors correspond to low first PC source terms. The manifold structure is evident here, as is the relevance of the position on the manifold for the first principal component source term. We proceeded with approximating the structure by a ‘local principal surface’ (Fig. 2), which effectively approximates the data by a mesh of tiny connected triangles (Einbeck and Evers, 2010).

The next, and most challenging step, is the regression of the PC source terms on the manifold. We explain the necessary methodology in the following section, and return to the combustion problem in Section 4.

3 Penalized regression on principal manifolds

For regression on principal surfaces (i.e., 2D manifolds), Einbeck and Evers (2010) suggested an algorithm which, in step 1, computes average responses within triangles, and in step 2, provides the fitted response in each triangle as the kernel-weighted average over the responses obtained in step 1. Here, we improve this (rather crude) method considerably by fitting piecewise linear functions on each triangle, “glued” together by a second order penalty penalizing differences in the fitted responses at the triangle edges.

More precisely, the regression algorithm starts with projecting the data onto the manifold. Each data point \mathbf{x}_i is projected onto the closest simplex

of the principal manifold (which in our case is a triangle). Denote this simplex by s_i . The projection of \mathbf{x}_i onto this triangle can then be expressed using the sides of the simplex as basis functions. Denote this coordinate vector of the projection of \mathbf{x}_i onto the j -th simplex by $\mathbf{c}^{(j)}(\mathbf{x}_i)$.

The method now assumes different regression models for each simplex, i.e. for simplex j

$$y_i = \mathbf{c}^{(j)}(\mathbf{x}_i)' \boldsymbol{\beta}^{(j)} + \epsilon_i \quad \text{for all } i \text{ such that the closest simplex } s_i = j.$$

Clearly, without additional penalty this model would be too parsimonious: neighbouring simplices would be allowed to have completely different regression functions. Thus a quadratic penalty is introduced which penalises the differences between predictions of neighbouring simplices at shared vertices. Denote by K the set of vertices (with coordinates \mathbf{v}_k) and by S_k the set of all simplices which contain the vertex k . Then the first quadratic penalty is

$$\sum_{k \in K} \sum_{j \in S_k} \left(\hat{y}_k^{(j)} - \bar{y}_k \right)^2,$$

where $\bar{y}_k = \frac{1}{|S_k|} \sum_{j \in S_k} \hat{y}_k^{(j)}$ and $\hat{y}_k^{(j)} = \mathbf{c}^{(j)}(\mathbf{v}_k)' \boldsymbol{\beta}^{(j)}$. This penalty however only shrinks the solution towards a continuous regression function. In order to obtain shrinkage towards a smooth regression function a second penalty is required. This penalty is based on the differences between the regression functions of neighbouring simplices. Define the opposite simplex $o(j, k)$ of simplex j w.r.t. vertex k as the simplex which shares all vertices with j , except k and one further vertex. Then the smoothness penalty can be written as

$$\sum_{k \in K} \sum_{j \in S_k} \left(\hat{y}_k^{(j)} - \hat{y}_k^{(o(j,k))} \right)^2$$

The problem can now be solved using one large penalised regression fit using $\mathbf{Z} = \mathbf{J} \square \mathbf{C}$ as design matrix, with $\mathbf{J}_{ij} = 1$ if $s_i = j$ and $\mathbf{J}_{ij} = 0$ otherwise, and $\mathbf{C} = (\mathbf{c}^{(s_1)}(\mathbf{x}_1)', \dots, \mathbf{c}^{(s_n)}(\mathbf{x}_n)')'$. The symbol \square denotes the row-wise Kronecker product (“box product”), i.e. the i -th row of \mathbf{Z} is the Kronecker

product of the i -th row of \mathbf{J} and the i -th row of \mathbf{C} . Using $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \\ \vdots \end{pmatrix}$

and rewriting the quadratic penalties from above as $\boldsymbol{\beta}' \mathbf{D}' \mathbf{D} \boldsymbol{\beta}$ and $\boldsymbol{\beta}' \mathbf{E}' \mathbf{E} \boldsymbol{\beta}$, the corresponding optimisation problem can be written as

$$\|\mathbf{Z}\boldsymbol{\beta} - \mathbf{y}\|^2 + \lambda \|\mathbf{D}\boldsymbol{\beta}\|^2 + \mu \|\mathbf{E}\boldsymbol{\beta}\|^2.$$

Though the matrices \mathbf{Z} , \mathbf{D} and \mathbf{E} can be very large, they are also very sparse, which allows for quick computations.

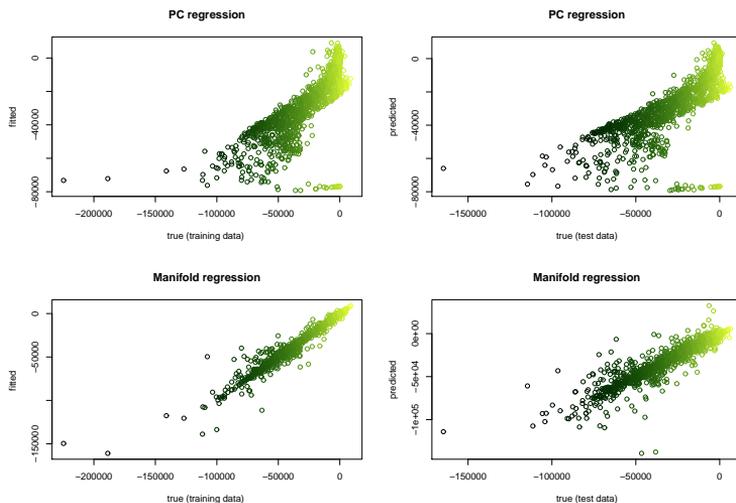


FIGURE 3. True versus fitted (predicted) responses, for training (left) and test data (right); each using linear PC regression (top), and manifold-based regression (bottom). The color scheme is the same as used in Fig. 1.

4 Results

The above technique, using $\lambda = \mu = 10^{-3}$, is now applied onto the system η , with the first component of s_η serving as response, y . The fitted regression output is visualized by color in Fig. 2. A test data set of size 4000 was used to benchmark the performance of this regression technique against competing methods. Fig. 3 compares plots of true versus fitted (predicted) values for manifold-based and linear ‘principal component’ regression (PCR), which indicate that the manifold is able to produce good predictions for both training and test data.

In this study, we also consider the nonparametric additive model (AM), multivariate adaptive regression splines (MARS), and the support vector machine (SVM), each of them using the first three PC scores as predictors. Results are provided in Fig. 4, which also includes a comparison with the localized manifold regression technique proposed by Einbeck & Evers (2010), using smoothing parameter $\lambda = 0.1$.

The clear improvement, in particular of the median prediction error, compared to all other techniques is evident. The two manifold-based regression approaches perform similarly, but the penalized version appears superior since it enables regression *within* the triangle, enabling extremely precise predictions especially at parts of the flame where variability is low. It should be noted that the SVM did actually win the comparison in terms of *mean* (rather than median) prediction error, since it produces less ‘very bad’ predictions, but, in turn, performs (by construction) not very well where

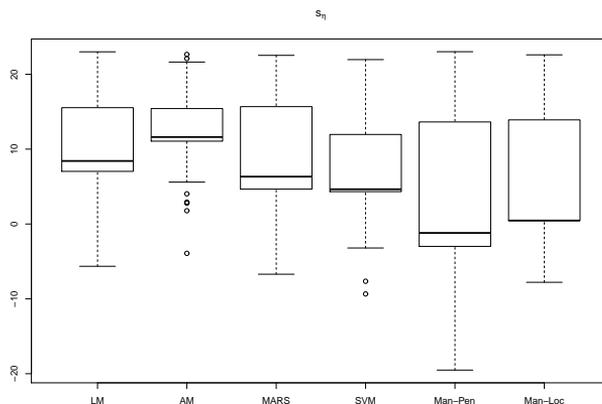


FIGURE 4. Log of squared prediction errors for a test data sample from the combustion data set, using, from left to right, PCR (LM), AM, MARS, SVM, penalized and localized regression on the manifold.

the information is very precise. We investigated this issue further and it appeared that those ‘very bad’ predictions for the penalized manifold regression relate to relatively ‘unimportant’ parts of the flame (burn-in process). Further improvement appears possible by refining the selection of smoothing parameters for the manifold estimation and regression, or by modifying the scaling used in the PCA step. Such issues are currently still under investigation.

Acknowledgments: This research was supported by scoping grant RF 060103 from the Durham Energy Institute.

References

- Einbeck, J. and Evers, L. (2010). Localized regression on principal manifolds. In: *Proceedings of the 25th International Workshop on Statistical Modelling*, Glasgow, pages 179–184.
- Punati, N., Sutherland, J.C., Kerstein, A.R., Hawkes, E.R., and Chen, J.H. (2011). An evaluation of the one-dimensional turbulence model: Comparison with direct numerical simulations of CO/H₂ jets with extinction and reignition, *Proceedings of the Combustion Institute*, **33**.
- Sutherland, J.C. and Parente, A. (2009). Combustion modeling using principal component analysis. *Proceedings of the Combustion Institute*, **32**, 1563–1570.

A dynamic spatio-temporal model to investigate the effect of movements of animals on the spreading of Bluetongue BTV-8 in Belgium

Chellafe Ensoy¹, Christel Faes¹, Marc Aerts¹

¹ I-Biostat, Hasselt University, Belgium

E-mail for correspondence: chellafe.ensoy@uhasselt.be

Abstract: When Bluetongue Virus Serotype 8 (BTV-8) was first detected in Northern Europe in 2006, several guidelines were immediately put into place with the goal to protect farms and stop the spreading of the disease. This however did not prevent further rapid spread of BTV-8 across Northern Europe, which has resulted to substantial economic losses, particularly in the sheep and cattle industry (Wilson and Mellor, 2009).

A better understanding of the BTV-8 transmission is needed to be able to define appropriate control guidelines. Using information on the 2006 Bluetongue outbreak in cattle farms in Belgium, a spatio-temporal transmission model was formulated, similar to the model proposed by Hooten et al (2010) for Influenza in North America. The model quantifies the local transmission of the disease between farms within a municipality, the short-distance transmission between farms across neighboring municipalities and the long-distance transmission as a result of the movement of animals. Different municipal-level covariates (i.e. farm density, land composition variables, temperature and precipitation) were assessed as possibly influencing each component of the transmission process.

The model allows to predict the dynamic spreading of the disease for different scenarios. This is especially useful in investigating the impact of movement (or lack of movement) of animals between farms in the transmission of bluetongue.

Keywords: Bluetongue, Movement, Spatio-temporal model, Dynamic model

1 Introduction

The livestock and poultry industry have been battling for decades the emergence and recurrence of various infectious animal diseases. Bluetongue (BT), which is a non-contagious, insect-borne infectious disease of ruminants, has become one of the most important diseases of livestock especially in Europe where a series of incursions took place, largely under the influence of climate change (Szmaragd et al, 2009).

When BTV-8 was first detected in the Netherlands and subsequently in Belgium, Luxemburg, Germany and France in 2006, several guidelines were immediately put into place: a 150 km surveillance zone was established around the first reported cases and a 20 km standstill zone was set up around the infected farms, within which, all ruminants must be kept inside at night, all movement of live animals on or off farms is prohibited, and the use of insecticide is compulsory in an effort to eradicate the *Culicoides* mites that carry the disease. In the protection and surveillance zones, strict controls must be carried out on all live animals and movement of live ruminants in or out of the zones is banned. This however did not prevent further rapid spread of BTV-8 across Northern Europe, which has resulted to substantial economic losses, particularly in the sheep and cattle industry (Wilson and Mellor, 2009).

To better understand the BTV-8 transmission, information on the 2006 Bluetongue outbreak in Belgium was used. A spatio-temporal transmission model was proposed to quantify the local transmission of the disease between farms within a municipality, the transmission between farms across neighbouring municipalities and transmission as a result of the movement/transport of animals. The model then allows to predict the dynamic spreading of the disease for different scenarios.

2 Materials and Methods

2.1 Data

The 2006 Bluetongue outbreak information for Belgium used in this study was extracted from the data collected by BT-DYNVECT for the bluetongue occurrences in the Northern Europe region. Figure 1 details the observed spatial and temporal trend of the BT outbreak in Belgium where the number of infected farms (farms with at least 1 reported case of infected animal) per municipality were counted. Out of the 40,141 farms in Belgium (spread across 576 municipalities), a total of 582 cases of infected farms were observed coming from 205 different municipalities.

Different covariates deemed as influential for the spread of BT were also investigated, namely:

- Farm Density per municipality (number of farms/sq.km.) and land area
- Land Usage Variables: *Compositional proportion of Pasture, Forest, and Urban areas relative to the Crop area*
- Average Weekly Temperature and Precipitation
- Animal transportation information

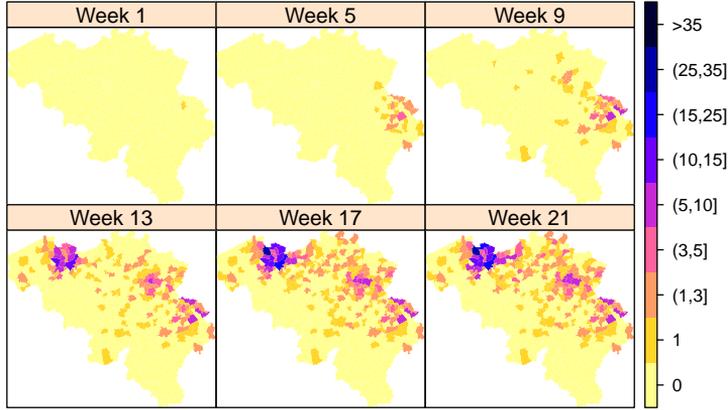


FIGURE 1. *Spatial trend of number of newly infected farms in Belgium for 2006. Week 1 refers to July 16-22, 2006.*

2.2 Infection Model

We model the infectious status of the farm and assumed that once a farm become infected, it is infectious until the winter period. Each farm was classified into either susceptible (no animals infected with BTV-8), infected or infectious class (at least one reported case of infected animal). The SI model for BT is a closed population model, where for a given time t at municipality i , the number of susceptible farms $S_{i,t}$ plus the number of infectious/infected farms $I_{i,t}$ sums up to the total number of farms for each municipality i , N_i . Thus, $S_{i,t} + I_{i,t} = N_i$. The susceptible component could then be rewritten as the difference of the total number of farms and the number of infectious cases, while the number of infectious farms at time point t , is just the sum of the total number of newly infected farms (Y) until time point t and is given by:

$$S_{i,t} = N_i - I_{i,t} \quad (1)$$

$$I_{i,t} = \sum_{k=1}^t Y_{i,k} \quad (2)$$

This number of newly infected farms can then be modeled as a binomial random variable which depends on the number of susceptible farms at the previous time point ($S_{i,t-1}$) and a parameter $\theta_{i,t}$. Thus, $Y_{i,t} \sim \text{bin}(\theta_{i,t}, S_{i,t-1})$

and

$$\text{logit}(\theta_{i,t}) = \begin{cases} (\beta_1 + \mathbf{X}\boldsymbol{\beta}_{int}) + \textit{between} + \textit{move} & \text{if } I_{i,t-1} = 0 \\ (\beta_2 + \mathbf{X}\boldsymbol{\beta}_{int}) + \textit{within} + \textit{between} + \textit{move} & \text{if } I_{i,t-1} > 0 \end{cases} \quad (3)$$

The parameter $\theta_{i,t}$ was formulated as a function of the previous infectious population and is composed of at most four additive terms representing the different transmission scenarios, similar to the method by Hooten et al (2010):

1. The first component of the model ($(\beta_1 + \mathbf{X}\boldsymbol{\beta}_{int})$ and $(\beta_2 + \mathbf{X}\boldsymbol{\beta}_{int})$) gives the background transmission which represents the general risk of infection per municipality. The risk changes depending on whether or not the municipality had an infection at the previous time point.
2. The local or within municipality transmission of BT is given by:

$$\textit{within} = (\mathbf{X}_W\boldsymbol{\beta}_W) I_{i,t-1} \quad (4)$$

3. Between municipality transmission represents the effect of the infectious state of neighbouring municipalities at previous time point ($I_{j,t-1}$), together with some municipal-level covariates. The binary weight $b_{i,j}$, represents the contiguity of municipalities i and j .

$$\textit{between} = \mathbf{X}_B\boldsymbol{\beta}_B \sum_{j=1}^N b_{i,j} \mathbf{I}(I_{j,t-1} > 0) \quad (5)$$

where $\mathbf{I}(\cdot)$ is the indicator function.

4. Transmission through animal transport, where the movement of animals from municipality j to municipality i is quantified through $a_{i,j}$, along with the infection status of municipality j where the movement originated ($I_{j,t-1}$) is given by:

$$\textit{move} = \mathbf{X}_A\boldsymbol{\beta}_A \sum_{j=1}^N a_{i,j} \mathbf{I}(I_{j,t-1} > 0) \quad (6)$$

When $I_{j,t-1} > 0$, random effects can be included in each of the terms to account for extra heterogeneity in the data.

3 Results

Results from model fitting revealed that all investigated covariates were found to influence (on varying degree) the transmission risk. Temperature

and precipitation, most especially affected significantly the risk of BT transmission within and between municipalities. This result was not surprising since other authors (i.e. Purse et al, 2004 and Szymaragd et al, 2009) have already reported these findings. Proportion of pasture, forest, and urban areas relative to the crop are also found to significantly affect the transmission risk, although not so in the within transmission, but more in the between transmission.

Figures 2 and 3 give the deterministic and stochastic prediction results from the fitted model. To compute for the predicted weekly values, observed data from the previous week was used in the deterministic prediction model. In the case of the stochastic prediction, only data until the 7th week of outbreak was used (79 observed cases) and the model was then allowed to predict the rest of the outbreak period. A total of 1000 simulations was done (depicted in gray lines in plots b and c in Figure 2), with the median stochastic prediction given by the black line. The model managed to capture fairly well both the temporal and spatial trend of the infection, with most observed points falling within the 90% bootstrap interval as given in Figure 2.

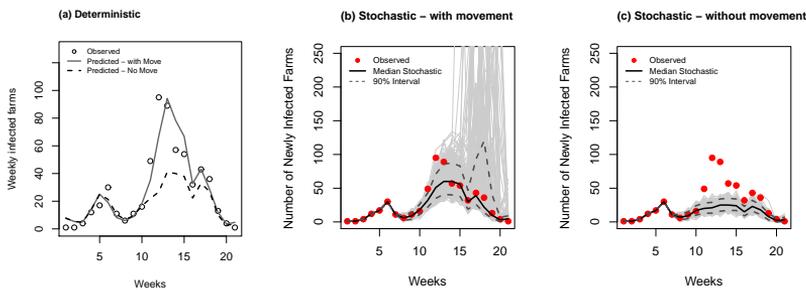


FIGURE 2. *Deterministic and Stochastic weekly number of infected farms with and without animal movements (gray lines are the predictions from 1000 stochastic simulations).*

To investigate the impact of movement of animals in the transmission of bluetongue, Figures 2 and 3 also shows the deterministic and stochastic prediction from the model with movement set to 0. For the stochastic prediction, only data until 7 weeks of the outbreak was used and thus, movement restriction was assumed to start from week 8. We can see in the temporal plots the reduction of the number of newly infected cases, and hence reduction in the number of cumulative infections per municipality when there is movement restriction. Maps of the spread of BT with and without animal movements from week 8 shows fewer cases of predicted BT infection especially in the East and West Flanders region where a high number of

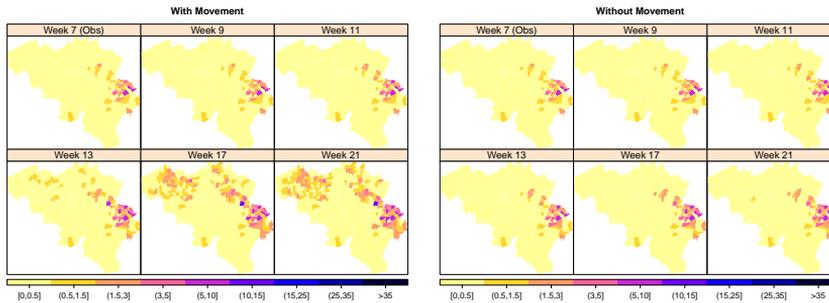


FIGURE 3. Median stochastic prediction of the spatial trend of the cumulative number of infected farms with and without animal movements.

incoming movements took place. This reduction in the predicted number of cases then suggests that animal movements had a significant impact in the spread of BT.

4 Conclusion

The dynamic model developed for the 2006 BT outbreak in Belgium managed to capture the spatial and temporal trend of the infection. By subdividing the model into the different transmission sources, it was shown that the transport of animals played a significant role in the spreading of the disease.

References

- Hooten, M., Anderson, J., and Waller, L. (2010). Assessing North American influenza dynamics with a statistical SIRS model. *Spatial and Spatio-temporal Epidemiology*, **1**, 177–185.
- Purse, B., Baylis M., Tatem, A., Rogers, D., Mellor, P., *et al.* (2004). Predicting the risk of bluetongue through time: climate models of temporal patterns of outbreaks in Israel. *Rev. sci. tech. Off. int. Epiz.*, **23(3)**, 761–775.
- Szmaragd, C., Wilson, A., Carpentre, S., Wood, J., Mellor, P. and Gubbins, S. (2009). A Modeling Framework to Describe the Transmission of Bluetongue Virus within and between Farms in Great Britain. *PLoS ONE*, **4(11)**, e7741.
- Wilson, A. and Mellor, P. (2009). Bluetongue in Europe: past, present and future. *Phil. Trans. R. Soc. B.*, **364**, 2669–2681.

A statistical model for spatial point aggregated data. The geostatistical potential model

Francesco Finazzi¹

¹ Dept. of IT and Mathematical Methods, University of Bergamo, Via Marconi 5, 24044 Dalmine BG, Italy.

E-mail for correspondence: `francesco.finazzi@unibg.it`

Abstract: This paper addresses the problem of estimating a spatially continuous random process from spatial point aggregated data which are supposed to carry information on the random process. Like the spatial point data, the spatial point aggregated data are referenced to precise points in space and like the area aggregated data they refer to sets of statistical units.

Since spatial point aggregated data exhibit interaction, neither the classic geostatistical models nor the classic spatial point process models can be adopted to estimate the realization of the underlying random process. Therefore, an approach based on a novel geostatistical model is developed.

As a case study, the estimation of the spatial market potential of a newspaper over the area of a city is considered. The sales data of the newsstands in the city are used to estimate the spatial market potential which is eventually analyzed in order to identify the areas with the highest market potential.

Keywords: Spatial point aggregated data; geostatistics; geomarketing

1 Introduction

The most commonly studied spatial data can be categorized as follows:

- Geostatistical data (temperature, pollutant concentration, etc.)
- Spatial point data (earthquake locations and magnitudes, location and diameter of the trees in a forest, etc.)
- Area aggregated data (number of poor people in a region, number of car accidents in a region, etc.)

Geostatistical data are measurements, collected at a finite number of locations in space, of an underlying continuous process that has a value at every location in a given spatial domain. The aim is usually to recover the realization of the underlying process as a continuous spatial surface and it

is attained by adopting geostatistical models and kriging techniques (see Diggle and Ribeiro, 2007). Spatial point data, on the other hand, arise as the realization of spatial point processes. In the most interesting cases, the information on the spatial location of the points comes with additional information about one or more features (the marks) of the statistical units at the points, with the convention that each point is related to exactly one statistical unit. The difference between geostatistical data and spatial point data is that the marks are defined only at the spatial location of the points, that is, it does not make sense to estimate the mark value at a new spatial location. Indeed, the aim is usually to study the spatial pattern of the points (see Illian, 2004) and its relationship with the mark values. Finally, area aggregated data are data related to disjoint sets of statistical units within different spatial areas. The main difference with the previous types of data is that area aggregated data are referenced with respect to areas rather than to precise points in space. Area aggregated data are analyzed by considering suitable statistical models such as the conditional autoregression (CAR) and the simultaneous autoregression (SAR) models (see Banerjee et al., 2003 for more details).

A fourth spatial data type worthy of being investigated is that of the *spatial point aggregated data*. Examples of spatial point aggregated data are the sales volume of spatially distributed stores, the number of hospital discharges at the hospitals of a region and the sales volume of a given drug at the chemists of a city. Spatial point aggregated data share properties of both the spatial point data and the area aggregated data. Like the spatial point data they are referenced to precise points in space and like the area aggregated data they refer to sets of statistical units. In order to clarify this aspect note that, in the hospital discharges example, the statistical units are not the hospitals themselves but the discharged patients. Even if each patient could be georeferenced in space, for instance with respect to her/his home address, due to privacy reasons the home address might not be available so that the information comes aggregated with respect to the hospitals.

The peculiarity of the spatial point aggregated data is that they are characterized by interaction phenomena. The sales volume of a given store with respect to a specific product, for instance, is related to the market potential of that product at the spatial location of the store but also to the presence of nearby stores selling the same product. The key aspect is that, given a set of stores, the choice of the customer (for a given product) is mutually exclusive. Like the geostatistical data, the spatial point aggregated data can be considered as measurements of an underlying spatially continuous process but, even when the measurements are conditioned with respect to the process, they are not independent. In this paper, the underlying spatially continuous process is called *potential* in order to stress that the observed spatial point aggregated data are related to a latent spatial random process and the higher the potential the higher the observed value.

When the aim is to estimate the realization of the potential, due to the above mentioned interaction, neither the classic geostatistical models nor the classic spatial point process models can be adopted. In order to solve the problem, a new model based on the geostatistical approach is proposed. As a case study, the model is applied to the estimation of the spatial market potential of a daily newspaper over the area of a city.

2 The geostatistical potential model

In this section, the geostatistical potential model (GPM) is introduced as the main statistical tool for the analysis of spatial point aggregated data. The GPM is described by the following hierarchy of equations:

$$\begin{aligned}
 y(\mathbf{s}; \mathcal{S}) &= h_\xi(u(\mathbf{s}), \mathbf{s}, \mathcal{S}) \\
 u(\mathbf{s}) &= q(\mathbf{s}) + \varepsilon(\mathbf{s}) \\
 q(\mathbf{s}) &= \mu + \mathbf{x}(\mathbf{s})\beta + \gamma w(\mathbf{s})
 \end{aligned}
 \tag{1}$$

At the first stage of (1), $h_\xi : \mathbb{R} \times \mathcal{D} \times \mathbb{S} \rightarrow \mathbb{R}$ is the *interaction function* which is parametrized by the parameter vector ξ . The set \mathbb{S} is the set of all finite spatial point patterns over \mathcal{D} . At the second stage, $\varepsilon(\mathbf{s})$ represents an error component which is assumed to be *i.i.d.* $N(0, \sigma_\varepsilon^2)$ and is supposed to capture both the measuring error and the model error. Finally, at the third stage, the potential $q(\mathbf{s})$ is modeled by three summands, where μ is the mean, $\mathbf{x}(\mathbf{s})$ is a vector of covariates, β is the vector of coefficient, $w(\mathbf{s})$ is a zero-mean latent Gaussian process and γ is a scale parameter. The covariance function of $w(\mathbf{s})$ is $cov(w(\mathbf{s}), w(\mathbf{s}')) = \rho_\theta(\mathbf{s}, \mathbf{s}')$, with $\rho_\theta(\mathbf{s}, \mathbf{s}')$ a valid correlation function parametrized by the vector θ . The model parameter vector is $\Psi = (\mu, \beta', \sigma_\varepsilon^2, \gamma, \theta', \xi')$.

The following family of interaction functions is adopted here:

$$h_\xi(u(\mathbf{s}), \mathbf{s}, \mathcal{S}) = u(\mathbf{s}) \cdot \left(1 + \sum_{\mathbf{s}' \in \mathcal{S}} f_\xi(\mathbf{s}, \mathbf{s}') \right)^{-1}
 \tag{2}$$

where $f_\xi(\mathbf{s}, \mathbf{s}') : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^+$ is a generic non-negative binary function. The function $f_\xi(\mathbf{s}, \mathbf{s}')$ can be any continuous function but, for practical applications, it should be monotonic and such that

$\lim_{\|\mathbf{s}-\mathbf{s}'\| \rightarrow 0} f_\xi(\mathbf{s}, \mathbf{s}') = 1$ and $\lim_{\|\mathbf{s}-\mathbf{s}'\| \rightarrow \infty} f_\xi(\mathbf{s}, \mathbf{s}') = 0$. For instance, $f_\xi(\mathbf{s}, \mathbf{s}') = f_\xi(\|\mathbf{s} - \mathbf{s}'\|) = \exp\left(-\frac{\|\mathbf{s}-\mathbf{s}'\|}{\xi}\right)$ where $\|\cdot\|$ is the Euclidean distance and ϕ is a parameter to be estimated defining the strength of the interaction.

3 Model estimation

Let $\mathbf{y} \equiv \mathbf{y}(\mathcal{S})$ be the $N \times 1$ vector of data collected at the sampling sites $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_N\}$. The measurement equation for the vector \mathbf{y} is

$$\mathbf{y} = \mathbf{G}(\mathbf{1}\mu + \mathbf{X}\beta + \gamma\mathbf{w} + \varepsilon) \quad (3)$$

where $\mathbf{1}$ is the $N \times 1$ vector of ones, $\mathbf{X} \equiv \mathbf{X}(\mathcal{S})$ is the $N \times b$ matrix of covariates, $\mathbf{w} \equiv \mathbf{w}(\mathcal{S})$ is the latent Gaussian process at \mathcal{S} with variance-covariance matrix $\Sigma_{\mathbf{w}} \equiv \Sigma_{\mathbf{w}}(\mathcal{S}, \theta)$ and $\varepsilon \equiv \varepsilon(\mathcal{S})$ is the measurement error at \mathcal{S} with diagonal variance-covariance matrix $\Sigma_{\varepsilon} = \sigma_{\varepsilon}^2 I_N$. Finally, $\mathbf{G} \equiv \mathbf{G}_{\xi}(\mathcal{S})$ is the $N \times N$ diagonal matrix whose diagonal vector is

$$\mathbf{g} = (g_{\xi}(\mathbf{s}_1; \mathcal{S} \setminus \mathbf{s}_1), \dots, g_{\xi}(\mathbf{s}_N; \mathcal{S} \setminus \mathbf{s}_N))$$

The model estimation problem is tackled here following the maximum likelihood (ML) approach. Being $w(\mathbf{s})$ a latent process and due to possible missing data, the expectation-maximization (EM) algorithm is adopted to find the ML estimate of Ψ .

Considering the approach described in Fassò and Finazzi (2011), the following closed form updating formulas have been derived:

$$\hat{\mu}^{(k+1)} = \frac{\text{tr} [(\hat{\mathbf{e}} + \mu^{(k)}\mathbf{1})(\mathbf{1})']}{N} \quad (4)$$

$$\hat{\beta}^{(k+1)} = [(\mathbf{X})' \mathbf{X}]^{-1} (\mathbf{X})' \cdot (\hat{\mathbf{e}} + \mathbf{X}\beta^{(k)}) \quad (5)$$

$$(\hat{\sigma}_{\varepsilon}^2)^{(k+1)} = \frac{1}{N} \text{tr} \left(\hat{\mathbf{e}} \cdot (\hat{\mathbf{e}})' + (\gamma^{(k)})^2 \hat{\mathbf{A}} \right) \quad (6)$$

$$\hat{\gamma}^{(k+1)} = \frac{\text{tr} [(\hat{\mathbf{e}} + \gamma^{(k)}\hat{\mathbf{w}})(\hat{\mathbf{w}})']}{\text{tr} [\hat{\mathbf{w}}(\hat{\mathbf{w}})' + \hat{\mathbf{A}}]} \quad (7)$$

where $\hat{\mathbf{e}} = (\mathbf{G})^{-1} \mathbf{y} - \mu^{(k)}\mathbf{1} - \mathbf{X}\beta^{(k)} - \gamma^{(k)}\hat{\mathbf{w}}$ and where $\hat{\mathbf{w}} = E_{\Psi^{(k)}}(\mathbf{w} | \mathbf{y})$ and $\hat{\mathbf{A}} = \text{Var}_{\Psi^{(k)}}(\mathbf{w} | \mathbf{y})$ are the estimated latent variable and the estimation variance, respectively. The remaining model parameters θ and ξ are updated by numerical optimization as they do not admit a close form formula.

4 Case study

The GPM is applied to estimate the market potential of a daily newspaper from sales data. The data represent the yearly average daily number of copies sold on working days by $N = 70$ newsstands located over the city of Bergamo, northern Italy. The newsstand spatial locations are shown in

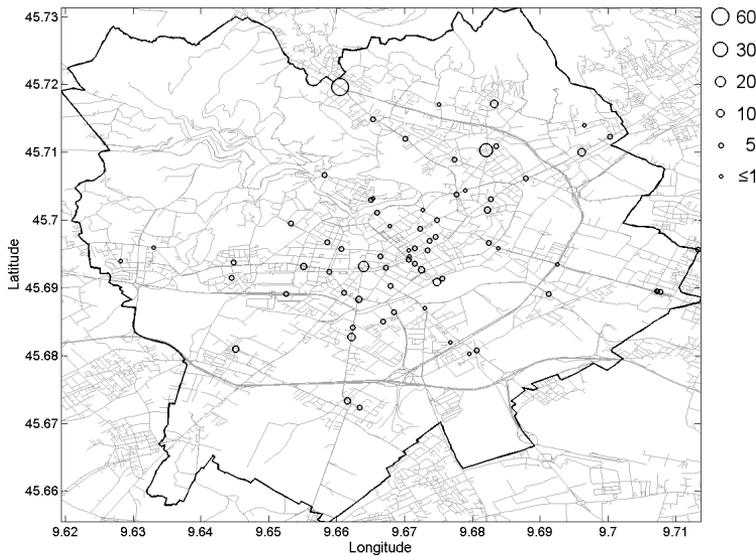


FIGURE 1. Newspaper sales data - yearly average daily number of copies sold on working.

Figure 1, along with the circle-plot of the average daily number of copies sold. For all the intents and purposes, the sales data can be considered as spatial point aggregated data as the newsstands are referenced to precise points in space and the sales volume can be directly related to the single buyers. Moreover, the market potential can be considered as a latent spatially continuous random process as it is well defined for each location in space.

As it is related to the number of copies sold, the population density for the city of Bergamo is considered as a covariate. The population density (not reported here) is normalized to the range $[0, 1]$ and it is available as a continuous spatial surface.

Table 1 reports the estimated model parameters and the respective bootstrap confidence intervals. Note that μ has been assumed to be equal to zero. Since they are not provided as a by-product of the EM algorithm, the confidence intervals has been obtained following the approach in Fassò and Cameletti (2009).

The image of Figure 2 shows the estimated market potential of the newspaper over the city of Bergamo. The maxima of the spatial surface represent the area of the city where it is profitable to have at least a newsstand. Indeed, the estimated market potential at a given location in space can be considered as the number of newspaper copies that would be sold by a newsstand, at that location, without other nearby newsstands.

TABLE 1. Estimated model parameter and 95% bootstrap confidence interval

	$\hat{\beta}$	$\hat{\sigma}_\varepsilon^2$	$\hat{\gamma}$	$\hat{\theta}$	$\hat{\xi}$
estimated	19.58	19.75	12.46	178.86	187.11
LCL	10.43	5.60	12.57	59.94	146.63
UCL	54.19	191.16	33.32	488.11	476.77

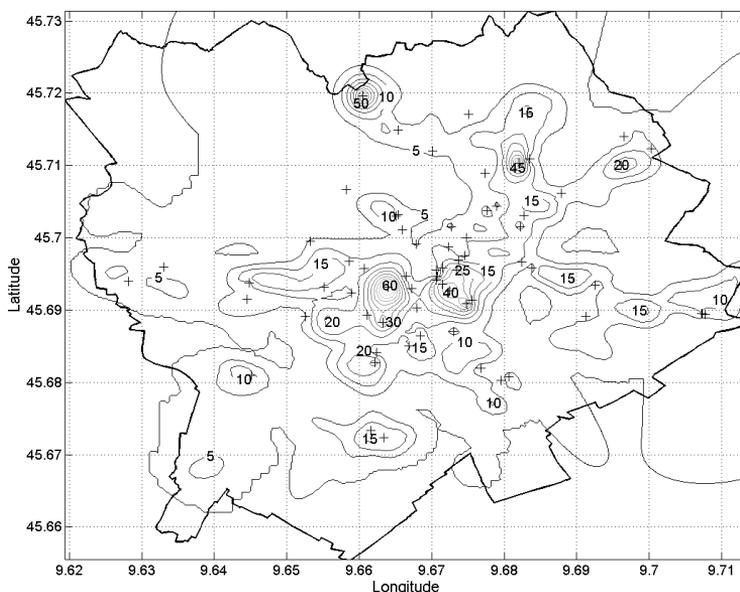


FIGURE 2. Estimated spatial market potential expressed as yearly average daily number of copies.

References

- Banerjee, S., Carlin, B.P. & Gelfand, A.E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC.
- Diggle, P. J. and Ribeiro, P. J. (2007). *Model Based Geostatistics*. Springer, New York.
- Fassò, A. and Cameletti, M. (2009) *A unified statistical approach for simulation, modelling, analysis and mapping of environmental data*. Simulation, 86, 139-154.
- Fassò, A. and Finazzi, F. (2011). *Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data*. Environmetrics 22 735-748
- Illian, J. (2008). *Statistical analysis and modelling of spatial point patterns*. Statistics in practice. John Wiley.

Smoothing parameter selection using the L-curve

Gianluca Frasso¹, Paul H. C. Eilers²

¹ Department of Mathematics and Statistics, Faculty of Economics, University of Naples Federico II, Italy

² Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands

E-mail for correspondence: gianluca.frasso@unina.it

Abstract: The L-curve method has been used to select the penalty parameter in ridge regression. We show that it is also very attractive for smoothing, because of its low computational load. Surprisingly, it also is almost insensitive to serial correlation.

Keywords: Whittaker smoother, P-splines; L-curve; Cross validation, serial correlation.

1 Introduction

P-splines combine a B-spline basis with many knots and a penalty on (higher order) differences of their coefficients (Eilers and Marx, 1996). A special case is the Whittaker smoother for data on a regular grid with a knot for every observation; the basis then is an identity matrix (Eilers, 2003). It is also known as the as the Hodrick-Prescott filter, which has gained popularity in econometrics (Hodrick and Prescott, 1997). There is only one parameter, λ , to tune the weight of the penalty and hence the smoothness of the result.

It is attractive to have an automatic procedure for setting λ . Common choices are leave-one-out cross-validation (LOO-CV) and AIC (Akaike's Information Criterion) or BIC (Bayesian Information Criterion). They all share two drawbacks: 1) they require the computation of the effective model dimension, and 2) they are sensitive to serial correlation in the noise around the trend.

We present an alternative approach, inspired by the L-curve method for ridge regression (Hansen, 1992; Hansen and O'Leary, 1993). This curve is a plot of the logarithm of the magnitude of the penalty term against the logarithm of the sum of squares of the residuals, parameterized by λ . There is no need to compute the effective model dimension, so using the L-curve

makes smoothing of long data series practical. Furthermore this criterion turns out to be robust to correlated noise.

2 The L-curve

The Whittaker smoother solves the following optimization problem:

$$\hat{z} = \underset{z}{\operatorname{argmin}}(\|y - z\|^2 + \lambda\|Dz\|^2) \quad (1)$$

where D is a matrix that forms differences of order 2, and the parameter λ tunes the smoothness of z . This leads to the linear system of equation $(I + \lambda D'D)\hat{z} = y$. It is a banded system and, using sparse matrices, it can be solved quickly: computation time is linear in the number of observations. The effective model dimension is the trace of $(I + \lambda D'D)^{-1}$, which takes a lot of time to compute (proportional to the cube of the number of observations).

We define the following quantities:

$$\psi = \log(\|y - \hat{z}\|^2); \quad \phi = \log(\|D\hat{z}\|^2).$$

Although not shown explicitly here, both ψ and ϕ are functions of λ . Their values are computed for a series of values of λ . If $\lambda = 10^\alpha$, we advise to use a grid for α from 0 to 8, in steps of 0.2. Then $\psi(\lambda)$ and $\phi(\lambda)$ define a parametric curve, which shows more or less an L-shape. Experience has shown that a value of λ in the corner of the L is a good choice. Hansen (1992) showed this to be the case for ridge regression, where the the corner of the L is very sharp, see Figure 2. In our smoothing applications it is less pronounced, but still very useful. Examples are shown in figures 1 and 2. The curvature of the parameterized curve is given by

$$\kappa = \frac{\psi' \phi'' - \psi'' \phi'}{[(\psi')^2 + (\phi')^2]^{3/2}} \quad (2)$$

In the corner κ is largest, so this formula can be used to search for the best λ . A simpler alternative uses the observation that in the corner the points that are used to plot the curve are closer to each other than on the legs of the L curve. So we search for the minimum of the squared step size $(\Delta\psi)^2 + (\Delta\phi)^2$ and take the geometric mean of the λ s at both ends of the step.

Figure 1 shows the performance of the L-curve when smoothing simulated data and it illustrates the two selection strategies. The noise is uncorrelated and here CV as well as AIC leads to essentially the same choice of λ .

Using simulations, we evaluate how the characteristics of the data impact on the shape of the L-curve. This issue is summarized in figure 2. The top-right panel shows the L-curve for a Whittaker smoother applied

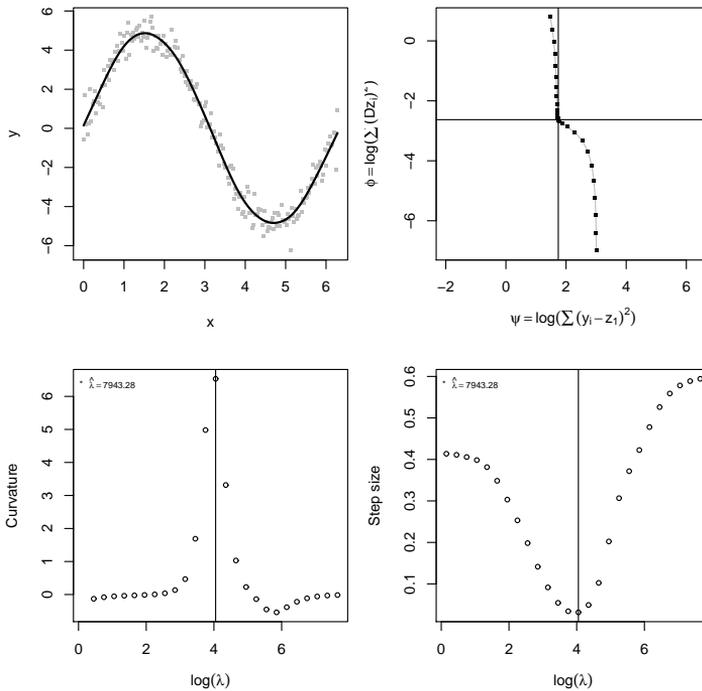


FIGURE 1. Whittaker smoothing of simulated data. The top-left panel shows data (dots) and smooth (line), and the top-right panel the associated L-curve. The lower-left panel shows the curvature and the lower-right the distance between adjacent points of L-curve.

to data simulated using the following scheme: $y = 10^{c_j} \sin(x_i) + N(0, 1)$ with $x_i = 1, \dots, 2\pi$ for $i = 1, \dots, 200$ and $c_j = 2.5, \dots, -2$ for $j = 1, \dots, 7$. It is clear that the convex region tends to disappear when noise gets stronger. Another way of presenting this aspect is shown in the bottom-left panel of figure 2. The results there were obtained using 200 observations, simulated as follows: $y = \sin(x) + N(0, \sigma_j)$ with $x = 1, \dots, 2\pi$ and $\sigma_j \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$.

The L-curve also works if we smooth data with autocorrelated noise. The bottom-left panel of figure 2 shows some results obtained using the Whittaker smoother on a set of data simulated as follows: $y = 3 \sin(x_i) + AR(1, \rho_j, \sigma = 1)$ with $x_i = 1, \dots, 2\pi$ for $i = 1, \dots, 200$ and $\rho_j = 0, \dots, 0.9$ for $j = 1, \dots, 7$ where ρ indicates the autocorrelation coefficient. Stronger correlation reduces the sharpness of the corner. In the next section we show an application to real data with very strong autocorrelation.

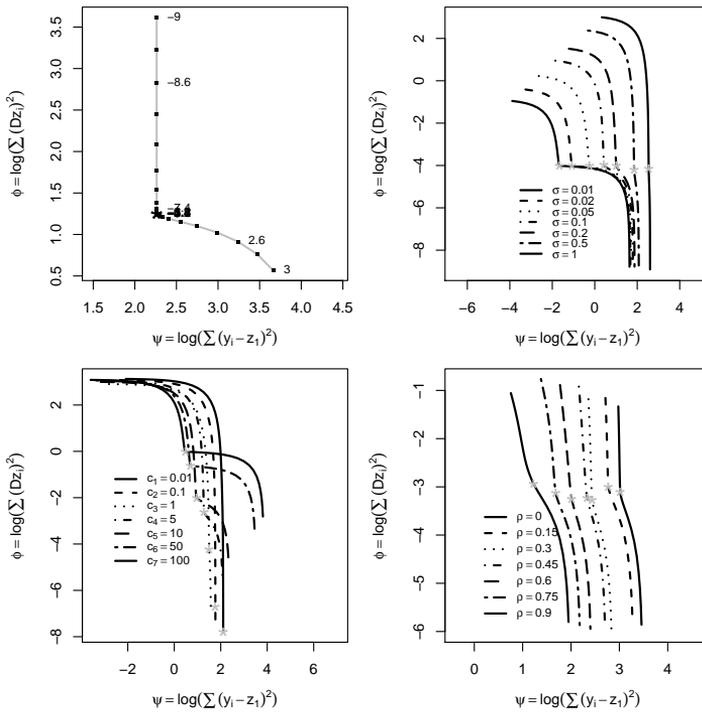


FIGURE 2. The top-left panel shows the L-curve obtained for a simulated ridge regression example. The top-right panel shows seven L-curves obtained for a sine curve, like in Figure 1, to which white noise of varying strength was added. In the bottom-left the strength of the noise is held constant and the amplitude of the sine is varied. The bottom-right panel shows seven L-curves for a Whittaker smoother estimated on data with an increasing autocorrelation of the noise. See the text for details of the simulations

3 A real data example

We use a monthly series for the price of orange juice (corrected for inflation) as an example (Stock and Watson, 2003). Figure 3 shows the results obtained using cross validation and the L-curve selection procedure. CV suggest very light smoothing, following the strong autocorrelation in the data. In contrast the L-curve leads to a reasonable trend.

There is no room to present more results, but we have applied the L curve to other real data (time series of sea levels and a space series of a ground wood surface), with similar pleasing results.

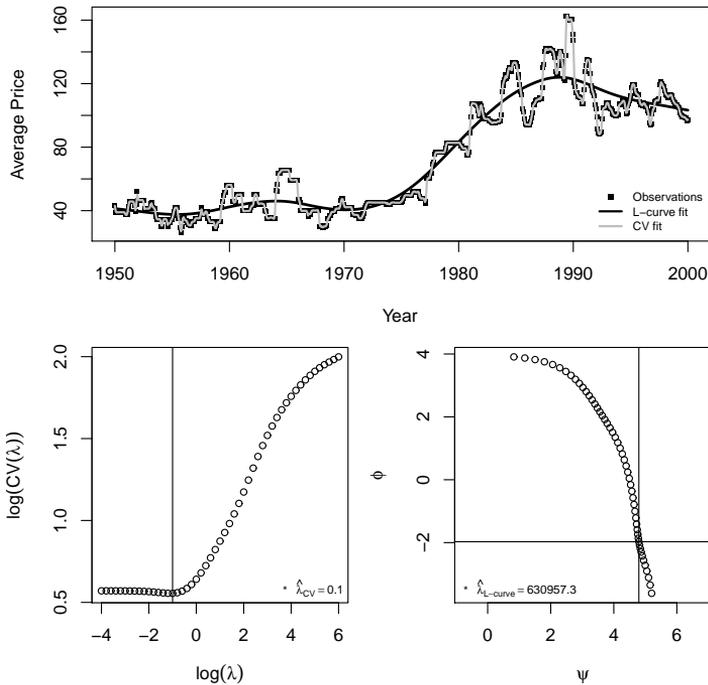


FIGURE 3. The upper panel shows the result obtained selecting the λ parameter of a Whittaker smoother using the cross validation and the L-curve methods. The lower panels show the cross validation and L-curve profiles and indicate the selected smoothing parameters. For both selection methods we considered $\log(\lambda) \in [-4, 6]$.

4 Discussion

A curve of the penalty term versus the residual sum of squares, on logarithmic scales, parameterized by λ , shows an L shape. In the corner of the L we find a good value of λ . This method shows excellent performance in practice, especially when there is strong autocorrelation.

We have no compelling explanations of why the L-curve works so well. Of course, the corner of an L-shaped curve is a special point, but it is not clear why it marks a good choice of smoothing parameter. The relative changes of both the penalty and the size of the residuals are small there, and approximately equal, and apparently that matters. The insensitivity to serial correlation in the noise is also hard to explain.

The L-curve criterion does not give reliable results in some extreme cases. It usually happens when smoothing data that approaches pure white noise on a flat trend.

There are many opportunities for further research. First results for spatial smoothing, generating an “L surface”, look promising. We will also study the applicability to the generalized linear model setting for smoothing of counts and binary data.

References

- Eilers, P.H.C., Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):pp. 115-121.
- Eilers, P.H.C. (2003). A perfect smoother. *Analytical chemistry*, 75(14):3631-3636.
- 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.
- Hodrick, R.J., Prescott, E.C. (1997). Postwar U.S. business cycles: an empirical investigation. *Journal of Money, Credit and Banking*, 29(1):pp. 1-16.
- Stock, J.H., Watson, M.W.(2003) *Introduction to econometrics*. Addison-Wesley.

Detecting aliasing in locally stationary textured images

Aimee N. Gott¹, Idris A. Eckley¹

¹ Lancaster University, UK

E-mail for correspondence: a.gott@lancaster.ac.uk

Abstract: Failing to sample an image at a sufficiently high rate can lead to aliasing, an effect which may distort the second-order properties of an image. Aliasing leads to high frequencies, that are unobservable in the spectral structure due to a low sampling rate, appearing as low frequency contributions to power. From a visual inspection it may not be possible to tell that an image is aliased. We consider how a locally stationary wavelet model may be used to determine whether an image is aliased at any location in the image and show the application of the method to textured data.

Keywords: aliasing; texture analysis; images; locally stationary

1 Introduction

Decreasing the sampling rate of an image is known to decrease the value of the Nyquist frequency, that is the highest observable frequency of an image. This means that we are no longer able to observe the highest frequency power contributions in the spectral structure of an image. Instead these highest frequencies wrap around and appear in the spectrum instead at lower frequencies. This is an effect known as aliasing. This distortion of the spectral structure which results from aliasing may have an adverse effect in several image processing applications, such as classification, which require accurate spectral estimates.

We consider an approach to detecting aliasing that is based on the locally stationary two-dimensional wavelet (LS2W) process model of Eckley et al. (2010). This approach allows us to consider images whose structure may change across locations of the image. As such an image may be aliased in some locations but not others. The approach we consider extends recent work on aliasing with locally stationary time series by Eckley and Nason (2011) to two-dimensions. This extension allows us to determine the presence of aliasing within a single realisation based upon the local wavelet spectrum.

Additionally we consider the application of this method to textured images. As a number of methods in texture analysis require spectral estimates we

wish to know that the estimates we use will be accurate and not corrupted by the effects of aliasing. We show that it is possible to apply the method discussed to textured images and detect the presence of aliasing in such images.

In this paper we give an overview of the LS2W model (section 2) and introduce the approach of Gott and Eckley (2012) for detecting aliasing in locally stationary images. Section 3 considers an application of the approach to textured images.

2 LS2W processes and aliasing

We begin by summarising the locally stationary two-dimensional (LS2W) modelling framework of Eckley et al. (2010) which we will shortly use to consider the effect of aliasing in the wavelet domain. This model allows us to consider images whose second order structure may vary across locations of an image. From Eckley et al. (2010) we give the following definition of an LS2W process.

Definition 1. Let $\mathbf{R} = (R_1, R_2)$ for $R_1 = 2^n$, $R_2 = 2^m$, $n, m \in \mathbb{N}$ and define $\mathbf{r} = (r_1, r_2)$, and $\mathbf{u} = (u_1, u_2)$ for $\mathbf{r}, \mathbf{u} \in \{1, 2, \dots, R_1\} \times \{1, 2, \dots, R_2\}$. Then a locally stationary two-dimensional wavelet (LS2W) process is defined such that,

$$X_{\mathbf{r}; \mathbf{R}} = \sum_{\ell} \sum_{j=1}^{\infty} \sum_{\mathbf{u}} \omega_{j, \mathbf{u}; \mathbf{R}}^{\ell} \psi_{j, \mathbf{u}}^{\ell}(\mathbf{r}) \xi_{j, \mathbf{u}}^{\ell},$$

where the sum over ℓ is the sum over the directions v , h and d .

Within this definition: (i) $\psi_{j, \mathbf{k}}^{\ell}$ are discrete wavelets, (ii) $\xi_{j, \mathbf{u}}^{\ell}$ are a mean zero random orthonormal increment sequence and (iii) $\omega_{j, \mathbf{u}; \mathbf{R}}^{\ell}$ are amplitudes measuring the contribution to $X_{\mathbf{r}}$ at location \mathbf{u} . Additional conditions relating to this definition are stipulated by Eckley et al. (2010). This includes a condition relating to the smoothness of the amplitudes so that they do not vary too much as a function of location, \mathbf{u} i.e. the controls ensure that the amplitudes are slowly varying across locations.

One way in which aliasing can be introduced is by subsampling an image, therefore reducing the sampling rate of the image $X_{\mathbf{r}}$. So let us consider an LS2W process $Y_{\mathbf{s}}$ which is a dyadically subsampled version of $X_{\mathbf{r}}$, that is $Y_{\mathbf{s}} = X_{2^p \mathbf{s}}$.

A measure of the local power (contribution to variance) of an LS2W process is given by the *local wavelet spectrum (LWS)*. We may estimate this in the form of the *local wavelet periodogram (LWP)*, which is given by $I_{j, \mathbf{u}}^{\ell} \equiv |d_{j, \mathbf{u}}^{\ell}|^2 \equiv |\sum_{\mathbf{r}} X_{\mathbf{r}} \psi_{j, \mathbf{u}}^{\ell}(\mathbf{r})|^2$. This however is a biased estimator of the LWS. For the Shannon wavelet, the expectation of the LWP for the LS2W process $X_{\mathbf{r}}$ can be shown to take the form,

$$\mathbb{E}(I_{j, [\mathbf{u}]}^{\ell}) = S_j^{\ell}(\mathbf{u}/\mathbf{R})2^{2j} + O\left\{\frac{1}{\min(R_1, R_2)}\right\}$$

while, as shown by Gott and Eckley (2012), the equivalent (asymptotic) expectation for the potentially aliased image, $Y_{\mathbf{s}}$, for the Shannon wavelet, takes the form,

$$E_{j,\boldsymbol{\alpha}}^\ell = \lim_{R_1, R_2 \rightarrow \infty} \mathbb{E}(I_{j,\boldsymbol{\alpha}}^\ell) = \sum_{\ell_1} \sum_{j_1=1}^p S_{j_1}^{\ell_1} \left(\frac{2^p \boldsymbol{\alpha}}{\mathbf{R}} \right) + S_{p+j}^\ell \left(\frac{2^p \boldsymbol{\alpha}}{\mathbf{R}} \right) \times 2^{2j}.$$

We focus here on the Shannon wavelet (see Chui (1997) for more details) as properties of this wavelet allow us to consider images which are not bandlimited, while use of the Daubechies wavelets (Daubechies, 1992) restrict us to bandlimited images.

It is clear from the above that the spectral estimate for the aliased image includes a bias term which takes the form of the sum of all the unobservable spectral information due to the sampling rate. This artefact takes the same form for all scales/directions of the LWP. We can use this feature to try and estimate and remove the effect of aliasing within the spectrum. This is achieved by considering the expressions for all scales/directions as a system of linear equations. Suppose we define \mathbf{E} to be a vector of the local variance of our image and the LWP estimates for all scales and directions so,

$$\mathbf{E} = \left(\text{Var} \left(\frac{2^p \boldsymbol{\alpha}}{\mathbf{R}} \right), E_{1,\boldsymbol{\alpha}}^v, \dots, E_{J,\boldsymbol{\alpha}}^v, E_{1,\boldsymbol{\alpha}}^h, \dots, E_{J,\boldsymbol{\alpha}}^h, \right. \\ \left. E_{1,\boldsymbol{\alpha}}^d, \dots, E_{J,\boldsymbol{\alpha}}^d \right)^T,$$

and \mathbf{S} to be the aliasing artefact along with the true LWS,

$$\mathbf{S} = \left(S_A = \left(\sum_{\ell_1} \sum_{j_1=1}^p S_{j_1}^{\ell_1} \right), S_{p+1}^v, S_{p+2}^v, \dots, S_{p+J}^v, S_{p+1}^h, \dots, S_{p+J}^h, \right. \\ \left. S_{p+1}^d, \dots, S_{p+J}^d \right)^T.$$

Then we may represent all the expected values as a system of linear equations of the form $\mathbf{E} = \mathbf{D}\mathbf{S}$, where \mathbf{D} is a correction matrix for the Shannon wavelet (see Gott and Eckley, 2012), so that we may obtain both the aliasing artefact and the LWS by taking

$$\mathbf{S} = \mathbf{D}^{-1}\mathbf{E}.$$

The first element of the vector \mathbf{S} is then the estimate of the aliasing artefact, S_A . If the image is not aliased then this will take the value zero for all locations of the image whereas where aliasing is present it will take some non-zero value.

We may determine whether a given image is aliased by implementation of a bootstrap test. Here, we take as our null hypothesis the assumption that the image is not aliased and as our test statistic the value of the aliasing artefact, S_A . By repeated simulation of images with an equivalent, non-aliased, second order structure, we may determine whether the value of the observed aliasing artefact is statistically significant.

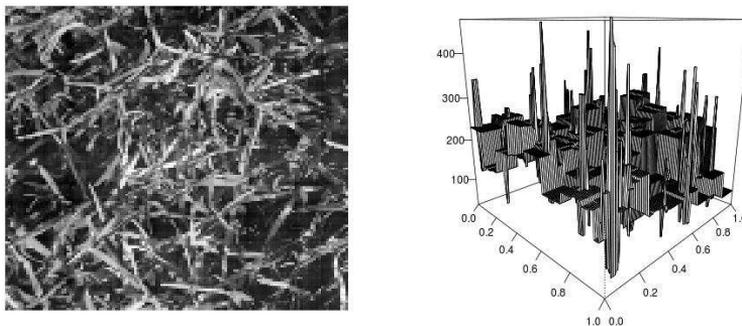


FIGURE 1. Texture image of grass (left) and aliasing estimate for texture image (right).

A similar form of the spectral structure described is also taken by white noise processes. That is, the wavelet spectrum of a white noise process takes some constant value at all scales/directions. As such a white noise process would also have power appearing in the aliasing artefact, S_A . Therefore in the hypothesis test described above we may only determine that an image is aliased or contains white noise.

3 Detecting aliasing in textured images

We now consider the application of the methods to a textured image shown in Figure 1 (left). As the assumptions of the LS2W process include that our image is zero mean we must first ensure that this holds for our textured image. As such we initially apply de-trending methods; here we use the method of Tukey (1977).

We may first of all estimate the aliasing artefact, S_A , for the image as discussed above. This estimate, for all locations of the image, is shown in the right hand plot of Figure 1. It is clear that most of the values for the aliasing artefact are approximately 150 to 200, with large spikes appearing across the image. From this plot alone it would appear that the image contains some form of aliasing or white noise.

We may apply the bootstrap test described to obtain a p-value for the aliasing of this image. In this case, with 200 bootstrap replicates, we obtain a p-value of zero. This is strong evidence that in this case the image is aliased or contains white noise, as would be expected based on the plot in Figure 1.

References

- Chui, C. K. (1997). *Wavelets: A Mathematical Tool for Signal Analysis*. SIAM, Philadelphia.
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.
- Eckley, I. A. and Nason, G. (2011). Alias detection and spectral correction for locally stationary time series. (*Under Revision*)
- Eckley, I. A., Nason, G., and Treloar, R. (2010). Locally stationary wavelet fields with application to the modelling and analysis of image texture. *Journal of the Royal Statistical Society (Series C)*, **59**(4):595–616.
- Gott, A.N. and Eckley, I.A. (2012) The detection of aliasing in images using locally stationary two dimensional wavelet processes. (*In Preparation*)
- Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley.

A general subhazard frailty model for multi-center competing risks data

Il Do Ha¹, Nicholas J. Christian², Jong-Hyeon Jeong³, Youngjo Lee⁴

¹ Department of Asset Management, Daegu Haany University, South Korea

² Department of Biostatistics, University of Pittsburgh, USA

³ Department of Biostatistics, University of Pittsburgh, USA

⁴ Department of Statistics, Seoul National University, South Korea

E-mail for correspondence: `idha@dhu.ac.kr`

Abstract: It is an important issue to investigate the potential heterogeneity in time-to-event (e.g. competing-risk event) between centers in multicenter randomized clinical trials because the treatment effect as well as the baseline risk may be changed over centers despite the use of standardized protocols. We propose a general subhazard-frailty modelling approach for investigating such heterogeneity in competing-risk data from multicenter trials, using hierarchical likelihood. The proposed method is illustrated using data from the B14 phase III breast cancer trials conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP), which consist of 2,817 patients from 167 centers.

Keywords: Competing risks; Frailty models; Hierarchical likelihood; Random treatment-by-center interaction; Subhazard distribution.

1 Introduction

In general, competing risks (CR) data arise when an individual can experience more than one type of event. That is, CR events are defined as events different from the event of interest, so that they hinder the observation of event of interest (Pintilie, 2006). For example, when a breast cancer patient dies due to causes unrelated to the disease, the death becomes a CR event. In the presence of CRs, the use of standard survival methods (e.g. Kaplan-Meier estimate or Cox's PH model) where the CR is treated as a censoring can lead to biased results (Putter et al., 2007). Thus, two broad classes of models for analyzing CR data have been developed under PH assumption; one is to model the cause-specific hazard of the different event types (Prentice et al., 1978), and the other is to model the subhazard (i.e. hazard function of subdistribution) for the event of interest (Fine and Gray, 1999). CR events can be often occurred within a cluster (e.g. center). In many applications involving CRs, individual events may be correlated within clusters,

due to unobserved shared factors across individuals. In this paper we focus on the CR events from multi-center trials. In CR setting the subhazard model by Fine and Gray (1999) directly describes the effect of a covariate on the probability of a particular cause of event, but it ignores correlation between event times within a center. For this Katsahian et al. (2006) and Christian (2011) have extended Fine-Gray model to a subhazard frailty model allowing one random-center effect. We are also interested in investigating the potential heterogeneity in outcomes between centers; for example, treatment effect as well as baseline risk (center effect) may be changed over centers, which can be treated as random effects. Thus, in this paper we extend a standard-frailty modelling approach (Ha, Sylvester, Legrand and MacKenzie, 2011) with various random components to a general subhazard-frailty modelling approach for investigating such heterogeneity in CR data from multicenter trials. For inference we develop the hierarchical likelihood (or h-likelihood; Lee and Nelder, 1996) approach, which obviates the need for an intractable integration over the frailties. The proposed method is illustrated using a practical data set. We also demonstrate model selection and show how to investigate heterogeneity over centers using prediction intervals for frailties of the individual centers.

2 A motivated data set

We re-examine the data from the B-14 randomized multicenter breast cancer trial conducted by the NSABP (Fisher et al., 1989, 1996). The 2817 eligible patients from 167 distinct centers were followed up for five years since randomization or until the first treatment failure, whichever came first. The number of patients per center varied from 1 to 241, with a mean of 16.9 and median of 8. The patients were randomized to one of two treatment arms, tamoxifen (1413 patients) or placebo (1404 patients). The average age of patients was 55 and the average tumor size was about 2 centimeters. The aim of this analysis is to investigate the effect of treatment on local or regional recurrence. For this we consider two event types. The first type is local or regional recurrence (Type I) and the second type is a new primary cancer, distance recurrence or death (Type II); only the event that occurs firstly is of interest in this analysis, so that the repeated event times are not considered. Here, Type I is an event of interest (314 patients; 11.15%), Type II is an event of CR (1303 patients; 46.25%), and no-events is censoring (1200 patients; 42.60%). Figure 1 presents the estimated cumulative incidence functions (CIFs) (Pintilie, 2006) for the two treatment arms. The tamoxifen group has lower CIFs compared to placebo group for both type I and type II. For type I the difference of CIFs of two arms seems to be large, whereas for type II it does not.

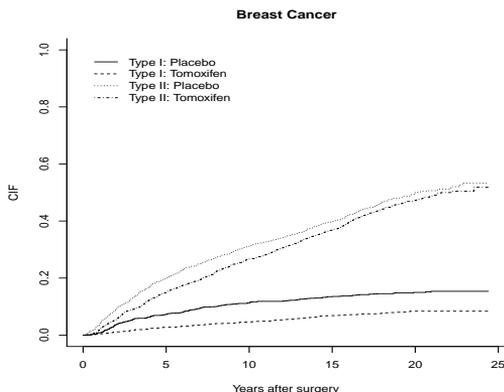


FIGURE 1. Estimated CIF for two types of events in breast cancer data.

3 The proposed model

Suppose that the data consist of censored time-to-event observations collected from q centers. We also assume that there are L distinct event types in each center. Let T_{ij} ($i = 1, \dots, q, j = 1, \dots, n_i, n = \sum_i n_i$) be the survival time for the j th observation in the i th center and let C_{ij} be the corresponding censoring time. Then observable data become $y_{ij} = T_{ij} \wedge C_{ij}$ and ϵ_{ij} . Here ϵ_{ij} takes a value from the set $\{1, 2, \dots, L\}$ representing causes of event, with the convention that $\epsilon_{ij} = 0$ under censoring. The CIF of event from cause $\epsilon_{ij} = 1$ is defined by $F_1(t) = Pr(T_{ij} \leq t, \epsilon_{ij} = 1)$, which represents the probability that an individual will experience an event of type 1 by time t . The corresponding hazard function of subdistribution (subhazard function) is also defined by

$$\begin{aligned} \lambda_1^s(t) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P\{t \leq T_{ij} \leq t + \Delta t, \epsilon_{ij} = 1 | T_{ij} \geq t \cup (T_{ij} < t \cap \epsilon_{ij} \neq 1)\} \\ &= -d \log\{1 - F_1(t)\} / dt. \end{aligned}$$

Following the motivated data, we consider the two event types ($L = 1, 2$). Thus, ϵ_{ij} has 0, 1 or 2; it is 1 for an event of interest and 2 for an event of CR. Below we propose a general subhazard frailty models allowing various random components (e.g. random center or random treatment effect) and their correlation, as in Ha et al. (2011). Denote by v_i an r -dimensional vector of unobserved log-frailties (random effects) associated with the i th center. Given v_i , the conditional subhazard function of T_{ij} is of the form

$$\lambda_{ij1}^s(t|v_i) = \lambda_{01}^s(t) \exp(\eta_{ij}), \tag{1}$$

where $\lambda_{01}^s(\cdot)$ is the unknown baseline subhazard function, $\eta_{ij} = x_{ij}^T \beta + z_{ij}^T v_i$ is the linear predictor for the log-hazard, and $x_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ and

$z_{ij} = (z_{ij1}, \dots, z_{ijr})^T$ are $p \times 1$ and $r \times 1$ covariate vectors corresponding to fixed effects $\beta = (\beta_1, \dots, \beta_p)^T$ and log-frailties v_i , respectively. We assume that the log-frailties v_i are independent and follow a multivariate normal distribution, $v_i \sim N_r(0, \Sigma)$. Here, the covariance matrix $\Sigma = \Sigma(\theta)$ depends on a vector of unknown parameters θ . The normal distribution has been used for modelling multi-component and correlated frailties.

4 Estimation procedure

We now how to derive the h-likelihood estimation procedure for fitting the proposed model (1). Since the functional form of $\lambda_0^s(t)$ is unknown, following Breslow (1972), we approximate the baseline cumulative subhazard function $\Lambda_{01}^s(t) = \int_0^t \lambda_{01}^s(u)du$ by a step function with jumps at the observed event times; $\Lambda_{01}^s(t) = \sum_{k: y_{(k)} \leq t} \lambda_{01k}^s$, where $y_{(k)}$ is the k th ($k = 1, \dots, D$) smallest distinct event time among y_{ij} 's, and $\lambda_{01k}^s = \lambda_{01}^s(y_{(k)})$. Let $\delta_{ij} = I(\epsilon_{ij} = 1)$ be event indicator. Following Ha et al. (2001) and Katsahian et al. (2006), the h-likelihood for subhazard frailty models (1) is defined by

$$h = h(\beta, v, \lambda_{01}^s, \theta) = \sum_{ij} \ell_{1ij} + \sum_i \ell_{2i}, \tag{2}$$

where $\sum_{ij} \ell_{1ij} = \sum_k d_{(k)} \log \lambda_{01k}^s + \sum_{ij} \delta_{ij} \eta_{ij} - \sum_k \lambda_{01k}^s \{ \sum_{(i,j) \in R_{(k)}} e^{\eta_{ij}} \}$, $\ell_{1ij} = \ell_{1ij}(\beta, \lambda_{01}^s; y_{ij}, \delta_{ij} | v_i)$ is the logarithm of the conditional density function for y_{ij} and δ_{ij} given v_i , $\ell_{2i} = \ell_{2i}(\theta; v_i) = -\frac{1}{2} [\log \det \{ 2\pi \Sigma_i(\theta) \}] - \frac{1}{2} v_i^T \Sigma_i(\theta)^{-1} v_i$ is that of the density function for v_i with parameters θ , and $\eta_{ij} = x_{ij}^T \beta + z_{ij}^T v_i$. Here, $\beta = (\beta_1, \dots, \beta_p)^T$, $v = (v_1^T, \dots, v_q^T)^T$, $\lambda_{01}^s = (\lambda_{011}^s, \dots, \lambda_{01D}^s)^T$, $d_{(k)}$ is the number of events of interest at $y_{(k)}$ and

$$R_{(k)} = R(y_{(k)}) = \{ (i, j) : y_{ij} \geq y_{(k)} \text{ or } (y_{ij} \leq y_{(k)} \text{ and } \epsilon_{ij} = 2) \} \tag{3}$$

is the risk set at $y_{(k)}$. Note that as compared to standard Cox model, the risk set $R_{(k)}$ comprises individuals who have not failed from any cause by $y_{(k)}$ but also those who have previously failed from competing causes. As the number of λ_{01k}^s 's can increase with the number of events, the function $\lambda_{01}^s(t)$ is potentially of high dimension. Accordingly, for estimation of (β, v) Ha et al. (2001) proposed the use of the profiled h-likelihood h^* from which λ_{01}^s is eliminated: $h^* = h|_{\lambda_{01}^s = \hat{\lambda}_{01}^s} = \sum_{ij} \ell_{1ij}^* + \sum_i \ell_{2i}$, where $\hat{\lambda}_{01k}^s(\beta, v)$ are solutions of the estimating equations, $\partial h / \partial \lambda_{01k}^s = 0$, for $k = 1, \dots, D$. Note here that $\sum_{ij} \ell_{1ij}^* = \sum_{ij} \ell_{1ij} |_{\lambda_{01}^s}$ does not depend on λ_{01}^s . Along the lines of Fine and Gray (1999) and Katsahian et al. (2006), the first term, $\sum_{ij} \ell_{1ij}^*$, of h^* is slightly modified using inverse probability of censoring weighting: $\ell_1^* \equiv \sum_{ij} \ell_{1ij}^* = \sum_{ij} \delta_{ij} \eta_{ij} - \sum_k d_{(k)} \log \{ \sum_{(i,j) \in R_{(k)}} w_{ij} \exp(\eta_{ij}) \}$ with a constant term eliminated, where

$$w_{ij} = w_{ij}(y_{(k)}) = I(y_{ij} \geq y_{(k)} \text{ or } \epsilon_{ij} = 2) \left\{ \frac{\hat{G}(y_{(k)})}{\hat{G}(y_{ij} \wedge y_{(k)})} \right\} \tag{4}$$

TABLE 1. Results for four subhazard models to Type I event of the data.

Model	Trt	Age	Tumor	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_{01}$ [$\hat{\rho}$]
	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\beta}_3$ (SE)			
Fine-Gray	-0.667 (0.119)	-0.026 (0.005)	0.082 (0.042)	—	—	—
Center	-0.672 (0.119)	-0.026 (0.005)	0.081 (0.042)	0.043	—	—
Indep	-0.704 (0.129)	-0.026 (0.005)	0.080 (0.043)	0.033	0.118	—
Corr	-0.658 (0.137)	-0.026 (0.005)	0.079 (0.043)	0.091	0.249	-0.108 [-0.721]

is the weight of center i at $y_{(k)}$, and $\hat{G}(\cdot)$ is the Kaplan-Meier estimate of the survival function for censoring times. Then Ha et al.'s (2011) procedure for standard frailty models without CRs can be straightforwardly extended to the subhazard frailty models (1), with an adjusted profile h-likelihood $p_{\beta,v}(h^*)$ for θ . Further quantities (e.g. prediction intervals of frailties) are directly applied. The data-directed simulation results have demonstrated that our procedure performs well (not shown).

5 Results

For the breast cancer data we consider the three covariates of interest: treatment (x_{ij1} is 1 for tamoxifen and 0 for placebo), tumor size (x_{ij2}) and age (x_{ij3}). Let v_{i0} and v_{i1} be random center effects and random treatment effects (i.e. random treatment-by-center interaction), respectively. Following Ha et al. (2011), we consider the four submodels of (1) for the Type I event, $\lambda_{1ij}^s(t|v) = \lambda_{01}^s(t) \exp(\eta_{ij})$ with η_{ij} allowing several frailty structures in models M2-M4; Here $(v_{i0}, v_{i1}) \sim BN$ means that $v_{i0} \sim N(0, \sigma_0^2)$, $v_{i1} \sim N(0, \sigma_1^2)$ and $\rho = \text{Corr}(v_{i0}, v_{i1})$; $(v_{i0}, v_{i1}) \sim IN$ also means BN with $\rho = 0$:

M1 (Fine-Gray): $\eta_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}$;

M2 (Center): $\eta_{ij} = v_{i0} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}$, with $v_{i0} \sim N(0, \sigma_0^2)$;

M3 (Indep): $\eta_{ij} = v_{i0} + (\beta_1 + v_{i1}) x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}$, with $(v_{i0}, v_{i1}) \sim IN$;

M4 (Corr): $\eta_{ij} = v_{i0} + (\beta_1 + v_{i1}) x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}$, with $(v_{i0}, v_{i1}) \sim BN$,

where ‘Center’ denotes the random center effect. Here M4 is our full model and the others are various simplifications of it by assuming null components, i.e. M1 ($v_{i0} = 0, v_{i1} = 0$), M2 ($v_{i1} = 0$) and M3 ($\rho = 0$). The models are fitted using SAS/IML and the results are summarized in Table 1. In all the four subhazard models the two fixed effects ($\beta_j, j = 1, 2$) are significant, except for β_3 . In particular, the use of tamoxifen significantly

reduces the risk of local or regional recurrence (type I event) as compared to patients who do not receive placebo. We also observe that overall, there are no substantial changes in the fixed-effects estimates, although the effect of main treatment (β_1) becomes slightly weaker due to the increased standard error when the two random components and/or their correlation are included as in M3 and M4. In M2-M4, the variances (σ_0^2 and σ_1^2) indicate the amount of variation between centers in baseline risk (i.e. center effect) and in the treatment effect, respectively. Here, the estimate of σ_1^2 is relatively larger than that of σ_0^2 . Furthermore, M4 explains the degree of dependency between two random components (i.e. the random center effect v_0 and the random treatment-by-center interaction v_1). The estimate of ρ ($\hat{\rho} = -0.721$) gives a negative value, indicating that the two predicted random components (\hat{v}_0 and \hat{v}_1) have a negative correlation. In particular, the estimate of β_1 in M4 is negative; we see that a decreasing value of v_{i1} corresponds to an increased treatment effect. Thus, the negative correlation leads to the conclusion that treatment confers more benefit in centers with a higher baseline risk. Though not reported here, 95% prediction intervals (Ha et al., 2011) for frailties of individual centers using a model M4 show overall homogeneity in treatment effect as well as baseline risk across 167 centers, leading that the treatment is shown to be effective. In addition, the model-selection criterion (Ha, Lee and MacKenzie, 2007) based on h-likelihood, $\text{hAIC} = -2p_{\beta,v}(h^*) + 2d$ where d is the number of dispersion parameters, selects M1 as a final model among models considered, which also confirms the homogeneity above.

Acknowledgments: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2009-0088978 and No. 2010-0021165).

References

- Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**, 496-509.
- Ha, I. D., Lee, Y., and Song, J.-K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233-243.
- Ha, I.D., Sylvester, R., Legrand, C., and MacKenzie, G. (2011). Frailty modelling for survival data from multi-centre clinical trials. *Statistics in Medicine* **30**, 2144-2159.
- Lee, Y., and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619-678.

Clustering in linear mixed models with Dirichlet process mixtures using EM algorithm

Felix Heinzl¹, Gerhard Tutz¹

¹ Department of Statistics, LMU Munich, Germany

E-mail for correspondence: `felix.heinzl@stat.uni-muenchen.de`

Abstract: In linear mixed models the assumption of normally distributed random effects is often inappropriate and unnecessary restrictive. The proposed Dirichlet process mixture assumes a hierarchical Gaussian mixture. In addition to the weakening of distributions assumptions the specification allows to estimate clusters of observations with a similar random effects structure identified. An Expectation-Maximization algorithm is given that solves the estimation problem and that exhibits advantages over in this framework usually used Markov chain Monte Carlo approaches. The method is evaluated in a simulation study and applied to lung function growth data.

Keywords: Dirichlet process mixture; mixed models; likelihood inference; EM algorithm.

1 Introduction

Linear mixed models (LMM) are a common tool for the modeling of longitudinal data. The classical model has the form

$$y_{ij}|b_i \stackrel{ind.}{\sim} N(x_{ij}^T\beta + z_{ij}^T b_i, \sigma^2), \quad i = 1, \dots, n, \quad j = 1, \dots, n_i, \quad (1)$$

where y_{ij} denotes the response observed for subject i at observation times t_{ij} with $t_{i1} < \dots < t_{ij} < \dots < t_{in_i}$. Population effects of covariates x_{ij} are collected in the parameter vector β whereas individual-specific effects of covariates z_{ij} are represented in the parameter vector b_i . The classical assumption in (1) is a Gaussian distribution for the random effects, i.e. b_i i.i.d. $N(0, D)$, see for example Verbeke and Molenberghs (2000) and Ruppert et al. (2003). While this choice is mathematically convenient, in applications it is often questionable for several reasons. The normal distribution is symmetric, unimodal and has light tails. Since the distributional assumption is made on unobserved quantities, it is typically hard to validate these properties based on estimates. Possible skewness and multimodality (arising, for example, from an unconsidered grouping structure in the data) may

be masked when checking the normal distribution in terms of estimated random effects. A finite mixture of normal distributions as a random effects distribution suggested, for example, by Verbeke and Lesaffre (1996), Verbeke and Molenberghs (2000), and Grün (2008) is much more flexible. One assumes

$$b_i \sim \sum_{h=1}^N \pi_h N(\mu_h, D), \quad (2)$$

where π_1, \dots, π_N are mixture weights. The number of mixture components is unknown and has to be chosen. A data driven choice of this number is desirable and could be achieved by a penalization of the mixture weights π_h . For example, Komárek and Lesaffre (2008) penalized differences between reparametrized weights. In contrast, Magder and Zeger (1996) used component specific covariance matrices subject to the constraint that their determinants are greater than or equal to some minimum value. Now we present a new penalization approach. The basic concept is to shrink the weights π_h towards zero in order to reduce the number of clusters. Therefore we consider a Dirichlet process mixture (DPM) for the random effects distribution and use the stick breaking procedure of the Dirichlet process (see Ferguson, 1973, for the theory behind the Dirichlet process and Sethuraman, 1994, for the stick breaking presentation of the Dirichlet process). The main advantage of Dirichlet processes is the cluster property: by using a DPM for the random effects distribution we obtain automatically a clustering of individuals. Under the assumption that the population can be described by few clusters we want to identify and interpret them. Since a Dirichlet process allows to specify a prior on probability measures, it has been mainly used in the Bayesian inference for density estimation and random effects models. For linear mixed models, Dirichlet process priors for random effects were first proposed by Kleinman and Ibrahim (1998).

We aim at establishing the Dirichlet process as a tool for frequentist modeling. Therefore, instead of using Markov chain Monte Carlo (MCMC) methods, which are usually applied for estimation in random effects models with Dirichlet processes (compare for example Heinzl et al., 2012), we extend the traditional Expectation-Maximization (EM) algorithm (Dempster et al., 1977) used in the heterogeneity model of Verbeke and Molenberghs (2000) and call it DPM-EM model. We will show that the EM algorithm has an essential advantage over MCMC methods, where Dirichlet processes are concerned. In summary, on the one hand our DPM-EM model is a regularization approach for the number of mixture components in (2). On the other hand our model is a method to obtain clustering of individuals in longitudinal data.

2 Model hierarchy

Collecting observations y_{ij} , $j = 1, \dots, n_i$, for individual i in the vector y_i , model (1) can be written in matrix notation as

$$y_i | b_i \stackrel{ind.}{\sim} N(X_i \beta + Z_i b_i, \sigma^2 I) \quad i = 1, \dots, n,$$

where I is the identity matrix and X_i and Z_i denote the individual design matrices constructed from covariates x_{ij} and z_{ij} , respectively. For the random effects distribution, we assume a hierarchical Gaussian mixture

$$\begin{aligned} b_i | \theta_i &\stackrel{ind.}{\sim} N(\theta_i, D), & i = 1, \dots, n, \\ \theta_i &\stackrel{i.i.d.}{\sim} G, & i = 1, \dots, n, \\ G &\sim DP(\alpha, G_0). \end{aligned} \quad (3)$$

Here, the Dirichlet process $DP(\alpha, G_0)$ is a distributional assumption for the unknown mixing distribution G . A special feature of the Dirichlet process is, that each realization of G is a discrete probability measure. So in the DPM specification, choosing a Dirichlet process for the θ_i , $i = 1, \dots, n$, creates ties among these and therefore forms clusters of subjects whereas each subject still has its own unique random effects value. In general, there are $k \leq n$ clusters and $\theta_1, \dots, \theta_n$ can be represented by cluster locations μ_1, \dots, μ_k and cluster allocation variables. Although in theory there is an automatic clustering structure induced by the Dirichlet process, some practical problems arise in the Bayesian context from using MCMC methods: One obtains a clustering of subjects within each iteration, but it is unclear how these can be merged into an universal clustering. Several operations exist to handle this (see for example Fritsch and Ickstadt, 2009), but due to the high number of possible clusterings, these methods are typically not feasible in larger problems. The advantage of the EM algorithm over MCMC methods is that the EM algorithm converges to fixed values, while MCMC methods converge to distributions. So with EM type algorithms the cluster property of the Dirichlet process can be used directly.

3 Application: Lung function growth

The practical use of the proposed method is investigated by considering the lung function growth of girls in Topeka (USA). These data are a subsample from the six cities study of air pollution and health in Dockery et al. (1983). The response variable is the logarithmic forced expiratory volume in one second ($fev1$). Our sample consists of 100 girls, with a minimum of two and a maximum of twelve observations over time. We use a linear mixed model with random intercepts and random slopes

$$\log(fev1)_{ij} | b_i \stackrel{ind.}{\sim} N(\beta_0 + b_{i0} + (\beta_1 + b_{i1})age_{ij}, \sigma^2),$$

and a DPM as random effects distribution (3). The clustering effect of the DPM-EM model can be seen from Figure 1. Here the axes represent the intercepts and slopes respectively. The square at coordinates (0,0) marks the population effect. All other icons are interpreted as deviations from the population effect. The thick big ones symbolize the cluster locations μ_h , the thin small ones the random effects b_i . Girls which assigned to the same cluster are marked with the same symbol and are arranged around the three cluster locations in the form of ellipses.

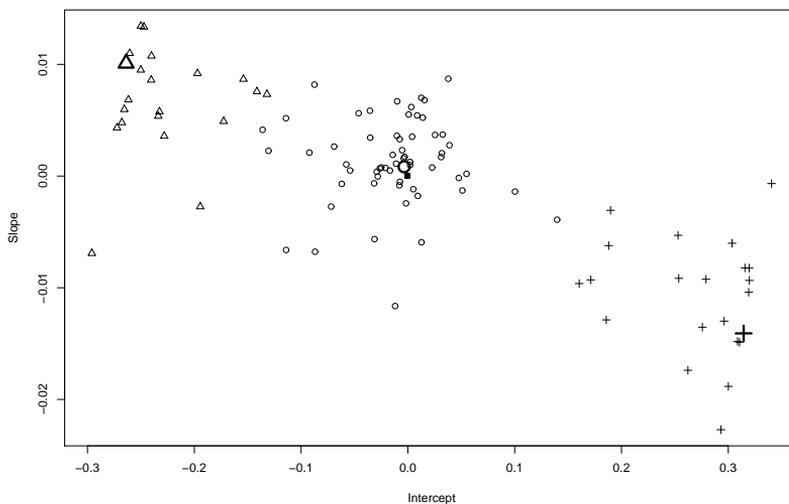


FIGURE 1. Cluster locations and corresponding random effects for lung function growth data

4 Conclusion

We introduced a linear mixed models with a DPM for the random effects distribution in order to penalize the number of clusters in the finite mixture of normal distribution. While models with Dirichlet processes are typically fitted by Bayesian methods like MCMC we used the EM algorithm because then the cluster property of the Dirichlet process can be used directly. So our method can be called an agglomerative clustering approach of individuals for longitudinal data. We detected in a simulation study that our approach outperforms the classical linear mixed model in the case of a underlying grouping structure. Applications of this DPM-EM algorithm were demonstrated by considering lung function growth data.

References

- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39**, 1–38.
- Dockery, D. W., C. S. Berkey, J. H. Ware, F. E. Speizer, and B. G. Ferris (1983). Distribution of fvc and fev1 in children 6 to 11 years old. *American Review of Respiratory Disease*, **128**, 405–412.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.
- Fritsch, A. and K. Ickstadt (2009). Improved criteria for clustering based on the posterior similarity matrix. *International Society for Bayesian Analysis*, **4**, 367–392.
- Grün, B. (2008). Fitting finite mixtures of linear mixed models with the EM algorithm. In P. Brito (Ed.), *Compstat 2008—Proceedings in Computational Statistics*, Volume II, Heidelberg, pp. 165–173. Physica Verlag.
- Heinzl, F., L. Fahrmeir, and T. Kneib (2012). Additive mixed models with Dirichlet process mixture and P-spline priors. *Advances in Statistical Analysis*, **96**(1), 47–68.
- Kleinman, K. and J. Ibrahim (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics*, **54**, 921–938.
- Komárek, A. and E. Lesaffre (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics and Data Analysis*, **52**, 3441–3458.
- Magder, L. S. and S. L. Zeger (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association*, **91**, 1141–1151.
- Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.
- Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217–221.
- Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Ordinal regression models for continuous scales

Gillian Z. Heller¹, Maurizio Manuguerra¹

¹ Department of Statistics, Macquarie University, Sydney, Australia

E-mail for correspondence: gillian.heller@mq.edu.au

Abstract: Ordinal regression analysis is a convenient tool for analyzing ordinal response variables in the presence of covariates. We extend this methodology to the case of continuous self-rating scales which measure subjects' perception of an intangible quantity, and cannot be handled as ratio variables because of inherent nonlinearity. We express the likelihood in terms of a function connecting the scale with an underlying continuous latent variable and approximate this function either parametrically or non-parametrically. We illustrate our method using a study on the impact of chemotherapy treatment on quality of life in advanced breast cancer patients.

Keywords: ordinal regression; nonparametric regression; Visual Analog Scale; Linear Analog Self-Assessment scale; quality of life; pain

1 Introduction

We propose regression models for outcomes which are intangible and difficult to measure on conventional scales, such as pain and quality of life. Continuous self-rating scales are referred to as Visual Analog Scales (VAS) in the pain literature and Linear Analog Self-Assessment (LASA) scales in quality of life studies. Subjects are typically given a linear scale of 100 mm and asked to put a mark where they perceive themselves. For convenience we refer to continuous scales of this type as VAS, and to the outcome as pain. Historically there has been controversy on the nature of VAS: whether it is ratio or ordinal; linear or nonlinear. "Linear" in this context is taken to mean that differences in pain between successive increments on the VAS are constant. The problem of non-interpretability of distances between measurements and the possibility of nonlinear behaviour, particularly at one or both extremes of the scale, is overcome by treating VAS measurements as ordinal rather than ratio data. We therefore refer to scales of this type as continuous ordinal.

Ordinal regression models are widely used for regression analysis of discrete ordinal responses Y within K ordered categories. The Y 's are considered as

coarse versions of an unobserved, continuous latent variable W , such that

$$Y = j \iff \alpha_{j-1} < W < \alpha_j, \quad j = 1, \dots, K$$

where the α_j 's are the correspondence on the latent variable scale of the category boundaries on the ordinal scale and $-\infty = \alpha_0 < \alpha_1 < \dots < \alpha_K = \infty$. Typically W is an intangible quantity such as pain, and $y = 1, 2, \dots, K$ codes for ordinal states such as *none*, *mild*, *moderate*, *severe*. To relate the cumulative probabilities $\gamma_j(x) = P(Y \leq j|x)$ to covariates $x = (x_1, \dots, x_p)'$ in the j th category, we assume that $W = -x'\beta + \epsilon$. When ϵ has the standard logistic distribution, this results in the cumulative logistic model (also called the proportional odds model) for Y :

$$\ln \left(\frac{\gamma_j(x)}{1 - \gamma_j(x)} \right) = \alpha_j + x'\beta, \quad j = 1, \dots, K - 1 \quad (1)$$

The ordinal regression model has been developed in the last two decades in order to incorporate additive and non-linear functional forms of predictors (Hastie and Tibshirani 1987) and spline-based smoothers of predictors (Yee and Wild 1996). In the applied literature, typically continuous ordinal responses are analysed using model (1) on a discretized version of the VAS responses, or simply by treating the VAS measurements as continuous responses in a normal regression model. Both of these approaches are less than satisfactory. More sophisticated methods have been proposed in the statistical literature. Bottai et al. (2010) apply the logistic transformation in order to overcome the difficulty inherent in formulating a model for a bounded response, and use quantile regression modelling on the transformed scores. Lesaffre et al. (2007) also apply the logistic transformation and consider models in which the transformed response has a normal distribution. We propose a generalization of the standard ordinal model which leads to a semiparametric ordinal regression model for continuous scales.

2 Regression model for continuous ordinal responses

Consider VAS measurements v which are sampled from a continuous response variable $V \in (0, 1)$, with density $f(v)$ and CDF $\gamma(v)$. The continuous ordinal response variable V can be taken to reflect the subjective perception of an underlying continuous latent variable W defined on the real line. The dependence between V and W is modelled by a smooth one-to-one function $g : (0, 1) \mapsto (-\infty, +\infty)$ that maps v on the VAS to $w = g(v)$ on the latent scale. This mapping is the link between the recorded perception of pain and an underlying metric. As for the standard ordinal model, covariates are modelled on the latent scale. Assuming $W = -x'\beta + \epsilon$,

$$\gamma(v|x) = P(V \leq v|x) = P(W \leq g(v)|x) = F(g(v) + x'\beta),$$

where $F(\cdot)$ is the CDF of ϵ . Inverting this translates to the generic ordinal regression model for continuous observations v :

$$F^{-1}(\gamma(v|x)) = g(v) + x'\beta .$$

We assume the standard logistic distribution for ϵ , but other distributions, such as the normal, can be used. The cumulative logistic ordinal model for continuous response variables is:

$$\ln \left(\frac{\gamma(v|x)}{1 - \gamma(v|x)} \right) = g(v) + x'\beta . \quad (2)$$

The function $g(v)$ in model (2) is the continuous analog of the discrete intercepts α_j in model (1), and its shape is informative of the change in perception of pain at different levels (as are the α_j). The linear component $x'\beta$ may incorporate fixed and random effects.

3 Model implementation

3.1 g function

In order to complete the specification of model (2), we need to define the form of the $g(v)$ function. g has to be capable of capturing the nonlinear behaviour of the ordinal measure. Any differentiable, increasing and “flexible enough” function which maps $(0, 1)$ to $(-\infty, +\infty)$ could be appropriate. This can be done using a parametric or non-parametric approach.

Parametric g function

Any inverse sigmoidal function could be appropriate. We choose g as the inverse of the generalized logistic function, which has the advantage of simplicity and mathematical tractability:

$$g(v) = M + \frac{1}{B} \log \left(\frac{Tv^T}{1 - v^T} \right) , \quad 0 < v < 1 \quad (3)$$

where M is the offset, B is the slope and T is the symmetry of the curve.

Non-parametric g function

We have used B-splines, with constraints to ensure monotonicity. To avoid overfitting of the data, a penalized likelihood approach has been used.

3.2 Predictors

A penalized likelihood approach for B-splines has been adopted for predictors which appear in the model as smooth functions.

3.3 Estimation and software

We have used Bayesian methodology, with model selection based on the DIC. Estimation of the B-spline coefficients has shown some instability, given the sparsity of data for high values of the predictor, and long burn-in phases and samples were necessary. The analyses were performed using the Metropolis-Hastings algorithm, implemented in R 2.8.1 using the *splines* and *MCMCpack* (Martin et al. 2009) libraries.

4 Application to quality of life data

4.1 Chemotherapy treatments in advanced breast cancer study

The ANZ 0001 trial is an unblinded randomized trial with three chemotherapy treatment arms ($n = 292$) (Manuguerra and Heller 2010). Quality of life (QOL) is assessed at each treatment cycle, from randomization until disease progression, when treatment is interrupted. Two treatments IC and CC are compared with the standard treatment CMF. The study aims to establish which treatment has a better impact on QOL, and in particular how this impact changes over chemotherapy cycle. QOL is assessed using LASA scales, in which high readings indicate bad QOL. No patients on CMF progressed beyond cycle 20, and for the other two treatments the data are sparse beyond this point. Among the several covariates available, only age and cycle number were found to be significant.

4.2 Model

The model for QOL v_{ij} for patient i on treatment k at cycle j is:

$$\log \frac{\gamma_{ij}}{1 - \gamma_{ij}} = g(v_{ij}) + x_i \beta + s_k(j) + b_i$$

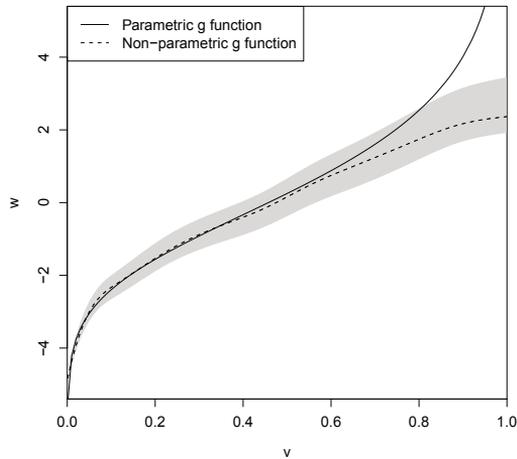
where x_i is age, $s_k(j)$ is a smooth term that depends on cycle number j and b_i are random effects sampled from $N(0, \sigma^2)$. We compare the parametric and non-parametric approaches for $g(v_{ij})$, using (3) and B-splines, respectively. B-splines are used for estimation of $s_k(j)$.

4.3 Results

Table 1 compares the parameters estimated with the parametric and non-parametric approaches. The two methods show similar results for β and σ , with a significant worsening of QOL with increasing age. In terms of DIC, the non-parametric model has an advantage over the parametric approach. In Figure 1 the estimates of the g function are shown. The accordance is good except in the region of worst QOL (right side of the scale), where fewer subjects have marked their perception. The fact that the non-parametric

TABLE 1. Chemotherapy impact on quality of life: parameter estimates.

	Parametric g function		Non-parametric g function	
	Median	95% CI	Median	95% CI
β	0.014	(0.009, 0.021)	0.022	(0.019, 0.023)
σ	0.210	(0.192, 0.232)	0.174	(0.158, 0.193)
Deviance		-1653		-2261
No parameters		3.4		14.0
DIC		-1648		-2244

FIGURE 1. Parametric and non-parametric estimates of the g function, with 95% percentile CI.

model has obtained a smaller DIC is probably due to a lack of flexibility of the parametric curve at this extreme. The two models have given similar results for the dependence of QOL on cycle number. In Figure 2, the result obtained by the non-parametric model is shown. CC has a clear advantage over IC, while it is not possible to give clear indications on CMF.

5 Discussion

We provide a regression framework for a response variable that is a recorded perception of an underlying latent variable which is difficult to observe or measure. The model is an extension of the ordinal regression model for discrete ordinal responses. Our choice of the inverse of the generalized logit function for g was based on its simplicity and flexibility. However, our

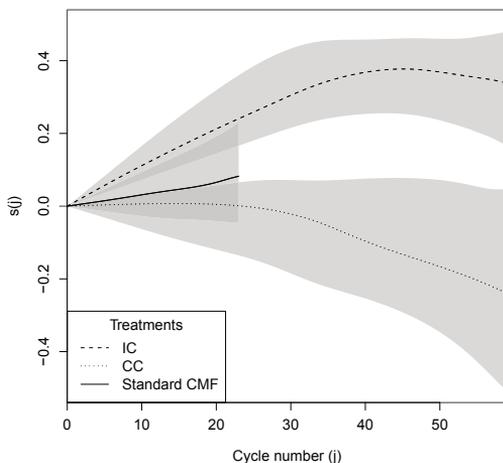


FIGURE 2. Dependence of QOL on cycle number, with 95% percentile CI.

method is not limited to this choice and the inverse of any sigmoid function may be used. Use of a nonparametric g function eliminates the need for a choice of function, at the expense of several degrees of freedom. In the case of a sparse data set, there will be obvious advantages to a judicious choice of parametric g function over the nonparametric approach.

References

- Bottai, M., B. Cai, and R. E. McKeown (2010). Logistic quantile regression for bounded outcomes. *Statistics in Medicine*, **29(2)**, 309–317.
- Hastie, T. and R. Tibshirani (1987). Non-parametric logistic and proportional odds regression. *Applied Statistics*, **36(2)**, 260–276.
- Lesaffre, E., D. Rizopoulos, and R. Tsonaka (2007). The logistic transform for bounded outcome scores. *Biostatistics*, **8(1)**, 72–85.
- Martin, A. D., K. M. Quinn, and J. H. Park (2009). *MCMCpack: Markov chain Monte Carlo (MCMC) Package*. R package version 1.0-4.
- Manuguerra, M. and G. Z. Heller (2010). Ordinal Regression Models for Continuous Scales. *International Journal of Biostatistics*, **6(1)**, 14.
- Yee, T. and C. Wild (1996). Vector generalized additive models. *JRSS Series B (Methodological)*, **58**, 481–493.

A joint marginalized multilevel model for longitudinal outcomes

Samuel Iddi¹, Geert Molenberghs^{1,2}

¹ I-BioStat, Katholieke Universiteit Leuven, Belgium

² I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium

E-mail for correspondence: Samuel.Iddi@med.kuleuven.be

Abstract: The shared-parameter model and its hierarchical or random-effect extensions are widely used joint modeling approaches to a combination of longitudinal continuous, binary, count, missing, and survival outcomes that naturally occurs in many clinical and other studies. It is well known that parameter estimates in a linear mixed model (LMM) for continuous repeated measures or longitudinal outcomes allow for a marginal interpretation, even though a hierarchical formulation is employed. This is not the case for the generalized linear mixed model (GLMM), i.e., for non-Gaussian outcomes. To derive marginally meaningful parameters for the binary models in a joint model, we adopted the marginal multilevel model of Heagerty (1999) and Heagerty and Zeger (2000), and formulated a joint marginal multilevel model for two longitudinal responses. This enables to (1) capture association between the two responses and (2) obtain parameter estimates that have a population-averaged interpretation for both outcomes. The model is fitted to data from a clinical trial in ophthalmology and results are compared with existing approaches. Estimates were found to be very close to those from single analysis per outcome but the joint model yields higher precision and allows for quantifying the association between outcomes. Parameters were estimated using maximum likelihood. The model is easy to fit using available tools such as the SAS procedure NLMIXED.

Keywords: Generalized estimating equation; Marginal multilevel model; Maximum likelihood estimation; Random effects model; Shared-parameter model.

1 Introduction

Joint modeling has received massive attention in recent years, owing to researchers' desire for more insight into their data with a single statistical model. The reason to find this type of analysis is because commonly researchers simultaneously record several kinds of outcomes in their studies. These outcomes are often of a mixed nature. Commonly found examples are combinations of continuous, binary, ordinal, survival, and missing outcomes. Continuous and binary outcomes often occur in longitudinal studies where one observes follow up measurements on patients. Frequently,

because studies are conducted in humans, data are incomplete owing to dropout or other reasons for missingness. This issue may also require careful attention. Single analyses per outcome are limited in that they do not provide answers to questions that take several or all outcomes simultaneously into account. An excellent review of various joint modeling approaches for longitudinal and time to event data can be found in Tsiatis and Davidian (2004). Molenberghs and Verbeke (2005) discuss a number of techniques that jointly model continuous and discrete outcomes.

It is generally known that the parameter estimates from a linear mixed model have a marginal interpretation even though a hierarchical formulation is employed. Such is not the case for the generalized linear mixed model for non-Gaussian outcomes. Heagerty (1999) proposed a now broadly known marginalized multilevel model (MMM) which enjoys the various strengths of marginal and conditional modeling techniques. Particularly, effects of covariates will have a direct marginal interpretation. We followed the modeling concepts of Heagerty (1999) and formulated a joint model for two longitudinal outcome. The generalized linear mixed model component in a shared-parameter model is replaced by the model of Heagerty (1999). Full maximum likelihood estimation with iterative numerical quadrature methods are adopted to obtain parameter estimates.

2 The Age Related Macular Degeneration Trial

The Age Related Macular Degeneration (ARMD) trial has been presented and studied by Buyse and Molenberghs (1998) and Molenberghs and Verbeke (2005). The data resulted as a product of a randomized multi-centric clinical trial for patients with ARMD, a condition associated with progressive loss of vision in the elderly. The aim of the trial was to compare experimental interferon- α to placebo. The outcome of the trial was the patients' visual acuity, which was measured at 4 follow-up visits (4, 12, 24, and 52 weeks). During each visit, patients were made to read lines of letters on standardized vision charts and the total number of letters that were correctly read was recorded as the patients' visual acuity. The full longitudinal profile was subjected to analysis in Molenberghs and Verbeke (2005). For illustration here, two distinct versions of visual acuity will be considered namely (1) change in visual acuity at the different time points, assumed to be normally distributed, after onset of treatment and (2) a binary variable indicating whether or not there is loss of vision at the various visit periods compared to baseline.

3 Methodology

Consider two longitudinal outcomes Y_{1ij} and Y_{2ik} , denoting the j th and k th measurement on the i th subject for continuous and binary type outcomes, respectively, ($i = 1, \dots, N$, $j = 1, 2, \dots, n_{1i}$, and $k = 1, 2, \dots, n_{2i}$).

This means that we need to develop an appropriate model for their joint distribution $f(\mathbf{Y}_{1i}, \mathbf{Y}_{2i})$. An attractive joint modeling techniques is the shared-parameter model (Molenberghs and Verbeke 2005), where an unobserved random variable is introduced, given which, the two outcomes are further assumed independent. In other words, the random effects are solely responsible for generating the association between the outcomes.

We begin by formulating a linear mixed model for the continuous outcomes Y_{1ij} . Next to this, a generalized linear mixed model is often specified for the binary outcome. We replace this model by Heagerty’s (1999) proposal, yielding what we refer to as a *joint marginalized multilevel model* (JOMMM). The logit-probit normal version is adopted so that analytical expressions can be derived for the joint distribution of the two responses. This means a logit link is used for the marginal specification of the model and a probit link for the conditional part. In doing so, we will retain the odds ratio interpretation of the marginal parameters while taking advantage of the computational ease emerging from the probit-normal relationship. The new model is completely spelled out as follows:

$$\begin{aligned} \mathbf{Y}_{1i}|\mathbf{b}_i &\sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i), \\ \text{logit}(\mu_{ik}^m) &= \mathbf{x}'_{ik}\boldsymbol{\xi}^m, \\ \Phi^{-1}(\mu_{ik}^c) &= \delta_{ik} + \mathbf{w}'_{ik}\mathbf{b}_i, \\ \mathbf{b}_i &\sim N(\mathbf{0}, \mathbf{D}). \end{aligned}$$

Here, \mathbf{w}'_{ik} is a vector of scale parameters and covariates (i.e., $\mathbf{w}'_{ik} = \Lambda\mathbf{z}'_{ik}$), $\mu_{ik}^m = E(Y_{2ik} = 1)$, $\mu_{ik}^c = E(Y_{2ik} = 1|\mathbf{b}_i)$ and

$$\delta_{ik} = \left(\sqrt{1 + \mathbf{w}'_{ik}\mathbf{D}\mathbf{w}_{ik}} \right) \Phi^{-1} \left\{ \text{expit}(\mathbf{x}'_{ik}\boldsymbol{\xi}^m) \right\}.$$

The term δ_{ik} is obtained from the integral equation:

$$\mu_{ik}^m = g(\mathbf{x}'_{ik}\boldsymbol{\xi}^m) = \int_b g(\delta_{ik} + \mathbf{w}'_{ik}\mathbf{b}_i) dF_b,$$

where $g(\cdot)$ is an inverse link function. This ensures that the fixed-effect parameters of the so-called marginal model will indeed possess a marginal interpretation. We can derive the joint marginal distribution by integrating out the random effect. Thus, the contribution of the i th subject to the

likelihood is given by

$$\begin{aligned}
 f_i(\mathbf{Y}_{1i} = \mathbf{y}_i, \mathbf{Y}_{2i} = 1) &= \frac{1}{(2\pi)^{\frac{n_i}{2}} |\mathbf{R}_i|^{\frac{1}{2}} |\mathbf{D}_i|^{\frac{1}{2}}} \times e^{-\frac{1}{2}[(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' \mathbf{R}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})]} \\
 &\times \prod_k | \mathbf{D}_i^{-1} + \mathbf{Z}'_i \mathbf{R}_i \mathbf{Z}_i + \mathbf{w}_{ik} \mathbf{w}'_{ik} |^{-\frac{1}{2}} \left[\frac{e^{\frac{1}{2} \left(\frac{r_2^2}{4r_1} - r_3 \right)}}{r_1^{\frac{1}{2}}} \right] \\
 &\times \Phi \left[\left(\delta_{ik} + \frac{r_2}{2r_1} \right) r_1^{\frac{1}{2}} \right],
 \end{aligned}$$

where

$$\begin{aligned}
 r_1 = r_{1(k)} &= I - \mathbf{w}'_{ik} \left[\left(\mathbf{D}_i^{-1} + \mathbf{Z}'_i \mathbf{R}_i \mathbf{Z}_i + \mathbf{w}_{ik} \mathbf{w}'_{ik} \right)^{-1} \right]' \mathbf{w}_{ik}, \\
 r_2 = r_{2(k)} &= \mathbf{w}'_{ik} \left[\left(\mathbf{D}_i^{-1} + \mathbf{Z}'_i \mathbf{R}_i \mathbf{Z}_i + \mathbf{w}_{ik} \mathbf{w}'_{ik} \right)^{-1} \right]' \mathbf{Q}', \\
 r_3 = r_{3(k)} &= -\frac{1}{4} \mathbf{Q} \left[\left(\mathbf{D}_i^{-1} + \mathbf{Z}'_i \mathbf{R}_i \mathbf{Z}_i + \mathbf{w}_{ik} \mathbf{w}'_{ik} \right)^{-1} \right]' \mathbf{Q}', \\
 \mathbf{Q} &= [(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' \mathbf{R}_i^{-1} \mathbf{Z}_i + (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' (\mathbf{R}_i^{-1})' \mathbf{Z}_i].
 \end{aligned}$$

4 Estimation

Parameters in the joint model are estimated using maximum likelihood, based on

$$L(\boldsymbol{\beta}, \boldsymbol{\xi}^m, D) = \prod_{i=1}^N f_i(\mathbf{Y}_{1i} = \mathbf{y}_i, \mathbf{Y}_{2i} = 1).$$

Even though this analytical joint marginal likelihood can be maximized, it is cumbersome to manipulate. Hence, it is more convenient to maximize the likelihood after employing numerical techniques to instead integrate out the random effect distribution. Gaussian and adaptive Gaussian quadrature are designed for such purpose, up to a pre-specified level of accuracy (Pineiro and Bates 2000). The standard errors of the parameter estimates are computed from the inverse Hessian matrix (second derivatives) at the estimates obtained numerically. Major statistical tools, such as the SAS procedure NLMIXED, are readily available for fitting the models specified in this paper. Other estimation techniques are discussed in Molenberghs and Verbeke (2005) for discrete and in Verbeke and Molenberghs (2000) for continuous outcomes.

5 Data Analysis

Results of fitting various approaches to the longitudinal continuous visual acuity and the binary vision-loss outcome in the ARMD data are presented. The covariate structure is maintained across models, for ease of comparison. The models have an intercept and the effect of treatment at each time point for both outcomes. Precisely, we assume the predictors to be: $\mu_{1ij} = \alpha_{0,j} + \alpha_{1,j}T_{ij}$ and $\text{logit}(\mu_{2ij}) = \beta_{0,j} + \beta_{1,j}T_i$, where $j = 1, 2, 3, 4$, and T_i is treatment allocation. In the corresponding conditional models, a random intercept, $b_i \sim N(0, d)$, was used. The single marginal models were fitted with a compound symmetry (exchangeable) variance or correlation structure. Also, the joint models were fitted using the shared-parameter model where the inflation factor (λ) was introduced in the model for the continuous sequence. Results of all models fitted are presented in Table 1. In summary, our proposed joint marginal multilevel model, produces parameter estimates similar to those of GEE and the MMM which gives an indication that the JOMMM model parameters will indeed have a marginal interpretation. Finally, estimates from the joint models tend to yield higher precision than those of the separate analyses.

6 Concluding Remarks

We have shown that a joint longitudinal model can be formulated where all parameters enjoy a marginal interpretation. This was achieved by incorporating the model of Heagerty (1999) into the shared-parameter or hierarchical joint models used to jointly model two longitudinal outcomes. The resulting model at the same time captures association between the two responses, and yields parameter estimates that have a population-averaged interpretation for both outcomes. Estimates were found to be close to those from single-outcome analyses but provided higher precision. The difference in precision could affect inferences. Thus, it is important to make use of such joint modeling approaches, which tend to provide unbiased and more precise estimates. Note that in the real data example, only a random intercept was used in the sub-models. This is not to say the model is merely restricted to one-dimensional random effect. It is indeed possible to add as many random effects as it is practicable. The user ought to be reminded that even for single response models, adding more random effects increases the complexity of the model and thus difficult or impossible to fit in some cases. In terms of implementation, the proposed method allows to efficiently make use of available resources, such as the SAS procedure NLMIXED. Codes are available at the authors' web pages.

Acknowledgments: The authors gratefully acknowledge the financial support from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

TABLE 1. *The Age Related Macular Degeneration (ARMD) Trial: Comparison of Joint and separate models for continuous and binary visual acuity sequences. Parameter estimates (standard errors).*

Effect	Continuous Sequence				Binary Sequence				
	Marg. Corr.	Sep. Hier.	J. Hier.	J. MMM	GEE	GLMM	MMM	J. Hier.	J. MMM
Int. 4	-3.23 (0.81)	-3.25 (1.30)	-3.27 (1.30)	-3.27 (1.30)	1.01 (0.24)	1.74 (0.42)	1.00 (0.24)	2.01 (0.46)	1.03 (0.23)
Int.12	-4.62 (1.07)	-4.62 (1.29)	-4.62 (1.29)	-4.62 (1.29)	0.91 (0.24)	1.56 (0.41)	0.91 (0.24)	1.82 (0.45)	0.93 (0.23)
Int.24	-8.37 (1.26)	-8.37 (1.29)	-8.37 (1.29)	-8.37 (1.29)	1.15 (0.25)	1.95 (0.43)	1.15 (0.25)	2.24 (0.47)	1.17 (0.25)
Int.52	-15.16 (1.64)	-15.16 (1.29)	-15.16 (1.29)	-15.16 (1.29)	1.65 (0.29)	2.76 (0.48)	1.63 (0.29)	3.11 (0.52)	1.63 (0.28)
Tr. 4	2.33 (1.06)	2.34 (1.76)	2.39 (1.76)	2.39 (1.76)	-0.42 (0.32)	-0.67 (0.54)	-0.37 (0.31)	-0.69 (0.59)	-0.36 (0.30)
Tr.12	2.34 (1.52)	2.34 (1.76)	2.34 (1.76)	2.34 (1.76)	-0.54 (0.31)	-0.88 (0.53)	-0.51 (0.31)	-0.93 (0.59)	-0.48 (0.30)
Tr.24	2.83 (1.84)	2.83 (1.76)	2.83 (1.76)	2.83 (1.76)	-0.52 (0.32)	-0.84 (0.54)	-0.50 (0.33)	-0.88 (0.60)	-0.47 (0.32)
Tr.52	4.12 (2.31)	4.12 (1.76)	4.12 (1.76)	4.12 (1.76)	-0.40 (0.38)	-0.61 (0.59)	-0.37 (0.37)	-0.62 (0.64)	-0.34 (0.36)
s.d.(res.)	-	8.42 (0.25)	8.22 (0.23)	8.23 (0.23)	-	-	-	-	-
s.d.(r.e.)	-	8.62 (0.55)	-	-	-	2.19 (0.27)	1.26 (0.15)	-	-
Infl.	-	-	-3.31 (0.34)	-5.87 (0.57)	-	-	-	-	-
-2logl.	5480	5168	-	-	773	773	-	-	-
Common parameters in Joint Models									
s.d.(r.e.)	2.66 (0.29)	1.50 (0.16)							
-2logl.	5745	5745							

References

Griswold, M.E. and Zeger, S.L. On marginalized multilevel models and their computation (November 2004). *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 99.*

Heagerty, P.J. (1999) Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688–698.

Heagerty, P.J. and Zeger, S.L. (2000) Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, **15**, 1–26.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. New York: Springer.

Pinheiro, J.C., and Bates, D.M. (2000) *Mixed Effects Models in S and S-Plus*. New York: Springer-Verlag.

Tsiatis, A.A., and Davidian, M. (2004) A joint modeling of longitudinal and time-to-event data: An overview. *Stat Sin*, **14**, 809–834.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Bayesian ODE-penalized B-spline model with Gaussian mixture as error distribution

Jonathan Jaeger¹, Philippe Lambert¹²

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: jonathan.jaeger@uclouvain.be

Abstract: In the standard Bayesian ODE-penalized B-spline approach, it is assumed that the error distribution is homogeneous Gaussian. But, in many applications, the normal assumption for the error distribution is not a realistic choice. The goal of this paper is to extend the standard Bayesian ODE-penalized B-spline approach to settings where the error term distribution can be described using a mixture of normals.

Keywords: Ordinary differential equations; ODE-penalized B-spline; Penalized Gaussian mixture.

1 Introduction

Ordinary differential equations (ODEs) are frequently used to model physical, chemical and biological processes.

Currently, the most commonly used estimation procedures rely on nonlinear least squares (Biegler et al., 1986). It uses minimization techniques for the estimation of the ODE parameters and a numerical solver for the approximation of the solution. These approaches are computationally intensive and often poorly suited for statistical inference. Lunn et al. (2002) propose a Bayesian framework for these NLS procedures. The possible use of prior information about the ODE parameters is a definite advantage, but the numerical approximation to the solution leads to a posterior distribution with no-closed form for the ODE parameters.

Alternative estimation methods of the state functions and the ODE parameters were proposed in Ramsay et al. (2007). It may be viewed as a generalization of the P-spline theory (Eilers and Marx, 1996) that involves some basis function expansion of each state function and a penalty term expressed using the set of differential equations. Jaeger and Lambert (2011) adapt this approach to a full Bayesian ODE-penalized B-spline setting when the ODEs are affine and the data distribution is assumed Gaussian. The two

major drawbacks of the frequentist ODE-penalized smoothing approach are overcome in the Bayesian framework: the selection of the ODE-adhesion parameter is now automatic and uncertainty measures about parameters can simply be obtained using MCMC. In addition, the possible use of prior information about the ODE parameters is a definite advantage.

The assumption of a Gaussian data distribution is often inappropriate in practice but is very convenient as it enables to marginalize the joint posterior distribution with respect to the spline coefficients and therefore to get rid of the inconvenient posterior correlation between the spline coefficients and the ODE parameters. To overcome this limitation, we model homogeneous non-normal data distribution using finite mixture of Gaussian distributions by adapting the approach of Komárek and Lesaffre (2008).

In Section 2, we remind some basis elements of the ODE-penalized B-spline approach. Section 3 describes the assumed model, i.e. the combination of the Bayesian ODE-penalized B-spline approach with the penalized Gaussian mixture approach. Section 4 proposes an application in pharmacokinetics.

2 Basics on ODE-penalized B-spline approach

We assume that changes in the states $\mathbf{x}(t) \in \mathbb{R}^d$ of a dynamic system are governed by a set of differential equations:

$$D\mathbf{x}(t) = f(\mathbf{x}, t, \boldsymbol{\theta}), t \in [0; T],$$

where f is a known affine function and $\boldsymbol{\theta} \in \mathbb{R}^q$ an unknown vector of parameters. It is assumed that only a subset $\mathcal{J} \subset \{1, \dots, d\}$ of the d state function \mathbf{x} are measured at time point t_{jk} , $j \in \mathcal{J}$, $k = 1, \dots, n_j$ with additive measurement error ϵ_{jk} . We denote by $y_{jk} = \mathbf{x}_j(t_{jk}) + \tau_j^{-1/2} \epsilon_{jk}$ the corresponding measurement. The objective is to jointly estimate the ODE parameters $\boldsymbol{\theta}$ and the state functions $\mathbf{x}(t)$ from $\{(t_{jk}, y_{jk}), j \in \mathcal{J}, k = 1, \dots, n_j\}$. Most often, one further imposes that $\mathbb{E}(\epsilon_j) = 0$ and $\mathbb{V}(\epsilon_j) = 1$ such that $x_j(t)$ and τ_j are respectively the conditional mean and the inverse-variance of y_j .

Each state function involved in the ODE is approximated by a B-spline basis function:

$$\tilde{x}_j(t) = (\mathbf{B}_j(t))^T \mathbf{c}_j.$$

The B-spline basis $\mathbf{B}_j(t)$ is chosen flexible enough to capture all the variation in the state function and the spline coefficients \mathbf{c}_j controls the shape of the approximation.

In order to capture all the variation and to solve the system of differential equations, Ramsay et al. (2007) propose to consider a large set of B-spline basis and to constraint the spline coefficients using an ODE model-based-penalty. For each differential equation of the system, a penalty term PEN_j

is introduced to assess the proximity of the approximation of the state function to the solution:

$$PEN_j(\tilde{\mathbf{x}}) = \int \{D\tilde{x}_j(t) - f_j(\tilde{\mathbf{x}}, t, \boldsymbol{\theta})\}^2 dt.$$

The full fidelity-to-ODE measure is then given by the sum of these penalties times some ODE-adhesion parameters γ_j :

$$PEN(\tilde{\mathbf{x}}|\boldsymbol{\gamma}) = \sum_{j=1}^d \gamma_j PEN_j$$

The ODE-adhesion parameters $\boldsymbol{\gamma}$ permit to weight and to control the fidelity to the ODE, i.e. to express the confidence that one has in the differential equations as descriptors of the dynamics of the state functions.

3 Model

3.1 Assumption on the error distribution

In order to allow distributional flexibility, the distribution of the error terms is modeled as a mixture of Gaussian distribution over a fixed grid of equidistant means centered around 0. For a given observed state function, the variances of the Gaussian components are fixed and common for all the mixture components:

$$\epsilon_{jk}|\boldsymbol{\pi}_j \sim \sum_{l=-L_j}^{L_j} \pi_{jl} \mathcal{N}(\mu_{jl}, \sigma_j^2).$$

One can rewrite the weights as:

$$\pi_{jl} = \frac{\exp(a_{jl})}{\sum_{k=-L_j}^{L_j} \exp(a_{jk})}, j \in \mathcal{J}, l = -L_j, \dots, L_j,$$

with identifiability constraints $a_{j,L_j} = 0$ for all $j \in \mathcal{J}$. With that parameterization, the constraint $\sum_{l=-L_j}^{L_j} \pi_{jl} = 1$ is automatically checked.

3.2 Prior for the spline coefficients and for the ODE-adhesion parameters

The ODE-penalty term is translated into a prior distribution for the spline coefficients:

$$p(\mathbf{c}|\boldsymbol{\gamma}, \boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}PEN(\tilde{\mathbf{x}}|\boldsymbol{\gamma})\right).$$

The prior distribution for the ODE-adhesion parameters is a gamma distribution $\mathcal{G}(a, b)$ with mean a/b and variance a/b^2 . As recommended in Lang and Brezger (2004) for standard P-spline model, we have two possibilities: either set a equal to 1 and b equal to a small quantity (e.g. 10^{-6}) or set $a = b$ equal to a small quantity. We will opt for the first specification as the corresponding density is finite at 0. This choice for the prior distributions of the ODE-adhesion parameters translates our prior confidence in the specification of the system of differential equations as description of the dynamic system.

3.3 Prior for the transformed weights and for the roughness penalty parameters

In order to capture the shape of the error density, Komárek and Lesaffre (2008) proposed to consider a large number of Gaussian components. The overfitting of the data and identifiability problems are avoided by putting a 3rd-order difference penalty on the transformed weights \mathbf{a}_j :

$$Q_j(\mathbf{a}_j|\lambda_j) = -\frac{\lambda_j}{2}\mathbf{a}_j^T \mathbf{D}_j^T \mathbf{D}_j \mathbf{a}_j,$$

where \mathbf{D}_j is the 3rd order difference penalty. The smoothness of the fitted error distributions is controlled by a roughness penalty parameters. As $a_{j,L_j} = 0$ for all $j \in \mathcal{J}$, this 3rd order penalty term is translated into a prior distribution for the transformed weights $\mathbf{d}_j = (a_{j,-L_j}, \dots, a_{j,L_j-1})^T$:

$$p(\mathbf{d}_j|\lambda_j) \propto \exp\left(-\frac{\lambda_j}{2}\mathbf{d}_j^T \mathbf{P}_j \mathbf{d}_j\right),$$

where \mathbf{P}_j is a sub-matrix of $\mathbf{D}_j^T \mathbf{D}_j$. As for the ODE-adhesion parameter, it is convenient to take a gamma prior distribution for the roughness penalty parameters.

3.4 Identification penalty

Additional constraints should be given on the transformed weights \mathbf{a} to force the error distribution to have conditional mean equal to zero and conditional variance equal to one. One can show that:

$$\mathbb{E}(\epsilon_j|\boldsymbol{\pi}_j) = \sum_{l=-L_j}^{L_j} \pi_{jl} \mu_{jl} = \mu_{\epsilon_j},$$

and

$$\mathbb{V}(\epsilon_j|\boldsymbol{\pi}_j) = \sigma_j^2 + \sum_{l=-L_j}^{L_j} \pi_{jl} \mu_{jl}^2 - \mu_{\epsilon_j}^2 = \sigma_{\epsilon_j}^2.$$

We adapt a proposition of Lambert (2011) by adding an identifiability penalty to the log-likelihood for all $j \in \mathcal{J}$

$$pen_{id,j} = -\kappa \left(\mu_{\epsilon_j}^2 + \left(\sigma_{\epsilon_j}^2 - 1 \right)^2 \right).$$

These identifiability penalties force the mean to be 0 and the variance to be equal to 1 when κ tends to infinity. Note also that σ_j^2 has to be less than one to ensure that $\sigma_{\epsilon_j}^2$ is not greater than 1.

3.5 Prior for the precisions and for the ODE parameters

For the conditional precision τ_j , $j \in \mathcal{J}$, of the vector of response \mathbf{y}_j , the prior can be chosen to be a gamma distribution. For the vector $\boldsymbol{\theta}$ of differential equation parameters, the chosen prior will depend on the context. Let us denote this prior distribution by $p(\boldsymbol{\theta})$.

4 Application

Theophylline is an anti-asthmatic agent administrated orally. Twelve subjects are given a oral dose at time 0 and 11 blood sample were taken on each subject the following 25 hours. A common model for the kinetics of Theophylline after oral administration is the one compartment model with first order absorption and elimination. The prior information available for the analysis of the Theophylline dataset concerns the initial condition of the state functions and the PK parameters. A total confidence is given on the initial condition, i.e. on the initial oral dose and on the null concentration of drug at the administration time. For the PK parameters, one has to force the constant of absorption to be greater than the constant of elimination. Figure 1 shows the pointwise posterior median of the estimated error density with 80% and 95% pointwise credibility intervals. It appears clearly that the error distribution is not Gaussian.

Acknowledgments: Financial support from the IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.

References

- Biegler, L., Damiano, J. and Blau, G. (1986). Nonlinear parameter estimation: A case study comparison: A case study comparison. *AIChE Journal*, **32**, 29–45.
- Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall.

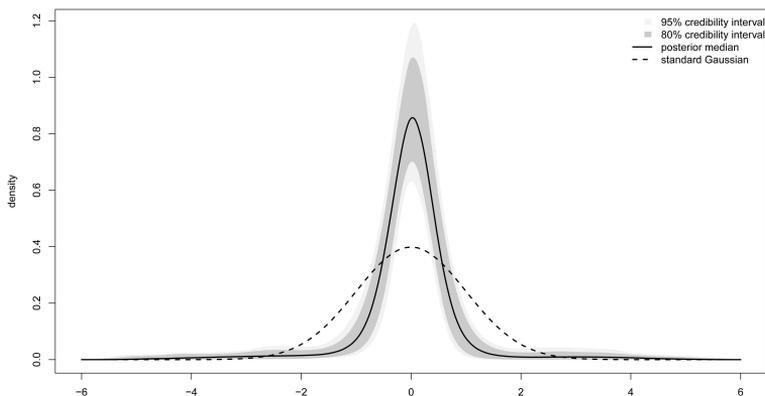


FIGURE 1. Fitted error density (solid curve is the pointwise posterior median and grey regions delimit pointwise 80% and 95% credible intervals), dotted curve is the Gaussian density.

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

Jaeger, J. and Lambert, P. (2011). Bayesian generalized profiling estimation in hierarchical linear dynamic systems. Technical Report 11001, IAP Statistics Network.

Komárek, A. and Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly interval censored data and flexible distributional assumptions. *Journal of the American Statistical Association*, **103**, 523–533.

Lambert, P. (2011). Nonparametric additive location-scale models for interval censored data. *Statistics and Computing*, doi: 10.1007/s11222-011-9292-6.

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

Lunn, D., Best, N., Thomas, A., Wakefield, J. and Spiegelhalter, D. (2002). Bayesian analysis of population PK/PD models: General concepts and software. *Journal of Pharmacokinetics and Pharmacodynamics*, **29**, 271–307.

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.

Modeling trait based ecological community assembly

Chaitanya Joshi¹, Daniel C. Laughlin², Peter M. Van Bodegom³, Zachary A. Bastow⁴, Peter Z. Fulé⁴

¹ Department of Statistics, University of Waikato, New Zealand.

² Department of Biological Sciences, University of Waikato, New Zealand.

³ Department of System Ecology, VU University, the Netherlands.

⁴ School of Forestry, Northern Arizona University, USA.

E-mail for correspondence: joshic@waikato.ac.nz

Abstract: Ecologists have long observed that phenotypic traits of species influence where they occur in the landscape (Schimper (1903), Grime(1979)) but only few attempts at modeling this phenomenon have been made (Shipley et al.(2006)). The maximum entropy (MaxEnt) model of trait-based community assembly was the first mathematical translation of environmental filtering (Shipley et al.(2006)). However, MaxEnt ignores phenotypic trait variation within species. To evaluate the importance of intraspecific trait variation in community assembly we have recently developed *Traitspace*, a new mathematical framework that explicitly models the filtering of individual-level plant traits through the environment. Rather than relying on trait means, we incorporate the full distribution of observed trait values. This approach allows species to overlap in trait space and allows individuals within species to differ. We use Traitspace to predict species relative abundance and also discuss the possible future developments of this model.

This theory is based on the premise that the traits that exist at a site were filtered by the environment, and that the relative abundance of each species is a function of its traits. Mathematically, this hierarchical dependence structure can be represented using a directed acyclic graph (DAG): $\mathbf{E} \longrightarrow \mathbf{T} \longrightarrow \mathbf{S}$, where E represents m -dimensional environmental gradients, T represents n -dimensional functional traits, and S is a vector of s unknown species S_1, \dots, S_s . We want to estimate the relative abundance of the i^{th} species at a given environment $P(S_i|E)$, by taking into account the information on the functional traits, i.e. by using $P(S_i|T, E)$. To do this, we first calibrate the conditional distribution of traits given species $\phi_{T|S_i}$ and also of traits given the environmental conditions $\phi_{T|E}$. We then use Bayesian methodology to estimate $P(S_i|T, E)$, and finally Monte Carlo integration to estimate $P(S_i|E)$.

Traitspace embraces the trait variation within species that has been a central focus of evolutionary biology. More accurate predictions could be obtained by including more environmental factors, including information regarding competition between the species and also which species established themselves first at

a particular location and Including stochastic processes such as dispersal and demography

Keywords: Bayesian Modeling; Monte Carlo Integration; Species Relative Abundance; Ecological Community Assembly.

Acknowledgments: This work was supported by a Joint Venture Agreement (08 – JV – 11221633 – 233) with the USDA Forest Service Rocky Mountain Research Station.

References

- Grime, J.P. (1979). *Plant Strategies and Vegetation Processes*.. Wiley, Chichester, UK.
- Schimper, A.F.W. (1903). *Plant-geography upon a physiological basis*.. Oxford: Clarendon Press.
- Shipley, B. and Vile, D. and Garnier, É. (2006). From Plant Traits to Plant Communities: A Statistical Mechanistic Approach to Biodiversity. *Science*, **314**, 812–814.

Flexible modeling of multivariate data by mixtures of Archimedean copulas

Göran Kauermann¹, Renate Meyer²

¹ Institut für Statistik, Ludwig-Maximilians-Universität München, Germany

² Department of Statistics, University of Auckland, New Zealand

E-mail for correspondence: `meyer@stat.auckland.ac.nz`

Abstract: Copulas allow for stochastic modelling of multivariate distributions with a flexibility well beyond that of the classical normal distribution. To further increase the versatility we propose the use of mixtures of different Archimedean copula families like Clayton, Frank, Gumbel, etc. Using a Bayesian approach, each family-specific parameter is modelled by imposing a prior distribution on the parameter. The mixture model itself is fitted in two ways: a fully Bayesian approach with MCMC-based posterior computation and a computationally much faster marginal likelihood approach using a penalized version of a classical quadrature which approximates the integrals. The performance of the new approach is evaluated on simulations and an example in the context of modelling the dependence structure of the log-returns of exchange rates.

Keywords: Archimedean Copula; Finite Mixture Model; Penalized Marginal Likelihood; Markov Chain Monte Carlo; Quadratic Programming.

1 Introduction

Realistic modeling of multivariate data requires complex statistical models with multivariate distributions. Sklar's Theorem (Sklar, 1959) enables a flexible modelling of multivariate distributions by separating the one-dimensional marginal cdf's from the dependence structure induced by the copula.

A *copula* C is a multivariate cdf on $[0, 1]^d$ whose marginals are all uniform on the interval $[0, 1]$. Concise mathematical treatments can be found in Joe (1997) and Nelsen (2006). Copulas and their applications have received growing interest in the last years, as for instance in financial econometrics (Embrechts, 2009) biostatistics (Bogaerts & Lesaffre, 2008) marketing (Danaher & Smith, 2011) and ecology, (Briggs et al., 2012). In this paper, we propose to further enhance the flexibility of modelling multivariate distributions by using *mixtures* of copulas. We extend the classical approach of mixture distributions using a finite number of components as reviewed in McLachlan and Peel (2000) and Marin (2005) to copulas.

Here we make use of Archimedean copulas of the form $C(u_1, \dots, u_p) = \varphi^{-1}(\varphi(u_1) + \dots + \varphi(u_p))$ where $\varphi(\cdot)$ is continuous and strictly decreasing from $[0, 1]$ to $[0, \infty)$ with $\varphi(1) = 0$, see Nelsen (2006). The function $\varphi(\cdot)$ is also called the generator function which is commonly parameterized by some low dimensional, often one-dimensional parameter θ , say. This paper is organized as follows. In Section 2, we propose to model the observations as a finite mixture of Archimedean copulas. Section 3 presents a Bayesian approach by specifying a prior distribution on the unknown parameters and implementing an MCMC approach for posterior computation. Section 4 presents a computationally faster frequentist alternative via a penalized marginal likelihood approach. The performance of the new frequentist approach is compared to the Bayesian approach in a case study in Section 5. We conclude with a discussion in Section 6.

2 Mixture Copula

We consider a set of Archimedean copula families labelled with index $j = 1, \dots, J$. Hence, index j refers to a particular generator function $\varphi_j(\cdot|\theta_j)$ so that $C_j(\mathbf{u}|\theta_j) = \varphi_j^{-1}(\sum_{i=1}^p \varphi_j(u_i|\theta_j)|\theta_j)$ is the resulting copula, where $\mathbf{u} = (u_1, \dots, u_p) \in [0, 1]^p$, where θ_j is low dimensional and for the sake of simplicity assumed to be univariate. To be specific, θ_j expresses the correlation induced by the copula as measured by Kendall's correlation coefficient, i.e.

$$\theta = 4 \int_0^1 \dots \int_0^1 C(u_1, \dots, u_p) dC(u_1, \dots, u_p) - 1.$$

Let $c_j(\mathbf{u}|\theta_j)$ denote the copula density relating to $C_j(\mathbf{u}|\theta_j)$. The joint copula itself is now constructed from a discrete mixture of the J different copula families. We assume that the j th copula family has a mixing probability $\pi_j \geq 0$, with $j = 1, \dots, J$ and $\sum_j \pi_j = 1$, which yields the following discrete mixture model

$$c(\mathbf{u}|\theta_1, \dots, \theta_J, \pi_1, \dots, \pi_J) = \sum_{j=1}^J \pi_j c_j(\mathbf{u}|\theta_j). \quad (1)$$

The sampling distribution in (1) depends on parameters $\theta = (\theta_1, \dots, \theta_J)$ and $\pi = (\pi_1, \dots, \pi_J)$.

3 Bayesian Mixture Model

A Bayesian model requires the specification of a prior distribution $f_\pi(\pi)$ and $f_\theta(\theta)$ for the parameters. In the absence of prior information on the copula parameters, noninformative uniform priors can be chosen for each

copula family parameter, e.g. Uniform(0,1) for the parameter of the Clayton copula that has parameter space $(0, 1]$. Alternatively, if prior information is available, an informative Beta prior distribution could be employed. We used the uniform distribution in all examples below and a Dirichlet(η_1, \dots, η_J) prior distribution for the mixture probabilities with $\eta_j = 1, j = 1, \dots, J$, so that the prior means of the π_j are all equal to $1/J$.

The posterior distribution is proportional to

$$f(\theta_1, \dots, \theta_J, \pi_1, \dots, \pi_J | \mathbf{u}) \propto \prod_{i=1}^n \left(\sum_{j=1}^J \pi_j c_j(\mathbf{u}_i | \theta_j) \right) \prod_{j=1}^J \pi_j^{\eta_j - 1} f_j(\theta_j). \quad (2)$$

We suggest the use of the Metropolis-Hastings algorithm to sample from the posterior distribution and make use of the R package `copula` (see Yan, 2007). Due to the evaluation of the density of a high-dimensional copula density in each iteration, the MH algorithm requires a considerable amount of computation time.

In the following section, we propose a penalized marginal likelihood approach with an empirical Bayesian flavour that circumvents the need for extensive and time-consuming MCMC simulations.

4 Penalized Marginal Likelihood Estimation

For the j th family, the *marginal* copula density can be calculated by integrating out the family-specific θ_j parameter through

$$c_j(\mathbf{u}) = \int_{\Theta_j} c_j(\mathbf{u} | \theta_j) f_j(\theta_j) d\theta_j. \quad (3)$$

The resulting marginal mixture copula density is then given by

$$c(\mathbf{u} | \pi_1, \dots, \pi_J) = \sum_{j=1}^J \pi_j c_j(\mathbf{u}) = \sum_{j=1}^J \pi_j \int_{\Theta_j} c_j(\mathbf{u} | \theta_j) f_j(\theta_j) d\theta_j. \quad (4)$$

Instead of specifying priors $f_j(\cdot)$ as in the previous section, we now pursue a frequentist approach and intend to estimate both $\boldsymbol{\pi}$ and $f_j(\cdot)$ appropriately based on the random sample $\mathbf{u}_1, \dots, \mathbf{u}_n$. That is, we do not treat $f_j(\cdot)$ as given priors for $j = 1, \dots, J$ but let these be part of the likelihood to be maximized. This is carried out by approximating the integral in (4) by a discrete quadrature formula as in Komárek and Lesaffre (2008) resulting in the log likelihood

$$l(\mathbf{w}) = \sum_{i=1}^n \log \left(\sum_{j=1}^J \sum_{k=1}^{K_j} w_{jk} c(\mathbf{u}_i | \theta_{jk}) \right) \quad (5)$$

with $\mathbf{w} = (w_{jk}, j = 1, \dots, J, k = 1, \dots, K_j), \sum_{k=1}^{K_j} w_{jk} = \pi_j$, and $\theta_{j0}, \theta_{j1}, \dots, \theta_{jK_j}, \theta_{jK_{j+1}}$ a set of equidistant knots covering the parameter space Θ_j . In this case we treat \mathbf{w} as a high dimensional parameter to be estimated from the data. To stabilize the resulting likelihood it is therefore advisable to impose some penalty on \mathbf{w} . Following the idea of Eilers and Marx (1996) we express the smoothness assumption by a difference penalty based on the q -th order difference matrix. To maximize the penalized likelihood subject to the linear constraints on \mathbf{w} , we use quadratic programming.

5 Case Study

A simulation study to demonstrate the applicability of the novel penalized marginal likelihood approach was performed and showed convincing performance. These results are not shown here because of space constraints. Instead, we compare this approach to a fully Bayesian approach when modelling the dependence structure of the log-returns of exchange rates as a mixture of copulas. We use the dataset of daily exchange rates of the US Dollar to the six currencies: Euro (EUR), Canadian Dollar (CAD), Swiss Franc (CHF), Japanese Yen (JPY), Singapore Dollar (SGD) and British Pound (GBP) from Jan 2000 to May 2011 obtained from the Federal Reserve System (www.federalreserve.gov). To fit and evaluate the model we exclude the last 400 observations. Table 1 provides Kendall's correlations of the log-returns. As marginal distributions we assume for simplicity t-distributions with 5 degrees of freedom and calculate $u_{ij} := F^{-1}(x_{ij})$ with $F(\cdot)$ as t-distribution and x_{ij} as daily log-return at timepoint i of the exchange to currency $j, j = 1, \dots, 6$ and $i = 1, \dots, 2349$.

	EUR	CAD	CHF	JPY	SGD	GBP
EUR	1.00	0.28	0.72	0.22	0.37	0.49
CAD	0.28	1.00	0.23	0.04	0.26	0.24
CHF	0.72	0.23	1.00	0.30	0.32	0.43
JPY	0.22	0.04	0.30	1.00	0.26	0.15
SGD	0.37	0.26	0.32	0.26	1.00	0.29
GBP	0.49	0.24	0.43	0.15	0.29	1.00

TABLE 1. Kendall's correlation coefficients for log-returns of daily exchange rates for six currencies to the US Dollar (for EUR and GBP we record negative changes).

Table 2 gives summary statistics of the posterior distribution of mixture weights π_c, π_f, π_g and π_I as well as the copula parameters obtained via the Metropolis-Hastings algorithm.

Parameter	Mean	Sd	MC error	2.5%	Median	97.5%
π_c	0.299	0.030	0.003	0.244	0.298	0.358
π_f	0.066	0.025	0.003	0.017	0.062	0.115
π_g	0.534	0.031	0.003	0.472	0.536	0.592
π_I	0.106	0.018	0.002	0.067	0.100	0.139
θ_c	0.282	0.027	0.002	0.236	0.279	0.345
θ_f	0.728	0.033	0.004	0.675	0.725	0.813
θ_g	0.539	0.014	0.001	0.509	0.540	0.567

TABLE 2. Summary statistics of 10,000 samples from the posterior distribution of the mixture weights π_c for 6-dim. Clayton, π_f for Frank, π_g for Gumbel copula and π_I for the independence density, and the corresponding copula parameters $\theta_c, \theta_f, \theta_g$.

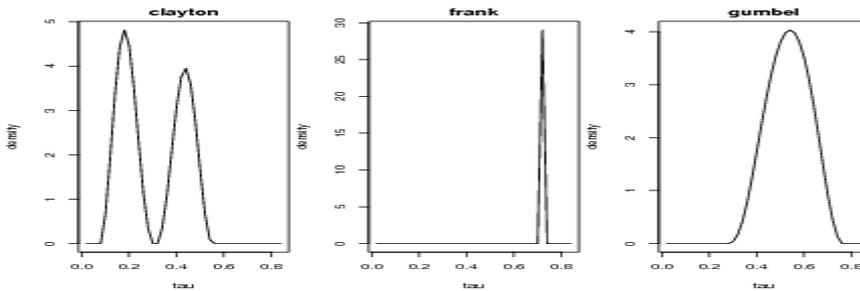


FIGURE 1. Penalized marginal likelihood estimates of densities $f_j(\theta)$ for the three mixture components for log returns of exchange rates.

The resulting penalized marginal likelihood estimates of the mixture probabilities are $\hat{\pi}_c = 0.335$, $\hat{\pi}_f = 0.018$, $\hat{\pi}_g = 0.579$ and $\hat{\pi}_I = 0.068$, so that the major mixture components are again Clayton and Gumbel copulas.

In Figure 1 we visualize the resulting fitted $f_j(\theta_j)$ for the dominating copulas Clayton, Frank and Gumbel. Interestingly, the penalized quadrature fit for the Clayton family is bimodal and the fitted prior for the Frank copula is spiky. To evaluate the model we take the 400 last observations not used for fitting and calculate the empirical lower and upper tail dependence resulting to $\hat{\lambda}_{\text{emp,L}} = 0.02$ and $\hat{\lambda}_{\text{emp,U}} = 1.65$, respectively. We compare these numbers to the tail dependencies taking the fitted mixture model. These result through $\hat{\lambda}_L = 0.01$ and $\hat{\lambda}_U = 1.28$, which give comparable numbers.

6 Discussion

The paper proposes the use of mixtures of Archimedean copulas as a flexible tool for modelling multivariate data. As a faster alternative to the Bayesian approach via MH-algorithms, we suggest an alternative frequen-

tist technique that is based on maximizing a penalized marginal likelihood. Testing on simulated as well as real data, the penalized marginal likelihood approach yields results that are comparable to those obtained from a fully Bayesian approach but can be obtained in a fraction of the time needed for MCMC algorithms to converge.

References

- Bogaerts, K., and Lesaffre, E. (2008). Modeling the association of bivariate interval-censored data using the copula approach. *Statistics in Medicine* **27** (30), 6379–6392.
- Briggs, J., Dowd, M., Meyer, R. (2012). Data assimilation in large scale spatio-temporal systems via a location particle smoother, Department of Statistics, University of Auckland, preprint.
- Danaher, P, and Smith, M.S. (2011). Modeling Multivariate Distributions Using Copulas: Applications in Marketing. *Marketing Science* **30**, 4–21.
- Eilers, P.H.C., and Marx, B.D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science* **11**, 89–121.
- Embrechts, P. (2009). Copulas: A personal view. *Journal of Risk and Insurance* **76** (3), 639–650.
- Joe, H. (1996). Families of m -variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In: *Distributions with Fixed Marginals and Related Topics*. Rüschendorf, L. and Schweizer, B. and Taylor, M.D.
- Komárek, A., and Lesaffre, E. (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random-effects distribution. *Computational Statistics and Data Analysis*, **52**, 3441–3458.
- Marin, J.M., Mengersen, K., and Robert, C.P. (2005). Bayesian modeling and inference on mixtures of distributions. In: Dey, D., Rao, C. (Eds.), *Handbook of Statistics*. North Holland, Amsterdam.
- McLachlan, G., and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.
- Nelsen, R. (2006). *An Introduction to Copulas*. Springer, Berlin.
- Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229–231.
- Yan, J. (2007). Enjoy the Joy of Copulas: With a Package copula. *Journal of Statistical Software* **21**, 1–21.

Nonparametric estimation of conditional Archimedean copula

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: Copulas enable to specify a p -variate distribution in two independent steps. The first one fixes the p marginal distributions while the second step determines the dependence structure of the marginal quantiles. When covariates are available, one usually assumes that they only affect the marginal responses without modifying the strength of dependence between them. Here, we extend these models by taking a nonparametric specification for the copula and by letting it change smoothly with covariates. Bayesian P-splines (Lang & Brezger 2004 ; Jullion & Lambert 2007) combined with carefully tuned MCMC algorithms are the key ingredients. The methodology is illustrated by analyzing the effect of age on the height and weight of boys and their dependence structure.

Keywords: Conditional copula; nonparametric Archimedean copula ; Bayesian P-splines.

1 Introduction

Sklar (1959) has shown that a multivariate distribution is fully characterized by its marginals and a copula. For ease of presentation, let us focus on the bivariate continuous case. If $H(y_1, y_2)$ denotes the joint distribution of $(Y_1, Y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$ with marginal distributions $F(y_1)$ and $G(y_2)$, then there exists a unique distribution function $C(\cdot, \cdot)$ on $[0, 1]^2$ (named *copula*) such that $H(y_1, y_2) = C(F(y_1), G(y_2))$ for all $(y_1, y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$.

Archimedean copulas were introduced by Genest & MacKay (1986). These are characterized by a continuous, strictly decreasing and convex function φ with $\varphi(1) = 0$ such that $C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$ for any (u, v) in $[0, 1]^2$. Popular choices for the generator are given in Table 1. The family defines the dependence structure with a parameter θ tuning the strength of dependence between the variates. It can be connected to a familiar rank-based concordance measure such as Kendall's tau using $\tau = 1 + 4 \int_0^1 \varphi(u)/\varphi'(u^+) du$.

TABLE 1. Examples of parametric Archimedean copulas.

Family	θ range	Generator	Kendall's tau
Frank	$(-\infty, +\infty)/\{0\}$	$-\log \frac{e^{-\theta u}-1}{e^{-\theta}-1}$	$1 - \frac{4}{\theta} \left(1 - \frac{1}{\theta} \int_0^\theta \frac{t}{e^t-1} dt\right)$
Clayton	$[-1, +\infty)/\{0\}$	$(u^{-\theta} - 1)/\theta$	$\theta/(\theta + 2)$
Gumbel	$[-1, +\infty)$	$(-\log u)^\theta$	$(\theta - 1)/\theta$

2 Nonparametric copula

Depending on the choice made for the parametric copula, different dependence structures will be obtained with, for example for the Gumbel copula, a loose dependence between smaller quantiles and a much stronger (positive) concordance for larger ones. Therefore, in a modelling exercise, besides a suitable specification of the margins, one must also select the copula family and the value of its dependence parameter. Alternatively, nonparametric forms for the copula could be preferred or used to help in the selection of a parametric one (Genest & Rivest, 1993): the use of kernels (Gijbels & Mielniczuk, 1990), local polynomials (see e.g. Chen & Huang, 2007) or splines (Lambert, 2007) have been advocated.

We propose to extend the specification in Lambert (2007) by combining the flexibility of splines with the ideas in Vandenhende & Lambert (2005) where (minus) the log of the generator of an Archimedean copula was written as a piecewise linear function of the log generator of the independence copula, $S(u) = -\log(-\log(u))$. We suggest to take as flexible "nonparametric" form for the Archimedean generator, $\varphi(u) = \exp\{-g(S(p)|\alpha)\}$, where $g'(s|\alpha) = \sum_k b_k(s)\alpha_k$ is a linear combination of cubic B-splines associated to a large number of equidistant knots within $(0,1)$ and to two further extreme knots at $\epsilon (=10^{-8}, \text{ say})$ and $1 - \epsilon$. Some constraints on α are required to have a valid generator. Combining the likelihood with a roughness penalty on r th order differences of

$$\{\log \alpha_2, \dots, \log \alpha_{K-1}\},$$

one can obtain a penalized likelihood estimate for α . Alternatively, a Bayesian P-splines approach can be used (Lang & Brezger, 2004 ; Jullion & Lambert, 2007).

3 Non parametric conditional copula

The effect of a covariate X on the bivariate distribution of Y_1 and Y_2 can be specified in different ways. In its simplest form, one might assume that X only affects the marginal distributions of $(Y_1, Y_2|X)$, in which case $H(y_1, y_2|x) = C(F(y_1|x), G(y_2|x))$. Then, the dependence structure, characterized by the copula, is not sensitive to the value of X .

More generally, one could assume that the dependence structure is also changing with X : $H(y_1, y_2|x) = C_x(F(y_1|x), G(y_2|x))$. 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_x was coined.

Assume that the copula is Archimedean. If it is parametric with generator $\varphi(\cdot|\theta)$, then a flexible relation between the copula parameter $\theta = \theta(x)$ and the covariates can be specified, see Acar *et al.* (2011) who work in a local polynomial framework. Alternatively, a function of θ could be written as a linear combination of B-splines.

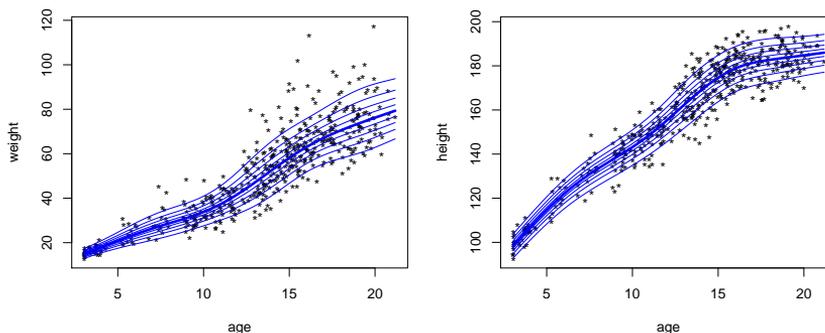
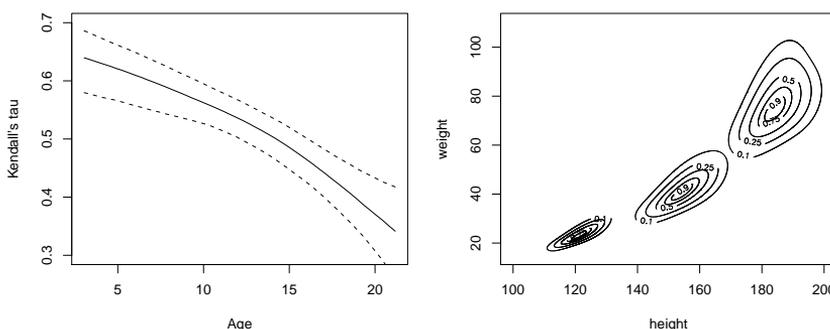
We propose to go one step further by starting from the nonparametric specification at the end of Section 2. Using the relation between Kendall's τ and the copula generator, one can show that multiplying the spline coefficients by some (suitable) constant divides the probability of non-concordance by the same amount. Hence, we propose to let the nonparametric generator of the Archimedean copula change in a smooth way with covariate x by taking $\varphi(\cdot|\alpha(x))$ where $\alpha_k(x) = \exp(\eta(x))\alpha_k$ with $\eta(x)$ expressed as a linear combination of B-splines on the range of the covariate values. If multiple covariates are available, an additive model can be considered, see Lambert (2011) for a recent example based on Bayesian P-splines. It induces a regression model for Kendall's tau.

4 Application

The data of interest relates to the growth of 490 boys in the Netherlands. Our goal is to study how height ($=Y_1$; 160.8 ± 24.89 cm) and weight ($=Y_2$; 51.6 ± 20.33 kg) (and their association) change with age ($=X$; 13.4 ± 4.47 yrs). Flexible additive location-scale regression models for $(Y_1|X)$ and $(Y_2|X)$ were fitted following Lambert (2011):

$$Y_j = f_j^\mu(X) + \exp(f_j^\sigma(X))\varepsilon_j,$$

where $f_j^\mu(X)$ and $f_j^\sigma(X)$ denote the conditional median and log interquartile range of Y_j given X , respectively, and ε_j has a distribution corresponding to a smooth unknown density (Lambert & Eilers, 2009). Rich B-splines bases were used to describe the unknown functional components in the models. The flexibility of these bases was counterbalanced by a roughness penalty using the prior of the spline parameters in a Bayesian framework (Jullion & Lambert, 2007). The fitted deciles of **weight** and **height** given **age** can be seen on Fig. 1. It reveals nonlinear patterns and an increasing heterogeneity of the responses with age. The evolution of the fitted

FIGURE 1. Fitted deciles for `weight` and `height` given `age`.FIGURE 2. Left: fitted Kendall's tau (solid) given `age` with 90% pointwise credible region. Right: contours of the fitted bivariate density of `(height,weight)` for (from left to right) `age` equal to 6, 12 and 20 years.

Kendall's tau can be seen on Fig. 2 (left panel): the concordance between height and weight strongly decreases with age. The contours of the fitted bivariate distributions of `(height,weight)` for boys aged 6, 12 and 20 years are plotted on the same figure (right panel). These distributions are smoothly changing with the covariate: it clearly illustrates that the dependence between the responses is getting looser as age increases.

5 Discussion

A new nonparametric procedure for the estimation of an Archimedean copula has been described. It has been extended to a conditional setting by letting the nonparametric generator of the copula change smoothly with a single covariate. The extension of this model to deal with multivariate covariates in an additive setting will be reported elsewhere.

References

- Acar, E.F., Craiu, R.V. and Yao, F. (2011) Dependence calibration in conditional copulas: a nonparametric approach. *Biometrics*, **67**: 445-453.
- Chen, S.X. and Huang, T.-Z. (2007) Nonparametric estimation of copula functions for dependence modelling *Canad. J. Statist.*, **35**: 265-282.
- Genest, C. and MacKay, J. (1986) Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données. *Canad. J. Statist.*, **14**: 145-159.
- Genest, C. and Rivest, L.-P. (1993) Statistical inference procedures for bivariate archimedean copulas. *J. Amer. Statist. Assoc.*, **88**, 1034-1043.
- Gijbels, I. and Mielniczuk, J. (1990) Estimating the density of a copula function. *Comm. Statist. Theory Methods*, **19**: 445-464.
- Jullion A. and Lambert P. (2007) Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics and Data Analysis* 51, 2542-2558.
- Lambert, P. (2011) Nonparametric additive location-scale models for interval censored data. *Statistics and Computing*, DOI: 10.1007/s11222-011-9292-6.
- Lambert, P. and Eilers, P.H.C. (2009) Bayesian density estimation from grouped continuous data. *Computational Statistics and Data Analysis*, **53**: 1388-1399.
- 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.
- Lang, S., Brezger, A. (2004) Bayesian P-splines. *J. Comput. Graphical Statist.*, **13**, 183-212.
- 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.

Upon closer inspection... Testing in comparative experiments

Joseph B. Lang¹

¹ Dept. of Statistics and Actuarial Science, Univ. of Iowa, Iowa City, IA USA

E-mail for correspondence: joseph-lang@uiowa.edu

Abstract: Standard tests of the “no-treatment-effect” hypothesis for a comparative experiment include permutation tests, the Wilcoxon rank sum test, two-sample t tests, and Fisher-type randomization tests. Practitioners are aware that these procedures test different no-effect hypotheses and are based on different modeling assumptions. However, this awareness is not always, or even usually, accompanied by a clear understanding or appreciation of these differences. Borrowing from the rich literatures on causality and finite-population sampling theory, this paper develops a modeling framework that affords answers to several important questions, including: exactly what hypothesis is being tested?, what model assumptions are being made?, and are there other, perhaps better, approaches to testing a no-effect hypothesis? The framework lends itself to clear descriptions of three main inference approaches: science-based, randomization-based, and selection-based. It also promotes careful consideration of model assumptions and targets of inference, and highlights the importance of randomization. Along the way, Fisher-type randomization tests are compared to permutation tests and a less well known Neyman-type randomization test. A small-scale simulation study compares the Neyman test to those of the other more familiar tests.

Keywords: Fisher vs. Neyman; No-treatment-effect hypotheses; Potential variables; Randomization-based inference; Science-based inference

1 Introduction

Sixty-four college students were enlisted to take part in a randomized comparative experiment to determine whether cell phone use while driving slows reaction times (Strayer and Johnston 2001). Of the 64 students, 32 were randomized to treatment 1 (drive while using a cell phone) and 32 were randomized to treatment 2 (drive without a cell phone). The reaction times, in milliseconds, for the 64 participants are recorded in Table 1. Is there a cell phone use effect? Generically, is there a *treatment effect*?

Standard tests of the “no-treatment-effect” hypothesis include permutation tests, the Wilcoxon rank sum test, two-sample t tests, and Fisher-type randomization tests. Practitioners are aware that these procedures test different “no-effect” hypotheses and are based on different modeling assumptions. However, this awareness is not always, or even usually, accompanied by a clear understanding or appreciation of these differences. This paper looks at each of these testing approaches and addresses the all

TABLE 1. Reaction Times (milliseconds).

Cell Phone:	636	623	615	672	601	600	542	554	543	520	609	559	595	565	573	554
	626	501	574	468	578	560	525	647	456	688	679	960	558	482	527	536
Control:	557	572	457	489	532	506	648	485	610	444	626	626	426	585	487	436
	642	476	586	565	617	528	578	472	485	539	523	479	535	603	512	449
Generically...																
Treatment 1:	$y_{1,1}$	$y_{1,2}$	\dots	$y_{1,32}$												
Treatment 2:	$y_{2,1}$	$y_{2,2}$	\dots	$y_{2,32}$												

important questions, exactly what hypothesis is being tested? and what model assumptions are being made? Along the way, we will have to confront several other questions such as, how is the definition of *treatment effect* operationalized?, what is the actual target of inference?, what is the role of randomization?, and are there other, perhaps better, approaches to testing a no-effect hypothesis?

To address these questions, we draw on ideas from the rich literature on causal analysis. In particular, we employ the useful concept of “potential variables.” Although the idea of potential variables can be traced back to Neyman (1923), Rubin, beginning with a series of papers on causal models in the 1970’s (see Rubin 2010 and references therein) is usually credited with more explicitly stating the potential variable model and extending it to both randomized and non-randomized design settings, with or without covariates (see Rubin’s causal model, Holland 1986).

To be clear, it is not the goal of this paper to summarize the vast literature on potential variables and causal modeling. Instead, the first goal is to exploit the benefits of hindsight to develop a modeling framework that supports clear descriptions and comparisons of the different testing approaches, and promotes careful consideration of the model assumptions and targets of inference. This modeling framework and associated notation draws clear distinctions between realizations and random variables, and between observed and unobserved data. It accommodates both treatment assignment and sampling from populations, and clearly differentiates between the two. Although the proposed model lends itself to generalizations in many directions, to simplify exposition, we will focus on the two-treatment comparative experiment setting.

The second goal of this paper is to address the question of availability of other testing approaches, besides the four common ones mentioned above. Toward this end, we revisit ideas introduced in Neyman (1923). Using the model structure introduced herein, we describe a less well known Neyman-type randomization test, which is qualitatively different than the Fisher-type randomization test (cf. Welch 1937, Rubin 2010). The Neyman-type randomization test, which uses a less restrictive “no-effect” hypothesis than Fisher’s, is based on a test statistic with the common form, (estimator

minus estimand)/(standard error of estimator). Neyman, with an eye on interval estimation rather than testing, derived the standard error with respect to a randomization distribution using tools from finite-population sampling theory. We argue that compared to Fisher-type randomization tests, the Neyman tests do have their advantages and disadvantages.

The third and final goal of this paper is to compare the operating characteristics of the five tests: the permutation test, the Wilcoxon rank sum test, the two-sample t -test, the Fisher-type randomization test, and the Neyman-type randomization test. Toward this end, we carried out a small-scale simulation study of the size and power of these five tests. Based on these comparisons, we make tentative recommendations on which test to use in different settings.

The balance of this paper highlights the main ideas used to achieve the aforementioned goals. We introduce potential variables and recast the data of Table 1 within this framework. A three-level sequential data generation model explicitly accommodates both random sampling and randomization. This model framework lends itself to explicit identification of three main targets of inference, a variety of “no-effect” hypotheses, and three main inference approaches—science-based, selection-based, and randomization-based. We give examples of tests for each of the inference approaches, some of them well known and some of them less well known. We analyze the cell phone data and report on a small-scale simulation study that compares the different testing approaches described herein.

2 Potential Variables

Going back to Neyman (1923), and following the lead of Rubin (e.g. 2005), we will view the data as observed values of a sample of “potential values.” Let $Y_{t,i}$ be the response for unit i when exposed to treatment t , where i is in population $\underline{P} = (1, \dots, N)$ and $t = 1, 2$. The response variables $Y_{1,i}$ and $Y_{2,i}$ are called potential variables because we can observe one or the other, but not both. Strictly speaking, it is not possible to observe the values of both potential variables because the same subject cannot be simultaneously exposed to both treatments. To the potential variable advocates, this is the “fundamental problem of causal analysis” (Holland, 1986).

The data in Table 1 can be viewed as observed values of a sample of the potential variable values. Let y_{t,s_j} be the response value for sampled subject s_j when exposed to treatment t . That is, y_{t,s_j} is a realization of Y_{t,s_j} . Of course, for each subject s_j , only one of the realizations, y_{1,s_j} or y_{2,s_j} , will be observed. From a potential variables viewpoint, the original data in Table 1 can be viewed as in Table 2.

TABLE 2. Reaction Time Data (Potential Values Viewpoint)

Treatment 1:	$\mathcal{Y}_{1.s_1}$,	$\mathcal{Y}_{1.s_2}$,	$y_{1.s_3}$,	\dots ,	$\mathcal{Y}_{\cancel{s}_{63}}$,	$y_{1.s_{64}}$
Treatment 2:	$y_{2.s_1}$,	$y_{2.s_2}$,	$\mathcal{Y}_{\cancel{s}_{63}}$,	\dots ,	$y_{2.s_{63}}$,	$\mathcal{Y}_{\cancel{s}_{64}}$

Only the 32 non- \times 'ed out values are observed for each treatment and $\underline{s} = (s_1, \dots, s_{64})$ is a sample from some population \underline{P} .

3 Data-Generation Models and Inference Goals

Let $\underline{Y} = (Y_{1.1}, \dots, Y_{1.N}, Y_{2.1}, \dots, Y_{2.N})$ be the vector of potential variables for the population \underline{P} and $\underline{y} = (y_{1.1}, y_{1.2}, \dots, y_{1.N}, y_{2.1}, \dots, y_{2.N})$ be the corresponding vector of realizations. To simplify vector component identification, define $\underline{x}.\underline{w} = (x_1.w_1, \dots, x_m.w_m)$, $k.\underline{x} = (k.x_1, \dots, k.x_m)$, and let $\underline{x}[\underline{b}]$ represent the collection of components in \underline{x} that have subscripts in \underline{b} . As examples, \underline{y} can be expressed as $\underline{y} = \underline{y}[1.\underline{P}, 2.\underline{P}]$. Similarly $\underline{y}[1.\underline{s}] = (y_{1.s_1}, \dots, y_{1.s_n})$ and $\underline{y}[\underline{t}.\underline{s}] = (y_{t_1.s_1}, \dots, y_{t_n.s_n})$. We will also use a notation for averages: $\bar{Y}[t.\underline{P}] = N^{-1} \sum_{i=1}^N \underline{Y}[t.i]$, $\bar{y}[t.\underline{P}] = N^{-1} \sum_{i=1}^N \underline{y}[t.i]$, and $\bar{y}[t.\underline{s}] = n^{-1} \sum_{j=1}^n \underline{y}[t.s_j]$.

The data-generation model is based on the following sequential generations:

$$\begin{array}{ll}
 \underline{y} \leftarrow \underline{Y} & \text{Here, } \underline{y} = (y_{1.1}, \dots, y_{1.N}, y_{2.1}, \dots, y_{2.N}) \\
 \underline{s} \leftarrow \underline{S} \mid (\underline{Y} = \underline{y}) & \text{Here, } \underline{s} = (s_1, \dots, s_n), s_j \in \underline{P}, s_j \neq s_{j'} \\
 \underline{t} \leftarrow \underline{T} \mid (\underline{Y} = \underline{y}, \underline{S} = \underline{s}) & \text{Here, } \underline{t} = (t_1, \dots, t_n), t_j \in \{1, 2\}.
 \end{array}$$

Borrowing from Rubin (2005), we will refer to the potential variables \underline{Y} and values \underline{y} as the “science,” to differentiate them from the “selection” variables $(\underline{S}, \underline{T})$ and values $(\underline{s}, \underline{t})$. The science portion describes how things behave in the two parallel worlds corresponding to the two treatments and the selection portion determines how we go about observing this behavior. Owing to the sampling and treatment randomization (the selection) and the fundamental problem of causal inference, we do not observe the entire $2N$ -dimensional vector of potential deviates \underline{y} (the science). Instead we observe only the n -dimensional sub-vector $\underline{y}[\underline{t}.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}]$. The inference goal of this paper can be stated succinctly as follows...

Inference Goal. Use the observed data $\underline{y}[\underline{t}.\underline{s}]$ from a comparative experiment to reduce uncertainty about one of the three targets: the vector $\underline{y}[1.\underline{s}, 2.\underline{s}]$, the vector $\underline{y}[1.\underline{P}, 2.\underline{P}]$, or the distribution of \underline{Y} .

4 “No-Treatment-Effect” Hypotheses

In a comparative experiment, a treatment effect can be based on head-to-head comparisons of the potential variables $\underline{Y}[1.i]$ and $\underline{Y}[2.i]$, or realizations thereof (cf. Cox 1958, Rubin 2010). Corresponding to each candidate

definition of a treatment effect is a “no-treatment-effect” hypothesis. As examples, we will consider $H_0^U : \underline{Y}[1.i] = \underline{Y}[2.i]$; $H_0^{EU} : \underline{Y}[1.i] \sim \underline{Y}[2.i]$; $H_0^{EU.1} : E(\underline{Y}[1.i]) = E(\underline{Y}[2.i])$; $H_0^{RU} : \underline{y}[1.i] = \underline{y}[2.i]$; $H_0^{RA} : \bar{y}[1.\underline{P}] = \bar{y}[2.\underline{P}]$; $H_0^{RU.s} : \underline{y}[1.s_j] = \underline{y}[2.s_j]$; $H_0^{RA.s} : \bar{y}[1.\underline{s}] = \bar{y}[2.\underline{s}]$.

5 Inference Approaches

Science-Based Inference. With the science-based approach, we condition on the selection and use

$$\underline{y}[t.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}] \mid (\underline{S} = \underline{s}, \underline{T} = \underline{t}) \sim \underline{Y}[t.\underline{s}] \mid (\underline{S} = \underline{s}, \underline{T} = \underline{t})$$

to carry out inferences about the distribution of \underline{Y} . The permutation and Wilcoxon rank sum tests are science-based approaches to testing H_0^{EU} . The two-sample t -tests are science-based approaches to testing $H_0^{EU.1}$.

Selection-Based Inference. With the selection-based approach, we condition on the science and use

$$\underline{y}[t.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}] \mid (\underline{Y} = \underline{y}) \sim \underline{y}[\underline{T}.\underline{S}] \mid (\underline{Y} = \underline{y})$$

to carry out inferences about $\underline{y}[1.\underline{P}, 2.\underline{P}]$. The Neyman-type selection test is a selection-based approach to testing H_0^{RA} .

Randomization-Based Inference. With the randomization-based approach, we condition on both the science and the sample and use

$$\underline{y}[t.\underline{s}] \leftarrow \underline{Y}[\underline{T}.\underline{S}] \mid (\underline{Y} = \underline{y}, \underline{S} = \underline{s}) \sim \underline{y}[\underline{T}.\underline{s}] \mid (\underline{Y} = \underline{y}, \underline{S} = \underline{s})$$

to carry out inferences about $\underline{y}[1.\underline{s}, 2.\underline{s}]$. The Fisher-type randomization test and the Neyman-type randomization test are randomization-based approaches to testing $H_0^{RU.s}$ and $H_0^{RA.s}$, respectively.

The less well known Neyman-type selection and randomization tests are based on Horvitz-Thompson (Horvitz and Thompson 1952) unbiased estimators of $\bar{y}[1.\underline{P}] - \bar{y}[2.\underline{P}]$ and $\bar{y}[1.\underline{s}] - \bar{y}[2.\underline{s}]$, respectively. Assumptions about $(\underline{Y}, \underline{S}, \underline{T})$ for the validity of each test approach are easily stated.

6 Empirical Results and Conclusions

Because the 64 participants in the cell phone study are not a random sample from any substantively interesting population, it is most reasonable to carry out randomization-based inference. The Fisher- and Neyman-type randomization p-values are 0.0074 and 0.0075. Because the unbiased estimate of $\bar{y}[1.\underline{s}] - \bar{y}[2.\underline{s}]$ is $51.59 > 0$, there is statistical evidence that reaction times are slower when cell phones are used, at least for this sample of 64. Results of the simulation study suggest that the Neyman-type randomization test has size close to the nominal target even when the sample sizes

are small, with one exception. The Normal approximation to the Neyman statistic is unreasonable when the sample sizes are small and there are extreme outliers. In all of the simulation scenarios, the Neyman-type randomization test had higher power than the Fisher version. On the basis of this limited simulation study, we recommend that practitioners give serious consideration to using the Neyman-type randomization test as an alternative to the Fisher-type randomization test, especially for moderate sample size when there are no extreme outliers.

In the binary response, comparative experiment setting, we give conditions under which *Fisher's exact test* for 2×2 tables is equivalent to the Fisher-type randomization test of $H_0^{RU.s}$ and the permutation test of H_0^{EU} . We note that the Neyman-type randomization test is also available for testing the no-treatment-effect hypothesis $H_0^{RA.s} : \bar{y}[1.\underline{s}] = \bar{y}[2.\underline{s}]$ in 2×2 tables. The simulation results suggest that this Neyman test for 2×2 tables may be somewhat more powerful than Fisher's exact test.

Acknowledgments: Research supported in part by NSF grant SES-1059955.

References

- Cox, D.R. (1958). *The Planning of Experiments*. New York: Wiley.
- Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945–968.
- Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments, essay on principles, section 9. *Roczniki Nauk Rolniczych Tom X* [in Polish]; English translation of excerpts by D.M. Dabrowska and T.P. Speed (1990), *Statistical Science*, **5**, 463–472.
- Rubin, D.B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association*, **100**, 322–331.
- Rubin, D.B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, **15**, 38–46. doi: 10.1037/a0018537
- Strayer, D.L. and Johnston, W.A. (2001). Driven to distraction: dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science*, **12**, 462–466.
- Welch, B.L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika*, **29**, 21–52.

Seasonal modulation smoothing mixed models for times series forecasting

Dae-Jin Lee¹, María Durbán²

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

Abstract: We propose an extension of a seasonal modulation smooth model with P -splines for times series data using a mixed model formulation. A smooth trend with seasonality decomposition can be easily obtained. Under the mixed model framework we extend the model to consider the forecasting of new future observations. We illustrate the methodology with monthly air pollution levels in a monitoring site in Europe from January 1990 to December 2002.

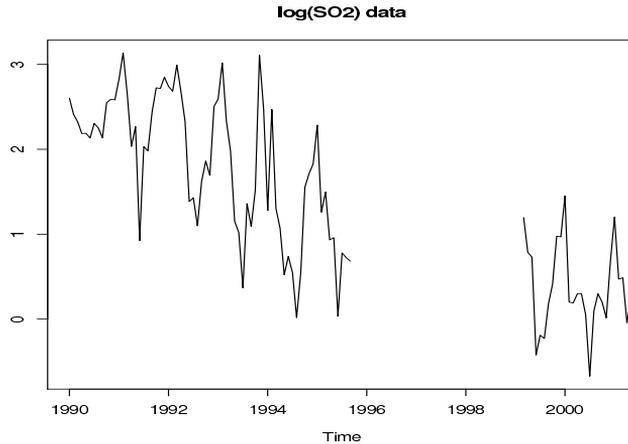
Keywords: P -splines; Mixed Models; times series forecasting; varying-coefficient models; harmonic regression

1 Introduction

Trend and seasonal estimation and decomposition are important tasks in the analysis of time series in many application areas, such as environmental sciences, economics and econometrics. Figure 1 shows the time series plot of monthly Sulphur dioxide (SO_2) concentration levels in logarithmic scale measured in a monitoring site in Europe from January 1990 to December 2002. The data presents two main characteristics: (i) the series shows clear evidence of a decreasing (possibly non-linear) trend plus a seasonal pattern, and (ii) there are some missing observations between October 1995 and March 1999. Consider a sequence of x_1, \dots, x_n time points, the response variable y can be modelled as:

$$y_i = t(x_i) + s(x_i) + \epsilon_i, \quad i = 1, \dots, n \quad (1)$$

where $t(\cdot)$ represents the temporal trend, $s(\cdot)$ the seasonal pattern, and ϵ_i is an error term. Eilers and Marx (2008) proposed a smooth model for (1) 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). Both components can be estimated in the generalized additive model frame-

FIGURE 1. Time series plot of $\log(SO_2)$ data.

work. They used the term *smooth modulation model* given by:

$$\mathbb{E}[y_i] = f(x_i) + \sum_{j=1}^J \{g_j(x_i) \cos(j\omega x_i) + h_j(x_i) \sin(j\omega x_i)\}, \quad (2)$$

where $f(\cdot)$ 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 J required for the seasonal component is usually taken as 1 or 2 to reduce the number of parameters to be estimated. For $J = 1$, we can rewrite model (2) in a compact matrix form:

$$\mathbb{E}[\mathbf{y}] = \mathbf{B}\boldsymbol{\theta}, \quad \text{with } \mathbf{B} = [\mathbf{B}|\mathbf{C}|\mathbf{S}\mathbf{B}], \quad (3)$$

where \mathbf{B} is a B -splines regression basis of size $n \times c$, and $\mathbf{C} = \text{diag}\{\cos(\omega x_i)\}$, and $\mathbf{S} = \text{diag}\{\sin(\omega x_i)\}$. The regression coefficients $\boldsymbol{\theta} = (\theta, \theta_c, \theta_s)'$, are penalized by a block-diagonal matrix, such that $\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}$, with

$$\mathbf{P} = \text{blockdiag}(P_1, P_2, P_3), \quad (4)$$

where $P_k = \lambda_k D_q' D_q$, for $k = 1, 2, 3$, is a $(c - q) \times c$ matrix of differences of order q . In practice, a single smoothing parameter for modulation sine and cosine terms is used, i.e. $\lambda_2 = \lambda_3$.

1.1 Smooth modulation mixed models

The representation of a penalized spline model as a mixed model has become very popular in recent years (see Ruppert et. al (2003) or Wood (2006)). The mixed model representation consists of reparameterizing the

B -spline basis \mathbf{B} and penalty \mathbf{P} , such that, the smooth model is written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Lambda}) \quad \boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{G}), \quad (5)$$

where $\mathbf{X}\boldsymbol{\beta}$ are the *fixed effects* term, and \mathbf{Z} is the matrix for the *random effects* with covariance matrix \mathbf{G} with variance components τ_k^2 . The error term $\boldsymbol{\epsilon}$ has covariance $\boldsymbol{\Lambda}$, e.g. $\boldsymbol{\Lambda} = \sigma^2\mathbf{I}$ for i.i.d errors. The smoothing parameters becomes $\lambda_k = \sigma^2/\tau_k^2$, and can be estimated by restricted maximum likelihood (REML). The order of the penalty q , denotes the fit when the smoothing parameters are very large ($\lambda_k \rightarrow \infty$), or the (parametric) *null model*, such that, taking $q_1 = 2$ for the trend, and $q_2 = q_3 = 1$ for the seasonal terms, it is straightforward to obtain, for $J = 1$, the fixed and random effects matrices as:

$$\mathbf{X} = [\mathbf{1}_n|x_i| \cos(\omega x_i) | \sin(\omega x_i)], \text{ and } \mathbf{Z} = [Z|C\check{Z}|S\check{Z}],$$

where $Z = B\Omega$, and $\check{Z} = B\check{\Omega}$ are the transformed B -spline bases for trend and modulation components. There are different alternatives for the transformation. For instance, we may consider $\Omega = D'_q(D_q D'_q)^{-1}$, with $q = 2$ and similarly for $\check{\Omega}$ with first order differences. Alternatively, we can also use a transformation based on the singular value decomposition of the penalty matrices in (4).

2 Forecasting with smooth modulation mixed models

In times series data, it is important to extrapolate or forecast the future observations. Currie et al. (2004) proposed a method for fitting and forecasting simultaneously with P -splines models when the coefficients are estimated using penalized least squares. For simplicity, we first illustrate the prediction of the trend component, as it results straightforward to include the prediction of the modulation component. Suppose given n observations of the response variable \mathbf{y} , we want to predict new n_0 values \mathbf{y}_0 at x_0 . The new vector of observations is $\mathbf{y}^* = (\mathbf{y}, \mathbf{y}_0)'$. We need a new B -spline basis \mathbf{B}^* constructed from a new set of knots that extends the original knots used to fit the observed data \mathbf{y} , and also includes a basis for the n_0 observations to forecast. Let us consider this new extended basis \mathbf{B}^* as

$$\mathbf{B}^* = \begin{bmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{B}_{(1)} & \mathbf{B}_{(2)} \end{bmatrix}, \text{ of size } n^* \times c^*, \quad (6)$$

where \mathbf{B} is the $n \times c$ basis used for fitting y , $\mathbf{B}_{(1)}$ and $\mathbf{B}_{(2)}$ are auxiliary B -spline basis for prediction up to $n^* = n + n_0$ values, of sizes $n_0 \times c$ and $n_0 \times c_0$ respectively. Figure 2 shows each of the blocks of the new basis (6). Since, we need to increase the number of knots to cover the new range of the covariate values x_0 , we also have to define a new penalty matrix, \mathbf{P}^*

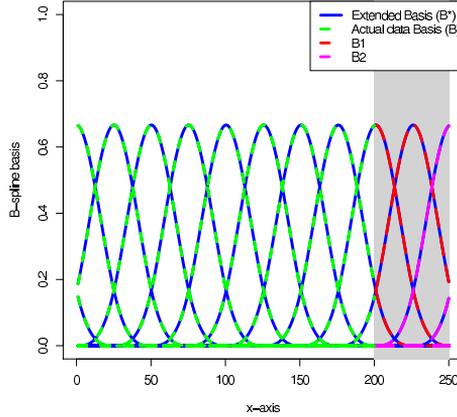


FIGURE 2. Visualization of the component of the B -spline basis \mathbf{B}^* of Eq. (6) for forecasting new values.

built from a difference matrix \mathbf{D}^* , defined by:

$$\mathbf{D}^* = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{D}_{(1)} & \mathbf{D}_{(2)} \end{bmatrix}. \quad (7)$$

Then, the new vector of coefficients is $\hat{\boldsymbol{\theta}}^*$ of length $c^* \times 1$. It can be proved that: (i) the new predicted values are $\hat{\mathbf{y}}_0 = \mathbf{B}_{(2)}\hat{\boldsymbol{\theta}}^*$; (ii) the $1, \dots, c$ coefficients of $\hat{\boldsymbol{\theta}}^*$, are exactly those obtain from in the estimation of the P -splines, i.e.: $\hat{\boldsymbol{\theta}}_{1, \dots, c}^* = \hat{\boldsymbol{\theta}}$, and (iii) the prediction coefficients: $\hat{\boldsymbol{\theta}}^* = (\hat{\boldsymbol{\theta}}, -\mathbf{D}_{(2)}^{-1}\mathbf{D}_{(1)}\hat{\boldsymbol{\theta}})'$.

In the mixed models case, we need to find a transformation matrix, in such a way that the actual and extended models are reparameterized as in (5), and the reparameterization used to fit the data is preserved. We consider:

$$\mathbf{T}^* = \begin{bmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{(2)}^{-1} \end{bmatrix}, \quad (8)$$

where $\boldsymbol{\Omega}$ is the transformation chosen to transform the original B -spline basis, and hence:

$$\mathbf{B}^*\mathbf{T}^* = \left[\begin{array}{c|cc} \mathbf{X} & \mathbf{Z} & \mathbf{0} \\ \mathbf{X}_0 & \mathbf{Z}_{(1)} & \mathbf{Z}_{(2)} \end{array} \right], \quad (9)$$

The prediction mixed model becomes:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{X}_0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}_0 \end{pmatrix} + \begin{pmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{Z}_{(1)} & \mathbf{Z}_{(2)} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}_0 \end{pmatrix} + \boldsymbol{\epsilon}, \quad (10)$$

where $\mathbf{X} = [1_n|x]$, and $\mathbf{X}_0 = [1_{n_0}|x_0]$, and $\mathbf{Z}_{(1)} = \mathbf{B}_{(1)}\boldsymbol{\Omega}$ and $\mathbf{Z}_{(2)} = \mathbf{B}_{(2)}\mathbf{D}_{(2)}^{-1}$. The predicted random effects are computed as

$$\hat{\boldsymbol{\alpha}}_0 = -\mathbf{D}_{(1)}\boldsymbol{\Omega}\hat{\boldsymbol{\alpha}}, \quad (11)$$

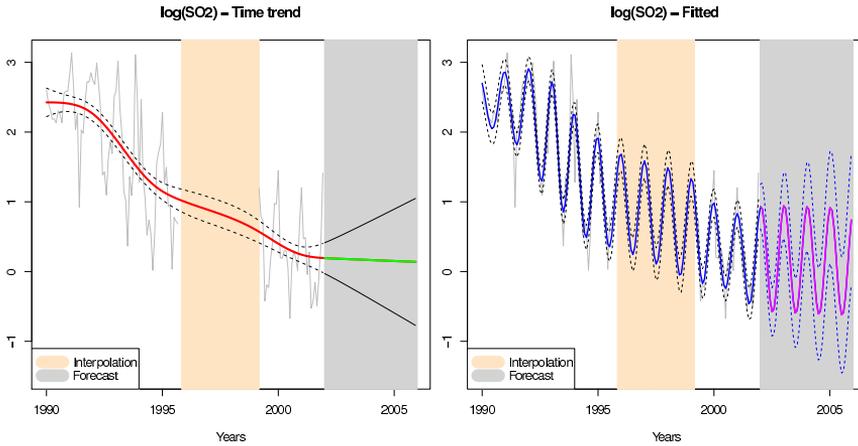


FIGURE 3. $\log(SO_2)$ times series data decomposition. Left: time trend, trend forecast and confidence bands. Right: fitted values (trend+modulation), forecast and confidence bands.

and the new predicted values are:

$$\hat{y}_0 = \mathbf{X}_0 \hat{\beta}_0 + \mathbf{Z}_0 \hat{\alpha}_0, \tag{12}$$

where

$$\mathbf{Z}_0 = \mathbf{B}_{(1)} \Omega - \mathbf{B}_{(2)} \mathbf{D}_{(2)}^{-1} \mathbf{D}_{(1)} \Omega. \tag{13}$$

In order to extend the prediction of the modulation component, we only need to extend the basis (6) and penalty (7) and include the varying-coefficients terms for the future observations x_0 . Hence, for the fixed effects, we add $\cos(\omega x_0)$, and $\sin(\omega x_0)$ and for the random effects, we have to consider a transformation matrix as in (8) with first order differences, and include the modulation varying-coefficient matrices $C_0 = \text{diag}\{\cos(\omega x_0)\}$ and $S_0 = \text{diag}\{\sin(\omega x_0)\}$ for the extended range x_0 into Equations (6)–(13). Figure 3 shows the forecasts of the $\log(SO_2)$ concentrations from January 2003 to December 2005. The left panel shows the fitted trend and forecast with corresponding confidence bands, and right panel shows the fitted and forecasted trend plus modulation components.

3 Concluding remarks

Smoothing techniques have become a very popular tool for the estimation of trends. However, for times series data, simultaneous smoothing and forecasting is still an open subject of research. In this work we proposed a mixed model formulation of the seasonal modulation model using varying-coefficient terms. This approach allows us to decompose the fitted curve

into time trend and seasonality. We extend this formulation to forecast future observations. The mixed model results for forecasting are equivalent to prediction with mixed models proposed in Welham et.al (2004), and provides a unified framework for smoothing and forecasting and also allows the incorporation of additional structures as for instance autorregressive (AR) residuals and estimate smoothing and correlation simultaneously as in Durbán (2003), and combine them for forecasting.

Acknowledgements

This research was funded by the Spanish Ministry of Science and Innovation (projects MTM 2008-02901, and MTM2011-28285-C02-02). The research of Dae-Jin Lee was funded by an NIH grant for the Superfund Metal Mixtures, Biomarkers and Neurodevelopment project 1PA2ES016454-01A2.

References

- Durbán, M. and Currie, I.D. (2003). A note on P-spline additive models with correlated errors *Computational Statistics*, 18, 251-262.
- 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., Marx, Gampe, J., Marx, B. D., and Rau, R. (2008). Modulation models for seasonal time series and incidence tables. *Statistics in Medicine*, 27:3430–3441.
- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient Models (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 60,271–293.
- Ruppert, D., Wand, M.P., and Carroll, R. J. (2003). Semiparametric Regression. *Cambridge University Press*.
- Welham, S., Cullis, B., Gogel, B., Gilmour, A., and Thompson R. (2004). Prediction in mixed linear models. *Aust. N. Z. J. Stat.*, 46:3, 325–347.
- Wood, S.N. (2006). Generalized Additive Models: an introduction with R. *Chapman & Hall/CRC*.

Cook's distance in polytomous logistic regression

Nirian Martín¹, Leandro Pardo²

¹ Department of Statistics, Universidad Carlos III de Madrid, Spain

² Department of Statistics and O.R., Universidad Complutense de Madrid, Spain

E-mail for correspondence: `nirian.martin@uc3m.es`

Abstract: The asymptotic distribution of the Cook's distance in the polytomous logistic regression was unknown until now. We found out that the asymptotic distribution is a linear combination of chi-square random variables. We develop an exhaustive approach for analyzing influential covariates and provide a new measure for the accuracy of predictions based on such a distribution. An illustrative example with a classical real data set, Liver enzyme data (Lesaffre and Albert (1989)), is presented.

Keywords: Multinomial Sampling; Cook's distance; Polytomous Regression Model.

1 Introduction

In many applications in bioassay, epidemiology, economics or social sciences, one encounters a qualitative response variables taking values in a set of unordered categories. Qualitative response models, specify the distribution of the discrete response variable as a function of explanatory variables. The most common case of nominal qualitative response is the binary outcome, for example in patient treatments we can consider two categories, 0 and 1, depending on the unsuccessful or successful results respectively. Polytomous outcomes appear for instance in a medical study of the long-term effects of radiation exposure on mortality, where 1 represents dead from cancer, 2 dead from cause other than cancer and 3 alive (the numbers are not meaningful and must be considered mere labels without ordering). The PLR is a special case of generalized linear model (Nelder and Wedderburn (1972)),

$$\tilde{\pi}(\mathbf{x}_i) = \mathbf{h}(\mathbf{X}_i\boldsymbol{\beta}), i = 1, \dots, I,$$

where:

- the i -th block associated with design matrix is $\mathbf{X}_i = \mathbf{I}_{J-1} \otimes \mathbf{x}_i^T$, with \otimes being the Kronecker product and $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ the vector of $p + 1$ explanatory variables associated with the i -th individual;

- the $(J - 1)(p + 1)$ -dimensional vector of unknown parameters is

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_{J-1}^T)^T, \quad (1)$$

and its subvectors $\boldsymbol{\beta}_j = (\beta_{0j}, \dots, \beta_{pj})^T$, $j = 1, \dots, J - 1$, are compound by an intercept parameter, β_{0j} , and regressor parameters $\beta_{1j}, \dots, \beta_{pj}$;

- $\tilde{\boldsymbol{\pi}}(\mathbf{x}_i) = (\pi_1(\mathbf{x}_i), \dots, \pi_{J-1}(\mathbf{x}_i))^T$ is the $(J - 1)$ -dimensional subvector of

$$\boldsymbol{\pi}(\mathbf{x}_i) = (\pi_1(\mathbf{x}_i), \dots, \pi_{J-1}(\mathbf{x}_i), \pi_J(\mathbf{x}_i))^T,$$

associated with the J -dimensional multinomial response variable with $n(\mathbf{x}_i)$ trials,

$$\mathbf{Y}_i = (Y_{i1}, \dots, Y_{i,J-1}, Y_{iJ})^T \stackrel{ind}{\sim} \mathcal{M}(n(\mathbf{x}_i), \boldsymbol{\pi}(\mathbf{x}_i));$$

- the multivariate response function (or inverse of the logit link function) is $\mathbf{h}(\eta_1, \dots, \eta_{J-1}) = (h_1(\eta_1), \dots, h_{J-1}(\eta_{J-1}))^T$, with

$$h_j(\eta_j) = \frac{\exp(\eta_j)}{1 + \sum_{l=1}^{J-1} \exp(\eta_l)}, \quad j = 1, \dots, J - 1.$$

It is assumed that the I vectors of explanatory variables are linearly independent and $I \geq p + 1$. Note that we have chosen the last category, J , to be the baseline category. Once we establish $\boldsymbol{\beta}_J$ to be a $(p + 1)$ -dimensional vector of zeros, $\boldsymbol{\beta}_J = \mathbf{0}_{p+1}$, the whole vector of probabilities, $\boldsymbol{\pi}(\mathbf{x}_i)$, is defined as $\boldsymbol{\pi}(\mathbf{x}_i, \boldsymbol{\beta}) = (\pi_1(\mathbf{x}_i, \boldsymbol{\beta}), \dots, \pi_{J-1}(\mathbf{x}_i, \boldsymbol{\beta}), \pi_J(\mathbf{x}_i, \boldsymbol{\beta}))^T$, where

$$\pi_j(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp\{\mathbf{x}_i^T \boldsymbol{\beta}_j\}}{1 + \sum_{l=1}^{J-1} \exp\{\mathbf{x}_i^T \boldsymbol{\beta}_l\}}, \quad j = 1, \dots, J. \quad (2)$$

The PLR is a useful tool for multiclass classification. For a prefixed vector of explanatory variables \mathbf{x}_i we can predict the value of the response variable for a unique observation as the most likely category, i.e. $\arg \max_j \pi_j(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, with $\hat{\boldsymbol{\beta}}$ being the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ and $\pi_j(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ the MLE of $\pi_j(\mathbf{x}_i, \boldsymbol{\beta})$.

For the binary logistics regression ($J = 2$) diagnostics, Pregibon (1981) is an essential reference, and Lessaffre and Albert (1989) extend the techniques of binary outcomes to polytomous responses. Cook's distance for the i -th vector of explanatory variables, \mathbf{x}_i , is a measure of the influence that \mathbf{x}_i has on the fit of the model. The diagnostics for PLR are analogous of the ones used for standard regression, for instance, the Cook's distance is directly related to the standardized residuals (outlier diagnostics) and leverages in both cases. The concern about the lack of a meaningful threshold for the Cook's distance is overcome in Section 3 by considering the asymptotic distribution of the Cook's distance. We propose the probability of having an influential vector of explanatory variables for measuring the accuracy of the prediction.

2 The distribution of Cook’s distance

It is well-known that influential observations in Polytomous Logistic Regression models are those points with greatly change on the results of the statistical analysis when omitted from the sample. Pregibon (Pregibon (1981)) proposed an influence measure associated with the i -th observation ($i \in \{1, \dots, I\}$) in the binary logistic regression. A natural extension of that measure for a PLR is given by

$$C^{(i)}(\hat{\beta}) = n(\hat{\beta} - \hat{\beta}^{(i)})^T \mathbf{X}^T \mathbf{W}_n(\hat{\beta}) \mathbf{X}(\hat{\beta} - \hat{\beta}^{(i)}). \tag{3}$$

where $n = \sum_{i=1}^I n(\mathbf{x}_i)$, $\mathbf{X}^T = (\mathbf{X}_1^T, \dots, \mathbf{X}_I^T)_{(J-1)(p+1) \times I(J-1)}$,

$$\mathbf{W}_n(\beta) = \bigoplus_{i=1}^I \frac{n(\mathbf{x}_i)}{n} \mathbf{V}_i(\beta), \tag{4}$$

$$\mathbf{V}_i(\beta) = (\mathbf{I}_{J-1}, \mathbf{0}_{J-1}) \left(\left(\bigoplus_{j=1}^J \pi_j(\mathbf{x}_i, \beta) \right) - \pi(\mathbf{x}_i) \pi(\mathbf{x}_i)^T \right) (\mathbf{I}_{J-1}, \mathbf{0}_{J-1})^T. \tag{5}$$

The influence measure $C^{(i)}(\hat{\beta})$ is the natural adaptation, to the context of PLR, of Cook’s distance for detecting influential observations in linear regression (Cook (1977)). Based on the similarity of $C^{(i)}(\hat{\beta})$ with the quadratic form, $C(\hat{\beta}) = n(\hat{\beta} - \beta_0)^T \mathbf{X}^T \mathbf{W}_n(\beta_0) \mathbf{X}(\hat{\beta} - \beta_0)$, where β_0 is the true value of the vector of unknown parameters, in Lessaffre and Albert (1989) it is suggested that the asymptotic distribution of $C^{(i)}(\hat{\beta})$ can be compared with a chi-square distribution with $(J - 1)(p + 1)$ degrees of freedom. A similar measure to the measure given in (3) was considered in Martín and Pardo (2009) for logistic regression models and its asymptotic distribution was obtained without using the arguments given by Pregibon (1981). In this section we shall obtain the asymptotic distribution of $C^{(i)}(\hat{\beta})$ in a more general setting and the binary logistic regression model is a particular case when $J = 2$.

Theorem 1 *Let $\hat{\beta}$ the MLE of parameter β based on all the explanatory variables and $\hat{\beta}^{(i)}$ the MLE of parameter β based on all the explanatory variables minus the i -th one ($i \in \{1, \dots, I\}$). Then the asymptotic distribution of the Cook’s distance for the PLR is*

$$C^{(i)}(\hat{\beta}) \xrightarrow[n \rightarrow \infty]{L} \sum_{j=1}^{J-1} \nu_j(\mathbf{x}_i, \beta_0) Z_j^2,$$

where $C^{(i)}(\hat{\beta})$ is (3), Z_j , $j = 1, \dots, J - 1$, are independent standard univariate normal random variables and

$$\nu_j(\mathbf{x}_i, \beta_0) = \frac{\alpha_j(\mathbf{x}_i, \beta_0)}{1 - \alpha_j(\mathbf{x}_i, \beta_0)}, \quad j = 1, \dots, J - 1, \tag{6}$$

with $\{\alpha_j(\mathbf{x}_i, \beta_0)\}_{j=1}^{J-1}$ being the set of eigenvalues of $\mathbf{H}_{ii}(\beta_0) = \lim_{n \rightarrow \infty} \mathbf{H}_{n,ii}(\beta_0)$,

$$\mathbf{H}_n(\beta_0) = \mathbf{W}_n(\beta_0)^{\frac{1}{2}} \mathbf{X} (\mathbf{X}^T \mathbf{W}_n(\beta_0) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_n(\beta_0)^{\frac{1}{2}}.$$

We would like to mention that there are some difficulties in the interpretation of the diagonal terms in the Hat matrix, $\mathbf{H}(\beta_0)$. In the two-group logistic regression ($J = 2$), the i -th diagonal element of the Hat matrix, $\mathbf{H}_{ii}(\beta_0)$, is a scalar and is called 'leverage' by analogy with linear regression. However, we highlight that some authors have questioned its appropriateness arguing that the same idea of variability through an infinitesimal increment cannot be considered (see Hosmer and Lemeshow (2000)). We proposed, in Martín and Pardo (2009), an alternative leverage measure, $h_{ii}(\beta_0)(1 - h_{ii}(\beta_0))^{-1}$, called 'potential function associated with the i -th vector of explanatory variables'.

Definition 2 For the PLR, the potential function associated with the i -th vector of explanatory variables is defined by

$$\nu(\mathbf{x}_i, \beta) = \frac{1}{J-1} \sum_{j=1}^{J-1} \nu_j(\mathbf{x}_i, \beta) = \frac{1}{J-1} \sum_{j=1}^{J-1} \frac{\alpha_j(\mathbf{x}_i, \beta)}{1 - \alpha_j(\mathbf{x}_i, \beta)}, \quad (7)$$

where $\{\alpha_j(\mathbf{x}_i, \beta)\}_{j=1}^{J-1}$ is the set of eigenvalues of $\mathbf{H}_{ii}(\beta)$.

3 Distribution based approach for Cook's distance

In this section, based on the distribution obtained in the previous section, we shall propose a threshold for considering a vector of explanatory variables to be influential and also an approach for measuring the accuracy of the predictions based on the probability of having a influential vector of explanatory variables.

Now we shall define some diagnostic tools for PLR introduced in Lessaffre and Albert (1989), residuals, Pearson residuals and an approximation for Cook's distance, the so-called confidence interval displacement, denoted by $\tilde{C}^{(i)}(\hat{\beta})$. We shall also show its asymptotic distributions. Since $Y_{iJ} = n(\mathbf{x}_i) - \sum_{j=1}^{J-1} Y_{ij}$, in order to remove the redundant components from the residuals we shall use matrix $(\mathbf{I}_{J-1}, \mathbf{0}_{J-1})$ and so for the PLR, the $(J-1)$ dimensional vector of residuals associated with the i -th explanatory variable \mathbf{x}_i is $\mathbf{r}(\mathbf{x}_i, \hat{\beta}) = (\mathbf{I}_{J-1}, \mathbf{0}_{J-1})(\mathbf{Y}_i - n(\mathbf{x}_i) \boldsymbol{\pi}(\mathbf{x}_i, \hat{\beta}))$. Since for the PLR, the chi-square test-statistic for the goodness of fit is defined as $X^2 = \sum_{i=1}^I X_i^2$, where $X_i^2 = \frac{1}{n(\mathbf{x}_i)} \mathbf{r}^T(\mathbf{x}_i, \hat{\beta}) \mathbf{V}_i^{-1}(\hat{\beta}) \mathbf{r}(\mathbf{x}_i, \hat{\beta})$ (see Gupta et al. (2008)) for more details), the vector of Pearson residuals associated with the i -th explanatory variable \mathbf{x}_i , denoted by $\mathbf{r}_i^*(\mathbf{x}_i, \hat{\beta})$, is such that $X_i^2 = \mathbf{r}^*(\mathbf{x}_i, \hat{\beta})^T \mathbf{r}^*(\mathbf{x}_i, \hat{\beta})$, i.e. $\mathbf{r}^*(\mathbf{x}_i, \hat{\beta}) = \frac{1}{\sqrt{n(\mathbf{x}_i)}} \mathbf{V}_i^{-\frac{1}{2}}(\hat{\beta}) \mathbf{r}(\mathbf{x}_i, \hat{\beta})$.

Proposition 3 The asymptotic distribution of the vector of Pearson residuals associated with the i -th vector of explanatory variables \mathbf{x}_i is given by

$$\mathbf{r}^*(\mathbf{x}_i, \hat{\beta}) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(\mathbf{0}_{J-1}, \mathbf{M}_{ii}(\beta_0)), \quad (8)$$

where $\mathbf{M}_{ii}(\beta_0) = \mathbf{I}_{J-1} - \mathbf{H}_{ii}(\beta_0)$.

The confidence interval displacement (CID) is defined in terms of the Pearson residuals as

$$\tilde{C}^{(i)}(\hat{\beta}) = \mathbf{r}^*(\mathbf{x}_i, \hat{\beta})^T \mathbf{M}_{ii}(\hat{\beta})^{-1} \mathbf{H}_{ii}(\hat{\beta}) \mathbf{M}_{ii}(\hat{\beta})^{-1} \mathbf{r}^*(\mathbf{x}_i, \hat{\beta}). \quad (9)$$

In the following theorem we shall establish an important relationship of the CID with respect to the Cook’s distance.

Theorem 4 *The asymptotic distribution of the confidence interval displacement, $\tilde{C}^{(i)}(\hat{\beta})$, as $n \rightarrow \infty$, is the same as $C^{(i)}(\hat{\beta})$ given in Theorem 1.*

Let $\tilde{X}^2(\mathbf{x}_i, \hat{\beta}) = \mathbf{r}^*(\mathbf{x}_i, \hat{\beta})^T \mathbf{M}_{ii}(\hat{\beta})^{-1} \mathbf{r}^*(\mathbf{x}_i, \hat{\beta})$, be the “standardized squared grouped residuals”, which is asymptotically distributed as a χ^2_{J-1} random variable. We shall consider the i -th vector of explanatory variables \mathbf{x}_i to be an outlier if $\tilde{X}^2(\mathbf{x}_i, \hat{\beta}) > \chi^2_{J-1, \alpha}$, where $\chi^2_{J-1, \alpha}$ is the quantile of order α for χ^2_{J-1} . Now we are going to give a new criterion for detecting leverage and influential observations. The asymptotic distribution of the type given in Theorem 1 is usually approximated by other one. In Satterthwaite (1946), for instance, the proposed approximation is $\bar{\nu}(\hat{\beta}) \chi^2_{\alpha, J-1}$, and we shall use it in the second part of the following definition.

Definition 5 *Once α significance level is prefixed for outliers, we propose the following criteria:*

i) $2\bar{\nu}(\hat{\beta})$ is a cutoff of $\nu(\mathbf{x}_i, \hat{\beta})$, for considering the i -th vector of explanatory variables to be a leverage, where

$$\bar{\nu}(\hat{\beta}) = \frac{1}{I} \sum_{h=1}^I \nu(\mathbf{x}_h, \hat{\beta});$$

ii) $\bar{\nu}(\hat{\beta}) \chi^2_{\alpha, J-1}$ is a cutoff of $C^{(i)}(\hat{\beta})$ (as well as $\tilde{C}^{(i)}(\hat{\beta})$), for considering the i -th vector of explanatory variables to be influential.

Since according to the criterion given in Definition 5 we have

$$\hat{p}(i, \alpha) \equiv \Pr(\mathbf{Y}_i \text{ influential}) = \Pr(\chi^2_{J-1} > \frac{\bar{\nu}(\hat{\beta})}{\nu(\mathbf{x}_i, \hat{\beta})} \chi^2_{J-1, \alpha}). \quad (10)$$

We propose considering \mathbf{x}_i such that $\hat{p}(i, \alpha) \leq 0.5$ as a potential \mathbf{x}_i with correct classification according to the Cook’s distance. In this sense in addition to the “rate of correct classification

$$RCC = \#(\arg \max_j Y_{ij} = \arg \max_j n(\mathbf{x}_i) \hat{\pi}_j(\mathbf{x}_i, \hat{\beta}))/I,$$

where $\hat{\mathbf{p}}_i = \frac{\mathbf{Y}_i}{n(\mathbf{x}_i)}$, we can use the rate of potential correct classification according to the Cook’s distance

$$RPCC = \#(\hat{p}(i, \alpha) \leq 0.5)/I,$$

where $\alpha = 0.05$, as measure of accuracy for the prediction.

4 Liver Enzyme Data

In a study of $I = 218$ patients with liver disease, $J = 4$ diagnostic groups for response variable were considered, acute viral hepatitis (1), persistent chronic hepatitis (2), aggressive chronic hepatitis (3) and post-necrotic cirrhosis (4). As explanatory variables, $p = 3$ the logarithmic transformation of enzymatic activities were taking into account, aspartate aminotransferase (AST), alanine aminotransferase (ALT) and glutamate dehydrogenase (GLDH). Figure 1 shows the Cook's distance values and associated threshold (left hand side) as well as an scheme of the procedure for obtaining $R_{PCC} = \frac{177}{218} = 81.19\%$ (right hand side), which is slightly different from $RCC = \frac{182}{218} = 83.49\%$.

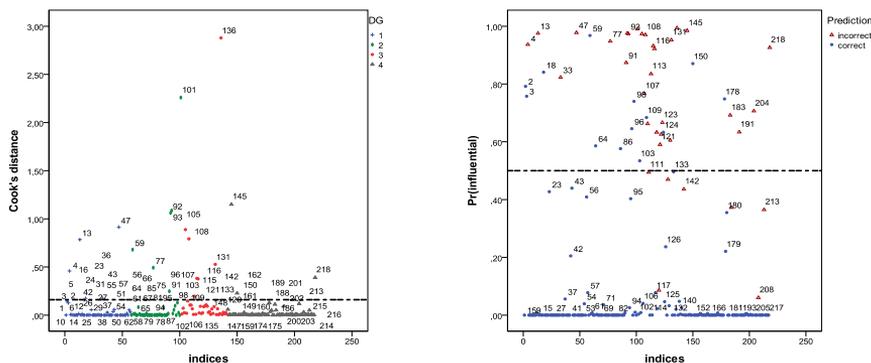


FIGURE 1. Cook's distances and probabilities of correct classification.

References

- Lesaffre, E. and Albert, A. (1989). Multiple-group logistic regression diagnostic. *Applied Statistics*, **38**, 425–440.
- Martín, N. and Pardo, L. (2009). On the asymptotic distribution of Cook's distance in logistic regression models. *Journal of Applied Statistics*, **36**, 1119–1146.
- Pregibon, D. (1981). Logistic Regression Diagnostics. *Annals of Statistics*, **9**, 705–724.

Fitting prediction intervals for BMI patterns in childhood by boosting quantile regression

Andreas Mayr¹, Torsten Hothorn², Nora Fenske²

¹ Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

² Department of Statistics, Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: andreas.mayr@imbe.med.uni-erlangen.de

Abstract: The construction of prediction intervals (PIs) for body mass index (BMI) patterns of individual children is problematic for standard parametric approaches, as the BMI distribution in childhood is typically skewed depending on age. A solution could be the usage of GAMLSS, but we present a more simple approach by directly modelling the borders of the PIs with non-parametric quantile regression, estimated by boosting. We conduct a simulation study before we fit PIs on data from a recent German birth cohort study with $n = 2007$ children.

Keywords: Longitudinal quantile regression; model based boosting; prediction intervals.

1 Introduction

We focus on obtaining reliable predictions for future BMI patterns of children. Prediction intervals (PIs) offer information on the expected variability by providing not only a point prediction but a covariate-specific interval which covers the future BMI for this individual child with high probability. We construct child-specific prediction intervals for the LISA study, a recent German birth cohort study with 2007 children. Data include up to ten BMI values per child from birth until the age of 10, as well as variables that are discussed to be potential early childhood risk factors for later obesity, such as breastfeeding, maternal BMI gain and smoking during pregnancy, parental overweight, socioeconomic factors, and weight gain during the first two years.

The distribution of BMI values is typically skewed and the degree of skewness depends on children's age (see Figure 1), which makes standard strategies to construct PIs relying on distributional and homoscedasticity assumptions problematic. One possibility to overcome these problems would be the usage of more sophisticated parametric approaches, as for example generalized additive models for location scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005). This model class has already been used for constructing PIs in combination with boosting (Mayr et al., 2012). However, the construction of PIs based on GAMLSS depends totally on the assumed distribution and the interpretation of covariate effects with respect to the

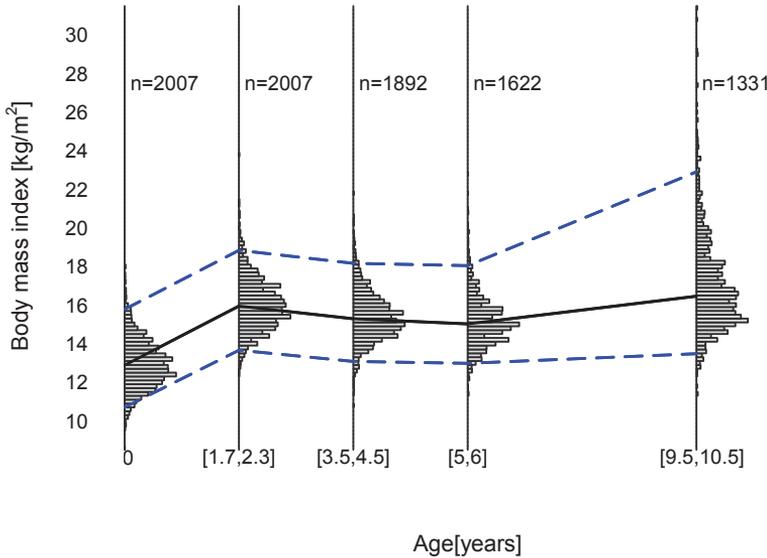


FIGURE 1. LISA birth cohort study: childhood BMI values up to the age of ten.

interval borders is not straightforward.

We avoid making distributional assumptions here by developing a new approach for constructing non-parametric prediction intervals based on quantile boosting.

2 Methods

2.1 Prediction intervals by conditional quantiles

The idea of using quantile regression to construct prediction intervals for new observations was presented by Meinshausen (2006). In contrast to standard regression analysis, quantile regression – thoroughly described by Koenker (2005) – does not estimate the conditional expectation of a random variable Y but the conditional quantile function $Q_\tau(Y|X = x) = q_\tau(x)$ for a given $\tau \in (0, 1)$ and a possible set of covariates $X = x$. Following the definition of quantiles the probability of the response Y being smaller than $q_\tau(x)$ is τ . The goal is therefore to estimate the conditional quantile function $\hat{q}_\tau(x)$ by quantile regression based on a training sample $(y_1, x_1), \dots, (y_n, x_n)$. For a new observation, the specific covariate combination x_{new} is plugged into $\hat{q}_\tau(x_{\text{new}})$. A prediction interval for y_{new} is then estimated by:

$$\widehat{\text{PI}}_{(1-\alpha)}(x_{\text{new}}) = [\hat{q}_{\frac{\alpha}{2}}(x_{\text{new}}), \hat{q}_{1-\frac{\alpha}{2}}(x_{\text{new}})] .$$

The resulting PI should cover a new observation y_{new} with probability $(1 - \alpha)$ while its length depends on x_{new} . There might be combinations of covariates that allow for a very precise prediction for y_{new} resulting in a narrow interval, whereas wide intervals imply that for a given x_{new} the prediction is more inaccurate.

2.2 Quantile boosting

In our approach, we determine conditional quantiles by additive quantile regression. For a fixed quantile $\tau \in (0, 1)$, the conditional quantile function is expressed by an additive predictor as follows:

$$q_\tau(\mathbf{x}_i) = \eta_{\tau i} = \beta_{\tau 0} + \sum_{j=1}^p f_{\tau j}(x_{ij}).$$

The index $i = 1, \dots, n$, denotes the individual, and $q_\tau(\mathbf{x}_i)$ stands for the τ -quantile of the response y_i conditional on its specific covariate vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$. The quantile-specific additive predictor $\eta_{\tau i}$ is composed of an intercept $\beta_{\tau 0}$ and a sum of different effects of p covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ on the quantile function. The functions $f_{\tau 1}, \dots, f_{\tau p}$ comprise linear effects, i.e. $f_{\tau j}(x_{ij}) = \beta_{\tau j} x_{ij}$, as well as non-linear effects whose functional form is not specified in advance. In fact, the additive predictor could also contain a wide variety of additional covariate effects, e.g. varying coefficient terms or spatial effects. Note that contrary to classical regression, there is no specific distributional assumption for the response. The only restriction is that the response must be continuous.

In general, the estimation of unknown parameters in quantile regression can be achieved by minimizing the *check function* ρ_τ as the appropriate loss:

$$\rho_\tau(y, \eta_\tau) = \begin{cases} \tau \cdot |y - \eta_\tau| & y > \eta_\tau \\ (1 - \tau) \cdot |y - \eta_\tau| & y \leq \eta_\tau. \end{cases}$$

We apply quantile boosting (Fenske et al., 2011) for the estimation of the additive quantile regression model: The minimization of the empirical risk is achieved by stepwise updating the predictor function η_τ . Therefore, *base-learners* are used, i.e. simple univariate regression models fitting the negative gradient of the empirical loss. The base-learners play a key role in the algorithm, since they define the kind of effects between each covariate and response. In our approach, we use simple linear models to represent linear covariate effects, penalized regression splines to represent non-linear effects and ridge-penalized models for the child-specific “random effects”. The advantage of quantile boosting is that the resulting predictor η_τ is strictly additive and interpretable, following the additive quantile regression model.

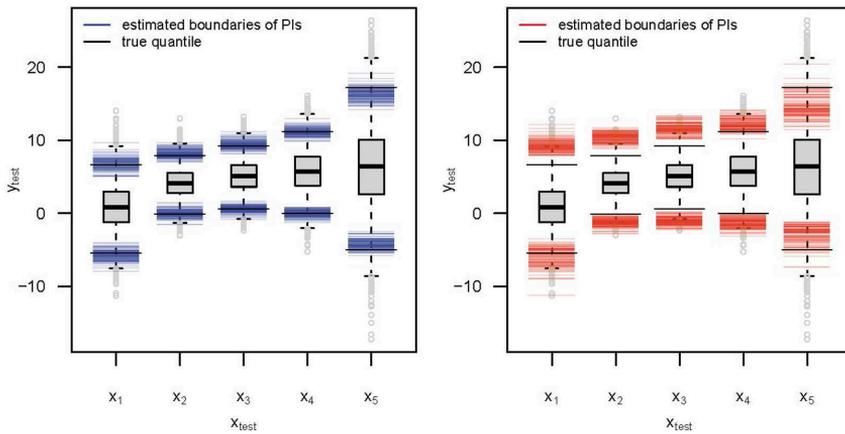


FIGURE 2. Resulting intervals for a non-linear and high-dimensional ($p > n$) setting, conditional on 5 arbitrarily chosen covariate combinations. Left: quantile boosting, right: quantile regression forest

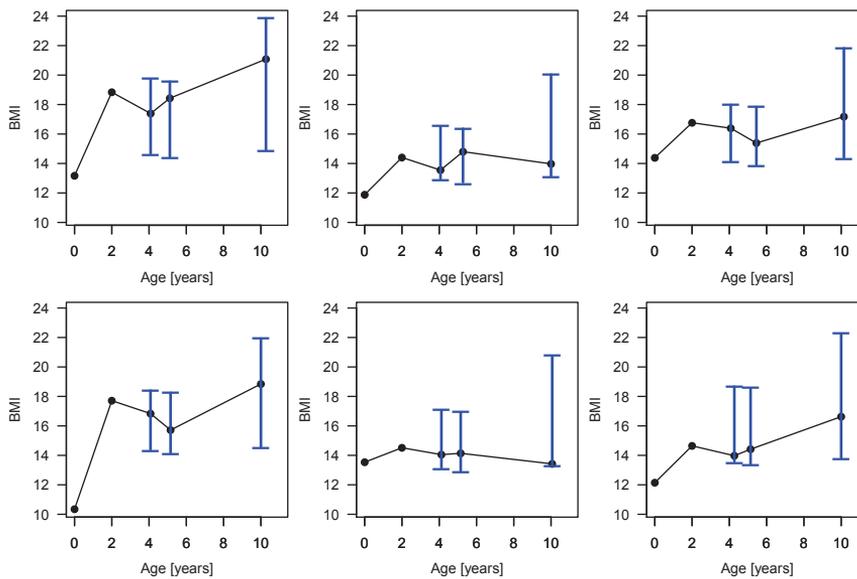


FIGURE 3. Resulting intervals for six children left out in the fitting process.

3 Results

In the work on this project, we discovered a severe pitfall in the validation of PIs: The correct validation of a method to fit PIs is based on conditional coverage (conditional on a new observation) rather than on sample coverage

(averaged over a new sample). For example, after constructing a 95% PI for the BMI of a child, we particularly expect the future BMI of this child with its exact measures to be covered with a probability of 95%. In frequentistic language, the BMI of 95% of children with exactly the same measures should be covered by the interval. That is a stronger claim than covering on average 95% of new samples with different measurements.

3.1 Simulation study

The simulation study was therefore designed to evaluate the conditional coverage rather than just averaging the coverage rates from new samples. We analyzed various settings including linear and non-linear effects as well as settings with more predictors than observations. As benchmark algorithm we chose the quantile regression forest approach, proposed by Meinshausen (2006).

Figure 2 presents results from one of these settings: The boxplots display the empirical distribution of the “future” observations for each of the test points $\mathbf{x}_1, \dots, \mathbf{x}_5$. The solid black lines are the true conditional quantiles and represent the true borders of a 95% PI for each test point. The shorter lines show the resulting estimated PI borders from 100 simulation runs of the two algorithms. Quantile boosting seems to work best in the center of the x -space, which is represented by test point \mathbf{x}_3 . For the other test points, the standard errors for the estimated quantiles get larger, yielding less accurate PIs. Quantile regression forest have more problems in fitting the correct conditional quantiles, which further explains why quantile boosting outperformed the benchmark algorithm with respect to the coverage rates.

3.2 BMI patterns

We used all information of the children at the age of two to predict their BMI patterns until the age of ten. To account for the longitudinal data structure, we included child-specific intercepts and slopes in the additive predictor. Figure 3 shows the resulting PIs for six randomly chosen children. Level and length of the PIs are child-specific, but the lengths of PIs at the age of ten are always larger than the lengths at earlier time points. From a methodological view, this absolutely reflects what we should expect from a valid method to fit PIs: The intervals do what they should, in reporting the increasing uncertainty in the prediction of BMI values until the age of ten based only on very limited information from the children in early childhood.

4 Conclusion

The aim of the present work was to construct prediction intervals for BMI patterns of individual children. We pursued this aim by applying quantile

boosting to directly model the borders of the PIs. As a result, we do not rely on any distributional assumptions.

In a simulation study we showed that our approach leads to accurate intervals that can outperform the benchmark algorithm. The resulting intervals for BMI patterns reported the increasing uncertainty of the prediction. It would be interesting to see if covariates explaining physical activity, nutrition and lifestyle habits of the children could help for getting smaller intervals as presented in this paper.

In conclusion, we think that quantile boosting is a promising approach to construct prediction intervals with correct conditional coverage in a non-parametric way. It can be applied to longitudinal settings and is therefore in particular suitable for the prediction of BMI patterns or similar data, where assumptions of standard parametric approaches are not fulfilled.

Acknowledgments: The authors thank Joachim Heinrich, Peter Rzehak and Heinz-Erich Wichmann from the Institute of Epidemiology, Helmholtz Zentrum München (German Research Center for Environmental Health) for providing the data, in this connection they also thank the LISA-plus Study Group for their work.

References

- 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.
- Koenker, R. (2005). *Quantile Regression*, New York: Cambridge University Press 2005.
- 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.

Conditional estimation of the bivariate distribution under dependent right censoring

Ana Moreira¹, Luís Machado¹

¹ Department of Mathematics and Applications, University of Minho

E-mail for correspondence: a.moreira.cris@gmail.com

Abstract: 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. Most of the times, one will be interested in describing the distribution of the joint gap times, the marginal distribution of the gap times but also the correlation structure among them. In recent years significant contributions have been made regarding this topic. However, most approaches assume independent censoring and do not account for the influence of covariates. This manuscript introduces two estimators that account for dependent censoring while including covariate information. A real data illustration is included.

Keywords: Beran estimator; Bivariate distribution; Conditional Survival; Dependent Censoring; Kaplan-Meier.

1 Introduction

Multi-state models are models for a stochastic process where individuals move from one state to another over the course of the study. The so called progressive three-state model plays a central role in the theory and practice of multi-state models (Andersen et al., 1993). In this model, individuals start in the “Alive and disease-free” state and subsequently move either to the “ill” stage and afterwards to the “dead” state. In the framework of the progressive three-state model we will have two gap (times between the consecutive events) times and an interesting quantity to calculate is the distribution of the joint gap times.,

The estimation of the bivariate distribution function is an issue that has received much attention recently. Among others, it was investigated by Lin et al. (1999) and de Uña-Álvarez and Meira-Machado (2008). The estimator proposed by Lin in 1999 uses inverse probability of censoring weighted (IPCW) based on the Kaplan-Meier estimator. On the other hand, the idea behind the estimator proposed by de Uña-Álvarez and Meira-Machado is the use of the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. However, the later estimator can

also be written as a sum of weights based on inverse probability of censoring. The two methods can be implemented using the R package `survivalBIV`. Further details about the methods (and other) and their implementation can be seen in the paper by Moreira and Meira-Machado (2012).

The main goal of this work is to provide new estimators for these targets that can account for dependent censoring. Furthermore, we will show that the methods discussed here can be adapted to include covariate information. In other words, we will show how to obtain estimators for the conditional bivariate distribution; the bivariate distribution given a continuous covariate under random censoring that could either be a baseline covariate or a current covariate that is observed for an individual.

The manuscript is organized as follows. In the next section, we introduce some notations and give the precise definitions of the bivariate distribution function $F_{12}(x, y)$. The methods will be introduced for the situation without covariates and its extension to include baseline covariate or a current covariate will be briefly explained. For illustration purposes we present in Section 3 some plots using real data.

2 Conditional Bivariate Distribution

2.1 Notation

Consider n independent and identically distributed pairs of successive failure (gap) times (T_{1i}, T_{2i}) , $1 \leq i \leq n$ with joint distribution function $F_{12}(x, y)$. These pairs of gap times are subject to univariate right-censoring at times C_i with distribution function $G(t) = P(C \leq t)$ and which we assume to be independent of (T_{1i}, T_{2i}) . Because of this we only observe $(\tilde{T}_{1i}, \tilde{T}_{2i}, \Delta_1, \Delta_2)$ where $\tilde{T}_{1i} = \min(T_{1i}, C_i)$, $\Delta_{1i} = I(T_{1i} \leq C_i)$, $\tilde{T}_{2i} = \min(T_{2i}, C_{2i})$, $\Delta_{2i} = I(T_{2i} \leq C_{2i})$ where $C_{2i} = (C_i - T_{1i})I(T_{1i} \leq C_i)$. Let $T = T_1 + T_2$ be the total time and put $\tilde{T} = \min(T, C)$. Since the censoring time is assumed to be independent of the process, the marginal distribution of the first gap time T_1 , say F_1 may be consistently estimated by the Kaplan-Meier estimator based on the (\tilde{T}_1, Δ_1) . Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on the $(\tilde{T}_i, \Delta_{2i})$'s. With this notation, the bivariate distribution is written as

$$F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y).$$

We are also interested in estimating the conditional distribution function: $F_{12}(x, y | Z)$, that can be computed for any times x and y , but conditional to some covariate value which we denote by Z . Again, following the notation we introduce above, the conditional bivariate distribution is written as

$$F_{12}(x, y | Z) = P(T_1 \leq x, T_2 \leq y | Z).$$

2.2 The Estimators

In this section we will introduce two estimators for the conditional bivariate distribution that can account for dependent censoring, $F_{12}(x, y | Z)$. Below we provide two competing non-parametric estimators of the bivariate distribution under (possibly dependent) right censoring. The first estimator is based on observations that are completely uncensored (i.e., fully observed till death)

$$\hat{F}_{12}(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{I(\tilde{T}_{1i} \leq x, \tilde{T}_{2i} \leq y) \Delta_{2i}}{\hat{G}(\tilde{T}_i)}.$$

This is known as the inverse-probability-of-censoring-weighting estimator (IPCW) which is a sum of iid terms $I(\tilde{T}_{1i} \leq x, \tilde{T}_{2i} \leq y)$ multiplied by Δ_{2i} weighted inversely by the probability that the failure time is observed (Satten and Datta, 2001).

The second estimator (Lin-based) is based on observations that were uncensored till a given time and was proposed by Lin et al. (1999):

$$\tilde{F}_{12}(x, y) = \frac{1}{n} \sum_{i=1}^n \frac{I(\tilde{T}_{1i} \leq x) \Delta_{1i}}{\hat{G}(\tilde{T}_{1i})} - \frac{1}{n} \sum_{i=1}^n \frac{I(\tilde{T}_{1i} \leq x, \tilde{T}_{2i} > y)}{\hat{G}(\tilde{T}_{1i} + y)}.$$

The methods introduced above can be the basis of two competing non-parametric regression estimators of the conditional bivariate distribution. In both estimators, local smoothing can be done by introducing regression weights that are either based on a kernel or a local linear regression. These quantities can be estimated in the following way. First estimate the d.f. of C given Z , G_Z . Let us start with the case where the d.f. G_Z is known. We propose to plug-in Beran's estimator \hat{G}_Z (Beran (1981)) and use the local linear estimator (LLE) or a Naradaya-Watson estimator (NNE), i.e.

$$\hat{F}_{12}(z; x, y) = \sum_{i=1}^n W_{1i}(z, b_n) \frac{I(\tilde{T}_{1i} \leq x, \tilde{T}_{2i} \leq y) \Delta_{2i}}{\hat{G}_{Z_i}(\tilde{T}_i)}$$

where $W_{1i}(z, b_n)$ are Naradaya-Watson weights or local linear weights. Note that if we assume independence between the censoring variable C and the covariate Z than, the estimator of the conditional distribution of C given Z , \hat{G}_Z , reduces to the honorable product-limit Kaplan-Meier estimator.

Alternative estimators (Lin-based) can be given for the conditional probabilities. In this case,

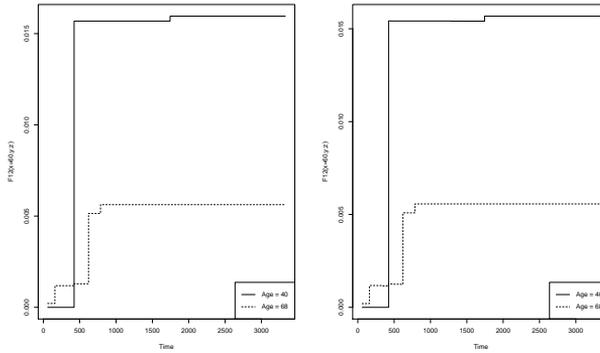


FIGURE 1. Conditional bivariate distribution for the Colon Cancer data for $age = 40$ and $age = 68$. (IPCW left hand-side and Lin-based right hand-side)

$$\tilde{F}_{12}(z; x, y) = \sum_{i=1}^n W_{1i}(z, b_n) \frac{I(\tilde{T}_{1i} \leq x) \Delta_{1i}}{\hat{G}_{Z_i}^0(\tilde{T}_{1i})} - \sum_{i=1}^n W_{1i}(z, b_n) \frac{I(\tilde{T}_{1i} \leq x, \tilde{T}_{2i} > y)}{\hat{G}_{Z_i}(\tilde{T}_{1i} + y)}$$

where G_Z^0 stands for an estimator of the conditional distribution $C | Z = Z_i$, for example, the based on the $(\tilde{T}_{1i}, 1 - \Delta_{1i})$'s.

3 Example of Application

Due to large number of people affected by cancer of colon, there is much demand for information on this disease. In a large percentage of the patients, the diagnosis is made at a sufficiently early stage when all apparent disease tissue can be surgically removed. These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer and it is available in the **R survival** package. From the total of 929 patients, 468 developed recurrence and among these 414 died. The covariate *recurrence* is a time-dependent covariate which can be expressed as an intermediate event which can be modeled using an progressive three-state model with states “Alive and disease-free”, “Alive with recurrence” and “dead”.

In this section we will present estimated bivariate distribution conditionally on current or past covariate measures such as *age*. These quantities were calculated using the method based on IPCW and Lin-based estimator.

For illustration purposes we show in Figure 1 the conditional bivariate distribution for patients with 40 years old and patients with 68 years old. This suggests that patients with 40 years old have a higher values of the bivariate distribution than patients with 68 years old. These and other results from the analysis of real data show that the estimates of the bivariate distribution function greatly depends on covariate information.

4 Conclusions

In this paper we present two estimators (Lin-based and IPCW) for the conditional bivariate distributions that respond for dependent censoring and for bivariate distribution without covariates. Both methods are based on local smoothing which is introduced using regression weights. Two different schemes of inverse censoring probability reweighting is used to deal with right censoring. In our study we perform simulations to compare the behavior of the two estimators in the presence (or not) of covariates. Results (not shown here) showed that the IPCW leads to better results for conditional bivariate distribution. In our simulations and for real data analysis we have used a common bandwidth selector, the **dpik** function from the **R KernSmooth** package, and normal kernels to estimate the conditional bivariate distribution. In our simulations we have used Nadaraya-Watson and Local Linear estimators for estimating the weights.

Acknowledgments: Luis F. Meira-Machado acknowledges financial support by Grant *PTDC/MAT/104879/2008* (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education. Ana Moreira acknowledges financial support by grant *SFRH/BD/62284/2009* of the Portuguese Ministry of Science, Technology and Higher Education. 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 a Tecnologia, within the Project Est - *C/MAT/UI0013/2011*.

References

- Andersen, P. K., Borgan, O., Gill, R. D., Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.
- Beran, R. (1981). *Nonparametric regression with randomly censored survival data*. Technical report, University of California, Berkeley.
- de Uña-Álvarez, J., Meira-Machado, L. (2008) A Simple Estimator of the Bivariate Distribution Function for Censored Gap Times. *Statistics and Probability Letters*, **78**, 2440–2445.
- 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.
- Moreira, A., Meira-Machado, L. (2012) survivalBIV: Estimation of the Bivariate Distribution Function for Sequentially Ordered Events Under Univariate Censoring. *Journal of Statistical Software*, March **46**:13.

Satten, G. A., Datta, S., and Robins, J. (2001). Estimating the marginal survival function in the presence of time dependent covariates. *Statistics & Probability Letters*, **54**, 397–403.

Autoregressive models with non-gaussian errors

David Morina^{1,2}, Pedro Puig², Jordi Valero³

¹ Centre Tecnològic de Nutrició i Salut, Spain

² Universitat Autònoma de Barcelona, Spain

³ Universitat Politècnica de Catalunya, Spain

E-mail for correspondence: david.morina@ctns.cat

Abstract: Suppose that Y_t follows a simple AR(1) model, that is, it can be expressed as $Y_t = \alpha Y_{t-1} + W_t$, where W_t is a white noise with mean equal to μ and variance σ^2 . There are many examples in practice where these assumptions hold very well. Consider $X_t = e^{Y_t}$. We shall show that the autocorrelation function (ACF) of X_t characterizes the distribution of W_t . Consequently, the knowledge of the empirical ACF of X_t can help to choose the distribution of the white noise W_t of the time series Y_t . This result can also be used to construct a goodness of fit test for the classical AR(1) model, where the white noise W_t is normally distributed. Several examples of application will be also shown.

Keywords: Time series; AR(1) models; Characterization of distributions.

1 Introduction

In this work, we are considering general AR(1) processes defined as

$$Y_t = \alpha \cdot Y_{t-1} + W_t, \quad (1)$$

where W_t is a white noise with mean μ and variance σ^2 , satisfying the condition of stationarity $|\alpha| < 1$. We also consider the auxiliary and stationary time series $X_t = e^{Y_t}$. Sometimes, it may be useful to exponentiate a series and obtain a new positive time series from which we can obtain information about the original one. This process can also be performed in reverse, as in many contexts, for a time series of positive terms that can be written as $X_t = e^{Y_t}$, the log transformation is useful in order to stabilize the variance. Moreover, for a huge range of time series of econometric indicators, a logarithmic transformation can improve the forecasting performance.

Several authors have studied autoregressive models with non-gaussian innovation or errors. McKenzie (1982) analyzes in detail the case with a gamma marginal for X_t . The paper by Sim (1994) is focused on parameter estimation and forecasting by means of several models with different marginal distributions for Y_t , such as exponential, logistic, hyperbolic secant

and others. Grunwald and Hyndman (1995) extend the AR(1) non-gaussian models and consider a nonparametric smoothing.

An important relation between the marginal distribution of X_t and Y_t (indicated by the random variables X and Y) is summarized in Corollary 1, on the basis of that of McKenzie(1982):

Lemma 1. (McKenzie, 1982) Writing $s = \alpha^k$, we have

$$\rho_{X_t}(k) = \frac{\mathbb{E}[X] (\mathbb{E}[X^{s+1}] - \mathbb{E}[X^s]\mathbb{E}[X])}{\mathbb{E}[X^s]Var[X]} \tag{2}$$

Corollary 1. The autocorrelation function of X_t can be written as

$$\rho_{X_t}(k) = \frac{\psi_Y(1)\psi_Y(\alpha^k + 1) - \psi_Y(1)^2\psi_Y(\alpha^k)}{\psi_Y(\alpha^k)\psi_Y(2) - \psi_Y(1)^2\psi_Y(\alpha^k)}, \tag{3}$$

where $\psi_Y(t)$ is the moment generating function of Y .

This result is the key because shows the relationship between the ACF of X_t and the moment generating function of Y .

2 Characterization of W_t

Taking logarithms and using some algebra on Corollary 1 we have

$$\log(\psi_Y(s + 1)) - \log(\psi_Y(s)) = \log \left(\frac{\psi_Y(2) - \psi_Y(1)^2}{\psi_Y(1)} \cdot \rho_{X_t}(s) + \psi_Y(1) \right). \tag{4}$$

Therefore, the cumulant generating function of Y , $\log(\psi_Y(s))$, can be found as the solution of (4). Because the cumulant generating function of a random variable is a convex function, using the following result we can prove that the difference equation (4) has a unique solution:

Theorem 2. (Kuczma et al., 1990) Let $I = [a, \infty)$, $-\infty \leq a < \infty$. If the function $h : I \rightarrow \mathbb{R}$ is concave (resp. convex) and satisfies the condition

$$\lim_{s \rightarrow \infty} [h(s + 1) - h(s)] = 0,$$

then the equation $f(s + 1) - f(s) = h(s)$ has a unique one-parameter family of convex (resp. concave) solutions $f : X \rightarrow \mathbb{R}$. These solutions can be constructed using the expression,

$$f(s) = c + (s - s_0)h(s_0) + \sum_{n=0}^{\infty} [(s - s_0)(h(s_0 + n + 1) - h(s_0 + n)) - (h(s + n) - h(s_0 + n))],$$

where $s_0 \in I$ is arbitrarily fixed and $c \in \mathbb{R}$ is an arbitrary constant.

Consequently, calling $f(s) = \log(\psi_Y(s))$ and

$h(s) = \log \left(\frac{\psi_Y(2) - \psi_Y(1)^2}{\psi_Y(1)} \cdot \rho_{X_t}(s) + \psi_Y(1) \right)$, we state the following characterizing theorem:

Theorem 3. Let $Y_t = \alpha Y_{t-1} + W_t$, and $Y_t = \log X_t$. Let $\rho_{X_t}(k)$ be the autocorrelation function of X_t , being $\rho_{X_t}(s)$ ($s = \alpha^k$) a concave function. Then the distribution of W_t is unique.

The statement of Theorem 2 can be developed for several scenarios, specially for the situation where W_t is normally distributed.

2.1 Some cases

The most common case in AR(1) models is when the white noise term follows a normal distribution, so we are going to explore this case in detail. Suppose that $W_t \sim N(\mu, \sigma^2)$. As Y_t follows the AR(1) process defined in (1), it is well known that the marginal distribution Y will be normal with mean $\frac{\mu}{1-\alpha}$ and variance $\frac{\sigma^2}{1-\alpha^2}$. As $Y_t = \log X_t$, we have that X follows a log-normal distribution with the corresponding parameters. In this case, we also have that

$$\text{Var}[X] = (e^{\frac{\sigma^2}{1-\alpha^2}} - 1)e^{2\frac{\mu}{1-\alpha} + \frac{\sigma^2}{1-\alpha^2}}. \tag{5}$$

Then, we can use Corollary 1 and some basic algebra to obtain the expression of the autocorrelation function of X_t ,

$$\rho_{X_t}(s) = \frac{e^{\frac{\sigma^2 s}{1-\alpha^2}} - 1}{e^{\frac{\sigma^2}{1-\alpha^2}} - 1}, \tag{6}$$

where we have written $s = \alpha^k$ for convenience. By Theorem 2, the ACF in (6) characterizes the distribution of W_t , allowing us to develop a goodness of fit test for the normality assumption on W_t . It is useful in order to choose an adequate distribution for the white noise term of the series Y_t , as we will see in the following section.

Although the gaussian error is the most common case, it is possible to consider other distributions for W_t , and therefore, to construct goodness of fit tests for other scenarios. For example, the case $X \sim \text{Gamma}$ is studied with detail in McKenzie (1982). In this case, we know that

$$\mathbb{E}[X^k] = \frac{\Gamma(\beta + k)}{\theta^k \Gamma(\beta)} \tag{7}$$

Using Corollary 1, we have that

$$\rho_X(k) = \frac{\Gamma(\beta + 1)}{\theta \Gamma(\beta)} \left(\frac{\frac{\Gamma(\beta + \alpha^k + 1)}{\theta^{\alpha^k + 1} \Gamma(\beta)} - \frac{\Gamma(\beta + \alpha^k)}{\theta^{\alpha^k} \Gamma(\beta)} \cdot \frac{\Gamma(\beta + 1)}{\theta \Gamma(\beta)}}{\frac{\Gamma(\beta + \alpha^k)}{\theta^{\alpha^k} \Gamma(\beta)} \cdot \frac{\beta}{\theta^2}} \right) \tag{8}$$

Using the properties of the Γ function and simplifying (8) one finally obtain that $\rho_X(k) = \alpha^k$.

A different model is constructed when the innovations are Gamma distributed, $W_t \sim \Gamma(\beta, \theta)$. Consider the $MA(\infty)$ expression of Y_t :

$$Y_t = \alpha Y_{t-1} + W_t = W_t + \alpha W_{t-1} + \dots + \alpha^k W_{t-k} + \dots \quad (9)$$

It can be easily shown that if $W_t \sim \Gamma(\beta, \theta)$, then $\alpha^k W_t \sim \Gamma(\beta, \alpha^k \theta)$. From here, the moment generating function of Y is,

$$\psi_Y(t) = \prod_{k=0}^{\infty} (1 - \alpha^k \theta t)^{-\beta} \quad (10)$$

From here the autocorrelation function of X_t can be calculated using Corollary 1, but it has not an explicit form and has to be evaluated numerically.

3 Goodness of fit

The results of the previous section can be used in order to construct a goodness of fit test for the classical AR(1) process, defining the null and alternative hypothesis as follows:

$$H_0 : W_t \sim N(\mu, \sigma^2) \quad H_1 : W_t \sim N(\mu, \sigma^2) \quad (11)$$

Given a time series Y_t , a new series is considered as $X_t = e^{Y_t}$. Therefore, the empirical autocorrelation function of X_t , $\hat{\rho}_{X_t}(k)$, is calculated in order to be compared with the theoretical ACF expressed in (6). However this ACF is not known in practice and must be estimated as well, by using the statistic,

$$\hat{\rho}_{H_0}(k) = \frac{e^{\frac{\hat{\sigma}^2 \hat{\alpha}^k}{1 - \hat{\alpha}^2}} - 1}{e^{\frac{\hat{\sigma}^2}{1 - \hat{\alpha}^2}} - 1}. \quad (12)$$

Here $\hat{\alpha}$ and $\hat{\sigma}^2$ are estimated from the original time series Y_t by means of a non parametric method like Yule-Walker equations. Therefore, a kind of Portmanteau-type test can be constructed by means of the following statistic:

$$\sum_{k=1}^3 (\hat{\rho}_{X_t}(k) - \hat{\rho}_{H_0}(k))^2. \quad (13)$$

Other tests can be constructed in a similar way, as in Chen and Deo (2004), Swanepoel and Doku (2003) or in a nonlinear context, in Cheng and Shuxia (2008).

In order to perform the test, we simulate several series with the same estimated parameters by bootstrapping the original series. To do that, we

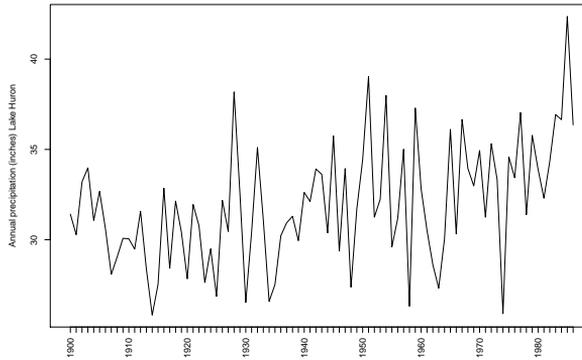


FIGURE 1. Lake Huron annual precipitation records

use the estimated process and a sample of the residuals instead of the non-observed values of W_t , and we calculate the statistic (13) for each series. This technique allows to obtain a $(1 - \alpha)$ confidence interval for the test statistic, and to compare it with the test statistic obtained using the original time series.

Note that this is a one-tailed test, so we will reject the null hypothesis if the test statistic calculated using the original series is greater than the $(1 - \alpha)$ th percentile of the distribution of the test statistic obtained using the bootstrap method.

3.1 Example

We have analyzed the annual precipitation over Lake Huron (Canada) from 1900 to 1986. The profile of this series is shown in Figure 1, and this will be considered as our X_t process which was previously analyzed as an AR(3) model in Hipel and McLeod (1994).

Our goal is to explore if the time series $Y_t = \log(X_t)$ comes from an AR(1) process with gaussian white noise. Using the Yule-Walker equations, the fitted parameters of Y_t are $\hat{\alpha} = 0.21$, $\hat{\mu} = 2.73$ and $\hat{\sigma}^2 = 0.01$.

Following the method described in the previous section, we have obtained that a 95% confidence interval for the test statistic is $(0, 0.086)$. Therefore, the null hypothesis cannot be rejected because the test statistic value is 0.05. Consequently, it seems that a normal distribution is a good choice for the distribution of the innovations W_t .

4 Data driven AR(1)

Given the autocorrelation function of X_t , $\rho_{X_t}(k)$, it is possible to reproduce the density of the error term W_t . If $\rho_{X_t}(k)$ is unknown but estimated from

the data, it is still possible to use some numerical techniques in order to estimate the density of W_t . Therefore, given a time series $X_t = e^{Y_t}$, the empirical autocorrelation coefficients can be computed and the autocorrelation function can be estimated, via polynomial or spline interpolation. Then, an estimation of the cumulant generating function of Y can be obtained as a solution of the difference equation (4) with $\psi_Y(0) = 0$ as the initial condition.

The relationship between the moment generating function of Y and the moment generating function of the error term W_t is the following:

$$\psi_W(s) = \frac{\psi_Y(s)}{\psi_Y(\alpha s)}. \quad (14)$$

Therefore, the probability density function of W_t , $f(t)$, can be recovered by Laplace transform inversion, according to the expression,

$$f(t) = \frac{1}{2\pi i} \lim_{T \rightarrow \infty} \int_{\gamma - iT}^{\gamma + iT} e^{st} \psi_W(s) ds, \quad (15)$$

References

- Chen, W.; Deo, R. (2004). A generalized Portmanteau goodness-of-fit test for time series models. *Econometric Theory*, **20**, 382–416.
- Cheng, F.; Shuxia, S. (2008). A goodness-of-fit test of the errors in nonlinear autoregressive time series models. *Statistics and probability letters*, **78**, 50–59.
- Grunwald, G.K. and Hyndman, R.J. (1998). Smoothing non-Gaussian time series with autoregressive structure. *Computational statistics and data analysis*, **28**, 171–191.
- Hipel, K. W.; McLeod, A. I. (1994). Time series modelling of water resources and environmental systems. *Elsevier*
- Kuczma, M.; Choczewski, B.; Ger, R. (1990). Iterative functional equations. *Encyclopedia of mathematics and its applications*, **32**.
- McKenzie, E. (1982). Product autoregression: a time-series characterization of the gamma distribution. *Journal of applied probability*, **19**, 463–468.
- Sim, C.H. (1994). Modelling non-normal first-order autoregressive time series. *Journal of forecasting*, **13**, 369–381.
- Swanepoel, C. J.; Doku, W. O. (2003) New goodness-of-fit tests for the error distribution of autoregressive time-series models. *Computational statistics and data analysis*, **43**, 333–340.

Smoothed score confidence interval for the breakpoint in segmented regression

Vito M. R. Muggeo

¹ Dipartimento Scienze Statistiche e Matematiche ‘S. Vianelli’, Università di Palermo, ITALY

E-mail for correspondence: vito.muggeo@unipa.it

Abstract: For the breakpoint parameter in segmented regression we consider confidence intervals based on the score statistic. Due to unsmoothness of the score, we propose to build the confidence intervals using its smoothed version under proper shape restrictions. Some simulations are presented to assess the finite sample performance of the proposed approach.

Keywords: segmented regression; break-point; shape restrictions; score; non-standard inference.

1 Introduction

Segmented regression is a useful tool to assess the effect of a quantitative covariate X on the response Y when there exists a threshold value where the effect of X changes gradually; for instance, Muggeo et al. (2009) model the fertility patterns for cohabitant women in Italy via segmented relationships of the hazard with respect to time since cohabitation. The segmented regression equation for the conditional mean $E[Y|x] = \mu$ is

$$\mu_i = w_i^T \gamma + \beta(x_i - \psi)_+ \quad i = 1, 2, \dots, n \quad (1)$$

where $(x_i - \psi)_+ = (x_i - \psi)I(x_i > \psi)$ and $w_i^T \gamma$ may include additional linear terms, such as other covariates, the model intercept, and the linear term for the segmented variable that represents the ‘left slope’ of the piecewise relationship. Likelihood-based inference for the breakpoint parameter ψ is difficult and challenging due to the non-regularities of model (1): the log likelihood is only piecewise differentiable with possibly local optima. While both maximization of the log likelihood and hypothesis testing are demanding, in this paper we are specifically interested in interval estimation for ψ ; previous proposals include likelihood based confidence intervals (Lerman, 1980), and the Wald statistic approach based on an approximate standard error for $\hat{\psi}$ (Muggeo, 2003).

2 Methods

2.1 Score and Information

To simplify notation we suppose that the linear parameters γ in model (1) are known; we will relax this unrealistic assumption later. Assuming a Gaussian distribution for the errors with relevant log-likelihood $\ell(\psi, \beta)$, the score vector $U = (U_\psi, U_\beta)^T$ has components

$$U_\psi(\psi, \beta) = \frac{\partial \ell}{\partial \psi} = \dot{\ell}_\psi = -\frac{\beta}{\sigma^2} \sum_i (y_i - \mu_i) I(x_i > \psi) \quad (2)$$

$$U_\beta(\psi, \beta) = \frac{\partial \ell}{\partial \beta} = \dot{\ell}_\beta = \frac{1}{\sigma^2} \sum_i (y_i - \mu_i)(x_i - \psi)_+. \quad (3)$$

Although U comes from a non regular model, it is still Normal since it is a linear combination of the y_i s. Moreover, as in regular problems, at the true parameter values it is $E[U] = 0$ and $\text{var}[U] = E[U^2] = -E[\dot{U}]$, where \dot{U} is the second derivative of the log-likelihood. Thus score-based inference is, at least in theory, feasible.

The relevant expected information matrix $\mathcal{I} = \text{var}(U)$ can be obtained by the minus second derivatives

$$\ddot{\ell}_{\psi,\psi} = -\frac{\beta^2}{\sigma^2} \sum_i I(x_i > \psi) \quad \ddot{\ell}_{\psi,\beta} = \frac{\beta}{\sigma^2} \sum_i (x_i - \psi)_+ \quad \ddot{\ell}_{\beta,\beta} = -\frac{1}{\sigma^2} \sum_i (x_i - \psi)_+^2.$$

We indicate the appropriate entries of the information matrix \mathcal{I} by $\mathcal{I}_{\psi\psi}$, $\mathcal{I}_{\psi\beta}$ and $\mathcal{I}_{\beta\beta}$. When only ψ is unknown the studentized score is simply $S_\psi = U_\psi / \sqrt{\mathcal{I}_{\psi\psi}} \sim \mathcal{N}(0, 1)$. More generally, when also the difference-in-slope parameter is unknown, β is replaced by the corresponding ML estimate given ψ , namely $\hat{\beta}_\psi$. The resulting score takes the form

$$U_{\psi|\beta} = U_\psi(\psi, \hat{\beta}_\psi) = -\frac{\hat{\beta}_\psi}{\sigma^2} \sum_i (y_i - \hat{\mu}_{i,\psi}) I(x_i > \psi) \quad (4)$$

with (conditional) variance $\text{var}(U_{\psi|\beta}) = \mathcal{I}_{\psi\psi} - \mathcal{I}_{\psi\beta} \mathcal{I}_{\beta\beta}^{-1} \mathcal{I}_{\beta\psi}$, which is always less than the unconditional $\text{var}(U_\psi)$. Note that when model (1) includes additional linear covariates w_i with unknown γ , the nuisance parameter vector is $(\gamma^T, \beta)^T$, the corresponding information $\mathcal{I}_{\beta\beta}$ is a matrix and $\mathcal{I}_{\beta\psi}$ is a vector, but the aforementioned formulas still apply.

The studentized Score statistic for ψ when nuisance parameters are replaced by their constrained ML estimates is

$$S_{\psi|\beta}(\psi) = \frac{U_{\psi|\beta}}{\text{var}(U_{\psi|\beta})^{\frac{1}{2}}} \quad (5)$$

that is approximately distributed according to a $\mathcal{N}(0, 1)$ for large samples.

Virtually we could use $S_\psi(\psi)$ or $S_{\psi|\beta}(\psi)$ to build confidence intervals for ψ in model (1); for instance in the case of nuisance parameter vectors, a $100(1 - \alpha)$ CI is given by $\{\psi : z_{\alpha/2} \leq S_{\psi|\beta}(\psi) \leq z_{1-\alpha/2}\}$ where the z s are the standard Normal quantiles. However, the shape of $S_{\psi|\beta}(\psi)$ does not allow to obtain univocally the roots of the equations $\{S_{\psi|\beta}(\psi) = z_{\alpha/2}\}$ and $\{S_{\psi|\beta}(\psi) = z_{1-\alpha/2}\}$ as $S_{\psi|\beta}(\psi)$ is not smooth neither monotone: to illustrate Figure 1 portrays the score function $S_{\psi|\beta}(\psi)$ coming from the same set of toy data $(x_i, y_i = \mu_i + \epsilon_i \sigma)$ having the same ϵ_i s but amplified by increasing values of the random variance σ .

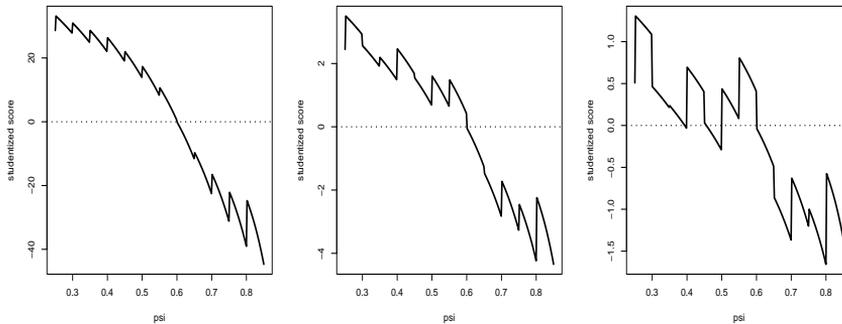


FIGURE 1. Profile score $S_{\psi|\beta}(\psi)$ coming from the same toy data $(x_i, y_i = \mu_i + \sigma \epsilon_i)$ with increasing values of σ from left to right.

We observe that for low values of noise σ the score function is reasonably smooth and monotone, at least locally around the maximum likelihood estimate $\hat{\psi}$; as σ increases $S_{\psi|\beta}(\psi)$ deteriorates, in that it becomes more wiggly reflecting uncertainty in the breakpoint estimate.

2.2 The smoothed Score

Based on patterns highlighted in Figure 1, we conjecture that, at least locally, the unsmoothness shape of $S_{\psi|\beta}(\psi)$ depends on the random noise of the observed data. Namely if we could observe data with a very low amount of noise, the resulting profile score $S_{\psi|\beta}(\psi)$ would be, at least locally, regular, i.e. smooth and monotone. Therefore from the actual $S_{\psi|\beta}(\psi)$ we seek ‘to extract’ a locally regular score $S_{\psi|\beta}^*(\psi)$ freed from random perturbations; $S_{\psi|\beta}^*(\psi)$ will allow straightforward computations of the endpoints of the confidence interval throughout $z_{\alpha/2} \leq S_{\psi|\beta}^*(\psi) \leq z_{1-\alpha/2}$. Bar-Lev et al. (2000) discuss that local rather global monotonicity can be sufficient to obtain confidence intervals for non-monotonic reparameterizations, and we speculate that a similar rationale applies here.

To obtain the ‘regular’ score function $S_{\psi|\beta}^*(\psi)$ we employ B -splines with proper shape restrictions: i) monotonicity, i.e. $S_{\psi|\beta}^*(\psi)$ should be non-increasing with respect to the ψ values; ii) $S_{\psi|\beta}^*(\hat{\psi}) = 0$, i.e. $S_{\psi|\beta}^*(\psi)$ should intersect the abscissa axis just at the maximum likelihood estimate $\hat{\psi}$. To fulfill both constraints we use two asymmetric penalty terms without putting any further penalty on the coefficients. Notice that we need local regularity, thus we smooth $S_{\psi|\beta}(\psi)$ within a pre-specified range (a, b) , say, of the covariate values. Thus to build the proposed score-based confidence intervals we propose the following steps:

1. fix a grid of values of ψ within a reasonable range (a, b) ;
2. compute the corresponding raw values of the studentized score $S_{\psi|\beta}(\psi)$;
3. obtain a locally smooth version $S_{\psi|\beta}^*(\psi)$ via B -splines with shape restrictions;
4. compute the endpoints of the confidence interval via the roots of the equations $\{S_{\psi|\beta}^*(\psi) = z_{\alpha/2}\}$ and $\{S_{\psi|\beta}^*(\psi) = z_{1-\alpha/2}\}$.

Figure 2 shows the application of the proposed approach along with the resulting confidence interval.

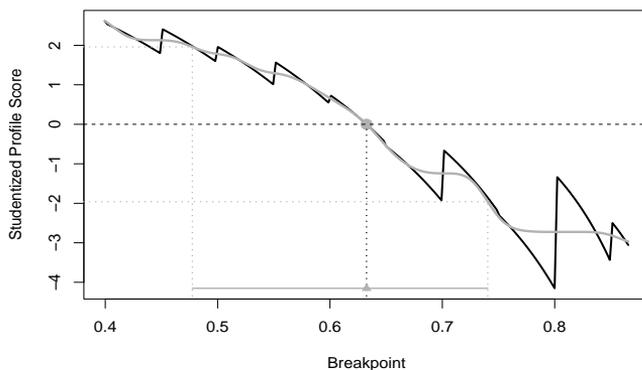


FIGURE 2. The studentized raw score $S_{\psi|\beta}(\psi)$ (black line) and corresponding ‘regular’ version $S_{\psi|\beta}^*(\psi)$ (grey line); the resulting 95% CI is displayed at the bottom.

To assess the finite sample performance of smoothed score interval estimator, Table 1 reports the coverage levels (CL) of the proposed approach along with the CL coming from the Wald statistic, i.e. $\hat{\psi} \pm z_{\alpha/2} \text{SE}(\hat{\psi})$, where $\text{SE}(\hat{\psi})$ is the approximate standard error from the Delta method

(Muggeo, 2003). Five sample sizes and two locations of the breakpoint are considered.

TABLE 1. Performance of the 95% CIs for the breakpoint ψ via the smoothed score and the Wald statistic in different scenarios: Coverage Levels (CL) based on 2000 replicates.

ψ	n	CL of the 95%CI	
		smooth Score	Wald
0.5	30	0.945	0.902
	50	0.957	0.910
	100	0.956	0.928
	500	0.948	0.941
	1000	0.960	0.944
0.8	30	0.987	0.837
	50	0.976	0.812
	100	0.981	0.793
	500	0.950	0.862
	1000	0.943	0.909

Symmetric confidence intervals from the Wald statistic are too narrow and provide CLs less than the nominal 0.95. This is due to underestimation of the variability of $\hat{\psi}$ (via $\text{SE}(\hat{\psi})$) and to the non Normal distribution of $\hat{\psi}$ especially when ψ is not in the middle of the covariate range. Unlike the Wald statistic the proposed approach performs reasonably well by producing asymmetric confidence intervals having CL approximately equal to or greater than the nominal level 0.95.

3 Conclusion

Here we have presented a new approach to build confidence intervals for the breakpoint by smoothing the studentized score statistic. The idea of smoothing objective functions is not new, and it has been applied in different contexts, for instance in censored quantile regression model (Pang et al., 2010) and for the accelerated hazard model (Zhang et al., 2011). Even for regular parameters, score based confidence intervals are less known and rather under-used, especially with respect to those based on the Wald statistic. However the score statistic typically depends on linear combination of the response random variables, leading to faster convergence to Normal distribution by the Central Limit Theorem, and the Fisher information is its *exact* variance. Therefore CIs based on the Score statistic are expected to perform reasonably well; this has been confirmed by our small-scale simulation study reported in the previous section. Comparisons with respect to likelihood-based CIs of Lerman (1980) even for non-Normal and/or non-independent observations represent scenarios to be investigated.

The proposed approach will be included in the next releases of the segmented package for R.

References

- Bar-Lev, S.K., Bshouty, D., Benjamin Reiser, B. (2000). Upper bounds for coverage probabilities of confidence intervals for nonmonotone parametric functions. *Journal of Statistical Planning and Inference*, **89**, 109-118.
- Lerman P.M. (1980). Fitting segmented regression models by grid search. *Applied Statistics*, **29**, 77-84.
- Muggeo, V.M.R. (2003). Estimating regression models with unknown breakpoints. *Statistics in Medicine*, **22**, 3055-3071.
- Muggeo, V.M.R., Attanasio M., Porcu M. (2009). A segmented regression model for event history data: an application to the fertility patterns in Italy. *Journal of Applied Statistics*, **36**, 973-988.
- Pang, L., Lu, W., Wang X.J. (2010). Variance estimation in censored quantile regression via induced smoothing. *Computational Statistics & Data Analysis*, **56**, 785-796.
- Zhang, J., Peng, Y., Zhao, O. (2011). A New Semiparametric Estimation Method for Accelerated Hazard Model. *Biometrics*, **67**, 1352-1360.

Nonparametric classification of noisy functions

Stanislav Nagy¹

¹ Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Praha, Czech Republic

E-mail for correspondence: nagy@karlin.mff.cuni.cz

Abstract: The classification task for data coming from certain subspaces of continuous functions will be discussed. The functions will be of noisy nature and no further assumptions about the distributions will be stated.

Special attention will be paid to depth-based classification and its possible generalisations. Several established depth functional classifiers will be compared. The outcoming drawbacks of these methods will be fixed by considering the derivatives of the smoothed versions of functions, although the observations don't have to be differentiable itself.

Thus, a new version of Fraiman-Muniz depth capable of measuring the centrality of a differentiable function is introduced. Its classification performance is compared to known classifiers and we show that proper derivative using in combination with DD-plot (depth-depth plot) techniques is a powerful tool not only for the classification of functional observations.

Keywords: Functional Data; Data Depth; Supervised Classification; DD-Plot.

1 Data Depth

Data depth is modern nonparametric tool for the analysis of multivariate, and recently also functional and general Banach-valued data. The notion of *halfspace depth* was introduced by Tukey (1975) as a powerful tool for the picturing of multivariate data and exploratory analysis. More than twenty years later, Liu et al. (1999) published a breakthrough article, where the concept of depth was rediscovered. Besides the investigation of simplicial depth, depth function was suggested to be exploited by means of constructing robust nonparametric tests in multidimensional setup. Since then, a wide variety of multivariate depth functions has emerged.

In general, **statistical depth function (depth)** is a mapping

$$D : S \times \mathcal{P}(S) \rightarrow \mathbb{R}^+ \quad (\rightarrow [0, 1]),$$

where $\mathcal{P}(S)$ denotes the set of all probability distributions on S . Depth defines linear semi-ordering (the same depth value may be assigned to several

points) in “center outward” sense. Depth should satisfy some conditions considered to be “reasonable” for depth functions. The desirable properties were stated and studied by Zuo and Serfling (2000). Popular are *halfspace*, *simplicial*, *spatial* or *zonoid* depths. However, those depths cannot be directly used for infinite-dimensional data and new definitions are needed.

2 DEPTH AND FUNCTIONAL DATA

We will focus on functional depths defined directly on the infinite dimensional space $\mathcal{C}[0, 1]$. Other possibilities, e.g., finitely dimensional depths for (coefficients of) *functional principal components* or the promising *integrated dual depth* are not discussed here.

The depths are based on *position of the graph* or *function values* of function x with respect to the underlying probability distribution P . We propose to include the graphs of derivatives into the depth calculation under the additional assumption on differentiability of functional data.

If the observed functions are not smooth enough, every measured function is suitably pre-smoothed in order to gain an additive information about the shape of the function. Alternatively, the function value may remain in the unsmoothed form, whereas the higher order derivatives may be computed from a pre-smoothed version of the function.

2.1 Examples of functional depth

Denote $X_1, \dots, X_n \in \mathcal{C}[0, 1]$ a *random sample coming from absolutely continuous distribution* (in the sense that all finite-dimensional distributions are absolutely continuous). Denote further X a random function (random variable $X : \Omega \rightarrow \mathcal{C}[0, 1]$) and P its *distribution* (probability measure on $\mathcal{C}[0, 1]$).

- **Fraiman-Muniz depth** (Fraiman and Muniz (2001)) for functional data is defined for any one-dimensional depth D as

$$FD(x, P) = \int_0^1 D(x(t), P_t) dt,$$

where P_t is the marginal distribution of $X(t)$ at $t \in [0, 1]$. Usual choice of D is the halfspace depth.

- **Band depth** (López-Pintado and Romo (2007,2009)) or the J -th order band depth for $J \geq 2$. Denote

$$\mathbf{B}(t, j) = [\inf\{X_1(t), \dots, X_j(t)\}, \sup\{X_1(t), \dots, X_j(t)\}]$$

where X_1, \dots, X_j forms a random sample from P and define

$$BD(x, P) = \frac{1}{J-1} \sum_{j=2}^J P[x(t) \in \mathbf{B}(t, j) \forall t \in [0, 1]],$$

Note that the depth $BD(x, P)$ of x may be lower than ε regardless of the Lebesgue measure of the set $\{t : x(t) \in \mathbf{B}(t, j)\}$. This property may be considered as a disadvantage of the band depth and therefore various other versions of band depth have been proposed.

It is not difficult to see that both depths, FD and BD are invariant to the permutation of coordinates, i.e., being $\tau : [0, 1] \rightarrow [0, 1]$ any measurable bijection, then $FD(x \circ \tau, P \circ \tau) = FD(x, P)$.

2.2 Depth including derivatives

The functional nature of data is considered more in the definition of **K -derivatives depth**. Let us first consider a probability distribution P on $\mathcal{C}^K[0, 1]$, the space of K -differentiable functions. Denote for $k = 0, \dots, K$ the joint k -order derivative of x as $x^{(0,k)}(t) = (x(t), x'(t), \dots, x^{(k)}(t))$ for $t \in [0, 1]$. Clearly $x^{(0,k)}(t)$ is a point in $k + 1$ dimensional space \mathbb{R}^{k+1} .

Denote further $\mathbf{S}_P^{(k+2)}$ a *simplex* (a closed convex hull) in \mathbb{R}^{k+1} given by random sample $X_1^{(0,k)}, \dots, X_{k+2}^{(0,k)}$ from distribution P where $k = 0, 1, \dots, K$. The K -derivatives depth is then defined as a version of simplicial depth where the multivariate observations are the function value and its derivatives. For any standard weights $w_0, \dots, w_K, \sum_{k=0}^K w_k = 1$, define

$$KSD(x, P) = \sum_{k=0}^K w_k \int_0^1 P[x^{(0,k)}(t) \in \mathbf{S}_P^{(k+2)}] dt.$$

Note, that for $K = 0$ it holds $KSD \equiv FD$. The main advantage of KSD depth is that the depth of a function may be small if the “shape” of the function is “outlying” in the random sample although the graph itself is not an “outlier”.

The KSD depth is detecting also “outliers” in derivatives. Therefore the depth may be quite different from FD or BD as can be seen in Figure 1 where just the first derivative is used in comparison to Fraiman-Muniz depth. On the other hand, BD depths give very similar results to FD for the most of smooth datasets.

3 Depth based classification of functional data

Let us consider a typical classification problem. There are probability distributions P_1, \dots, P_l on space $\mathcal{C}[0, 1]$ smoothed into space $\mathcal{C}^K[0, 1]$ such that X follows distribution P_i with a prior probability π_i . The goal is to assign the smoothed version of a new observation to one of the distributions such that the average probability of *misclassification* is small. The only information about P_i is available in the form $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_l)$ of random samples (*training sets*) $\mathbf{Y}_i = (Y_{i,1}, \dots, Y_{i,n_i})$ from P_i for $i = 1, \dots, l$. Also π_i is

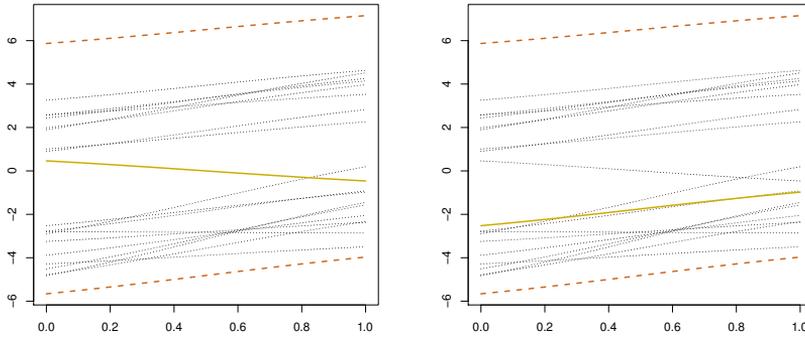


FIGURE 1. *FD* halfspace depth on the left hand side and *KSD* depth, $K = 1$, $w_0 = w_1 = 1/2$ on the right hand side. The solid line function is the functional median (i.e. observation with the highest depth value), the dashed observations are zero depth functions. The apparent outlier is wrongly identified as the median by the depth not including derivatives.

estimated as $\hat{\pi}_i = n_i/n$, where $n = \sum_{i=1}^l n_i$. The classifier is a function $\delta : (x, \mathbf{Y}) \mapsto \{1, \dots, l\}$, where $x \in \mathcal{C}^K[0, 1]$ (it may be also considered $\delta(x, \mathbf{Y}) = 0$, in the case when the observation x is unclassifiable).

The simplest classification depth-based rule is the **highest depth (HD) classifier**, i.e.,

$$\delta(x, \mathbf{Y}) = \{i : D(x, P_i) \geq D(x, P_j), j = 1, \dots, l\}.$$

This classifier works well in finite dimension for unimodal elliptically symmetric distributions with equal priors different in location only (in other words, *under very restrictive conditions*), when it is asymptotically optimal Bayes classifier, see Ghosh and Chaudhuri (2005).

DD-plot classifier was proposed recently by Li et al. (2010) for multivariate data. The idea for $l = 2$ is following:

1. Calculate $D(x, P_i)$ for all $i = 1, 2$.
2. Plot couples $(D(Y_{i,j}, P_1), D(Y_{i,j}, P_2))$ for $j = 1, \dots, n_i$ and $i = 1, 2$ (the *DD-transformations of data*).
3. Classify x to P_2 if $D(x, P_2) > p(D(x, P_1))$, p being an appropriate polynomial of order less than given q . Here p is chosen such that the misclassification rate for the training set \mathbf{Y} is smallest possible.

Note that the highest depth classifier is a special case of the DD-plot classifier with the linear rule $p(a) = a$. Hence, the classification rule is $D(x, P_2) > D(x, P_1)$.

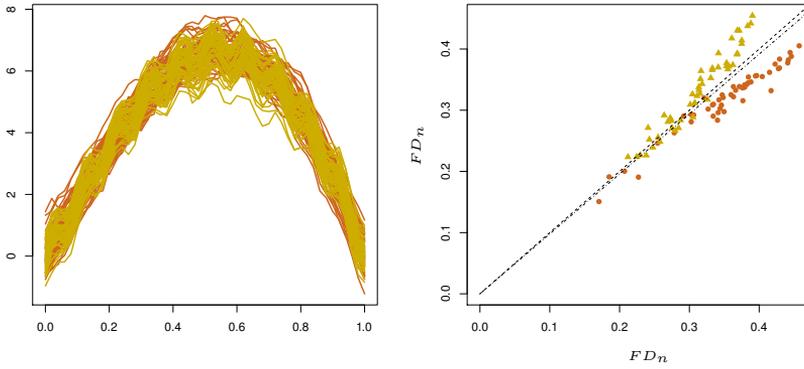


FIGURE 2. FD halfspace depth classification. Very poor resolution of the two DD-transformed clusters.

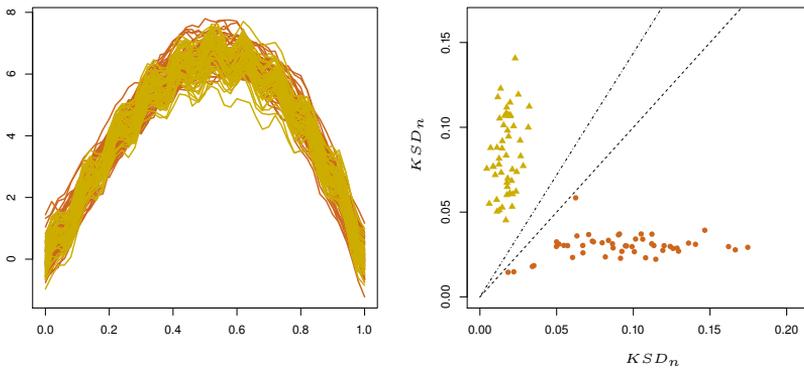


FIGURE 3. Simplicial KSD depth, $K = 1, w_0 = 0, w_1 = 1$. Superior to FD depth, excellent resolution of clusters due to the incorporation of the first derivative.

This approach is applicable for functional data as well.

4 Classification results

The functional depths and DD–plot analysis methods described above will be compared in a short supervised classification simulation study. In every setup, two independent training samples of Gaussian processes of size 50 will be generated and 50 test functions will be classified.

As a result, we show that proper (pre-smoothed versions') derivatives utilizing, even if the process trajectories are not smooth, can improve the pattern recognition substantially, as can be seen in Figures 2 and 3. On the left hand side we can see the training samples, on the right hand side the DD-transformations of the test sample. The dashed line in the DD-plot represents the highest depth decision rule and the dot-dashed line the linear DD-plot classifier.

Similarly other models, such as mean function shift, variance difference, and various combinations of them can be considered in order to stress out the importance of DD-plot analysis methods as well as derivatives incorporating. It appears to be clear that although the DD-plot classifier is quite good its performance may be improved (or vice-versa) by using suitable depth function.

Acknowledgments: The work was supported by the grant SVV 265315/2012.

References

- Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, **10**, 419–440.
- Ghosh, A.K. and Chaudhuri, P. (2005). On Maximum Depth and Related Classifiers. *Scandinavian Journal of Statistics*, **2**, 1467–1469.
- Li, J., Cuesta-Albertos, J.A., Liu, R.Y. (2010). DD-Classifier: Nonparametric Classification Procedure Based on DD-plot. *To appear in Journal of the American Statistical Association*
- Liu, R.Y., Parelius, J.M, Singh, K (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Ann. Stat.*, **27**, 783–858.
- López-Pintado, S. and Romo, J. (2007). Depth-based inference for functional data. *Comput. Stat. Data Anal.*, **51**, 4957–4968.
- López-Pintado, S. and Romo, J. (2009). On the Concept of Depth for Functional Data. *Journal of the American Statistical Association*, **104**, 718–734.
- Tukey, J.W. (1975). Mathematics and the picturing of data. In: *Proc. int. Congr. Math., Vol. 2*, Vancouver, pp. 523–531,
- Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Ann. Stat.*, **28**, 461–482.

Statistical models for deficient count data

Gerhard Neubauer¹

¹ University of Technology, Graz, Austria

E-mail for correspondence: ge.neubauer@aon.at

Abstract: Counting events is a basic data collection technique in any research area. A structure or system designed to observe and record the occurrence of a certain event may make errors. If events that occur are not recognized by the system we speak under-reporting. And reporting events although they never occurred is known as over-reporting. The most prominent example of under-reporting are crime data: often crime figures are lower than the actual number of crimes. Over-reporting is less considered, but insurance companies suspect that not every claim for a stolen article is based on theft. Thus models that allow for both kinds of error are of special interest. We present an overview on models for under-reporting that are based on randomized binomial sampling schemes. Further we show how these models can be extended to cover over-reporting and finally give the results of application to real data.

Keywords: Under- and over-reporting, regression, crime data

1 Introduction

Deficient reporting or register systems are likely to occur in various fields of application. Examples are criminology, epidemiology and production. The official number of crimes is likely to be lower than the actual number for various reasons. Whenever a counting system makes this error we speak of under-reporting. Reporting events that never took place is another kind of error such a system can make, and this is known as over-reporting. Insurance companies suspect that not every claim for a stolen article is based on theft. Typical articles are bicycles or skiing equipment. For health data like cause of death, errors may occur due to errors in diagnosis. Any misclassification will cause both errors in the counting system: an under-reported count in the true category, and an over-reported count in the actual category. For industrial production it is most important to get accurate feedback on the product failure rate from the user. Estimates are often based on warranty claims. For different reasons not every claim is actually made (under-reporting) and not every claim made is justified (over-reporting). Therefore considering both kinds of error is of great importance when modelling count data.

2 Models for over- and under-reporting

Cameron and Trivedi (1998) present a layout for over- and under-reporting based on the crosstabulation of the binary variables

$$E = \left\{ \begin{array}{l} 1 \text{ if an event occurs} \\ 0 \text{ otherwise} \end{array} \right\}, \text{ and } R = \left\{ \begin{array}{l} 1 \text{ if a record is made} \\ 0 \text{ otherwise} \end{array} \right\}$$

and the intention is to model the dependence between the two Bernoulli variables E (event) and R (recording). Table 1 shows this crosstable to the left. In the main diagonal we have the correct counts C_0 and C_1 , and the off-diagonal entries O and U denote the number of over- and under-reported cases. In statistical testing the O and U are known as type 1 and type 2

TABLE 1. Two approaches for modelling over- and under-reporting

	Bivariate Bernoulli sampling			Two independent samplings		
	R = 0	R = 1		R = 0	R = 1	
E = 0	C_0	O	N_0		O	
E = 1	U	C_1	N_1	U	C_1	N
	$N - Y$	Y	N		Y	

error. In the margins we have Y , the observed count of reported events and the three unknown quantities N , N_0 and N_1 . When counting attributes of persons or animals, the total number of individuals N is known. For epidemiological data N is the population size, N_1 is the number of subjects showing a certain disease and N_0 is the number of subjects without this disease. Counts of events (e.g. crimes) are not necessarily related to a population. Hence the quantity C_0 does not exist and the crosstable approach is not useful.

For situations where C_0 does not exist and over-reporting does not occur, models for under-reporting based on a binomial sampling have been proposed by Neubauer et al. (2011). In this case the crosstable in Table 1 is reduced to the second row. The models rely on assuming a conditional Binomial distribution for the observed counts, i.e. $Y|N, P \sim \text{Binomial}(N, P)$, and appropriate mixing distributions for N and P . Examples for marginal models are

- $Y \sim \text{Beta-Binomial}(\lambda, \pi, \theta)$, from the assumptions
 $P \sim \text{Beta}(a, b)$ and $N = \lambda$ (fixed parameter),
- $Y \sim \text{Negative Binomial}(\lambda\pi, \omega)$, from the assumptions
 $N \sim \text{Negative Binomial}(\lambda, \omega)$ and $P = \pi$ (fixed parameter),

- $Y \sim \text{Beta-Poisson}(\lambda, \pi, \theta)$, from the assumptions
 $P \sim \text{Beta}(a, b)$ and $N \sim \text{Poisson}(\lambda)$,

where $E(P) = \pi = a/(a + b)$ and $\theta = a + b$.

For cases where C_0 does not exist but over-reporting may occur, we propose a model where the observed count Y is considered a sum of two independent samplings (see right side of Table 1). Let $Y = C_1 + O$, where C_1 is the number of reports that were given when an event occurred, and O is the number of reports that were given despite that no event occurred. Hence Y is the sum of true reports and over-reports and to model Y we need to specify an under-reporting model for C_1 and a count model for O . Following Neubauer et al. (2011) we assume that C_1 is generated by under-reporting, i.e. $C_1|N, P \sim \text{Binomial}(N, P)$ where N is the unknown number of events and P is the reporting probability. Further we assume that $O \sim \text{Poisson}(\alpha)$. Now the distribution of Y is found as convolution of $p_1(C_1)$, the marginal distribution of C_1 , and $p_0(O)$, where $E(Y) = E(C_1) + E(O)$ and $\text{Var}(Y) = \text{Var}(C_1) + \text{Var}(O)$, as $\text{Cov}(C_1, O) = 0$ by assumption. Specific convolutions are derived from specific marginal models for C_1 . For all models we have $E(Y) = \lambda\pi + \alpha$ and $\text{Var} = \mu\phi + \alpha$, with $\phi = (1 - \pi)(\lambda + \theta)/(1 + \theta)$ for the Beta-Binomial, $\phi = (1 - \pi)^{-1}$ for the Negative-Binomial, and $\phi = 1 + \lambda(1 - \pi)/(1 + \theta)$ for the Beta-Poisson convolution. To allow for greater flexibility in modelling a regression approach is used in the mean decomposition. For a sample of data $y_t, t = 1, \dots, T$ we use $E(Y_t) = \mu_t = \lambda_t\pi + \alpha = \exp(x_t'\beta)\pi + \alpha$, where x_t is a vector of known regressors and β is a vector of unknown parameters. The above setting specifies a family of convolutions, that can be extended if the assumption $O \sim \text{Poisson}(\alpha)$ is replaced by a more general count distribution, and a great variety of possible models can be easily generated.

3 Application to real data

For the analysis of real data we use the under-reporting models and one of the mentioned convolution models, namely the convolution of Negative-Binomial and the Poisson distribution, which is known as Delaporte distribution. The under-reporting models have been implemented in R, and the Delaporte distribution is available in the R package `gamlss` (Stasinopoulos and Rigby, 2007). For model selection we use the non-nested testing approach of Allcroft and Glasbey (2003).

3.1 Bicycle Theft Data

The data used here are weekly counts of bicycle theft in an Austrian city (pop. > 100000) for a period of four years. Figure 1 shows the data together with the results of the analysis. The Beta-Binomial, the Negative Binomial and the Beta-Poisson model were estimated and tested against each other.

The test result is in favour of the Negative Binomial model which has a log likelihood value of $\ell = -747.96$ and an estimated reporting probability of $\hat{\pi} = 0.61$. The model indicates that about 40 % of bicycle thefts are not reported to the police. In absolute values we have $\hat{E}(Y) = 36$, $\hat{E}(U) = 23$ and $\hat{E}(N) = 59$. The main result of the Delaporte model is an estimated over-reporting rate of 43 % and a reporting probability estimate $\hat{\pi} = 0.76$ at a log likelihood value of $\ell = -747.99$. In absolute values we have $\hat{E}(O) = 15$, $\hat{E}(C_1) = 20$, $\hat{E}(U) = 5$ and $\hat{E}(N) = 25$. So the two models have quite different results and interpretation. Using the non-nested test for model choice we get a non conclusive result. Both models fit the data equally well. Although we cannot base it on a formal test we favour the Delaport model. It allows for both errors and thus offers a greater insight in the data.

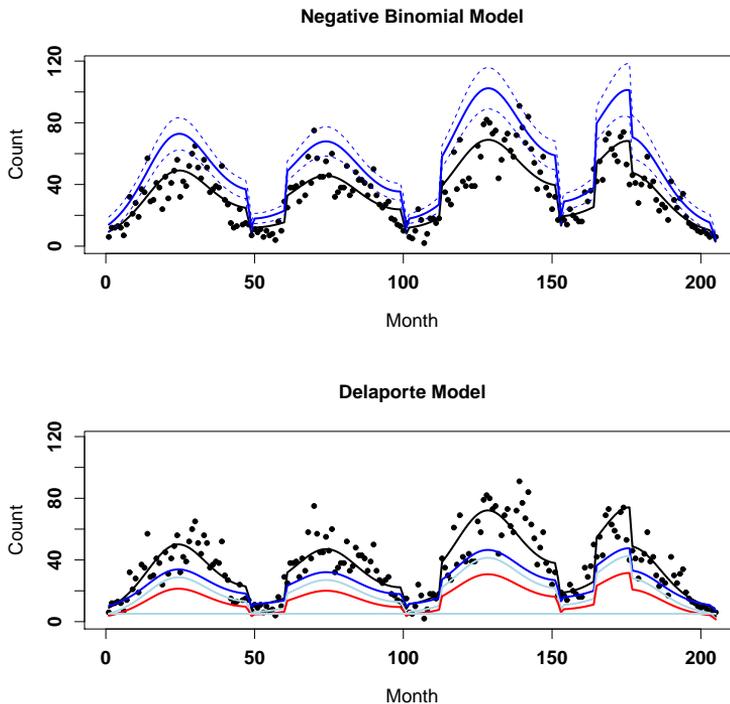


FIGURE 1. Bicycle theft data and estimates

Top: Estimated mean (black), estimated total number of bicycle theft with confidence interval (blue).

Bottom: Estimated mean (black), estimated total number of bicycle theft (blue), components of the total number (light blue), over-reported cases (red).

3.2 Heart Attack Data

The data used here are monthly counts of hospitalizations due to a heart attack in Styria for a period of nine years. Experts presume that there are two reasons why heart attacks are not seen in hospitals: (i) people die before reaching a hospital, and (ii) slight heart attacks are hardly even noticed by the person concerned. Besides that there is a general concern about the quality of hospital discharge data. Therefore we apply our methods to the hospital data. Again the Beta-Binomial, the Negative Binomial and the beta-Poisson model were estimated and tested against each other and again the test result is in favour of the Negative Binomial model which has log likelihood value of $\ell = -419.23$ and an estimated reporting probability of $\hat{\pi} = 0.56$. The model indicates that about 45 % of heart attacks are not seen in hospitals, which is quite a lot. In absolute values we have $\hat{E}(Y) = 163$, $\hat{E}(U) = 129$ and $\hat{E}(N) = 292$. The Delaporte model estimates are essentially the same as the Negative Binomial model estimates. So there we have no evidence for over-reporting, but the estimated under-reporting is quite high. As fatal heart attacks are a possible reason for under-reported hospital data we consider also cause of death data. Figure 2 shows the hospital data as black dots and the estimated mean as black line. The blue dots depict the sum of the hospital data and the cause of death data. And the blue lines show the estimated total number of heart attacks together with pointwise confidence intervals. We see that the confidence interval for the total number of heart attacks covers the sum of the two data sources very well. Note that the estimates are based on the hospital data only.

4 Conclusion

We propose a method for the estimation of over- and under-reporting. It is based on modelling count data as the sum of two independent components, where one component comprises the under-reporting and the other the over-reporting aspect. We apply the method to two data sets with counts of bicycle theft and counts of heart attacks. The analysis of the crime data is inconclusive. Both approaches give contradicting results and moreover a non-nested test does not distinguish between the two models. In the analysis of the hospital data both methods coincide, and moreover the estimated total number of heart attacks is validated by additional cause of death data.

References

- Allcroft, D.J. and Glasbey, Ch.A. (2003) A simulation-based method for model evaluation, *Statistical Modelling*, Vol. **3**, pp. 1-14.
- Cameron, A.C. and Trivedi, P.K. (1998) *Regression analysis of count data*. Cambridge: Cambridge University Press.

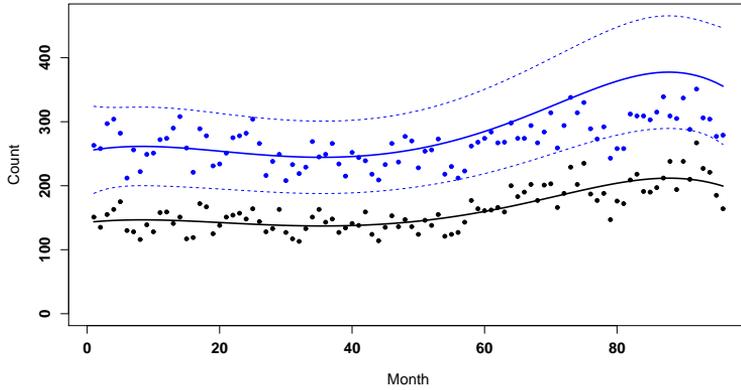


FIGURE 2. Heart attack data and estimates

Hospital data as black dots; sum of hospital data and cause of death data as blue dots. Estimated mean (black solid), estimated total number of heart attacks with confidence interval (blue dashed).

Neubauer, G., Djuras, G. and Friedl, H. (2011) Models for Underreporting: A Bernoulli Sampling Approach for Reported Counts. *Austrian Journal of Statistics*, 40, 85-92.

Stasinopoulos, D.M. and Rigby, R.A. (2007) Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, Vol. 23, Issue 7, Dec 2007,

Testing goodness-of-fit of the Accelerated Failure Time model with time-varying covariates

Petr Novák^{1,2}

¹ Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic

² Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic

E-mail for correspondence: novakp@karlin.mff.cuni.cz

Abstract: The accelerated failure time model presents a way to easily describe and interpret survival regression data. We assume, that each observed unit ages internally faster or slower, depending on the covariate values. It is desirable to check if observed data fit the model assumptions, therefore we present a goodness-of-fit testing procedure based on modern martingale theory. We work with the generalized model introduced in Cox & Oakes (1984) and studied in Lin & Ying (1995), which allows for covariates that change through time. We focus on particular important situations where time-varying covariates are used, such as when an additional factor is added during the observation or when the influence of one covariate gradually increases. On simulated data we estimate the empirical properties of the test.

Keywords: Accelerated Failure Time model; Survival Analysis; Goodness-of-fit.

1 Introduction

Let us observe survival regression data representing time to failure with possible right censoring. We want to describe the dependence of the failure time distribution on available, possibly time-dependent covariates. Suppose T_i^* , $i = 1, \dots, n$, are the real failure times, $Z_i(t) = (Z_{i1}(t), \dots, Z_{ip}(t))^T$ covariates and β_0 the vector of real parameters. Denote C_i the censoring times, $T_i = \min(T_i^*, C_i)$ the times of the end of observation and $\Delta_i = I(T_i^* \leq C_i)$ noncensoring indicators. Suppose T_i^* and C_i are independent for all i .

We assume that T_i^* are independent, continuous with d.f. $F_i(t)$, density $f_i(t)$, hazard function $\alpha_i(t)$ and $A_i(t) = \int_0^t \alpha_i(s) ds$ the cumulative hazard. The classic Accelerated Failure Time model (AFT, see Buckley & James,

1979) was suggested as the log-linear transformation

$$\log T_i^* = -Z_i^T \beta_0 + \epsilon_i,$$

with ϵ_i iid. It can be generalized to accommodate time-varying covariates (see Cox & Oakes, 1984 and Lin & Ying, 1995) by taking T_i^* as the solution of

$$e^{\epsilon_i} = \int_0^{T_i^*} e^{Z_i^T(s)\beta} ds =: h_i(T_i^*, \beta).$$

The function $t \rightarrow h_i(t, \beta)$ can be understood as the backward time transformation from virtual to observed age.

The data may be represented as counting processes. Denote the noncensored-failure and the at-risk indicators as $N_i(t) = I(T_i \leq t, \Delta_i = 1)$, $Y_i(t) = I(t \leq T_i)$, intensities $\lambda_i(t) = Y_i(t)\alpha_i(t)$ and cumulative intensities $\Lambda_i(t) = \int_0^t \lambda_i(s)ds$. All functions and processes are studied on an interval $t \in [0, \tau]$, where $\tau < \infty$ is some point beyond the last observed survival time. It can be shown, that under the model assumptions $M_i(t) := N_i(t) - \Lambda_i(t)$ are martingales with respect to $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), Z_i, 0 \leq s \leq t, i = 1, \dots, n\}$ (see Fleming & Harrington, 1992). The log-likelihood for the data can be written as

$$l(t) = \sum_{i=1}^n \int_0^t (\log(\alpha_i(s))dN_i(s) - Y_i(s)\alpha_i(s)ds).$$

Let $U(t, \beta) = \frac{d}{d\beta}l(t)$ be the score process. The parameter estimates are taken as solutions to score equations $U(\tau, \beta) = 0$.

To obtain reliable estimates, the model assumptions must be met. However, the data can deviate from the model, for example if the dependence is different than log-linear or if we neglected one or more covariates. We present a goodness-of-fit test for the AFT model based on martingale approach and resampling techniques. We then study the empirical properties of the test in various situations on simulated examples.

2 The test statistics

We define the transformed counting processes as

$$\begin{aligned} N_i^*(t, \beta) &= N_i(h_i^{-1}(t, \beta)) = \Delta_i I(h_i(T_i, \beta) \leq t) \\ Y_i^*(t, \beta) &= Y_i(h_i^{-1}(t, \beta)), \quad Z_i^*(t, \beta) = Z_i(h_i^{-1}(t, \beta)). \end{aligned}$$

Denote

$$S_0^*(t, \beta) = \sum_{i=1}^n Y_i^*(t, \beta), \quad S_1^*(t, \beta) = \sum_{i=1}^n Y_i^*(t, \beta)Z_i^*(t, \beta),$$

$$E^*(t, \beta) = \frac{S_1^*(t, \beta)}{S_0^*(t, \beta)}, \quad \hat{A}_0(t, \beta) = \int_0^t \frac{dN_{\bullet}^*(s, \beta)}{S_0^*(s, \beta)}.$$

The score process has the form

$$U(t, \beta) = \sum_{i=1}^n \int_0^t Q(s)(Z_i^*(s, \beta) - E^*(s, \beta))dN_i^*(s, \beta),$$

with $Q(s) = (\frac{s\alpha'_0(s)}{\alpha_0(s)} + 1)$. Unbiased consistent estimators of β can be obtained also with $Q_1(s) \equiv 1$ (Lin et al, 1998). We work with time-transformed martingales

$$M_i^*(t, \beta) = N_i^*(t, \beta) - \int_0^t Y_i^*(s, \beta)dA_0(s, \beta)$$

and their estimates $\hat{M}_i^*(t, \beta)$ obtained by inserting $\hat{A}_0(t, \beta)$. The score process can be replicated with $G_i, i = 1, \dots, n$ as *iid* standard normals. Let

$$U^G(t, \beta) = \sum_{i=1}^n \int_0^t (Z_i^*(s, \beta) - E^*(s, \beta))d\hat{M}_i^*(s, \beta)G_i.$$

Take $\hat{\beta}^*$ as the solution of the equation $U(\beta) = U^G(\hat{\beta})$. Then $\sqrt{n}(\hat{\beta} - \beta_0)$ has asymptotically the same distribution as $\sqrt{n}(\hat{\beta} - \hat{\beta}^*)$ (Lin et al, 1998).

Let $w_i = f(Z_i(t_0))$ be p-dimensional weight vectors with a bounded function f of the covariates at a fixed time-point t_0 . Let $U_w^G(t, \beta)$ and $S_w^*(t, \beta)$ be the same as $U^G(t, \beta)$ and $S^*(t, \beta)$, respectively, with w_i inserted in place of Z_i . Denote $H(t, \beta) = \frac{\partial}{\partial \beta} \left(-h_i(h_i^{-1}(t, \beta), \beta_0) \right)$ and

$$f_N(t) = \frac{1}{n} \sum_i \Delta_i w_i f_0(t)H(t, \beta_0), \quad f_Y(t) = \frac{1}{n} \sum_i w_i g_0(t)H(t, \beta_0),$$

where $f_0(t)$ and $g_0(t)$ are the baseline densities of e^{ϵ_i} and $h_i(T_i, \beta_0)$, respectively. Let \hat{f}_N and \hat{f}_Y be their empirical counterparts with kernel estimates $\hat{f}_0(t)$ and $\hat{g}_0(t)$. We consider the test process

$$W(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \hat{M}_i^*(t, \hat{\beta}).$$

Under certain regularity assumptions, if the model holds, then the process $W(t)$ has the same limiting zero-mean Gaussian distribution as the process

$$\hat{W}(t) = \frac{1}{\sqrt{n}} U_w^G(t, \hat{\beta}) - \sqrt{n} \left(\hat{f}_N(t) + \int_0^t \hat{f}_Y(s)d\hat{A}_0(s, \hat{\beta}) \right)^T (\hat{\beta} - \hat{\beta}^*)$$

$$-\frac{1}{\sqrt{n}} \int_0^t S_w^*(s, \hat{\beta}) d(\hat{A}_0(s, \hat{\beta}) - \hat{A}_0(s, \hat{\beta}^*)),$$

which can be resampled with components from above (see Novák, 2011). For testing whether the AFT model fits the data well, we choose appropriate weights w_i , compute the values of the process $W(t)$ from the observed data and replicate $\hat{W}(t)$ (i.e. $200\times$, $500\times$ or $1000\times$). As the test statistics we can take

$$\sup_{t \in [0, \tau]} |W(t)| \quad \text{or} \quad \sup_{t \in [0, \tau]} \left| \frac{W(t)}{\sqrt{\widehat{\text{var}}W(t)}} \right|$$

with a suitable variance estimator. The hypothesis that the data follow the AFT model is rejected, if the statistics computed from the data exceeds $(1 - \alpha)\%$ of the statistics from the replicated $\hat{W}(t)$. We can compute the supremum also only over some subset of $[0, \tau]$.

3 Special cases

If some constant influence is added for each individual at a different time, say s_i , we can take:

$$Z_i(t) = \begin{cases} 1 & t > s_i \\ 0 & t \leq s_i. \end{cases}$$

Then $h_i(t, \beta) = \min(t, s_i) + e^{\beta(t - s_i)^+}$, which means that at the time s_i , the individual i starts to age internally faster or slower by the factor of e^{β} . If the influence added at time s_i is different for each individual, we can take:

$$Z_i(t) = \begin{cases} Z_i & t > s_i \\ 0 & t \leq s_i, \end{cases}$$

Then $h_i(t, \beta) = \min(t, s_i) + e^{Z_i\beta(t - s_i)^+}$. At the time s_i , the individual starts to age internally faster or slower by the factor of $e^{Z_i\beta}$. If the effect of one covariate increases gradually, it can be modelled as:

$$Z_i(t) = \log(1 + t)Z_i$$

or with some other strictly increasing non-negative function instead of $\log(1 + t)$.

4 Simulation study

We want to estimate the empirical properties of the test in various situations. We simulate data from the three presented models to estimate the empirical level of significance and again with one confounding covariate for each case satisfying $e^{\epsilon_i} = \int_0^{T_i^*} e^{Z_i(t)\beta_1 + X_i\beta_2} dt$ to estimate the empirical power. We take $\beta_1, \beta_2 = 1$, $Z_i, X_i \sim N(3, 1)$, $s_i \sim LN(4, 1)$ and log-normal

baseline distribution $e^{\epsilon_i} \sim LN(\mu = 3, \sigma^2 = 1)$. 500 samples were generated, with sample sizes 200 and 1000 of both non-censored (NC) and with about 1/4 censored (C). $\hat{W}(t)$ was replicated 200×, for testing we use $\sup_{t \in [0, \tau]} \left| \frac{W(t)}{\sqrt{\text{var}W(t)}} \right|$ with the variance estimated from the replicated processes.

The variance-standardising helps to improve the empirical power, because if the model does hold only for higher or lower failure times, the non-standardised supremum statistics may not be able to detect the deviation in the other part. See example on Figure 1, where the deviation from the model for lower failure times would remain undetected by the non-standardised statistics.

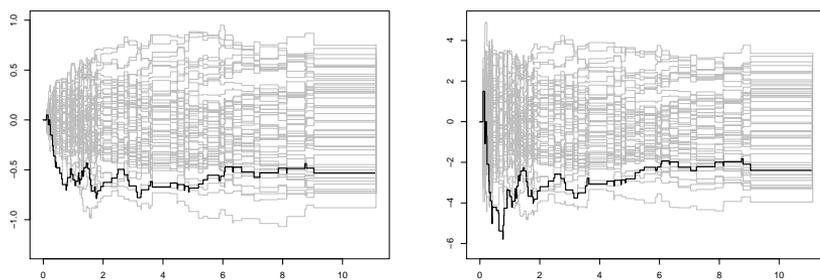


FIGURE 1. Left: The observed process $W(t)$ (bold) and its 50 replications $\hat{W}(t)$ under the hypothesis of the AFT model. Right: the variance-standardised version.

The empirical level of significance was always below 5%. For the results for empirical power, see Table 1. The empirical power is higher with larger sample. With censoring, the power diminishes somewhat. The power is reasonably high in most cases even with $n = 200$, sometimes even equal to one. Some weight choices give considerably better results than others.

TABLE 1. The empirical power against the respective model with a confounding covariate.

Model type	Weights	Empirical power			
		$n = 200$		$n = 1000$	
		NC	C	NC	C
$I(t > s_i)$	$I(s_i \leq \text{med}(s))$	0.554	0.482	1	0.970
	$I(Z_i \leq \text{med}(Z))$	0.190	0.242	0.868	0.774
$Z_i I(t > s_i)$	$I(s_i \leq \text{med}(s))$	1	1	1	1
	Z_i	0.216	0.270	0.954	0.888
$\log(1+t)Z_i$	$I(Z_i \leq \text{med}(Z))$	0.694	0.5	0.846	0.642
	$Z_i I(Z_i \leq \text{med}(Z))$	0.352	0.353	0.664	0.442

5 Conclusions & Outlook

We presented a goodness-of-fit procedure usable to check whether the data fit the AFT model with possibly time-varying covariates. It covers a large set of time transformations. More situations can be explored, as are models with more jumps or with transient effects. Empirical properties for various settings (baseline distribution, covariate distribution, alternative of an entirely different model etc.) and weight selection could be further studied.

Acknowledgments: This work was supported by grants SVV 265315 and MŠMT ČR 1M06047.

References

- Buckley, J. and James, I.R. (1979). Linear regression with censored data. *Biometrika*, **66**, 429–436.
- Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.
- Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
- Lin, D.Y., Wei, L.J., and Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika*, **85**, 605–618.
- Lin, D.Y. and Ying, Z. (1995). Semiparametric inference for the accelerated life model with time-dependent covariates. *Journal of Statistical Planning and Inference*, **44**, 47–63.
- Novák, P. (2011) Goodness-of-fit test for the AFT model based on martingale residuals, Research Report 2309, ÚTIA AV ČR.

Estimation of multi-state model parameters from panel data: A comparison of different methods

Ekaterina Ogurtsova¹

¹ Max Planck Institute for Demographic Research, Rostock, Germany

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

Abstract: We compare three methods to estimate the transition rates of a multi-state model from panel data: Maximization of the likelihood assuming piece-wise constant transition rates, maximization via the EM-algorithm and direct numerical solution of the Komogorov differential equations. We assess the performance of the three approaches in a simulation study and show results for an application.

Keywords: Multi-state models; Panel data; Transition rates; Markov models.

1 Introduction

Continuous-time multi-state models are a widely used to describe processes in which individuals go through different stages over time. Examples are disease progression in medical studies or changes in socio-economic position – e.g. marital status or labor force participation – over the life-course. Besides the state-space, which gives the set of potential states, the transition rates are the key parameters in a multi-state model.

Ideally, to estimate these transition rates, complete event-histories (all transitions, exact event times) are available. However, in practise the individuals often are only observed at discrete and possibly irregular intervals so that we may miss some intermittent transitions as well as the exact transition times. Such an observation scheme leads to so called panel data.

Several methods have been proposed to estimate transition rates in multi-state models from panel data. We will compare three different approaches in a simulation study with several scenarios: Direct likelihood maximization, estimation via the EM-algorithm, and a method that solves the Kolmogorov differential equations numerically. Finally, we apply the methods in two studies of health transitions in old age.

2 Multi-state models, transition rates

In this paper we describe an individual life-course by a multi-state Markov model in continuous time. At each time $t \in [0, \infty)$ the individual is in one

of M states, which is denoted by X_t . The state-space $S = \{1, \dots, M\}$ collects all potential states and transitions between the states are governed by the transition rates $\lambda_{ij}(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(X_{t+\Delta t} = j | X_t = i)$. From the transition rates $\lambda_{ij}(t)$ the transition probabilities $P_{ij}(t, u) = P(X_u = j | X_t = i)$ can be derived. If we denote

$$\Lambda(t) = \left(\lambda_{ij}(t) \right) \in \mathbb{R}^{M \times M}$$

with $\lambda_{ii}(t) = - \sum_{j \neq i} \lambda_{ij}(t)$ and $\mathbf{P}(t, u) = (P_{ij}(t, u))_{ij}$, then the transition probabilities satisfy the so called Kolmogorov forward differential equations (Gardiner, 1985)

$$\frac{d\mathbf{P}(t, u)}{du} = \mathbf{P}(t, u)\Lambda(u). \quad (1)$$

If the transition rates are independent of t , i.e., $\lambda_{ij}(t) \equiv \lambda_{ij}$, the process is time-homogeneous and the $P_{ij}(t, u)$ only depend on the difference $u - t$.

3 Estimation from panel data

Estimation of the transition rates $\lambda_{ij}(t)$ is commonly done by maximizing the likelihood, which depends on the observation scheme. For panel data, when the individual is only observed at K discrete time points $t_1 < t_2 < \dots < t_K$, at which it is in the states $X_{t_k} = s_k \in S$, the contribution of this individual to the likelihood function is

$$L = P(X_{t_1} = s_1, \dots, X_{t_K} = s_K) = \prod_{k=1}^K P_{s_{k-1}, s_k}(t_{k-1}, t_k). \quad (2)$$

The transition rates $\lambda_{ij}(t)$ enter in the likelihood function indirectly via the solutions of the Kolmogorov equations (1). For most forms of $\Lambda(t)$, these equations are analytically intractable. For time-homogeneous transition rates, however, the formal solution of the Kolmogorov equations is $\mathbf{P}(t) = \exp\{t\Lambda\}$, but even then the exponential operation may be hard to evaluate numerically. Piecewise-constant rates are a simple alternative, for which the likelihood is still tractable (Jackson, 2011). Here the $\lambda_{ij}(t)$ are step-functions with some pre-chosen cut-points.

Of the three approaches to estimate time-inhomogeneous transition rates in multi-state models, which we study here, two are based on the assumption of piecewise-constant rates, while the third one assumes smooth transition rates that are modelled by B -splines.

The piecewise-constant rates with break-points at $\tau_l, l = 0, \dots, L$, and $\tau_0 = 0, \tau_{L+1} = \infty$, can be written as

$$\lambda_{ij}(t) = \sum_{l=0}^L \lambda_{ij}^l I_{[\tau_l, \tau_{l+1})}(t) \quad (3)$$

with $\lambda_{ij}^l \geq 0$. For notational simplicity we assume that all rates share the same break-points.

The first approach is based on the direct maximization of the likelihood with individual contributions (2). If only one break-point $\tau_l \in (t_{k-1}, t_k)$ then ((2) is equal to

$$L = \prod_{k=1}^K P_{s_{k-1}, s_k}(t_{k-1}, t_k) = \prod_{k=1}^K \sum_{r \in M^*} P_{s_{k-1}, s_r}(t_{k-1}, \tau_l) P_{s_r, s_k}(\tau_l, t_k), \quad (4)$$

where M^* contains all states out of M that can possibly be visited between s_{k-1} and s_k . Equation (4) can be easily extended to more than one cut-point in (t_{k-1}, t_k) .

Alternatively, the EM-algorithm can be used in the piecewise-constant setting, as suggested by Lindsey and Ryan(1998). If complete event-histories are observed, the MLE of λ_{ij}^l is given as the ratio of N_{ij}^l , the number of events (transitions $i \rightarrow j$), and T_i^l , the total time at risk, in interval (τ_l, τ_{l+1}) : $\hat{\lambda}_{ij}^l = N_{ij}^l / T_i^l$. Consequently, in the E-step one calculates $E(N_{ij}^l)$ and $E(T_i^l)$ based on the data and the current parameter values $\lambda = (\lambda_{ij}^l)_{i,j,l}$:

$$E(N_{ij}^l | \text{data}, \lambda) = \frac{(\lambda_{ij}^l)^{\delta_k} \exp \left\{ - \sum_{r=0}^{l-1} \lambda_{ij}^r \Delta_r \right\} \cdot \left\{ \exp [\lambda_{ij}^l \Delta_l] - 1 \right\}}{\sum_{l: \tau_l \in [t_{k-1}, t_k]} (\lambda_{ij}^l)^{\delta_k} \exp \left\{ - \sum_{r=0}^l \lambda_{ij}^r \Delta_r \right\}}$$

$$E(T_i^l | \text{data}, \lambda) = \frac{\Delta_l \cdot (\lambda_{ij}^l)^{\delta_k} \exp \left\{ - \sum_{r=0}^l \lambda_{ij}^r \Delta_r \right\}}{\sum_{l: \tau_l \in [t_{k-1}, t_k]} (\lambda_{ij}^l)^{\delta_k} \exp \left\{ - \sum_{r=0}^l \lambda_{ij}^r \Delta_r \right\}},$$

where $\delta_k = 1$, if the individual experiences an event in (t_{k-1}, t_k) and zero otherwise, and $\Delta_l = \tau_l - \tau_{l-1}$.

A different approach has recently been proposed by Titman (2011). Here the solution of the Kolmogorov equations is obtained numerically and used to compute the likelihood function. The transition rates are assumed to be smooth and are modelled by quadratic B -splines. If $B_r^{ij}(t)$ is a B -spline basis with knots $\tau_r^{ij}, r = 1, \dots, R$, then

$$\lambda_{ij}(t) = \lambda_{ij} \sum_{r=0}^R \alpha_r^{ij} B_r^{ij}(t),$$

with coefficients $\alpha_r^{ij} \geq 0$ and $\alpha_0^{ij} = 1$. Time-homogeneity of $\lambda_{ij}(t)$ is achieved if $\alpha_0^{ij} = \alpha_1^{ij} = \dots = \alpha_R^{ij} = 1$. When $\tau_r^{ij} = t_r$ for all i, j , the model may be thought of as a smoothed version of a model with piecewise-constant rates and break-points at the t_r .

4 Simulation study

We consider a three-state illness-death model, which includes two alive states (*healthy* and *disabled*) and the absorbing state *dead*, see Figure 1. We use the model in two forms, with and without the possibility of a backward transition from the state *disabled* to *healthy*. The process time is the age of the individual.

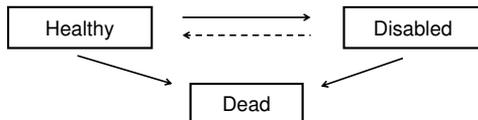


FIGURE 1. Multi-state illness-death model without or with (dashed arrow) potential recovery from the state “disabled”.

The MicMac microsimulation software (Zinn, 2009) was used to simulate the individuals in the study sample. Their behavior is determined by transition rates that are assumed either constant or varying with age. Because of space limitations we will focus on the scenarios with time-varying rates here, which are shown in Figure 2. The initial population mimics the age structure of the Dutch population in 2004 and includes about 400,000 individuals. The values of death transition rates are taken from the Eurostat projection and the rates for the onset of disability and the recovery are modelled as a linear growth and a slow exponential decline, respectively. The age range is $[0, 101]$. The simulation was run for 40 years.

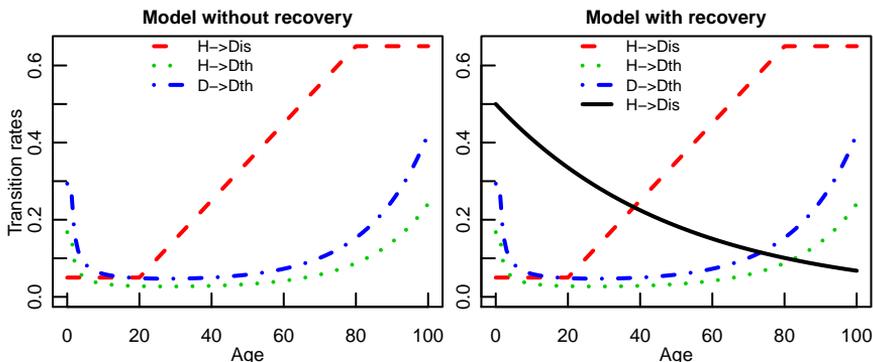


FIGURE 2. Transition rates for the simulation study.

We create panel samples, each with five waves, under several different assumptions.

5 Result of the simulation study

For each simulated data set we compare three estimation procedures: the maximum likelihood estimation (MLE), the direct solution of the Kolmogorov differential equations (DSofDE), and the EM-algorithm.

In Figure 3 we show the relative deviations for different distances between the panel waves and over the age groups, each with width ten years between 0 and 100. Increasing interval lengths between panel observations have particularly strong influence on the results of the EM-algorithm. Moreover, the procedure strongly underestimates the rates for transitions to disability and recovery rates, while overestimating death rates.

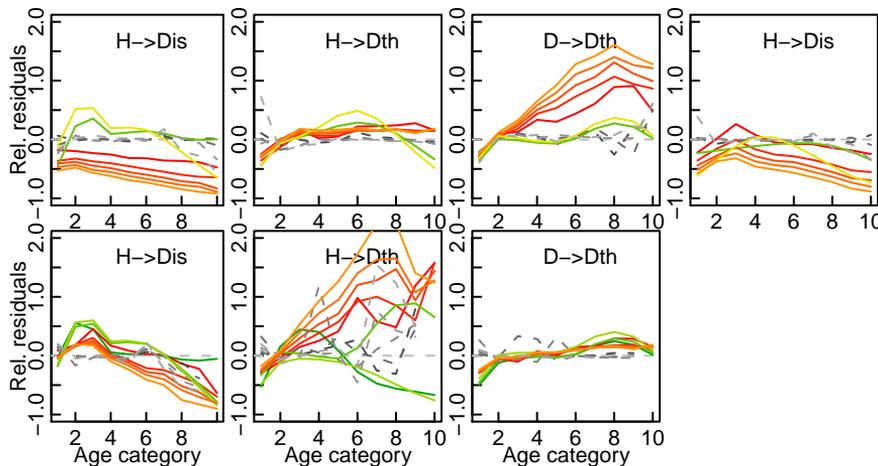


FIGURE 3. Relative deviations from the original rates for different lengths of intervals (with lighter colors display longer intervals) and over the age groups. Top row: model with recovery. Bottom row: model without recovery. The dashed lines represent the MLE-estimation, red solid lines the EM-algorithm, and green solid lines give the estimates based on the numerical solution of the Kolmogorov equations.

6 Application

To demonstrate the procedures in a practical application we estimated the illness-death model shown in Figure 1 for two panel data sets: The US Health and Retirement Study (HRS) and the English Longitudinal Study of Ageing (ELSA). The HRS sample was collected from 1992 to 2008, covering 10 waves. The ELSA data sample was drawn from private household in England and is composed of 4 rounds made in 2002 – 2009.

We analyze the model with recovery. The disability status is defined by the Katz basic activities of daily living (ADL). Only individuals who were aged 55 or older and who participated at least in two interviews were included. The results are shown in Figure 4 for females. We estimated the model both for time-homogeneous transition rates and for age-dependent rates (piecewise constant). The estimated age-trajectories are consistent with previous findings. Transition rates to and from disability are higher in ELSA than in the HRS, while death rates are lower.

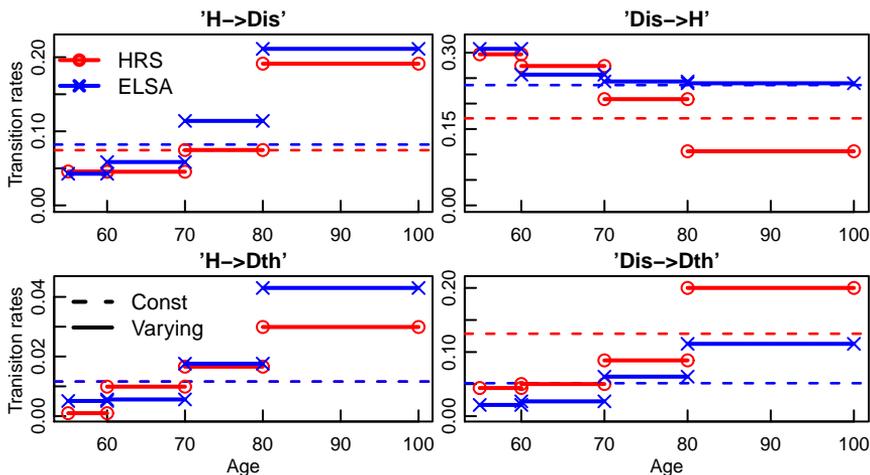


FIGURE 4. Estimates for females in the HRS (red lines, circles) and ELSA (blue lines, crosses) for the illness-death model with recovery; MLE estimates. Dashed lines: time-constant rates. Solid lines: piecewise-constant transition rates.

References

Gardiner, C. (1985) *Stochastic Methods. A Handbook for the Natural and Social Sciences*. New-York: Springer.

Jackson, C. (2011). Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*, **28(8)**, 1–29.

Lindsey, J., Ryan, L (1998). Tutorial in Biostatistics Methods for Interval-Censored Data. *Statistics in Medicine*, **17**, 219–238.

Titman, A. (2011). Flexible Non-homogeneous Markov Models for Panel Observation Data. *Biometrics*, **67(3)**, 780–787.

Zinn, S., Gampe, J., Himmelspach, J. and Uhrmacher, A.M. (2009). MIC-CORE: a tool for microsimulation. *Proceedings of the 2009 Winter Simulation Conference*.

Nonparametric circular density estimation for temperature cycles

Maria Oliveira Perez, Rosa M. Crujeiras, A. Rodríguez-Casal¹

¹ Department of Statistics and Operations Research, University of Santiago de Compostela (Spain)

E-mail for correspondence: maria.oliveira@usc.es

Abstract: A new plug-in rule procedure for bandwidth selection in kernel circular density estimation is introduced. The performance of this proposal is checked throughout a simulation study considering a variety of circular distributions exhibiting multimodality, peakedness and/or skewness. The plug-in rule behaviour is also compared with other existing bandwidth selectors. The method is applied to analyze a real data example belonging to the International Polar Year project. SiZer for circular data is proposed and applied for real data analysis.

Keywords: bandwidth selection; circular density; kernel estimator; von Mises distribution.

1 Introduction

The International Polar Year addresses as one of the main subjects the quantification and understanding of the environmental change in the polar regions, in particular, monitoring the retreat of glaciers is in the scope of this project. Measurement stations were placed in glacier Monte Alvear (Argentina), recording temperatures hourly at different depths. The occurrence of changes in cycles of temperature (from frosting to defrosting and viceversa) are important for the analysis of the mobility in the glacier's surface. The hours where a cycle change has occurred constitute a sample of *circular data*, coming from an unknown circular distribution, that must be estimated in order to determine the cycle change behaviour.

Circular data appear in a large variety of disciplines, such as ecology, meteorology or environmental sciences, being its analysis approached from parametric and nonparametric perspectives. Since the data motivating this work did not seem to follow any simple parametric distribution model, the kernel density estimator introduced by Hall et al. (1987) will be considered for its analysis. Like any other smoothing procedure, a bandwidth or smoothing parameter must be chosen. A rule of thumb has been recently introduced by Taylor (2008), but the performance of this selector may be extremely poor in some distribution settings involving multimodal-

ity, peakedness or skewness. A new plug-in procedure for selecting the smoothing parameter in kernel circular density estimation will be introduced, jointly with an exploratory tool for detecting significant features. This paper is organized as follows. Section 2 is devoted to the introduction of the kernel density estimator for circular data, revising existing bandwidth selection procedures and introducing the new method. The performance of the new rule is checked in a simulation study in Section 3. The analysis of temperature cycle changes in Monte Alvear is carried out in Section 4.

2 Circular kernel density estimation

Given a random sample of angles $\Theta_1, \Theta_2, \dots, \Theta_n \in [0, 2\pi)$ from a density f , the kernel circular density estimator of f is defined as:

$$\hat{f}(\theta; \nu) = \frac{1}{n} \sum_{i=1}^n K_\nu(\theta - \Theta_i), \quad 0 \leq \theta < 2\pi,$$

where K_ν is the circular kernel function with concentration parameter $\nu > 0$ (see Di Marzio et al., 2009). As a circular kernel, the von Mises distribution can be considered. The von Mises distribution, $vM(\mu, \kappa)$, is a symmetric unimodal distribution characterized by a mean direction $\mu \in [0, 2\pi)$, and a concentration parameter $\kappa \geq 0$, with probability density function

$$g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu)\}, \quad 0 \leq \theta < 2\pi,$$

where I_r denotes the modified Bessel function of order r . With this specific kernel, the density estimator is given by:

$$\hat{f}(\theta; \nu) = \frac{1}{n(2\pi)I_0(\nu)} \sum_{i=1}^n \exp\{\nu \cos(\theta - \Theta_i)\}, \quad 0 \leq \theta < 2\pi. \quad (1)$$

A critical issue is the choice of the smoothing parameter ν for (1), with large values leading to highly variable (undersmoothed) estimators and small values showing an opposite behaviour. The SiZer map introduced by Chaudhuri and Marron (1999) for linear data, is a graphical tool that explores simultaneously a wide range of bandwidths, seeking for significant features. An adaptation to the circular setting, the CSiZer, will be used in Section 4 for the data analysis.

The bandwidth parameter is usually selected in order to minimize some error criterion, such as the mean integrated squared error (MISE, $MISE(\nu) = \mathbb{E}(\int (\hat{f} - f)^2)$). The asymptotic expression for the MISE (AMISE) is derived by Di Marzio et al. (2009). For the circular kernel estimator (1), the AMISE(ν) when $\nu \rightarrow \infty$ and $\sqrt{\nu}n^{-1} \rightarrow 0$ is given by:

$$AMISE(\nu) = \left\{ \frac{1}{16} \left[1 - \frac{I_2(\nu)}{I_0(\nu)} \right]^2 \int_0^{2\pi} [f''(\theta)]^2 d\theta + \frac{I_0(2\nu)}{2n\pi (I_0(\nu))^2} \right\}, \quad (2)$$

where f'' denotes the second-order derivative of the target density to be estimated, which measures the curvature of f . A *rule of thumb* was proposed by Taylor (2008). Assuming that the data follow a von Mises distribution with concentration parameter κ , the bandwidth minimizing the AMISE can be estimated by

$$\hat{\nu}_{RT} = \left[\frac{3n\hat{\kappa}^2 I_2(2\hat{\kappa})}{4\pi^{1/2} I_0(\hat{\kappa})^2} \right]^{2/5}, \tag{3}$$

where $\hat{\kappa}$ is obtained by maximum likelihood. This selector performs satisfactorily in fitting unimodal symmetric distributions, without highly peaked modes but its behaviour can be dramatically misleading in the presence of antipodal modes and/or skewed distributions (see Section 3).

An alternative route would be to plug-in a more flexible distribution family as a reference density in the AMISE. For that purpose, a mixture of von Mises can be considered. A finite mixture of M von Mises distributions, $vM(\mu_i, \kappa_i)$ with proportions $\alpha_i, i = 1, \dots, M$, has density:

$$g(\theta) = \sum_{i=1}^M \alpha_i \frac{\exp\{\kappa_i \cos(\theta - \mu_i)\}}{2\pi I_0(\kappa_i)}, \quad \text{with} \quad \sum_{i=1}^M \alpha_i = 1. \tag{4}$$

The proposed plug-in bandwidth selector, $\hat{\nu}_{PI}$, is obtained as follows (see Oliveira et al. 2012 for further details):

- Step 1. Based on the sample information, select the number of mixture components M for the reference distribution.
- Step 2. Estimate the parameters in the von Mises mixture (4), $(\mu_i, \kappa_i, \alpha_i)$, for $i = 1, \dots, M$ and compute the integral $\int (\hat{g}''(\theta))^2 d\theta$. Plug-in this quantity in the AMISE expression (2) to get $\widehat{\text{AMISE}}(\nu)$.
- Step 3. Minimize $\widehat{\text{AMISE}}(\nu)$ and obtain $\hat{\nu}_{PI}$.

For Step 1, the selection of the number of mixture components in the reference distribution can be done by AIC, considering different numbers of mixtures. Maximum likelihood estimation via EM algorithm is used for Step 2 (see Banerjee et al., 2005).

Other alternatives to smoothing parameter selection are the cross-validation rules proposed by Hall et al. (1987). The likelihood cross-validation bandwidth $\hat{\nu}_{LCV}$ is obtained by maximizing:

$$LCV(\nu) = \prod_{i=1}^n \hat{f}_{-i}(\theta_i; \nu), \tag{5}$$

where \hat{f}_{-i} denotes the circular kernel density estimator (1) leaving out the i -th observation. Our empirical experiments show a more stable behaviour of this selector compared with the classical least-squares cross-validation method (see also Taylor, 2008).

3 Simulation study

A variety of circular distributions displaying multimodality, skewness and/or peakedness have been tried: von Mises (M1), wrapped skew-normal (M2), mixture of two von Mises (M3), mixture of von Mises and wrapped Cauchy (M4), mixture of Cardioid and wrapped Cauchy (M5) and mixtures of three, four and five von Mises (M6, M7 and M8). See Figure 1 for plots. Technical details on these distribution models can be found in Mardia and Jupp (2000) and Pewsey (2000).

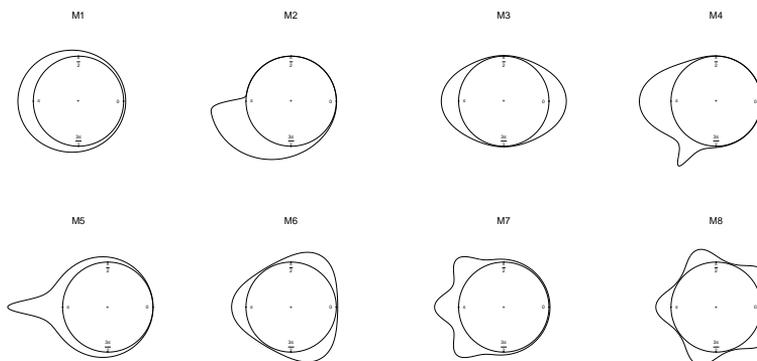


FIGURE 1. Circular density models.

For each distribution model, 1000 random samples of sizes $n = 250$ were generated. In Table 1, the average integrated squared errors (denoted by $MISE(\nu)$ for simplicity) of the circular kernel density estimator (1), considering different bandwidth selectors are shown. As a benchmark, the minimum average ISE has been computed for a broad grid of bandwidth parameters, denoted in the table by $MISE(\nu_0)$.

For simple models such as model M1, all bandwidth selectors show a similar behaviour. The performance of $\hat{\nu}_{RT}$ in an asymmetric distribution (M2) is far from satisfactory. For mixtures, the performance of $\hat{\nu}_{RT}$ is quite poor, specially in highly peaked (M5) models and models with antipodal modes (M3). The plug-in rule $\hat{\nu}_{PI}$ provides good results for all the models, whereas $\hat{\nu}_{LCV}$ seems to be a competitor except for models M4 and M5. Simulations were also obtained for other models and sample sizes and results showed that, for moderate and large sample sizes, the proposed plug-in selector is the best for most models, and in general, it is always a good alternative (see Oliveira et al., 2012).

TABLE 1. Average integrated squared error for different bandwidth selectors, MISE ($\times 100$), and standard deviations ($\times 100$, in parentheses). Bandwidth selectors: $\hat{\nu}_{RT}$ (rule of thumb), $\hat{\nu}_{PI}$ (plug-in rule), $\hat{\nu}_{LCV}$ (likelihood cross-validation). $MISE(\nu_0)$: benchmark average integrated squared error. Sample size: $n = 250$.

$n = 250$	$MISE(\nu_0)$	$MISE(\hat{\nu}_{RT})$	$MISE(\hat{\nu}_{PI})$	$MISE(\hat{\nu}_{LCV})$
M1	0.2568	0.3201 (0.2211)	0.3499 (0.2866)	0.3610 (0.2891)
M2	1.3422	2.1665 (0.5032)	1.6544 (0.7329)	1.5842 (0.6379)
M3	0.5762	10.6753 (0.1786)	0.5986 (0.3400)	0.5976 (0.2917)
M4	1.3545	2.0187 (0.4325)	1.5816 (0.6363)	2.0316 (0.6941)
M5	1.8929	6.6517 (0.7443)	2.2035 (0.8596)	3.2325 (1.2213)
M6	0.6766	6.4797 (0.0016)	0.7358 (0.3678)	0.7368 (0.3278)
M7	1.1325	2.9559 (0.2953)	1.3480 (0.6393)	1.4273 (0.5726)
M8	1.1141	7.8224 (0.0117)	1.1355 (0.3753)	1.1473 (0.3857)

4 Data analysis

The kernel circular density estimator has been applied to explore the distribution of changes in cycles of temperatures at ground level in Monte Alvear. The dataset includes 350 hourly cycle temperature changes. Fits for the circular kernel density estimator (1) are shown in Figure 2 (left). The rule of thumb, $\hat{\nu}_{RT}$, provides an oversmoothed estimate, close to the uniform circular distribution, although a mode at 11 a.m. is identified by the estimators with $\hat{\nu}_{LCV}$ and $\hat{\nu}_{PI}$.

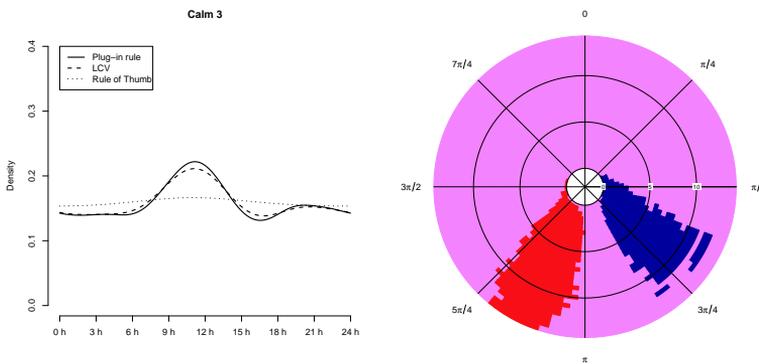


FIGURE 2. Left panel: Kernel density estimator for cycle changes in temperatures in Monte Alvear with plug-in, likelihood cross-validation and rule of thumb selectors. Right panel: Circular SiZer (CSiZer) map.

In order to check if the mode is significant, a modification of the SiZer map is proposed. SiZer (SIGNificance of ZERO crossings of smoothed estimates) is a visualization method based on nonparametric curve estimates. SiZer

involves testing the significance of the mean of $f'(\theta, \nu)$ based on a confidence interval (if 0 lies in the interval, the slope of the smoothing is not significant, whereas positive and negative intervals will indicate increasing and decreasing trends, identified by blue (black in black and white versions) and red (dark grey) colors. With suitable modifications, this technique can be adapted to circular data (CSiZer), plotting an annulus where the width of the ring is determined by the range of bandwidths considered. In Figure 2 (right), the CSiZer for the temperature cycle changes, shows that for values of the smoothing parameter between approximately 0 and 12 (including the values of the three selectors) the mode is significant (changes from blue to red), although it could not be fully appreciated for the rule of thumb in the density plot (Figure 2, left panel).

Acknowledgments: The authors want to acknowledge Prof. Augusto Pérez-Alberti for providing the data. This work has been supported by Project MTM2008-03010 from the Spanish Ministry of Science and Innovation.

References

- Banerjee, A., Dhillon, I.S., Ghosh, J. and Sra, S. (2005). Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, **6**, 1345–1382.
- Chaudhuri, P. and Marron, J. S. (1999), SiZer for exploration of structures in curves, *Journal of the American Statistical Association*, **94**, 807823.
- Di Marzio, M., Panzera A. and Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statistics & Probability Letters*, **79**, 2066–2075.
- Hall, P., Watson, G.P. and Cabrera, J. (1987). Kernel density estimation for spherical data. *Biometrika*, **74**, 751–762.
- Mardia, K.V. and Jupp, P.E. (2000). *Directional Statistics*. Wiley, New York.
- Oliveira, M., Crujeiras, R. M. and Rodríguez–Casal, A. (2012). A plug–in rule for bandwidth selection in circular density estimation. (*Submitted*). arXiv:1202.6076v1.
- Pewsey, A. (2000). The wrapped skew–Normal distribution on the circle. *Communications in Statistics - Theory and Methods*, **29**, 2459–2472.
- Taylor, C. C. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, **52**, 3493–3500.

Poisson mixture regression for Bayesian inference on large over-dispersed transportation origin-destination matrices

Konstantinos Perrakis ¹, Dimitris Karlis ², Mario Cools ³ ⁴,
Davy Janssens ¹, Geert Wets ¹

¹ Transportation Research Institute, Hasselt University, Belgium

² Department of Statistics, Athens University of Economics and Business, Greece

³ Centre for Information, Modeling and Simulation, Hogeschool-Universiteit Brussel, Belgium

⁴ Research Foundation Flanders, Belgium

E-mail for correspondence: `konstantinos.perrakis@uhasselt.be`

Abstract: Statistical modeling of origin-destination (OD) matrices is advocated as a viable alternative to traditional transportation models. To this end, Poisson mixture models are utilized in order to model a large over-dispersed OD matrix derived from the 2001 Belgian travel census. Emphasis is placed on a novel Bayesian application of the Poisson-inverse Gaussian model. As shown the model has desirable attributes both in its marginal and in its hierarchical form.

Keywords: OD matrix; Poisson mixtures; Poisson-inverse Gaussian.

1 Introduction

Consider an area which can be divided into m zones and let T_{od} denote the number of trips from zone of *origin* o to zone of *destination* d , where $o, d = 1, 2, \dots, m$. The OD matrix \mathbf{T} , is then

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} & \dots & T_{1m} \\ T_{21} & T_{22} & \dots & T_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ T_{m1} & T_{m2} & \dots & T_{mm} \end{bmatrix}.$$

In an alternative notation, the matrix \mathbf{T} can be represented by a vector \mathbf{y} with elements y_i for $i = 1, 2, \dots, n$ and $n = m^2$, namely $\mathbf{y} = (y_1, y_2, \dots, y_n)^T = (T_{11}, T_{12}, T_{13}, \dots, T_{mm})^T$. Within the traditional transportation modeling framework, OD modeling is incorporated in the *four-step model*, a sequential procedure which involves the independent modeling phases of (a) trip-generation, (b) trip-distribution, (c) modal-split and (d) traffic-assignment.

OD estimation depends on step (a) and is handled in step (b). Modeling procedures within step (b) include *growth-factor*, *gravity*, *intervening-opportunities* and *direct-demand* modeling approaches (see e.g. Ortúzar and Willumsen, 2001). Historically, the development of these models depended to a large degree on the availability of OD data which for most cases originated from travel surveys. Collecting travel survey data has clear financial advantages – in comparison to travel census studies for instance – but it also has a downside as travel-survey OD matrices are subjected to considerable sources of error (see e.g. Stopher and Greaves, 2007). Due to that fact, OD regression studies have been limited within the field of transportation. The focus of our research is on cases where reliable OD information is available and the aim is to demonstrate how traditional transportation modeling can be potentially replaced by statistical modeling approaches. Some of the merits of the proposed approach are presented in Perrakis et al. (2012). In this paper the Poisson mixture modeling approach is investigated further.

2 Data

The OD matrix from the 2001 Belgian travel-census study contains information about the departure and arrival locations for work and school related trips of the approximately 10 million Belgian residents. The application area is the northern region of Flanders which roughly accounts for 60% of the population and 44% of the country's surface area. The analysis is for the 308 Flemish municipalities and the resulting OD matrix contains 94864 cells. The explanatory variables are six dummy variables and twelve covariates. The set of covariates includes variables such as employment ratio, population density, relative length of road networks, distance etc. Due to the particularity of the OD problem some of the covariates are used in pairs, i.e. twice; one time for origin-zones and one time for the destination-zones, which results to a total set of 25 explanatory variables.

3 Poisson mixture models

With Poisson mixture models we assume that the OD flows y_i are i.i.d. Poisson realizations and that the rate of the Poisson distribution is $\lambda_i = \mu_i u_i$ for $i = 1, 2, \dots, n$, where μ_i relates to the vector of $p + 1$ unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ and the set of explanatory variables $\mathbf{x}_i = (x_{i0}, x_{i1}, \dots, x_{ip})^T$ through the log-link function $\log \mu_i = \boldsymbol{\beta}^T \mathbf{x}_i$, and u_i is a random component – interpreted as a multiplicative random effect – which is attributed with a density $g_1(u_i)$. The Poisson mixture modeling formulation is summarized as $y_i \sim \text{Pois}(\lambda_i)$ with $\lambda_i = \mu_i u_i$ and $\mu_i = e^{\boldsymbol{\beta}^T \mathbf{x}_i}$, and further $u_i \sim g_1(u_i)$ with $E(u_i) = 1$. The density g_1 is known as the mixing density. Alternatively, from a generalized linear mixed model

(GLMM) perspective the model can be expressed as $y_i \sim Pois(\lambda_i)$ with $\log \lambda_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i$, $\varepsilon_i \sim g_2(\varepsilon_i)$ and $E(\varepsilon_i) = 0$, where ε_i is an additive random error term, namely an observation random effect or random intercept as it is most commonly known. The Poisson likelihood is the conditional likelihood given the unobserved random effect vector $\mathbf{u} = (u_1, u_2, \dots, u_n)^T$. Integration over \mathbf{u} results to the marginal sampling likelihood, that is $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\mu}, \mathbf{u})g_1(\mathbf{u})d\mathbf{u}$. The Poisson-mixture and GLMM formulations are equivalent, but the resulting intercept estimates and the interpretations of marginal means are different due to the identifiability constraints (Lee and Nelder, 2004). From a Bayesian perspective the models are also referred to as hierarchical Poisson models since the mixing density is actually a first-level prior distributions.

In particular we investigate the performance of the Poisson-gamma (PG), Poisson-lognormal (PLN) and Poisson-inverse Gaussian (PIG) models in their multiplicative random effect form. The PG model arises when g_1 is a gamma distribution and is the most frequently used model leading to a marginal negative binomial distribution (see e.g. Lawless, 1987). The PLN is the predominant alternative mainly due to its distinct historical development as a GLMM based on the assumption that g_2 is a normal distribution. The density g_1 is lognormal, consequently. When g_1 is inverse Gaussian, the PIG emerges which is the less known model among the three. Despite the fact that the theoretical properties of this model have been thoroughly explored (e.g. Dean et al., 1989) its usage is still limited. To our knowledge this is a first Bayesian application of the model. The model formulation is based on non-informative priors and is as follows

$$\begin{aligned} y_i | \boldsymbol{\beta}, u_i &\sim Pois(e^{\boldsymbol{\beta}^T \mathbf{x}_i} u_i), \\ \boldsymbol{\beta} &\sim N_{p+1}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \text{ with } \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = n(\mathbf{X}^T \mathbf{X})^{-1} \sigma_{\boldsymbol{\beta}}^2 \text{ and } \sigma_{\boldsymbol{\beta}}^2 = 10^3, \\ u_i &\sim IG(1, \zeta) \text{ and} \\ \zeta &\sim Gamma(a, a) \text{ with } a = 10^{-3}. \end{aligned}$$

The multivariate normal prior for regression parameters has a g -prior structure, analogue to one of the benchmark priors in Fernández et al. (2001) for normal linear models. The inverse Gaussian prior follows the parameterization of Holla (1967), so that a-priori $E(u_i) = 1$ and $Var(u_i) = \zeta^{-1}$. The gamma hyperprior with shape and rate 10^{-3} is also diffuse with mean equal to 1 and a variance of 1000. The regression prior is the same for the PG and PLN models only that in these models we have that $u_i \sim Gamma(\theta, \theta)$ with $\theta \sim Gamma(a, a)$ and $u_i \sim LN(-\sigma_u^2/2, \sigma_u^2)$ with $\sigma_u^2 \sim InvGamma(a, a)$, respectively. The three posterior distributions are not of known form. For the PIG model it can be easily shown that a-posteriori $u_i \sim GIG(y_i - 1/2, 2e^{\boldsymbol{\beta}^T \mathbf{x}_i} + \zeta, \zeta)$ and $\zeta \sim Gamma(a + n/2, a + \sum_i (u_i - 1)^2/2u_i)$, where $GIG(\cdot)$ is the generalized inverse Gaussian distribution. For the PG model the only known conditional is that of u_i which is $Gamma(y_i + \theta, e^{\boldsymbol{\beta}^T \mathbf{x}_i} + \theta)$ while for the multiplicative PLN none of the conditionals is known. Thus,

the PIG in its hierarchical form is the easiest to fit. Marginally, the PIG likelihood involves a modified Bessel function, nevertheless computer routines for computation of PIG probabilities are available. In striking contrast, the marginal PLN likelihood is intractable and thus numerical integration is usually employed.

For the problem at hand, the size of the OD dataset constitutes hierarchical fitting a rather daunting task. Therefore, the parameters of scientific interest are estimated through the marginal structures. Predictive inference on the other hand is based on hierarchical structures. The latter is easily achievable for the PG and PIG models, which have conjugate distributions for the random effects, but is not straightforward for the PLN model. An independence-chain Metropolis-Hastings (M-H) algorithm is employed on the marginal forms of the three models with a multivariate normal proposal for the regression vector and a gamma proposal for the dispersion parameter of each model centered at the corresponding ML estimates. Runtime for the PLN model was considerably longer due to numerical integration for calculation of probabilities from the PLN distribution.

4 Results

The posterior estimates for the parameters of scientific interest (not presented here) reveal that all regression parameters have statistically significant effects. Interestingly, the posterior means of the PLN and PIG models are closer, especially for the intercept estimate; the corresponding posterior means are $\beta_0^{PG} = 4.034$, $\beta_0^{PLN} = 6.124$ and $\beta_0^{PIG} = 6.841$.

Information criteria results for Bayesian versions of AIC and BIC (based on the posterior means of deviance) as well as marginal and hierarchical DIC (the latter only for PG and PIG models) are summarized in Table 1.

TABLE 1. The values of AIC, BIC and DIC for the three models.

Criterion	PG	PLN	PIG
AIC	281519.2	279386.9	278469.1
BIC	281774.6	279642.3	278724.5
DIC (marginal)	281492.2	279362.4	278442.2
DIC (hierachical)	224141.4	-	224146.1

Marginally, the three criteria provide more support to the PLN and PIG models which might also explain why the posterior means of these two models are closer. The result is partially anticipated since the PLN and PIG allow for longer tails and are in theory more appropriate for cases of highly positive-skewed count data (Willmot, 1997). In addition, all three criteria indicate that the PIG distribution is the most appropriate marginal sampling distribution.

On the other hand, distinguishing a “better” hierarchical model is not as clear as the DIC value for the hierarchical PG model is just slightly lower than the corresponding value of the PIG model. This does not provide much evidence regarding which is the most appropriate predictive model.

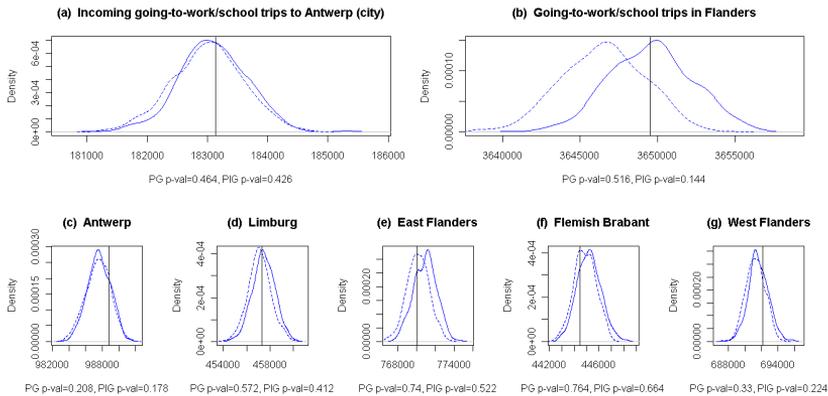


FIGURE 1. Kernel predictive distributions for work/school trips from the PG model (straight line) and the PIG model (dashed line) for (a) incoming trips to the city of Antwerp, (b) total number of trips in Flanders and intra-zonal trips for the five Flemish provinces; (c) Antwerp, (d) Limburg, (e) East Flanders, (f) Flemish Brabant and (g) West Flanders. The vertical black lines indicate the observed quantities.

Overall goodness-of-fit predictive tests based on absolute $\sum_i |y_i - E(y_i|\beta, u_i)|$ and squared $\sum_i (y_i - E(y_i|\beta, u_i))^2$ distances result in respective Bayesian p-values equal to 0.276 and 0.468 for the PG model and equal to 0.424 and 0.512 for the PIG model. Both models provide satisfactory p-values in general, although PIG predictions seem to replicate the data better for small deviations of expected values from observed data. We additionally implement case-specific predictive tests on aggregated levels. Examples of such tests – which are of substantial interest for transportation planning and forecasting – are presented in Figure 1. In general all p-values in Figure 1 are within acceptable limits, nevertheless there is a striking difference between the predictive distributions of Figure 1b for the total number of trips within Flanders. Although Bayesian p-values are not formally comparable across models, they are useful when examining goodness-of-fit from many different aspects. A closer examination reveals that both models significantly overestimate the number of 0-valued cells, especially the PG model. That provides a strong indication that the PG distribution in Figure 1b is more well-centered not due to consistent predictions but due to overestimating more the number of 0-valued cells. This finding also implies that future research on zero-inflated versions is interesting.

5 Conclusions

Statistical OD modeling is advocated as a viable alternative to traditional trip-generation and trip-distribution modeling. To this end, we propose that Poisson mixtures and Bayesian methods provide a suitable framework for modeling large, over-dispersed OD datasets when the focus of interest is not only on parameter estimation but also on short-term prediction. In particular, the performance of the PG, PLN and PIG models was evaluated on a Flemish OD matrix from the 2001 Belgian travel census. The PIG model was found not only to provide the best marginal fit, but that it also has desired distributional properties very much alike the PG model and unlike the rather cumbersome PLN model.

References

- Dean, C., Lawless, J. F. and Willmot, G.E. (1989). A mixed Poisson inverse Gaussian regression model. *Canadian Journal of Statistics*, **17**, 171-181.
- Fernández, C., Ley, E. and Steel, M.F.J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, **100**, 381-427.
- Holla, M.S. (1967). On a Poisson-inverse Gaussian distribution. *Metrika*, **11**, 115-121.
- Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, **15**, 209-225.
- Lee, Y. and Nelder, J.A. (2004). Conditional and marginal models: another view. *Statistical Science*, **19**, 219-228.
- Ortúzar, J. de D. and Willumsen, L.G. (2001). *Modeling Transport*. Chichester: John Wiley and Sons.
- Perrakis, K., Karlis, D., Cools, M., Janssens, D., Vanhoof, K. and Wets, G. (2012). A Bayesian approach for modeling origin-destination matrices. *Transportation Research, Part A*, **46**, 200-212.
- Stopher, P.R. and Greaves, S.P. (2007). Household travel surveys: where are we going? *Transportation Research, Part A*, **41**, 367-381.
- Willmot, G.E. (1990). Asymptotic tail behaviour of Poisson mixtures with applications. *Advances in Applied Probability*, **22**, 147-159.

Fine-scale downscaling of environmental covariates in biodiversity modelling

Iain Proctor^{1,2}, Rognvald I. Smith^{1,2}, E. Marian Scott²

¹ Centre for Ecology and Hydrology, Edinburgh, U.K.

² Department of Mathematics and Statistics, University of Glasgow, U.K.

E-mail for correspondence: ipro@ceh.ac.uk

Abstract: Downscaling techniques are used widely to predict the value of meteorological variables at spatio-temporal positions. This paper describes the fine-scale downscaling of environmental predictands using Perfect Prognosis and prediction distributions, prior to their inclusion in the modelling of a biodiversity response. A local relationship is formed between altitude as a predictor and each predictand, allowing for estimation of the uncertainty in model covariates. The method is applied in the hierarchical modelling of plant species data from the 1998 U.K. Countryside Survey.

Keywords: Downscaling; Biodiversity; Spatial Modelling.

1 Introduction

The Countryside Survey (CS) was set up in 1978 with the intention to record an ‘ecological snapshot’ of the state of U.K. habitats at a certain point in time. As part of achieving this end, a wide-scale vegetation survey is conducted roughly every decade. It is comprised of nearly 600 1km² sites, within which various plots are sampled. The CS used a stratified random sampling technique to ensure a range of different habitats would be surveyed across Great Britain (Bunce et al. (1999)).

Extensive analyses have been conducted on CS data to assess the response of individual species to environmental change (see e.g. Smart et al. (2003)). The aim of this analysis is to create a framework to describe how plant species community dynamics are affected in response to environmental pressures. As our response in this case, we take a biodiversity measure of flora from across the United Kingdom, using data from the CS sampling plots. In the non-downscaled models, covariate values of rainfall and nitrogen deposition are assumed to be uniformly distributed across their respective grid squares. The aims of the analysis reported here are two: firstly to predict the rainfall and nitrogen deposition at the location of the CS plot, rather than using the grid square value; secondly to estimate the uncertainty associated with these predictions. This is necessary in or-

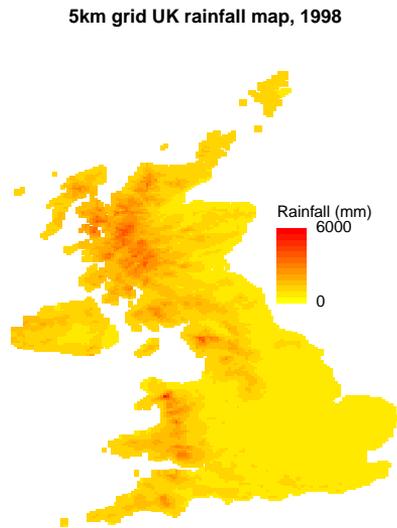


FIGURE 1. The U.K. 5km grid rainfall map for 1998. There is a general trend of increasing annual estimates from the South-East to the North-West of the country. This gradient is interactive with altitude, as areas with higher mean elevation have a greater estimate.

der that correct inferences are made as to the relationships between these explanatory covariates and the biodiversity responses.

2 Method

2.1 Downscaling

Perfect Prognosis is the use of real-world statistical relationships between observed values of a predictand and selected variables (see e.g. Klein (1971)). Here, a similar regression relationship for the downscaling of rainfall data is outlined: for every 1km^2 CS survey site i , there exists a 5km^2 square with a non-downscaled rainfall value which here is termed R_{i1} . There are (up to) eight other 5km^2 squares adjacent to it. Thus, there are n terrestrial 5km^2 squares associated with each survey site, where $n \leq 9$. Figure 1 depicts a map of these estimated rainfall values. The estimates of rainfall and mean altitude for each 5km^2 region are therefore interpreted as belonging to a specific survey site. For each site, linear regression is then applied to the

rainfall values, using mean altitude as an explanatory variable. A prediction of rainfall is needed for each plot k at each site i . So, for every site i , the regression equation is formed as follows:

$$R_{ij} = \alpha_i + \beta_i B_{ij} + \epsilon_{ij}, \quad (1)$$

where R_{ij} is the estimated rainfall and B_{ij} the mean altitude in grid square j associated with site i , α and β are parameters and ϵ_{ij} is the Gaussian distributed error term. In order to obtain new predictions for rainfall at the 14m^2 sampling plot locations, a finer-scale altitude map is used. Each more precise altitude value, termed z_{ik} , is used to predict the rainfall at the plot location k in site i , where $E(q_{ik})$ is the expected rainfall. The prediction, q_{ik} is calculated by the following equation:

$$q_{ik} = \alpha_i + \beta_i z_{ik} + \zeta_{ik}, \quad (2)$$

where α_i and β_i are as estimated in equation 1, $\zeta_{ik} = SE_{\hat{q}(z_{ik}^*)} t_i$, and $SE_{\hat{q}(z_{ik}^*)}$ is the standard error, given the prediction \hat{q} at altitude z_{ik} . The t_i values are randomly sampled from the t-distribution with $n - 2$ degrees of freedom. These randomly sampled values of the q_{ik} in equation 2 are inserted into the model in place of the non-downscaled R_{i1} value. This same procedure is applied to the 1km^2 gridded estimates for total annual nitrogen deposition (denoted C_{i1}) in the same manner, by regressing against their respective mean altitudes and downscaling using the parameter estimates.

3 Application

3.1 Data

Within each 1km^2 CS survey site, up to 5 quadrat plots of 14m^2 are positioned randomly. A subset of 1107 sampling plots are selected as the model responses from 337 survey sites. An univariate index response is calculated for each plot surveyed: the Shannon-Wiener biodiversity index:

$$S = - \sum_{h=1}^H \rho_h \ln \rho_h \quad (3)$$

where ρ_h is the proportion of species h in the plot. The covariates of rainfall and nitrogen deposition, after being downscaled, are then inserted into an generalised additive model with a Gaussian error structure, to which the biodiversity response is fitted. The initial fixed effects included in the model are as follows:

$$S_{ik} \sim s(P_{ik}) + s(Q_i : BH_{ik}) + s(F_i : BH_{ik}) + s(A_i) + s(E_{ik}, N_{ik}) + BH_{ik}, \quad (4)$$

TABLE 1. The results of the non-downscaled and downscaled models are shown here. *ND* denotes the chosen model for the non-downscaled data, *D* is the same model, but with downscaled covariates Q and F in place of R and C respectively. Model *FD* is the chosen model using the downscaled data.

Model	Covariates	AIC	Dev. expl
FND	$s(P) + s(E,N) + s(R:BH) + s(D:BH) + BH$	1355	52%
D	$s(P) + s(E,N) + s(Q:BH) + s(F:BH) + BH$	1385	50%
FD	$s(P) + s(E,N) + BH$	1381	48%

where S_{ik} is the Shannon-Wiener index at plot k in site i . The ‘ $s(\cdot)$ ’ denotes a spline of one or two covariates. Interaction between covariates is denoted by ‘:’. The explanatory covariates are denoted thus: Easting (E) and Northing (N) define the positioning of each plot, accurate to 100m. Coarse altitude data (A) is recorded as the mean of each 5km² square. The downscaled covariates are denoted as Q for rainfall and F for total nitrogen deposition, in place of R and C respectively. The previous index value (P) is the Shannon-Wiener index at the same plot in the previous CS survey in 1990. Each plot is classified as having exactly one Broad Habitat (BH), using the species assemblage to inform which habitat is present there. There are seven distinct habitat types used.

This hierarchical model is run 1000 times, with results obtained for each separate run. The choice of model covariates is made via backward selection, using the criteria of the medians of the AIC and covariate p-values, according to the model output.

4 Results

As noted, the initial model is as in equation 4. The final chosen model, using the downscaled rainfall and nitrogen deposition data, is:

$$S_{ik} \sim s(P_{ik}) + s(E_{ik}, N_{ik}) + BH_{ik} \quad (5)$$

Table 1 compares the model results, using gridded and downscaled data. In comparison with the final non-downscaled data model, *FND* in table 1, model *FD* contains neither rainfall nor nitrogen deposition as a covariate. Due to very large p-values obtained for both downscaled covariates, they were removed from the model. The change in deviance explained between models *FND* and *D*, which was fitted using downscaled prediction, is approximately 2%, but the AIC value is much greater, indicating a relatively poorer goodness of fit in the downscaled model. Similar results are obtained for model *FD*, which contains neither downscaled covariate. Figure 2 shows the median AIC for model *D* computed after each model run. The graph shows this median to be approaching a stable value, and to vary

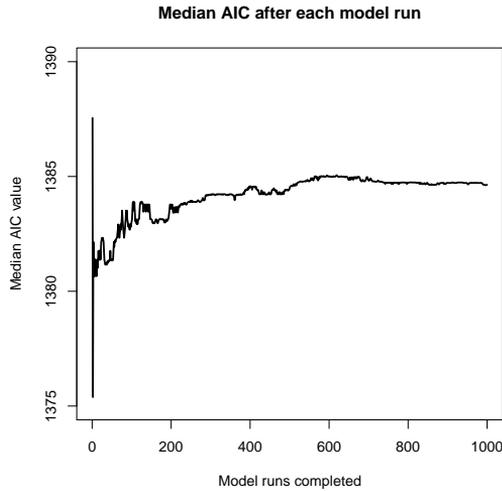


FIGURE 2. The graph shows the value of the median AIC value after each model run for the downscaled data model D in table 1.

within a small band; similar results were obtained for all models containing downscaled covariates. Thus, the number of runs is considered sufficient to obtain a stable result.

5 Conclusion

Using this approach for downscaling the covariate data resulted in the removal of rainfall and nitrogen deposition from this biodiversity model. It is due to the sampled ζ_{ik} error terms, which are added to the predicted covariate values causing random scatter to perturb the covariate predictions. This leads to the correlation between the model response and the downscaled covariates of interest to attenuate to zero. The lack of a significant relationship between the response data and these downscaled covariates does not rule out the possibility that these environmental pressures do indeed impact upon floral biodiversity. This framework is applicable to other modelling scenarios, where explanatory covariates which contain unknown measurement error can be downscaled using fine-scale predictors, allowing the covariate uncertainty to be estimated. As shown in this example, when this local uncertainty is estimated, covariate significance can change greatly.

Acknowledgments: Special thanks to NERC for funding this research and CEH for providing required data.

References

- Bunce, R. G. H., Smart, S. M., van de Poll, H. M., Watkins, J. W., and Scott, W. A. (1999). *Measuring change in British vegetation. ECO-FACT Volume 2*. Institute of Terrestrial Ecology.
- Klein, W. H. (1971). Computer Prediction of Precipitation Probability in the United States. *Journal of Applied Meteorology*, **10**, 903–915.
- Smart, S. M., Robertson, J. C., Shield, E. J., and van de Poll, H. M. (2003). Locating eutrophication effects across British vegetation between 1990 and 1998. *Global Change Biology*, **9**, 1763–1774.

Modelling the movement of dusky kob in the Sundays River

Leendert Punt ¹, Linda M. Haines ², Christien Thiaart²

¹ Old Mutual Finance, Cape Town, South Africa

² University of Cape Town, Cape Town, South Africa

E-mail for correspondence: `linda.haines@uct.ac.za`

Abstract: This paper is concerned with the modelling of the movement of dusky kob in the Sundays River in South Africa over a three-month time period using within-river and weather data as explanatory variables. A novel approach based on a discrete choice framework is introduced and shown to be effective in modelling the data.

Keywords: Movement ecology; Spatio-temporal data; Discrete choice models.

1 Introduction

Zoologists from Rhodes University in South Africa were interested in identifying environmental factors which influence the movement of fish in their local rivers. To this end they monitored the movement of 23 dusky kob over a short stretch of the Sundays River in the Eastern Cape over a period of just over a year using a sophisticated transmitter-receiver system. The aim of the present study is to develop a model for a subset of their data taken over a three month period in 2008 in an attempt to provide some statistically-based answers to their broad research question.

2 Data

23 dusky kob (*Argyrosomus japonicus*) from the Sundays River were tagged with transmitters which had a unique identification code and which emitted signals at irregular intervals varying from 20 to 60 seconds. 16 data-logging receivers were moored at approximately 1km intervals along a 16km stretch of the river, starting at the river mouth. The receivers were omnidirectional with a range of approximately 500 to 600m and provided a very comprehensive coverage of the movement of the fish. In addition 5 stationary data-loggers were positioned along the 16km stretch of the river and were programmed to record the temperature and depth of the water at those points at hourly intervals. Weather data in the form of wind speed,

wind direction, atmospheric pressure and rainfall were acquired from the Coega weather station situated some 12km from the river. Data recorded on only one of the 23 kob for the three-month time period from the 1st of June to the 1st of September, 2008, were used in the present study.

The spatio-temporal nature of the data was captured by discretizing both space and time to yield a simple two-dimensional grid-based representation. Thus the 16km of river under study was viewed as a one-dimensional stretch divided into 16 intervals or boxes of approximately 1km in length, each corresponding to the range of an individual receiver. In addition time was divided into 60 minute intervals giving a total of 4320 such time stamps. The location of the fish at a particular time stamp was then taken as that specified by the receiver which last recorded a signal during that time interval. The within-river data, that is water temperature and water depth, were only available at 5 locations along the 16km stretch of the river. Values of these variables at each of the receiver locations which define the 16 boxes along the river were therefore obtained by invoking linear interpolation. Data obtained from the Coega weather station were not amenable to interpolation and were thus taken as holding uniformly along the full extent of the 16km stretch of the river at a specified time stamp. Rainfall data were extremely sparse, with a total of less than 10mm of rain falling in the 3-month period of interest, and were therefore omitted from the study.

3 The Model

The movement of the dusky kob was modelled within the discrete choice framework (Cameron and Trivedi (2005); Train (2009)). Specifically the kob was regarded as a decision maker and, while at a particular time stamp and irrespective of its present location, must choose one of the 16 boxes along the river into which to move. The choice is considered as being influenced by both the within-river variables and the external weather data during that particular time stamp.

The basic logit model, which underpins the more general discrete choice model, can now be formulated in terms of the probability that the dusky kob chooses box i at time stamp t as

$$P_{ti} = \frac{e^{V_{ti}}}{\sum_{i=1}^{16} e^{V_{ti}}} \quad \text{for } i = 1, \dots, 16 \text{ and } t = 1, \dots, 4032,$$

where V_{ti} is the utility the kob associates with the particular box and time stamp and depends on the explanatory within-river and out-of-river variables. Water temperature and water depth vary across the river boxes, or alternatives in the terminology of discrete choice models, and the representative utility can thus be expressed as

$$V_{ti}^{(c)} = T_{ti} \gamma_T + D_{ti} \gamma_D$$

where T_{ti} and D_{ti} represent water temperature and water depth at time stamp t and location i and γ_T and γ_D are the attendant unknown parameters, respectively. This utility therefore specifies a conditional logit model. The weather data do not depend on the river alternatives but only on the particular time stamp. The representative utility is therefore specified using a multinomial logit model as

$$V_{ti}^{(m)} = \alpha_i + WS_t \beta_{i,WS} + WD_t \beta_{i,WD} + AP_t \beta_{i,AP}$$

where α_i is an intercept term, WS_t , WD_t and AP_t denote wind speed, wind direction and atmospheric pressure respectively, and the attendant parameters $\beta_{i,WS}$, $\beta_{i,WD}$ and $\beta_{i,AP}$ vary across the alternatives. The mixed logit model combines the conditional and multinomial specifications and has overall representative utility

$$\begin{aligned} V_{ti} &= V_{ti}^{(c)} + V_{ti}^{(m)} \\ &= \alpha_i + T_{ti}\gamma_T + D_{ti}\gamma_D + WS_t \beta_{i,WS} + WD_t \beta_{i,WD} + AP_t \beta_{i,AP}. \end{aligned}$$

4 Results

The three models, that is the conditional, the multinomial and the mixed logit, were all fitted to the data under study. The explanatory variables were introduced into the models one at a time and, in order to select the “best” model, three relative measures, the likelihood ratio index, McFadden’s R^2 and Akaike’s Information Criterion (AIC), were considered (Train (2009)). In fact the first two measures provide no more information than the log-likelihood and thus selection was based on the AIC. In absolute terms it is also important to appraise the goodness-of-fit of a model. The Percentage Correctly Predicted (PCP) measure is, as the name suggests, the percentage of times a chosen alternative is the alternative with highest fitted probability and has been used to assess the goodness-of-fit of discrete choice models. The measure is however not entirely meaningful since it requires that the choice of the decision maker be predicted exactly (Train (2009)). Two measures of goodness-of-fit were therefore devised in this study, a relaxed variant of the PCP that accommodates a number of locations with highest fitted probabilities, and a measure specifically devised for the present setting.

The best fitting conditional and multinomial models were those which incorporated all the relevant explanatory variables. For the combined or mixed logit setting, the best model was that based on water depth and the three weather variables but only marginally so. On balance therefore the mixed model comprising all the explanatory variables under consideration was selected for interpretation and was shown to provide a good fit to the data. The success of this model indicates that the movement of the dusky kob is motivated by water temperature, water depth, wind speed, wind

direction and atmospheric pressure. More specifically the dusky kob was observed to move to the river mouth when the atmospheric pressure was high, and this feature is well-modelled by the mixed logit formulation.

5 Conclusion

It is clear from this study that the discrete choice framework is an effective and efficient means of modelling the movement of dusky kob in the Sundays River. The approach would undoubtedly increase in accuracy as explanatory variable information increases in both resolution and relevance. More broadly, the discrete choice model could well provide a powerful tool in a novel context, that of movement ecology (Nathan et al (2008); Cagnacci et al (2010)), for describing statistically the movements of animals as observed in the dynamic spatio-temporal environment defined by an experiment.

Acknowledgments: The authors would like to thank Paul Cowley and Amber Childs of the Ichthyology Department of Rhodes University for making the data available to us and the University of Cape Town and the National Research Foundation of South Africa for financial support. The work forms part of the M.Sc. dissertation of the first author and was completed while he was a research student in the Department of Statistical Sciences at the University of Cape Town.

References

- Cameron, A.C. and Trivedi, P.K. (2005). *Microeconometrics : Methods and Applications*. Cambridge: Cambridge University Press.
- Cagnacci F., Boitani L., Powell R.A., and Boyce M.S. (2010). Preface to the Special Issue : Challenges and opportunities of using GPS-based location data in animal ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 2155.
- Nathan, R., Getz W.M., Revilla E., Holyoak M., Kadmon R., Saltz D., and Smouse P.E. (2008). Movement Ecology Special Feature : A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences*, **105**, 19052–19059.
- Train, K.E. (2009). *Discrete Choice Methods with Simulation*. Second Edition. Cambridge: Cambridge University Press.

Piecewise transition models with random effects for unequally-spaced measurements

Reyhaneh Rikhtehgaran ^{1,2}, Iraj Kazemi ¹, Geert Verbeke ³,
Wim de Kort ⁴, Emmanuel Lesaffre ^{3,5}

¹ Department of Statistics, University of Isfahan, Isfahan, Iran

² Department of Mathematical Sciences, Isfahan University of Technology, Iran

³ L-Biostat, Catholic University of Leuven, Leuven, Belgium

⁴ Sanquin Blood Bank, Southeast Region, Nijmegen, the Netherlands

⁵ Department of Biostatistics, Erasmus MC, Rotterdam, the Netherlands

E-mail for correspondence: r_rikhtehgaran@stat.ui.ac.ir

Abstract: In this paper, we consider the analysis of unequally-spaced longitudinal data using transition regression models with random effects. We assume that such data have been generated via either a diffusion and/or a stabilization process. We adapt current transition models for fitting such longitudinal data, but focus here on the stabilization case. The initial conditions problem which usually arises in transition models with random effects is addressed here. The usefulness of the proposed model is assessed on a large data base of longitudinal hemoglobin values collected from blood donations by a Dutch private organisation.

Keywords: Blood donations; Initial conditions problem; Longitudinal data; Open-BuGs.

1 Introduction

In many applications of longitudinal studies, the usual random effects approach is unable to completely capture the general association structure among responses. Transition regression models with random effects extend the random effects structure by including in the linear predictor a lagged response variable (Heckman, 1991).

This paper aims to analyze longitudinal data on hemoglobin values obtained from donors visiting Sanquin at Nijmegen, a private blood bank organization in the Netherlands, during the years 2000-2004 (Baart et al., 2011). This analysis hopes to help in the practical organization of blood donations. Indeed, to ensure a high quality of the donated blood, blood donors who present themselves with a hemoglobin (Hb) level below the cutoff values 8.4 and 7.8 (mmol/L) for males and females, respectively, are rejected for donation. Such a rejection might demotivate the donors for further participation, but too many rejected donations also impact the

practical organization of the blood bank. In this paper, we develop a statistical model that aims to predict future Hb values in order to better plan the next visit. To this end, we adapted the current transition models in order to deal with the specific structure of the data.

We focus on a subset of the data set which excludes all measurements after donation was rejected (because of a too low Hb level or other reasons). Restricting to this subset creates a more homogeneous group of Hb values. Indeed, donation implies a subsequent drop in Hb level but does not occur in case donation is rejected. Consequently, analyzing a mix of Hb levels resulting from both actual donation and rejection would imply a more complicated statistical model.

2 Specification of transition random-effects models

We first treat longitudinal data where measurements are equally spaced over time. Let y_{i1}, \dots, y_{iT_i} be $T_i \geq 2$ repeated responses collected for the i th individual and \mathbf{x}_{it} be the corresponding vector of covariates. Consider the following transition model

$$y_{it} = \lambda + \gamma y_{i,t-1} + \mathbf{x}'_{it} \boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \quad i = 1, \dots, n \quad (1)$$

where $y_{i,t-1}$ is a lagged response variable and λ , γ and $\boldsymbol{\beta}$ are unknown regression parameters. The conventional assumptions are that the random individual effects $\alpha_i \stackrel{\text{iid}}{\sim} N(0, \sigma_\alpha^2)$ and the errors $\varepsilon_{it} \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$, are independent for all i and t , and are independent of covariates.

3 The initial conditions problem

In fitting a transition model, the initial conditions problem (ICP) usually occurs because the individual effects α_i that capture the unobserved heterogeneity are correlated with the initial state y_{i0} . Ignoring this association might lead to seriously biased parameter estimates, even for large samples (Kazemi and Davies, 2002). One solution to this problem is to consider

$$y_{i0} = \lambda_0 + \mathbf{x}'_{i0} \boldsymbol{\beta}_0 + u_{i0}, \quad i = 1, \dots, n, \quad (2)$$

where \mathbf{x}_{i0} represents the vector of initial period covariates supposed to be uncorrelated with u_{i0} . It is further assumed that the $u_{i0} \stackrel{\text{iid}}{\sim} N(0, \sigma_0^2)$, $\text{corr}(u_{i0}, \alpha_i) = \rho_{0\alpha}$ and $\text{corr}(u_{i0}, \varepsilon_{it}) = 0$ for all i, t . The ML estimates can be achieved analytically as is shown by Kazemi and Crouchley (2006).

4 A piecewise exponential transition model

In practice, measurement occasions are often unequally spaced over time as in the longitudinal study on Hb levels. In most transitional data analyses,

the effect of previous outcomes on the current state varies over time. For this reason, we introduce an extension of the model (1) as follows

$$y_{it} = \lambda + h(\gamma, \Delta d_{it}) y_{i,t-1} + \mathbf{x}'_{it} \boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \tag{3}$$

where $h(\gamma, \Delta d_{it})$ is a function of γ and Δd_{it} , and Δd_{it} is the time lag between the t th and $(t-1)$ th measurements of the i th individual. We distinguish between two processes for this relationship. The first process expresses a straightforward extension of Equation (1). Namely, the regression coefficient γ , regressing y_{it} on $y_{i,t-1}$, is replaced by a function that depends on the time lag Δd_{it} . This process is coined as the *diffusion process* when $0 < \gamma < 1$. For this process, a pragmatic choice of function $h(\gamma, \Delta d_{it})$ is $\gamma^{\Delta d_{it}}$. It is computationally useful to consider the reparameterization $\gamma = e^{-\varphi}$ where $\varphi \geq 0$, implying that $h(\gamma, \Delta d_{it}) = e^{-\varphi \Delta d_{it}}$. Thus, when the Δd_{it} becomes larger, the effect of previous outcomes on the current state becomes smaller. However, our longitudinal study suggests that there might also be a second process playing a role. For the hemoglobin longitudinal study, the stabilization process expresses that the hemoglobin level immediately drops after donation. Thereafter it rises slowly again to arrive at its initial level during some weeks. For the stabilization process, we assume the following piecewise function

$$\begin{aligned} h(\gamma, \Delta d_{it}) &= e^{-\varphi_1 - \varphi_2(t_c - \Delta d_{it})} && \text{if } \Delta d_{it} \leq t_c \\ &= e^{-\varphi_1} && \text{if } \Delta d_{it} > t_c, \end{aligned} \tag{4}$$

where φ_1 indicates the effect of previous outcomes on the current states and $\varphi_2 (\geq 0)$ is the influence of time intervals on this effect. When Δd_{it} is small, the association between y_{it} and $y_{i,t-1}$ is weak. The cutoff value t_c must also be estimated from the data.

5 The analysis of longitudinal study on Hb levels

5.1 Models without handling the ICP

Since the Hb level is measured before blood donation, decreases afterwards and rises again slowly every day to arrive at its initial level, we focus on the stabilization process. This does not preclude that the diffusion process also takes place, but fitting both processes asks too much from the data. Let hb_{it} be the t th measurement for the Hb level of the i th donor. The covariates are age, the average BMI over the period of donation and a binary indicator of the season of donation at time t defined as to one in the cold period and zero otherwise. Then, for $i = 1, \dots, n$ ($n = 1466$ for males and $n = 712$ for females) and $t = 1, \dots, T_i$, the proposed model is given by

$$hb_{it} = \beta_0 + h(\gamma, \Delta d_{it}) hb_{it-1} + \beta_1 season_{it} + \beta_2 age_i + \beta_3 bmi_i + \alpha_i + \varepsilon_{it}, \tag{5}$$

TABLE 1. Bayesian estimation results for males with model comparison criteria.

Parm.	With the ICP			Without the ICP
	M1	M2	M3	M3E
Subsequent equation				
λ	8.914(0.108)	7.331(0.135)	7.340(0.119)	7.690(0.149)
β_1	0.065(0.008)	0.074(0.008)	0.072(0.008)	0.069(0.008)
β_2	-0.004(0.001)	-0.003(0.001)	-0.003(0.001)	-0.003(0.001)
β_3	0.026(0.004)	0.022(0.003)	0.022(0.003)	0.023(0.003)
φ_1	-	-	1.733(0.053)	1.988(0.093)
φ_2	-	-	0.033(0.007)	0.045(0.010)
t_c	-	-	3.871(0.184)	3.885(0.180)
γ	-	0.177(0.011)	-	-
Initial equation				
λ_0	8.883(0.139)	8.884(0.139)	8.885(0.142)	8.918(0.141)
β_{01}	0.074(0.030)	0.073(0.030)	0.073(0.030)	0.052(0.023)
β_{02}	-0.003(0.002)	-0.003(0.002)	-0.003(0.002)	-0.003(0.002)
β_{03}	0.026(0.005)	0.026(0.005)	0.026(0.005)	0.025(0.005)
Variance components				
σ_ε^2	0.168(0.002)	0.173(0.002)	0.172(0.002)	0.170(0.002)
σ_α^2	0.182(0.008)	0.116(0.007)	0.116(0.006)	0.132(0.007)
σ_0^2	0.335(0.012)	0.335(0.012)	0.335(0.012)	0.325(0.012)
$\rho_{0\alpha}$	0	0	0	0.719(0.016)
Model comparison criteria				
AIC	18093.0	17848.2	17810.0	17037.0
BIC	18174.1	17936.7	17913.2	17147.6
RAD	654.0	601.7	601.1	546.5

* Bayesian standard deviations are given in parentheses.

where $h(\gamma, \Delta d_{it})$ is specified by Equation (4). To compare the various random-effect models, we fit the usual linear mixed model by ignoring the lagged response in Equation (5) (model M1), the linear transition mixed model without accounting for unequally-spaced measurements by assuming $h(\gamma, \Delta d_{it}) = \gamma$ (model M2), and the non-linear transition mixed model which considers the time-lag between measurements and is specified by Equations (4) (model M3). We also assume the following model for the initial response

$$hb_{i0} = \lambda_0 + \beta_{01}season_{i0} + \beta_{02}age_i + \beta_{03}bmi_i + u_{i0}, \quad i = 1, \dots, n, \quad (6)$$

where $corr(u_{i0}, \alpha_i) = 0$. This equation does not influence the estimation results as long as there is no relation between this equation and Equation (5). Equation (6) is considered in order to have the same number of observations to compare the results with those in Section 5.2.

TABLE 2. Bayesian estimation results for females with model comparison criteria.

Parm.	With the ICP			Without the ICP
	M1	M2	M3	M2E
Subsequent equation				
λ	7.860(0.114)	6.342(0.200)	6.449(0.200)	6.943(0.208)
β_1	0.053(0.013)	0.061(0.013)	0.061(0.013)	0.058(0.013)
β_2	0.006(0.001)	0.005(0.001)	0.005(0.001)	0.005(0.001)
β_3	0.016(0.004)	0.013(0.004)	0.013(0.004)	0.014(0.004)
φ_1	-	-	1.727(0.130)	-
φ_2	-	-	1.476(1.842)	-
t_c	-	-	1.738(0.746)	-
γ	-	0.192(0.022)	-	0.115(0.023)
Initial equation				
λ_0	7.908(0.141)	7.913(0.140)	7.905(0.142)	7.927(0.137)
β_{01}	0.067(0.038)	0.067(0.039)	0.067(0.038)	0.060(0.031)
β_{02}	0.005(0.002)	0.004(0.002)	0.005(0.002)	0.004(0.002)
β_{03}	0.017(0.005)	0.017(0.005)	0.018(0.005)	0.017(0.005)
Variance components				
σ_ε^2	0.144(0.004)	0.153(0.004)	0.152(0.004)	0.148(0.004)
σ_α^2	0.142(0.009)	0.085(0.009)	0.089(0.009)	0.107(0.010)
σ_0^2	0.261(0.014)	0.261(0.014)	0.261(0.014)	0.250(0.014)
$\rho_{0\alpha}$	0	0	0	0.714(0.025)
Model comparison criteria				
AIC	5741.7	5667.6	5678.8	5347.9
BIC	5810.4	5742.6	5766.3	5429.1
RAD	232.3	205.2	294.1	190.5

* Bayesian standard deviations are given in parentheses.

Because of the complexity of model M3, we opted for a Bayesian analysis of the data making use of the Markov chain Monte Carlo (McMC) methods. Computations were done using the OpenBUGs software (Lunn et al., 2009). Vague priors were taken for all parameters. The estimates of models were obtained by removing 15,000 initial burn-in samples. The posterior means and SEs were then based on 40,000 samples with a thinning factor equal to five (and thus were based on 200,000 iterations).

Parameter estimates are reported in Tables 1 and 2 for males and females, respectively. Model performance is evaluated by standard model comparison criteria, such as AIC, BIC and $RAD = \sum_{i,t} \left| \frac{hb_{it} - \widehat{hb}_{it}}{\widehat{hb}_{it}} \right|$ where \widehat{hb}_{it} is the predicted value of hb_{it} . These criteria are computed based on marginal likelihoods. These results are also shown in Tables 1 and 2.

It is clear that, for both males and females, the transition mixed models M2 and M3 fit better the hemoglobin responses than mixed model M1.

Upon closer inspection of the results, it appeared that the main difference of models M2 and M3 pertains to the measurements with a relatively short time (less than the cutoff value). Model M3 shows a better performance than model M2 for males, while this performance is reversed for females. The reason for this is that among women there are only 32 out of 3816 observations with a time lag of less than four months, while for men this value is 4953 out of 11762 observations. Knowing that, from a clinical point of view, a few weeks are usually sufficient for recovery we conclude that the parameter φ_2 is difficult to estimate among women which is also confirmed by its large estimated standard deviation. Therefore there is no need to introduce the first term of Equation (4) for females.

5.2 Models that handle the ICP

Based on the results obtained from the previous subsection, we chose model M3 for males and model M2 for females for the extended models. To take the ICP into account now $\text{corr}(u_{i0}, \alpha_i) = \rho_{0\alpha}$ in Equation (6) is estimated from the data. Parameter estimates and the corresponding model comparison criteria for these extended models, M3E and M2E, are reported in the last columns of Tables 1 and 2, for males and females, respectively. These criteria indicate that models that take into account the ICP are preferred over the other models. Indeed, as seen by the RAD values, predictions based on model M3E for males and M2E for females are superior.

References

- Baart, A.M., de Kort, W.L.A.M., Moons, K.G.M., Vergouwe, Y. (2011). Prediction of low haemoglobin levels in whole blood donors. *Vox Sanguinis*, **100**, 204-211.
- Heckman, J.J. (1991). Identifying the hand of past: distinguishing state dependence from heterogeneity. *The American Economic Review*, **81** (2), 75-79.
- Kazemi, I. and Crouchley, R. (2006). Modelling the initial conditions in dynamic regression models of panel data with random effects. Ch 4, In Baltagi, B.H., ed. *Panel Data Econometrics, Theoretical Contributions and Empirical Applications*, Elsevier, Amsterdam, Netherlands.
- Kazemi, I., Davies, R.B. (2002). The asymptotic bias of MLEs for dynamic panel data models. In Stasinopoulos, M., Touloumi, G., eds. *Statistical Modelling in Society*, Proceedings of the 17th IWSM, 391-395.
- Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. (2009). The BUGS project: evolution, critique and future directions (with discussion). *Statistics in Medicine*, **28**, 3049-3082.

A pseudo-adaptive Gaussian quadrature rule for fitting joint models for longitudinal and time-to-event data

Dimitris Rizopoulos¹

¹ Department of Biostatistics, Erasmus Medical Center, the Netherlands

E-mail for correspondence: d.rizopoulos@erasmusmc.nl

Abstract: Joint models for longitudinal and time-to-event data have recently attracted a lot of attention in the literature. The main difficulty in fitting these models arises from the requirement for numerical integration with respect to the random effects. This paper offers a solution to this problem by basing the fit of the model on a pseudo-adaptive Gauss-Hermite rule. The idea behind this rule is to use information for the shape of the integrand by separately fitting a mixed model for the longitudinal outcome. Simulation studies show that the pseudo-adaptive rule performs excellent in practice, and is considerably faster than the standard Gauss-Hermite rule.

Keywords: Gauss-Hermite rule; Numerical integration; Random effects; Time-dependent covariates; Survival analysis.

1 Introduction

Joint models for longitudinal and survival data constitute an attractive modeling paradigm for accounting for dropout in longitudinal studies as well as handling endogenous time-dependent covariates in a survival analysis setting (Tsiatis and Davidian, 2004). The key component in this type of models is a set of random effects that is assumed to induce the association between the longitudinal responses and the event time process. Even though the use of random effects greatly facilitates the formulation of joint models, the need for numerical integration over these random effects renders their estimation a rather computationally demanding task. In this paper we propose an alternative numerical integration approach, specially suited to joint models, that is based on the re-scaling idea behind the adaptive Gauss-Hermite rule. In particular, we first fit the mixed effects model for the longitudinal outcome and extract information regarding the location and scale of the posterior distribution of the random effects given the longitudinal responses for each subject. This information is then used to appropriately re-scale the subject-specific integrands in the definitions of the log-likelihood and score vector of the joint model. The key advantage of

this approach is that typically, after this re-scaling, a very small number of quadrature points (e.g., 3 to 5) is required to achieve an approximation error of the same or even smaller magnitude as the standard Gauss-Hermite rule with a moderate number of quadrature points (e.g., 15 to 21). The computation gains are evident, especially when high-dimensional random-effects structures are utilized.

2 Joint Modeling Framework

2.1 Submodels' Specification

Let T_i denote the observed failure time for the i th subject ($i = 1, \dots, n$), which is taken as the minimum of the true event time T_i^* and the censoring time C_i , i.e., $T_i = \min(T_i^*, C_i)$. Furthermore, we define the event indicator as $\delta_i = I(T_i^* \leq C_i)$, where $I(\cdot)$ is the indicator function that takes the value 1 if the condition $T_i^* \leq C_i$ is satisfied, and 0 otherwise. Thus, the observed data for the time-to-event outcome consist of the pairs $\{(T_i, \delta_i), i = 1, \dots, n\}$. For the longitudinal responses, let $y_{ij} = \{y_i(t_{ij}), j = 1, \dots, n_i\}$ to denote the value of the longitudinal outcome at visit times t_{ij} .

To account for the fact that the longitudinal marker is an endogenous time-dependent covariate measured with error, we will assume that the risk for an event depends on the true and unobserved value of the marker at time t , denoted by $m_i(t)$. In particular, to quantify the effect of $m_i(t)$ on the risk for an event we use a relative risk model of the form:

$$h_i(t \mid \mathcal{M}_i(t), w_i) = h_0(t) \exp[\gamma^\top w_{i1} + \alpha \{w_{i2} \times m_i(t)\}], \quad t > 0, \quad (1)$$

where $\mathcal{M}_i(t) = \{m_i(u), 0 \leq u < t\}$ denotes the history of the true unobserved longitudinal process up to time point t , $h_0(\cdot)$ denotes the baseline risk function, and $w_i^\top = (w_{i1}^\top, w_{i2}^\top)$ is a vector of baseline covariates. The first part w_{i1} , with corresponding regression coefficients vector γ , denotes the main effects of the baseline covariates, whereas the second part w_{i2} denotes possible interaction terms between the longitudinal marker $m_i(t)$ and some of the baseline covariates (i.e., usually w_{i2} will contain a subset of the elements of w_{i1}). The baseline risk function is assumed piecewise-constant, i.e., $h_0(t) = \sum_q \xi_q I(v_{q-1} < t \leq v_q)$, with $0 = v_0 < v_1 < \dots < v_Q$ denoting a split of the time scale, with v_Q being larger than the largest observed time, and ξ_q denotes the value of the hazard in the interval $(v_{q-1}, v_q]$.

To estimate $m_i(t)$ and successfully reconstruct the complete longitudinal history $\mathcal{M}_i(t)$, we utilize the available measurements of each subject and postulate a suitable mixed-effects model. Here we focus on normal data and therefore we employ a linear mixed-effects model to describe the subject-specific evolutions in time:

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= x_i^\top(t)\beta + z_i^\top(t)b_i + \varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2), \quad b_i \sim \mathcal{N}(0, D) \end{aligned} \quad (2)$$

where β denotes the vector of the unknown fixed effects, b_i denotes a vector of random effects, $x_i(t)$ and $z_i(t)$ denote row vectors of the design matrices for the fixed and random effects, respectively. The measurement error terms $\varepsilon_i(t)$ are assumed mutually independent, independent of b_i , and normally distributed with mean zero and variance σ^2 . To produce a good estimate of $\mathcal{M}_i(t)$ care should be given in the specification of $x_i(t)$ and $z_i(t)$. Therefore, in studies in which patients exhibit nonlinear longitudinal profiles, such as in the PBC dataset, it is advisable to consider a flexible representation for $x_i(t)$ and $z_i(t)$ using, e.g., B-splines of time (Rizopoulos et al., 2009; Ding and Wang, 2008).

2.2 Estimation

The main estimation methods that have been proposed for joint models are (semiparametric) maximum likelihood and Bayes using MCMC techniques have proposed a conditional score approach in which the random effects are treated as nuisance parameters, and they developed a set of unbiased estimating equations that yields consistent and asymptotically normal estimators. Here we focus in the maximum likelihood method for joint models as the one of the more traditional approaches.

Maximum likelihood estimation in joint models is based on the maximization of the log-likelihood corresponding to the joint distribution of the observed time-to-event and longitudinal outcomes $\{T_i, \delta_i, y_i\}$. To define this joint distribution we will assume that the vector of time-independent random effects b_i underlies both the longitudinal and survival processes. This means that these random effects account for both the association between the longitudinal and event outcomes, and the correlation between the repeated measurements in the longitudinal process (conditional independence). The maximum likelihood estimates (MLEs) are typically obtained using standard maximization algorithms, such as the expectation-maximization (EM) or the Newton-Raphson algorithms. The key component for applying either of these two algorithms in joint models is the score vector of the observed data log-likelihood function $\ell(\theta) = \sum_i \log p(T_i, \delta_i, y_i; \theta)$. Under the conditional independence assumptions and the submodels' specification presented in Section 2.1, $\ell(\theta)$ can be written as

$$\begin{aligned} \ell(\theta) &\propto \sum_i \log \int \prod_{q=1}^Q \{\xi_q \mathcal{R}_i(T_i)\}^{D_{iq}} \exp\left\{-\xi_q \int_{\Omega_{iq}} \mathcal{R}_i(s) ds\right\} & (3) \\ &\times (\sigma^2)^{-n_i/2} \exp\left\{-\|y_i - X_i\beta - Z_i b_i\|^2 / 2\sigma^2\right\} \\ &\times \det(D)^{-1/2} \exp(-b_i^\top D^{-1} b_i / 2) db_i, \end{aligned}$$

where $\theta = (\theta_t^\top, \theta_y^\top, \theta_b^\top)^\top$ denotes the full parameter vector, with θ_t denoting the parameters for the event time outcome, θ_y the parameters for the

longitudinal outcomes, θ_b the unique parameters of the random-effects covariance matrix, y_i is the $n_i \times 1$ vector of longitudinal responses of the i th subject, $D_{iq} = \delta_i I(v_{q-1} < T_i \leq v_q)$ is the event indicator for the q th interval for the piecewise-constant baseline risk function, $\mathcal{R}_i(t) = \exp[\gamma^\top w_{i1} + \alpha\{w_{i2} \times m_i(t)\}]$, and $\Omega_{iq} = \{s : \min(T_i, v_{q-1}) < s \leq \min(T_i, v_q)\}$. Moreover, it is useful to note that the score vector corresponding to (4) can be rewritten in the form

$$\begin{aligned} S(\theta) &= \frac{\partial \ell(\theta)}{\partial \theta^\top} = \sum_i \frac{\partial}{\partial \theta^\top} \log \int p(T_i, \delta_i | b_i; \theta) p(y_i | b_i; \theta) p(b_i; \theta) db_i \\ &= \sum_i \int A(\theta, b_i) p(b_i | T_i, \delta_i, y_i; \theta) db_i, \end{aligned} \quad (4)$$

where $A(\theta, b_i) = \partial\{\log p(T_i, \delta_i | b_i; \theta) + \log p(y_i | b_i; \theta) + \log p(b_i; \theta)\} / \partial \theta^\top$. Note that the observed data score vector is expressed as the expected value of the complete data score vector with respect to the posterior distribution of the random effects. This implies that if the score equations corresponding to (4) are solved with respect to θ , with $p(b_i | T_i, \delta_i, y_i; \theta)$ fixed at the θ value of the previous iteration, then this corresponds to an EM algorithm, whereas if the score equations are solved with respect to θ considering $p(b_i | T_i, \delta_i, y_i; \theta)$ also a function of θ , then this corresponds to a direct maximization of the observed data log-likelihood.

3 Pseudo-adaptive Gauss-Hermite Quadrature rule

The integrals involved in the specification of the score vector (4) do not have a closed-form solution, and therefore a numerical method must be employed for their evaluation. A standard choice is the Gauss-Hermite rule that approximates the integral by a weighted sum of integrand evaluations at pre-specified abscissas (Pineiro and Bates, 1995). The quality of this approximation is improved as the number of quadrature points is increased. A critical aspect that also greatly influences the quality of the Gauss-Hermite approximation is the location of the quadrature points with respect to the location of the main mass of the integrand. That is, if $g(b) = A(\theta, b)p(b | T_i, \delta_i, y_i; \theta)$ is concentrated around a point far from zero, or if the spread in $g(b)$ is quite different from the spread of the weight function $\exp(-b^\top b)$, then applying the standard Gaussian-Hermite rule directly to $g(b)$ can give a very poor approximation, even for large K (Pineiro and Bates, 1995). To solve this problem the adaptive Gauss-Hermite rule has been proposed that appropriately centers and scales the integrand in each iteration of the optimization algorithm. This rule typically requires much less quadrature points to obtain an approximation error of the same magnitude compared to the standard Gauss-Hermite, but is much more computationally demanding due to the requirement to locate \hat{b}_i for each subject and in each iteration.

To decrease this computational burden we propose here to use a pseudo-adaptive Gauss-Hermite quadrature. The motivation behind this rule comes from the properties of the posterior distribution of the random effects $p(b_i | T_i, \delta_i, y_i; \theta)$ whose mode \hat{b}_i and second order derivative \hat{H}_i we need to determine. Written in the log-scale, this density is proportional to

$$\sum_{j=1}^{n_i} \log p\{y_i(t_{ij}) | b_i; \theta_y\} + \log p(b_i; \theta_b) + \log p(T_i, \delta_i | b_i; \theta_t, \beta),$$

from which we observe that as n_i increases, the leading term is the logarithm of the density of the linear mixed model $\sum_j \log p\{y_i(t_{ij}) | b_i; \theta_y\}$. This is quadratic in b_i and will resemble the shape of a multivariate normal distribution. In particular, using a variant of the Bayesian central limit theorem and under general regularity conditions, we obtain that as $n_i \rightarrow \infty$, $p(b_i | T_i, \delta_i, y_i; \theta) \xrightarrow{P} \mathcal{N}(\tilde{b}_i, \tilde{H}_i^{-1})$, where $\tilde{b}_i = \operatorname{argmax}_b \{\log p(y_i | b; \theta_y)\}$ and $\tilde{H}_i = -\partial^2 \log p(y_i | \tilde{b}_i; \theta_y) / \partial b \partial b^\top$. In practice, this suggests that as n_i increases, it is sufficient to re-center and re-scale the integrand for each subject by utilizing only the information that comes from the mixed-effects model for the longitudinal outcome. Thus, instead of the standard transformation used in the adaptive Gauss-Hermite rule, we propose to first fit the linear mixed-effects model, extract the empirical Bayes estimates and their covariance matrix, and use the transformation

$$\begin{aligned} E\{A(\theta, b_i) | T_i, \delta_i, y_i; \theta\} \\ \approx 2^{q/2} |\tilde{B}_i|^{-1} \sum_{t_1 \dots t_q} \pi_t A(\theta, \tilde{r}_t) p(\tilde{r}_t | T_i, \delta_i, y_i; \theta) \exp(-\|b_t\|^2), \end{aligned} \quad (5)$$

where $\tilde{r}_t = \tilde{b}_i + \sqrt{2} \tilde{B}_i^{-1} b_t$, b_t denote the quadrature points, \tilde{B}_i the Choleski factor of \tilde{H}_i , and $\tilde{\theta}_y$ are the maximum likelihood estimates from the linear mixed model fit. This procedure is very similar with the adaptive Gauss-Hermite rule, but we implement it only *once*, at the start of the optimization, and we do not further update the quadrature points afterwards. The computational advantages are twofold: First, we can use fewer quadrature points than we would have used in the standard Gauss-Hermite rule, and second, we can avoid the computationally demanding relocation of the quadrature points at each iteration of the adaptive Gauss-Hermite rule.

4 Analysis of the PBC Dataset

We illustrate the proposed numerical integration rule in the primary biliary cirrhosis (PBC) study conducted by the Mayo Clinic from 1974 to 1984 (Murtaugh et al., 1994). The dataset includes 158 patients randomized to D-penicillamine and 154 to placebo, for which we are interested in the association between the longitudinal serum bilirubin levels and the risk

for death. To measure the strength of the association between these two outcomes we fit an appropriate joint model to the available data. In a previous analysis of the same dataset by (Ding and Wang, 2008) it has been noted that the serum bilirubin profiles are rather nonlinear for some of the patients. Thus, to allow for more flexibility we expand the random-effects vector and postulate a separate B-spline basis for the time effect for each patient. This formulation results in a four-dimensional random-effects vector that requires a four-dimensional numerical integration for each patient during each iteration of the maximization algorithm. For the survival process we use the relative risk model

$$h_i(t \mid \mathcal{M}_i(t), w_i) = h_0(t) \exp[\gamma_1 \mathbf{D-pnc}_i + \alpha_1 m_i(t) + \alpha_2 \{m_i(t) \times \mathbf{D-pnc}_i\}],$$

where the baseline risk function $h_0(\cdot)$ is assumed piecewise constant in $Q = 7$ intervals. Primary interest is in parameters α_1 and α_2 that measure the association between the true value of serum bilirubin at time t and the risk for an event at the same time point for the two treatment groups. We observed that serum bilirubin is strongly related with the risk for death, with each one unit increase of the current value of log serum bilirubin is associated with a 4.1-fold increase (95% CI: 3.03; 5.46) in a patient's risk in the placebo group, and a 3.4-fold increase (95% CI: 2.63; 4.41) in a patient's risk in the D-penicillamine group. With respect to computing time, the pseudo-adaptive rule was several orders of the magnitude faster than the standard Gauss-Hermite quadrature. In addition, the finite sample performance of the pseudo-adaptive rule has been empirically corroborated with simulation studies, which showed excellent performance in practice with respect to both parameter estimates and standard errors.

References

- Ding, J. and Wang, J.-L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, **64**, 546–556.
- Murtaugh, P. et al. (1994). Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, **20**, 126–134.
- Pinheiro, J. and Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.
- Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B*, **71**, 637–654.
- Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, **14**, 809–834.

Incorporating prior knowledge in Bayesian modeling of sparse networks

Veronika Ročková¹, Emmanuel Lesaffre^{1,2}

¹ Department of Biostatistics Erasmus MC Rotterdam, The Netherlands

² L-BioStat, Catholic University Leuven, Belgium

E-mail for correspondence: v.rockova@erasmusmc.nl

Abstract: We adopted the Bayesian networks approach to modeling conditional independence relations between two sets of genomic variables through sparse estimation of a system of linear regressions. Sparsity is a natural requirement achievable through variable selection/shrinkage priors within the hierarchical Bayesian framework. We propose a hierarchical prior construction for Bayesian shrinkage estimation in linear regression, where the shrinkage behavior is locally influenced by a stochastic linear combination of multiple prior association scores. We developed a generalized EM algorithm, which constitutes a computationally fast alternative to the MCMC strategies, which is beneficial in high-dimensional applications.

Keywords: Adaptive Bayesian LASSO; Bayesian Networks; EM algorithm.

1 Introduction

Deciphering gene regulatory mechanisms is one of the most important research topics in genomics. Improved understanding of these complex mechanisms necessitates integration of different sources of genomic data. Bayesian learning provides a methodological benefit in such data integration, as it enables flexible incorporation of prior association structures implied by molecular characteristics of the DNA sequences and various biochemical interactions.

We considered the contributions of such an integrative analysis in the context of predicting functional relationships between microRNAs and genes in acute myeloid leukemia. MicroRNAs are short non-coding messenger RNAs that target their protein coding counterparts through complementary base-pairing. Several computational algorithms have been proposed to predict the microRNA targets. These are primarily based on biological considerations with a limited or no incorporation of experimental data. These algorithms give rise to multiple prior association networks that can be used in the analysis.

2 Motivating Dataset

We analyze data collected at the Department of Hematology at Erasmus Medical Center in Rotterdam on patients with acute myeloid leukemia. The 212 cases were analyzed for (1) the expression of $M = 153$ microRNAs using quantitative reverse-transcription-polymerase chain reaction, (2) gene expression using Affymetrix Human Genome Gene-Chips. Prior association networks were generated from 4 existing prediction algorithms (`miRanda`, `TargetScan`, `Pita`, `PicTar`). Only $G = 898$ genes that were associated with survival and at the same time targeted by at least one of the prediction algorithms were selected for the analysis. Each of the 4 prior networks is represented by a $G \times M$ matrix of prior binary association scores, where the 0/1 pattern encodes the predicted associations.

3 Motivation

Suppose we have N independent realizations of a gene expression vector assembled in G vectors $\mathbf{y}_1, \dots, \mathbf{y}_G$ of size $N \times 1$. Denote \mathbf{X} a $N \times M$ matrix of microRNA expressions measured on the same set of samples. We assume that the vectors of gene expression values and the columns of \mathbf{X} have been centered to have a zero mean. One possible strategy to elucidate relationships between the two sets of variables is to consider a series of individual sparse regressions, treating the microRNAs as independent and genes as dependent variables. In the analogy to the Meinshausen and Bühlmann (2006) approach we might consider the following independent LASSO regression problems:

$$\min_{\beta_i} \left(|\mathbf{y}_i - \mathbf{X}\beta_i|^2 + \lambda_i \sum_{j=1}^M |\beta_{ij}| \right), \quad (i = 1, \dots, G). \quad (1)$$

It is natural to assume that if i -th gene and j -th microRNA are a priori likely to be associated, the regression coefficient β_{ij} should be shrunk to a lesser extend. This can be achieved by considering a weighted LASSO version:

$$\min_{\beta_i} \left(|\mathbf{y}_i - \mathbf{X}\beta_i|^2 + \lambda_i \sum_{j=1}^M w_{ij} |\beta_{ij}| \right), \quad (i = 1, \dots, G), \quad (2)$$

where the weights w_{ij} are related to the prior association scores and as such regulate the shrinkage intensity for each individual coefficient. The disadvantage of this approach is that the weights are assumed to be deterministic. It is not straightforward to combine multiple prior scores in one deterministic summary as the multiple biological priors may not be informative to the same extend. It is desirable to (1) regulate flexibly the extend to which each prior network contributes to the analysis and (2) quantify the level of agreement between the prior networks and the data at hand. In

addition, we may prefer to induce regularization on the whole model rather than only on the individual regressions. These requirements have lead us to the proposal, which is presented in the next section.

4 Method

Denote $\widetilde{\mathbf{X}}$ a $GM \times q$ matrix of assembled prior scores from q prediction databases, where the $[(i - 1)M + j]$ -th row contains scores about the association between i -th gene and j -th microRNA. Denote $\widetilde{\mathbf{x}}_i$ the i -th row of the matrix $(\mathbf{1}_{GM}, \widetilde{\mathbf{X}})$. Our hierarchical Bayesian model for microRNA targets is then the following:

$$\begin{aligned} \mathbf{Y}_i | \mathbf{X}, \boldsymbol{\beta}_i, \sigma_i^2 &\sim N_N(\mathbf{X}\boldsymbol{\beta}_i, \sigma_i^2 \mathbf{I}_N), \quad (i = 1, \dots, G), \\ \boldsymbol{\beta}_i | \boldsymbol{\tau}_i, \sigma_i^2 &\sim \prod_{j=1}^M N(0, \sigma_i^2 \boldsymbol{\tau}_{ij}^2), \quad (i = 1, \dots, G), \\ \boldsymbol{\tau}_1^2, \dots, \boldsymbol{\tau}_G^2 | \boldsymbol{\lambda}_1^2, \dots, \boldsymbol{\lambda}_G^2 &\sim \prod_{i=1}^G \prod_{j=1}^M \lambda_{ij}^2 \exp(-\lambda_{ij}^2 \boldsymbol{\tau}_{ij}^2), \\ \lambda_{ij}^2 | \mathbf{b} &\sim \Gamma(a, \widetilde{\mathbf{x}}'_{(i-1)M+j} \mathbf{b}), \\ b_l &\sim \Gamma(\alpha, \gamma), \quad (l = 1, \dots, q + 1), \\ \sigma_i^2 &\sim \text{IGamma}(c, d), \end{aligned}$$

where $\Gamma(a, b)$ and $\text{IGamma}(a, b)$ denote the gamma and inverse gamma distributions with shape a and scale b . This hierarchical model is a joint shrinkage model for G separate regressions. The intercept b_1 can be regarded as a *global* shrinkage hyper-parameter, whereas the remainder of the linear predictor induces *local* shrinkage on the individual regression coefficients. Higher values of the linear predictor correspond to heavier shrinkage and therefore stronger evidence *against* the association. Note that we are using reversed binary coding, where the prior score equals 1 whenever there was no prior association and 0 otherwise. The coefficients \mathbf{b} are estimated together with the remaining parameters, enabling posterior inference about the degree of informativeness of the respective prior networks. Finally, we assume that the prior scores have a positive shrinkage impact and therefore we assign independent gamma priors on the coefficients in \mathbf{b} .

5 Computation

For the sake of illustration, we outline the computation assuming $G = 1$, i.e. we omit the subscript i from the notation. The algorithm extends naturally to multiple responses using block parameter updates. Instead of involved MCMC calculations to obtain the full posterior, we extend the EM algorithm of Griffin and Brown (2012) for posterior mode finding, treating the latent variances in $\boldsymbol{\tau}^2$ as missing data. The conditional

expectation of the complete log posterior distribution given the observed data and current values is

$$\begin{aligned} Q\left(\boldsymbol{\beta}, \mathbf{b}, \sigma \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}\right) &= E_{\tau^2} \left[\log p(\boldsymbol{\beta}, \mathbf{b}, \sigma, \boldsymbol{\tau}^2 \mid \mathbf{y}) \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}, \mathbf{y} \right] \\ &= C(\boldsymbol{\tau}^2) + Q_1\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}\right) \\ &\quad + Q_2\left(\mathbf{b} \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}\right), \end{aligned}$$

where

$$\begin{aligned} Q_1\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}\right) &= -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} - \frac{1}{2\sigma^2} \sum_{j=1}^M \beta_j^2 E_{\tau_j^2} \left(\frac{1}{\tau_j^2} \right) \\ &\quad - \frac{N + M + 2c + 2}{2} \log(\sigma^2) - \frac{d}{\sigma^2}, \\ Q_2\left(\mathbf{b} \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}\right) &= \sum_{j=1}^M \left[\log(a \tilde{\mathbf{x}}_j' \mathbf{b}) - (a + 1) E_{\tau_j^2} \left. \log(1 + \tau_j^2 \tilde{\mathbf{x}}_j' \mathbf{b}) \right| \right] \\ &\quad + \sum_{l=1}^{q+1} \left[(\alpha - 1) \log b_l - \frac{b_l}{\gamma} \right] \end{aligned}$$

and $E_{\tau^2}(\cdot)$ denotes the conditional expectation $E_{\tau^2}(\cdot \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}, \mathbf{y})$. Note that the expected complete log posterior is separable with respect to \mathbf{b} and $(\boldsymbol{\beta}, \sigma)'$, which facilitates the subsequent M-step. Whereas the solutions for $\boldsymbol{\beta}$ and σ can be found analytically, the maximization with respect to \mathbf{b} is complicated by the unavailability of the conditional expectation $E_{\tau^2} \left. \log(1 + \tau_j^2 \tilde{\mathbf{x}}_j' \mathbf{b}) \right|$ in closed form. In the spirit of a generalized EM algorithm, instead of finding the value that globally maximizes the function $Q_2\left(\mathbf{b} \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}\right)$ we choose $\mathbf{b}^{(k+1)}$ such that

$$Q_2\left(\mathbf{b}^{(k+1)} \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}\right) \geq Q_2\left(\mathbf{b}^{(k)} \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}\right). \quad (3)$$

Condition (3) is sufficient to guarantee the monotonicity property. The update $\mathbf{b}^{(k+1)}$ that satisfies (3) can be found by maximizing a surrogate function

$$\begin{aligned} M_2\left(\mathbf{b} \mid \boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k)}, \sigma^{(k)}\right) &= \sum_{j=1}^M \left[\log(a \tilde{\mathbf{x}}_j' \mathbf{b}) - (a + 1) E_{\tau_j^2} \left(\frac{\tau_j^2}{1 + \tau_j^2 \tilde{\mathbf{x}}_j' \mathbf{b}^{(k)}} \right) \tilde{\mathbf{x}}_j' \mathbf{b} \right] \\ &\quad + \sum_{l=1}^{q+1} \left[(\alpha - 1) \log b_l - \frac{b_l}{\gamma} \right], \end{aligned}$$

which follows from the minorization-maximization argument using the monotonicity property of conditional expectation and the fact that the linear Taylor approximation is a global underestimator of the convex function $\log(1 + \tau_j^2 \tilde{\mathbf{x}}_j' \mathbf{b})$. All the conditional expectations involved in the E-step then have a closed form solution.

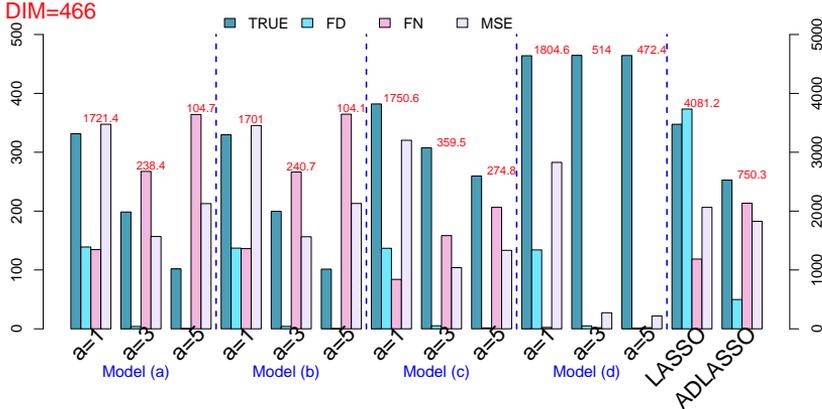


FIGURE 1. FD/FN/TRUE are average false/non/true discoveries. MSE is average mean squared error for β 's. Values are compared to the scale on the left apart from FD (scale on the right). Average estimated dimension is above the quadruplet of columns.

6 Simulation Study

We set $G = 900, M = 150$ and $N = 200$ to have similar dimensionalities as in our dataset. The rows of the predictor matrix \mathbf{X} were drawn independently from $N_M(\mathbf{0}, \Sigma_X)$, where $\Sigma_X = (\rho_{ij})_{i,j=1}^M$ with $\rho_{ij} = 0.5^{|i-j|}$. The true network with 466 associations was generated using an algorithm for random networks. Nonzero coefficients were sampled from the uniform distribution $U[-1, -0.5]$ (microRNAs negatively correlate with their targets). Response vectors were at each of the 100 replications generated independently from $N_N(\mathbf{X}\beta_i, \sigma^2 I_N)$ with $\sigma = 3$. We consider 4 settings: (a) no prior association matrices included, (b) two random prior association matrices $G \times M$ (Phi correlation coefficients $|r_\Phi| < 0.01$), (c) two random and two mildly informative matrices (r_Φ correlations 0.29 and 0.41) (d) two random matrices and true (perfect) prior. For each of the settings we consider 3 values of the shape parameter a . This parameter regulates the heaviness of the tails of the marginal prior distribution, where large values may cause unwanted bias in estimating large effects and too small values imply under-shrinkage of the noise coefficients. We set $c = d = 0, \alpha = 1$ and $\gamma = 10$. The results (Figure 1) are compared to the (adaptive) LASSO approach, where the penalty parameters were selected by cross-validation.

7 Application

Based on our simulation study, we select the shape parameter $a = 5$ (we want to minimize the number of false discoveries). We obtained a network with 548 associations (280 of which indicate negative correlation and correspond to potential true targets) and the following estimates $\hat{\mathbf{b}} : \hat{b}_{\text{intercept}} =$

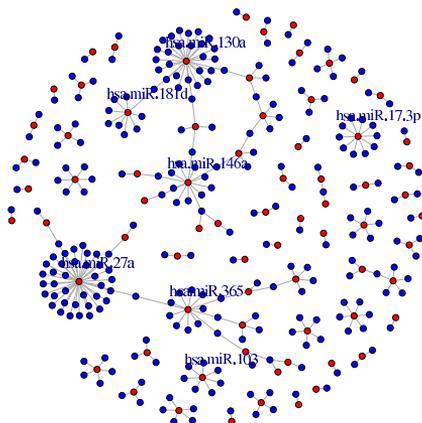


FIGURE 2. Estimated network, microRNAs/genes depicted as red/blue nodes. Only nodes with at least one nonzero association plotted.

7.43×10^3 , $\hat{b}_{\text{pita}} = 2.24 \times 10^{-1}$, $\hat{b}_{\text{pictar}} = 1.03 \times 10^6$, $\hat{b}_{\text{miranda}} = 7.38 \times 10^4$, and $\hat{b}_{\text{target}} = 1.43 \times 10^5$. Out of the 4 prediction algorithms, PicTar carried the most influential information. The plot of the network consisting of 280 negative associations is depicted in Figure 2. In contrast, using the LASSO method we obtained 15 099 interactions (8 040 of which indicative of negative correlation).

8 Discussion

We developed a method for Bayesian shrinkage estimation in linear regression that incorporates external knowledge and demonstrated its utility in estimating sparse graphical models. Our model shares similarities with the proposal of Stingo et al. (2010) who employed spike and slab priors and used MCMC for the estimation. We have opted for absolutely continuous priors and EM algorithm, which offers substantial computational benefit.

References

- Griffin, A. and Brown, S. (2012). Bayesian hyper-LASSOS with non-convex penalization. *Australian & New Zealand Journal of Statistics*, **53**, 423–442.
- Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, **34**, 1436–1462.
- Stingo, F., Chen, Y., Vannucci, M., Barrier, M., and Mirkes, P. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *Annals of Applied Statistics*, **4**, 2024–2048.

A semi-parametric joint model for two sequential times to events and one longitudinal covariate

Carles Serrat¹, Jaime-Abel Huertas², Guadalupe Gómez³

¹ Dept. Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Barcelona, Spain

² Dept. Estadística, Universidad Nacional de Colombia, Bogotá, Colombia

³ Dept. Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail for correspondence: `carles.serrat@upc.edu`

Abstract: The present work proposes a joint model for two sequential times to events together with longitudinal information, as an extension of the joint model by Wolfsohn and Tsiatis (1997) for one time to event and one longitudinal variable. The model is applied to the TIBET, a clinical trial in which antiretroviral therapy interruptions guided by CD4 counts and plasma HIV-1 RNA levels in chronically HIV-1-infected patients are under evaluation. Details on the modelling strategy and the resulting estimates are given.

Keywords: Joint Modelling; Longitudinal Data; Sequential Times; Survival Analysis.

1 Introduction and the Motivating Clinical Trial

This research is motivated by the TIBET clinical trial (Ruiz *et al.*, 2007) and we want to model two sequential times to event with longitudinal measurements. T_1 is the time to re-initiate *HAART* therapy and T_2 is the time to suspend therapy from the first re-initiation, and the longitudinal measurements are the levels of CD4 cell counts each four weeks. The goal is to use the longitudinal measurements as a marker for the times to event. The joint model will allow us to give prognosis for a time to event given covariates, the longitudinal process and the previous event time.

2 Notation

The variables of interest for each subject $i = 1, \dots, n$ followed over an interval $[0, \tau)$ are $\{T_{1i}, T_{2i}, R_i(u), 0 \leq u \leq \tau, X_i\}$, where T_{1i} and T_{2i} are event times, $\{R_i(u), 0 \leq u \leq \tau\}$ is the longitudinal response trajectory

for all times $u \geq 0$ and $X_i = [X_{1i}^T \ X_{2i}^T]^T$ is a vector of baseline (time 0) covariates, X_{1i} with influence over T_1 , and X_{2i} over T_2 .

We will consider only a situation where T_1 and T_2 may be right censored by the censoring times C_1 and C_2 respectively, so instead of T_{ji} we observe (Y_{ji}, δ_{ji}) , $j = 1, 2$, where $Y_{ji} = \min\{T_{ji}, C_{ji}\}$ and $\delta_{ji} = I(T_{ji} \leq C_{ji})$ which indicates whether Y_{ji} is an uncensored right value of T_{ji} . On the other hand, for some set of times t_{ij} , $j = 1, \dots, n_i$, instead of the true values $R_i(t_{ij})$ we observe $Z_i(t_{ij})$, then the observed data for subject i is $O_i = \{X_i, Y_i, \delta_i, Z_i, \tilde{t}_i\}$, where $\tilde{t}_i = (t_{i1}, \dots, t_{in_i})^T$, $Z_i = (Z_i(t_{i1}), \dots, Z_i(t_{in_i}))^T$, $Y_i = (Y_{1i}, Y_{2i})$, and $\delta_i = (\delta_{1i}, \delta_{2i})$.

3 Semi-parametric Joint Model of Two Sequential Times to Event and One Longitudinal Variable

We approach the problem of two sequential times to event with a sequence of conditional distributions (Lawless, 2003, Section 11.3). A natural choice for the survival model is to consider a distribution for a time to event given previous observed event times. Moreover, the sequence of conditional distributions for the times to event is jointly modeled with a mixed model, adapting the model of Wulfsohn and Tsiatis (1997).

3.1 Linear Trend for the Longitudinal Variable

First, we assume that the longitudinal variable is monotone not increasing (or viceversa) with linear trend as we show in the Figure 1 part (a). This is the simplest case for the longitudinal variable that we consider. Due to the existing association between the longitudinal and survival processes, there is a high probability that the longitudinal trend changes with the occurrence of the first event. Nevertheless we treat it because might happen cases where the monotonous trend over time persists. In the TIBET clinical trial we found that the GPT enzyme has linear trend over T_1 and T_2 , so we need the methodology with longitudinal linear trend to analyze whether the effect of the enzyme in the times to event is significant.

We analyze our proposal linking the longitudinal and survival sub-models with the current value, and assuming a linear trend for the longitudinal data without fix part. A particular joint model analyzed, is

$$Z_{ij} = b_{0i} + b_{1i} t_{ij} + e_i(t_{ij}) \tag{1}$$

$$\lambda_{T_1}(t_1 | b_i, X_{1i}; \eta_1, \beta_1) = \lambda_{1,0}(t_1) \exp\{\eta_1 X_{1i} + \beta_1(b_{0i} + b_{1i}t_1)\} \tag{2}$$

$$\lambda_{T_2}(t_2 | t_{1i}, b_i, X_{2i}; \eta_2, \beta_2, \gamma) = \lambda_{2,0}(t_2) \exp\{\eta_2 X_{2i} + \beta_2(b_{0i} + b_{1i}(t_{1i} + t_2)) + \gamma t_{1i}\}, \tag{3}$$

where η_1 and η_2 are vectors of parameters associated to baseline covariates, β_1 and β_2 are parameters of association between the longitudinal and

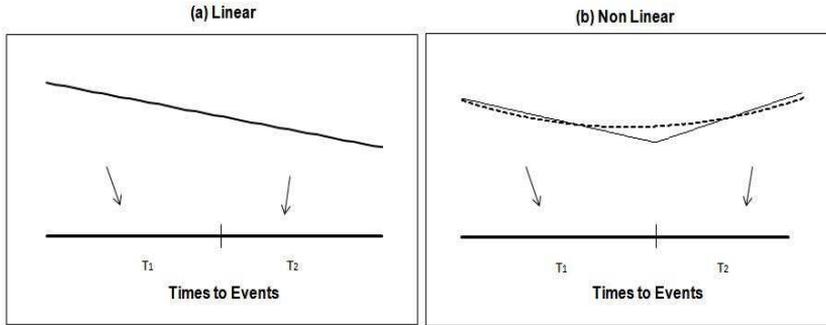


FIGURE 1. Two types of trends for the longitudinal data over T_1 and T_2

survival processes, γ describes the relation among the times to event, and both baseline risks $\lambda_{1,0}(\cdot)$ and $\lambda_{2,0}(\cdot)$ are left unspecified, and different. For the likelihood function we apply the same assumptions made by Wulfsohn and Tsiatis (1997). The assumption of non-informative censoring extend to this case of censoring process. The errors e_i are assumed to be mutually independent, normally distributed with mean 0 and variance σ_e^2 , and independent of b_i and conditionally independent of all other variables given (b_i, X_i) . If we assume that, given random effects and covariates, Z , T_1 , and $T_2 | T_1$, are all independent, then the observed likelihood is

$$L(\Omega) = \prod_{i=1}^n \int_{b_i} \left\{ \prod_{j=1}^{n_i} f(z_{ij} | b_i; \sigma_e^2) \right\} f(Y_i, \delta_i | b_i, X_i; \psi_{T|b}) f(b_i; B, \Gamma) db_i, \quad (4)$$

where $\Omega = (\psi_{T|b}, B, \Gamma, \sigma_e^2)$ and $\psi_{T|b} = (\eta_1, \eta_2, \beta_1, \beta_2, \gamma, \lambda_{1,0}, \lambda_{2,0})$. The vector of random effects $b_i = [b_{0i} \ b_{1i}]^T$ is taken to be normally distributed with mean B and covariance matrix Γ .

The algorithm by Wulfsohn and Tsiatis (1997) is extended in many cases of joint models with longitudinal and survival data. In our case the model is extended and applied in two parts. The first step models Z and $T_1 + T_2$ using the Wulfsohn and Tsiatis’s algorithm ignoring the sequence of the times to event, and the second step fits T_1 and $T_2 | T_1$ likely the modelling was direct and jointly with Z . We named this method as EM modified algorithm for joint model with longitudinal information and sequential times to event.

3.2 Non Linear Trend for the Longitudinal Variable

When the first time to event occurs the conditions in the statistical units may change, affecting the evolution of the longitudinal variables. For our TIBET dataset, the CD4 cell counts changes its trend when the therapy is

restated. Figure 1 part (b) shows how the trend may be linear in two pieces or parabolic. We focus the analysis for this two cases of non linear trend.

a)

$$Z_{ij} = b_{0i} + (b_{1i} t_{ij} + b_{2i}(t_{ij} - t_{1i})I(t_{ij} \geq t_{1i})) + e_i(t_{ij}) \quad (5)$$

$$\lambda(t_1 | b_i; \beta_1) = \lambda_{1,0}(t_1) \exp\{\beta_1(b_{0i} + b_{1i}t_1)\} \quad (6)$$

$$\lambda(t_2 | t_{1i}, b_i; \beta_2, \gamma) = \lambda_{2,0}(t_2) \exp\{\beta_2(b_{0i} + b_{1i} \cdot t_{1i} + (b_{1i} + b_{2i})t_2) + \gamma t_{1i}\}. \quad (7)$$

b)

$$Z_{ij} = b_{0i} + b_{1i} t_{ij} + b_{2i} t_{ij}^2 + e_i(t_{ij}) \quad (8)$$

$$\lambda(t_1 | b_i; \beta_1) = \lambda_{1,0}(t_1) \exp\{\beta_1(b_{0i} + b_{1i}t_1 + b_{2i}t_1^2)\} \quad (9)$$

$$\lambda(t_2 | t_{1i}, b_i; \beta_2, \gamma) = \lambda_{2,0}(t_2) \exp\{\beta_2(b_{0i} + b_{1i}(t_{1i} + t_2) + b_{2i}(t_{1i} + t_2)^2) + \gamma t_{1i}\}, \quad (10)$$

4 Application to the TIBET Dataset

In the TIBET dataset T_1 is the time to the first restart of therapy, and T_2 is the time from the first restart of therapy to the suspension of it. We have 74% observed cases for T_1 and 68% observed cases for T_2 .

The evolution of the CD4 is decreasing until the first time to event, and then is increasing, so we model it with a two piecewise based on (5)-(7), and with a parabolic trend based on (8)-(10). The analysis of these models only aims to compare them in order to determine the best alternative to model the longitudinal variable for prognosis. Table 1 shows the results for both models fitted with the EM modified algorithm, and Figure 2 shows some fitted cases with these models.

The selected joint model is as follow, and the results obtained with the EM modified algorithm are shown in Table 2.

$$Z_{ij} = b_{0i} + (b_{1i} t_{ij} + b_{2i}(t_{ij} - t_{1i})I) + e_i(t_{ij}) \quad (11)$$

$$\lambda(t_1 | b_i, VL_i; \eta_1, \beta_1) = \lambda_{1,0}(t_1) \exp\{\eta_1 VL_i + \beta_{11}(b_{0i} + b_{1i}t_1) + \beta_{12}b_{1i}\} \quad (12)$$

$$\lambda(t_2 | t_{1i}, b_i, VL_i; \eta_2, \beta_2, \gamma) =$$

$$\lambda_{2,0}(t_2) \exp\{\eta_2 VL_i + \beta_{21}(b_{0i} + b_{1i}(t_{1i} + t_2) + b_{2i}t_2) + \beta_{22}(b_{1i} + b_{2i}) + \gamma t_{1i}\}. \quad (13)$$

The influence of the intercepts b_0 and $b_0 + b_1 t_1$ in T_1 and T_2 respectively, are not significative. It is logic since patients start the trial without therapy with good and similar conditions, and the restart of therapy is due to the threshold reached in the levels of CD4 and viral load.

TABLE 1. Joint models for T_1 and T_2 with two different model for CD4. It is assumed semi parametric form in the hazard risks.

<i>Parameter</i>	Two Piecewise			Parabolic		
	<i>Estimate</i>	<i>s.e.</i>	<i>p - value</i>	<i>Estimate</i>	<i>s.e.</i>	<i>p - value</i>
<i>Mixed</i>						
B_0	25.8263	0.4569	< 0.0001	26.5564	0.4709	< 0.0001
B_1	-0.0502	0.0041	< 0.0001	-0.0732	0.0092	< 0.0001
B_2	0.1248	0.0080	< 0.0001	0.0007	0.0001	< 0.0001
σ_{11}	20.8764	2.9524	< 0.0001	22.1763	3.1362	< 0.0001
σ_{12}	-0.1186	0.0221	0.6483	-0.1964	0.0477	0.7388
σ_{13}	0.0166	0.0364	< 0.0001	0.0001	0.0003	< 0.0001
σ_{22}	0.0017	0.0002	< 0.0001	0.0085	0.0012	< 0.0001
σ_{23}	-0.0016	0.0004	< 0.0001	-4E-5	6E-6	< 0.0001
σ_{33}	0.0063	0.0009	< 0.0001	3E-7	4E-8	< 0.0001
σ_e^2	6.1463	0.1881	< 0.0001	5.5382	0.1695	< 0.0001
<i>Survival T_1</i>						
β_1 (Assoc.)	-0.1881	0.0361	<0.0001	-0.1060	0.0335	0.0016
<i>Survival T_2</i>						
β_2 (Assoc.)	0.0465	0.0414	0.2614	-0.0002	0.0386	0.9958
γ (T_1)	-0.0172	0.0063	0.0063	-0.0178	0.0072	0.0134

TABLE 2. Survival result for the joint model for T_1 and T_2 with two piecewise mixed model for the CD4 evolution, based in EM modified algorithms. It is assumed semi parametric form in the hazard risks.

<i>Parameter</i>	<i>Estimate</i>	<i>s.e.</i>	<i>p - value</i>
<i>Survival T_1</i>			
β_{11} ($b_{0i} + b_{1i}t_1$)	-0.2044	0.0405	< 0.0001
β_{12} (b_{1i})	-14.3298	3.5632	< 0.0001
η_1 (VL_i)	0.6521	0.1858	0.0004
<i>Survival T_2</i>			
β_{21} ($b_{0i} + b_{1i}(t_{1i} + t_2) + b_{2i}t_2$)	0.0257	0.0430	0.5500
β_{22} ($b_{1i} + b_{2i}$)	6.7904	2.4090	0.0048
η_1 (VL_i)	-0.0169	0.2302	0.9414
γ (t_{1i})	-0.0210	0.0079	0.0079

The slope is the only significant random effect in T_2 , and due to the fact that the effect of the viral load pre-therapy is diluted in T_2 , then we have the slope of the longitudinal variable along T_2 and the observed values of T_1 , the only significant covariate in the survival model for T_2 .

The negative sign for $\hat{\gamma}$ indicates that for long times to restart therapy, we have long times to suspend therapy.

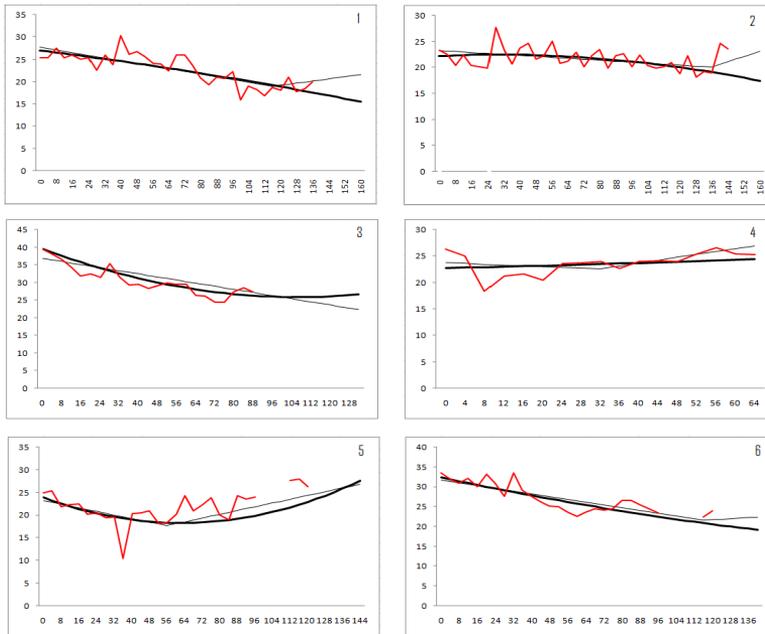


FIGURE 2. Evolution of the CD4 with models based in two piecewise and parabolic trend, for some cases of the TIBET dataset

Acknowledgments: This work has been partially supported by the grant MTM2008-06747-C02-01 from the Ministerio de Ciencia e Innovación of Spain. Authors are indebted to the GRASS group for fruitful discussions and to the Fundació Lluita contra la SIDA for providing the data.

References

- Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd edition. *Wiley* Hoboken.
- Ruiz, L., Paredes, R., Gómez, G., Romeu, J., Domingo, P., Pérez-Alvarez, N., Tambussi, G., Llibre, J.M., Martnez-Picado, J., Vidal, F., Fumaz, C.R., Clotet, B., and Group, T.I.B.E.T.S. (2007). Antiretroviral therapy interruption guided by CD4 cell counts and plasma HIV-1 RNA levels in chronically HIV-1-infected patients. *AIDS*, **21**(2), 169–178.
- Wulfsohn, M.S., and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

Modelling the rank order of Web search engine results

Michael G. Schimek¹, Marcus Bloice¹

¹ Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, 8036 Graz, Austria

E-mail for correspondence: michael.schimek@medunigraz.at

Abstract: In this paper we examine data collected by Web search engines. Such devices produce very long lists of distinct items (URLs) representing locations of profile pages in rank order. Our goal is to obtain objective and reproducible information from Web resources that are less browser or user profile dependent. For this purpose at least two search engine results for a given query term are required. These results can be represented by individual rankings, i.e. ordered lists of Web sites containing relevant findings. The statistical task is the identification of common top-ranking items from two or more lists with the goal of constructing a set of consolidated items of high consensus between the involved search engines. Thus informative partial lists can be obtained which the user would wish to consider in detail. Up to now there has not been a strategy to estimate items of top-relevance from ranking data. We take advantage of a moderate deviation-based inference procedure due to Hall and Schimek (2012) to solve this problem. We have developed R and Java software that integrates top-ranked search engine results. An example based on Bing, Yahoo! and Google query findings is presented.

Keywords: Consensus Ranking; Query Term; Ranked List; Top- k List; Web Search Engine.

1 Motivation

Web analytics are becoming increasingly popular. Many tasks are in the domain of statistics rather than computer science such as data mining and its speciality, Web mining (the usual tools for both are clustering and classification; see e.g. Schoier and Schimek, 2003). In content mining of the Web we examine data collected by search engines. Such devices produce very long lists of distinct items (URLs) representing locations of profile pages in rank order. For a while the primary goal was to develop efficient algorithms that provide most complete results in as little time as possible. Now these immense investments need to pay off for companies offering such services. As an immediate consequence, content that can be found is variable beyond what can be explained by differences in technology because of

prefiltering by the search engine providers, reflecting their business, mostly advertising, interests. Nowadays, the commercial value of a company owning a Web service is closely associated with its ability to connect user query terms with the type of match and its potential to constitute consumption activities. Also, content providers have effectively reverse engineered search engine techniques, thus pushing selected results artificially high. As a direct consequence, the algorithms are black boxes. For a search engine user with educational, research or academic motifs the results he/she can obtain are neither objective nor reproducible any more because they depend on the internet protocol address (for host identification and location addressing) as well as the user profile unless an anonymous access to the Web is taken. To prevent customers from an unfiltered usage of Web content, Google, the company with the highest penetration of the market, has stopped providing APIs among other measures to disclose principally free information. Another strategy is to control the number of protocolled searches one can perform or the length of the ordered list of hits. As a consequence a new field of research has emerged that is concerned with the multiple use of Web query systems (see e.g. White et al., 2008).

2 The statistical task

In this paper our interest is contrary to usage mining where the browser data of users are examined. We wish to obtain information from Web resources that are as much as possible browser and user profile independent. For this purpose multiple, at least two, search engine results for a given query term are required. These results can be represented by individual rankings, i.e. ordered lists of Web sites containing relevant findings. The statistical task is the identification of common top-ranking items from two or more lists with the goal of constructing a set of consolidated items of high consensus between the involved search engines. Thus informative partial lists of length k (notion of top- k lists) can be obtained which the user could consider in detail. To our knowledge, there has not been a strategy to estimate items of top-relevance from ranking data.

In our data-driven approach a general decrease of the probability for consensus rankings with increasing distance from the top rank position is assumed. This assumption is a built-in feature of search engine technology. Instead of analyzing the URL lists directly we calculate zero-one stream data from the pairs of rankings representing the discordance of items with respect to their rank positions. Moreover, we introduce a distance measure which implicitly defines the data stream. What we aim at is the estimation of the point of information degradation into noise. We take advantage of Hall and Schimek's (2012) contribution in which they have developed an efficient inference procedure to solve this problem. For our algorithm, apart from some tuning parameters, nothing but the data stream input is required to identify and integrate top-ranked objects. This fact reduces

substantially the computational burden and thus very long ranked lists can be processed.

Why is statistical modelling involved? Theoretically speaking, for a given set of N items, arbitrary ranked lists can be constructed by successive permutations. However, this fact does not help in practice when we have to analyze realizations of such lists because they comprise irregularities in terms of position shifts, inverted orderings, missing assignments, etc. As a consequence, a unique top- k list or a complete set of top-conforming items does not exist. What can be done is to fit a domain of models that adequately represents the above entities. The involved models are controlled via tuning parameters.

3 The distance-based data stream input

Let us assume an arbitrary fixed set of N URLs in the Web space. The ranking of URLs is always from 1 to N and the rank assignment in each list is independent of the assignment in the other lists (i.e. independent ranking mechanisms). Of course there can be missing items in one or more lists.

Let us define a sequence of indicators, where $I_j = 1$ if the ranking, given by the second search engine to the item ranked j by the first search engine, is not more than δ index positions distant from j , and otherwise $I_j = 0$. The parameter δ is the admissible shift in index positions of a particular item o in one list, say τ_i , with respect to the other list, say τ_j . This means that we assume concordance (i.e. $I = 1$) for an arbitrary item when its rank positions in τ_i and τ_j are maximal δ index values apart. Further, let us assume (i) independent Bernoulli random variables I_1, \dots, I_N , with $p_j \geq \frac{1}{2}$ for each $j \leq j_0 - 2$, $p_{j_0-1} > \frac{1}{2}$, and $p_j = \frac{1}{2}$ for $j \geq j_0$; (ii) a “general decrease” of p_j for increasing j that does not need to be monotone.

There are exploratory strategies how to select the distance parameter δ from empirical data (for details see Schimek and Budinská, 2010). However, for rank data of a known structure such as query results the δ -value can be prefixed.

4 The modelling of rank order data

Our task is to model ℓ input lists under the condition of missing assignments (incomplete rankings) because a specific item might not be listed in each of the query results. The non-parametric procedure due to Hall and Schimek (2012) provides an estimate of the point of degeneration j_0 for each pair of lists, where $j_0 - 1 = k$ is the length of the top list. It allows for various types of rank irregularities, missing rank assignments, and list lengths in the magnitude of thousands of objects. The estimation of \hat{j}_0 is achieved via a *moderate deviation*-based approach. In theoretical analysis of the

probability that an estimator, computed from a pilot sample size ν , exceeds a value z , the deviation above z is said to be a moderate deviation if its associated probability is polynomially small as a function of ν , and to be a large deviation if the probability is exponentially small in ν . In regular cases, the values of $z = z_\nu$ that are associated with moderate deviations are

$$z_\nu \equiv (C \nu^{-1} \log \nu)^{1/2},$$

where $C > \frac{1}{4}$. The null hypothesis H_0 that $p_k = \frac{1}{2}$ for ν consecutive values of k , versus the alternative H_1 that $p_k > \frac{1}{2}$ for at least one of the values of k , is rejected if and only if $\hat{p}_j^\pm - \frac{1}{2} > z_\nu$. The quantities \hat{p}_j^+ and \hat{p}_j^- represent estimates of p_j computed from the ν data pairs I_i for which i lies immediately to the right of j , or immediately to the left of j , respectively. Under H_0 , the variance of \hat{p}_j^\pm equals $(4\nu)^{-1}$ hence, we can evaluate the above inference procedure in practice. As pointed out already, apart from the pilot sample size ν and the constant C (the latter defaults to 0.251), the obtained model for the consolidated (top- k) list items also depends on the distance δ .

The complex decision problem is solved via an iterative algorithm, adjustable for irregularity in the rankings. The overall estimate \hat{k}^* for the ℓ lists τ_ℓ is calculated in the following way: The inference procedure is executed for all possible pairs $L = (\ell^2 - \ell)$ of lists τ_ℓ , thus we obtain L values \hat{k}_j ($j = 1, 2, \dots, L$). The overall top- k list length is then defined by $\hat{k}^* = \max_j(\hat{k}_j)$. Having obtained such an overall index \hat{k}^* , we arrive at truncated lists τ^* that can be integrated as well as represented by a special aggregation plot.

5 A graphical summary of consolidated results

For the integration of the truncated lists τ^* of individual lengths \hat{k}_j , we have further developed the aggregation map, first introduced by Schimek and Budinská (2010). This mapping tool produces list aggregation plots controlled by the two tuning parameters δ and ν . Again, this dependence reflects our modelling approach accounting for the fact that consolidated query term results can never be unique.

The extended aggregation map can be characterized as follows: We define an index $p = 1, 2, \dots$ and combine $\ell - 1$ aggregation levels (groupings of truncated lists) in one display: For each group of $\ell - p$ truncated lists down to the smallest group consisting of just one pair of lists, we (i) select an arbitrary reference list L^0 under the condition that it comprises $\max_i(\hat{k}_i)$ items among all pairwise comparisons in the group of rankings, (ii) print the symbols of its $\max_i(\hat{k}_i)$ items (i.e. URLs) vertically from the highest to the lowest rank position, and (iii) add the aggregation information for all remaining $\ell - p$ rankings (pairwise list combinations) in the group, ordered according to descending list length.

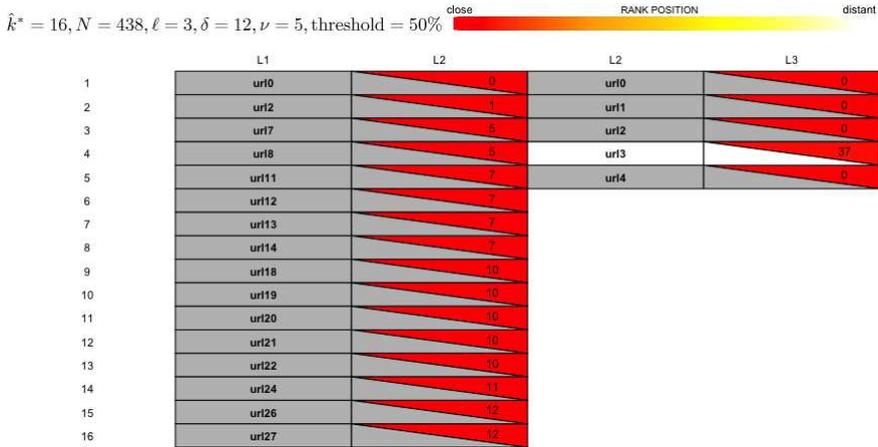


FIGURE 1. Aggregation map result from 3 Web search engines.

The aggregation information per symbol, item, and group consists of three measures represented by colored triangles and rectangles, respectively, outlined in array format: (a) The **membership** of an individual item in the top- k lists, *yes* is denoted by the color ‘grey’ and *no* by the color ‘white’. (b) The **distance** d of the rank of an individual item $o \in L^0$ from its position in the other list, is denoted by a triangle color scaled from ‘red’ *identical* to ‘yellow’ *far distant*. An additional integer value gives the numerical distance between the item’s rank positions, a negative sign means ranked lower, and a positive sign means ranked higher in L with respect to L^0 . (c) The rectangular of a symbol takes on the color ‘grey’ when the **percentage** of $d \leq \delta$ across the columns of a group is above some prespecified threshold, and ‘white’ otherwise. An example of such a plot is given in FIGURE 1.

6 An application

For demonstration purpose we have analyzed the results from three Web search engines for the query ‘GLIM statistics’ (a well-known term in the statistical modelling community). GLIM in statistics stands for the predecessor of this conference as well as the package ‘Generalized Linear Interactive Modelling’. The query ‘GLIM’ without ‘statistics’ would have produced far too many irrelevant hits. Our Java software allows us to convert Bing and Yahoo! query results into a format we can use for the integrative analysis described here. For the third search engine, Google, it was a time-consuming manual task (intended by Google to prevent users from research use of their data) to produce a ranking we could further analyze. For our example we used ranked lists each $N = 438$ URLs long. For the

consolidation of these three query results we applied an extended version of the R package `TopKLists` (Schimek et al., 2011). The user can perform the modelling of rank order data in an interactive fashion using sliders to control the distance, the pilot sample size, and the threshold. In FIGURE 1 we display an aggregation map result for an optimal choice of tuning parameters. The estimated overall top- k list length is 16. As can be seen, lists L2 and L1 have 16 URLs in common but all three top-lists together only url0. When the lists L3 and L1 are considered separately, they share 5 URLs, however, url3 with $d=37$ for a given $\delta = 12$ is below the threshold of 50%.

This short demonstration should at least allow for a glimpse at what can be gained from our approach in terms of consolidated reproducible query results including a clear indication of how many Web addresses are relevant for detailed investigation. Last but not least, quality differences between search engines can be worked out using our approach.

References

- Hall, P. and Schimek, M. G. (2012) Moderate deviation-based inference for random degeneration in paired rank lists. To appear in *Journal of the American Statistical Association*.
- Schimek, M. G. et al. (2011) Package “TopKLists” for rank-based genomic data integration. *Proceedings of the IASTED International Conference Computational Bioscience* (CompBio 2011), 434-440.
- Schimek, M. G. and Budinská, E. (2010) Visualization techniques for the integration of rank data. *COMPSTAT 2010. Proceedings in Computational Statistics* (e-book ISBN 978-3-7908-2603-6). Physica, Heidelberg, 1637-1644.
- Schoier, G. and Schimek, M. G. (2003) *On the analysis of Web Access Logs: Identifying dense clusters*. In Verbeke, G. et al. (ed) *Proceedings of the 18th International Workshop on Statistical Modelling*, 391-396.
- White, R. W. et al. (2008) Enhancing web search by promoting multiple search engine use. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. [no page numbers].

Modeling latent curves for genotype by environment interaction

Sabine K. Schnabel^{1,2}, Fred A. van Eeuwijk^{1,2}, Paul H. C. Eilers^{1,3}

¹ Biometris, Wageningen University and Research Centre, The Netherlands

² Centre for BioSystems Genomics, Wageningen, The Netherlands

³ Erasmus Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: `sabine.schnabel@wur.nl`

Abstract: In plant research data for a population of plants is often collected through repeated field and greenhouse trials under different environmental conditions. An appropriate model for this type of data should also deal with genotype by environment interaction. While the environments in field trials can frequently be characterized by geographic or meteorological conditions, it is equally likely that they are ordered according to a latent property. Therefore the order of the environments can often be unknown or unclear. We propose to use smooth latent curves to estimate this underlying order. The method is illustrated with simulated and empirical data.

Keywords: Latent curves; genotype by environment interaction; smoothing; P -splines

1 Introduction

A typical result of breeding trials with plants is a table in which for each genotype (G) a characteristic property (a so-called phenotype) is recorded for a number of environments (E). Generally, this kind of table cannot be fitted well by an additive model with effects for G and E. Hence we speak of genotype-by-environment (GxE) interaction. In general, the resulting data is given in a GxE table of phenotypic means. Most methods use centered data: row-wise, column-wise or double-centered.

In the literature one finds several ways to proceed which are essentially all based on the addition of multiplicative components. References to this model go back as far as Fisher and MacKenzie in 1923. Here, we explore another approach that assumes a smooth latent environmental gradient.

Suppose we have a set of n smooth curves –possibly with added noise– and we sample all of them at m positions given by the vector x . In the following we collect the data in an m by n matrix Y . Given this matrix and x , it is easy to estimate the curves by any smoothing method that works on each

column of Y separately. Permuting the rows randomly can do no harm, as long as the elements of x are permuted in the same way.

Imagine that the rows are permuted indeed and that we lost the vector x . How can we reconstruct it in a decent way? We propose a model that can perform this task. For a given GxE table it provides a latent gradient, the estimated x , that can be interpreted as an unknown environmental characteristic.

2 The model

Assume that x is given and we will smooth one column of Y denoted as y . An attractive choice is to use P -splines, which minimize

$$S_j = \sum_i (y_i - \sum_k a_k B_k(x_i))^2 + \lambda \sum_k (\Delta^2 a_k)^2 \quad (1)$$

Here, $B_k(x_i)$ is one of a set of (cubic) splines. The set is large and based on equally spaced knots. To tune smoothness a difference penalty is applied to the coefficients – see Eilers and Marx (1996) for details. If we consider all columns of Y , the objective function is the same for each of them, but we have to index the coefficients for the columns, to get a_{jk} . The overall objective function becomes

$$S = \sum_j \sum_i (y_{ij} - \sum_k a_{jk} B_k(x_i))^2 + \lambda \sum_j \sum_k (\Delta^2 a_{jk})^2. \quad (2)$$

Notice that the function is separable, i.e. we can smooth each column of Y separately.

Now assume that the coefficients are given as $A = [a_{jk}]$. We do not know x , but an approximation \tilde{x} to it. Therefore we want to compute a vector of corrections u minimizing

$$S^* = \sum_j \sum_i (y_{ij} - \sum_k a_{jk} B_k(\tilde{x}_i + u_i))^2. \quad (3)$$

We try to get as good a fit as possible given the coefficients A by shifting the positions of the rows of Y .

If the corrections u are small, we can use the following first order approximation:

$$B_k(\tilde{x}_i + u_i) \approx B_k(\tilde{x}_i) + u_i B'_k(\tilde{x}_i), \quad (4)$$

where $B'_k(\tilde{x}_i)$ is the first derivative of the k th spline evaluated at \tilde{x}_i . It is easy to see that minimization of S^* in (3) leads to regression of the residuals

$$r_{ij} = y_{ij} - \sum_k a_{jk} B_k(\tilde{x}_i) \quad (5)$$

on the derivatives

$$g_{ij} = \sum_k a_{jk} B'_k(\tilde{x}_i) \quad (6)$$

This is again a separable problem leading, for each i , to

$$u_i = \sum_j r_{ij} g_{ij} / \sum_j g_{ij}^2. \quad (7)$$

Derivatives of cubic B -splines are easily computed by combining quadratic B -splines (using the same knots), differences of the coefficients and a correction for the knot distance.

Now we have the building blocks for an iterative algorithm:

- Fit the P -spline using the current estimates \tilde{x} .
- Update \tilde{x} .
- Repeat until convergence.

The scale and location of x is arbitrary, so we have to choose a normalization. Our choice is to scale and shift it after each iteration, so that the minimum is 0 and the maximum is 1.

The iterative procedure is straightforward, but the crucial step is the choice of the starting values for \tilde{x} . We have experimented with two approaches. One is to compute the singular value decomposition of Y and use the singular vector connected to the largest singular value (after proper normalization). In simulations this approach gave mixed results. An alternative is to use a random start. This seems to work well, but tens of trials may be needed. The random start vector minimizing the final S^* is chosen.

Here, the role of the penalty with smoothing parameter λ is minor, in the sense that it prevents singularities when fitting the splines. Without the penalty, an unfortunate choice of \tilde{x} might lead to missing support for one or more B -splines.

3 Some results

Figure 1 shows the results of a simulation with eight curves in 15 hypothetical environments. Some of the simulated curves are almost linear, others show curvature. We permuted the rows in the originally simulated data set. A set of 50 random starts was tried. The knot distance for the B -spline basis is 0.05 and $\lambda = 1$. Convergence is linear and not too slow: after 50 iterations the size of the updates in u is of the order of 10^{-5} .

Gregorius and Namkoong (1986) use a small data set with five environments and information about the stem strength of six different types of pines. The results are presented in Figure 2. In these data the environments

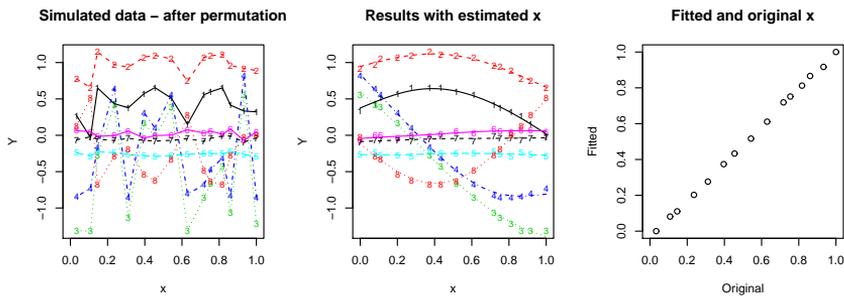


FIGURE 1. Simulated data with permuted x -axis (left), results from latent curve modeling (middle), original versus fitted x (right).

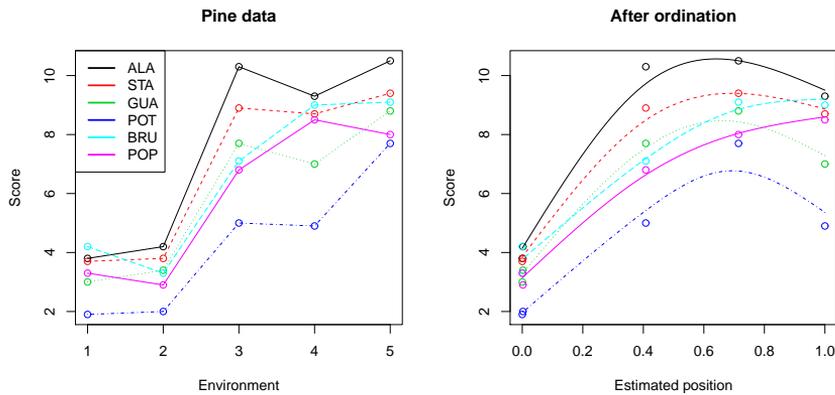


FIGURE 2. Data from Gregorius and Namkoong (1986). Original data (left), results with ordinated environments (right).

were already quite ordered, but the principle of the method is illustrated nevertheless.

Additionally we applied the technique to well-known textbook data from the plant breeding literature (Kleinhofs et al. 1993). A double haploid barley cross with 150 lines has been evaluated in 16 different environments and years. First analyses show promising results. In order to provide more guidelines for breeders for choosing well performing genotypes we suggest to estimate the relative performance of the genotypes. This second step can be done using performance measurement based on expectiles as suggested by Schnabel and Eilers in 2009.

4 Conclusion

By using smooth latent curves to describe an unknown environmental gradient we propose a new approach to model genotype by environment interaction in plant breeding trials. The order of the environments can be unknown when they are ordered along a latent gradient. The result of our proposed method is an order of the environments which was initially unknown. These complete data can be used in further analysis to gain more insight about the relationship of the phenotypic trait and the environmental conditions. Additionally characteristics of the fitted curves might be associated with the genetic background of the plants. The model can also be extended to accommodate missing data through a weighting scheme. Results will be reported elsewhere.

References

- Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.
- Fisher, R.A. and MacKenzie, W.A. (1923). Studies in variation II. The manurial response in different potato varieties. *Journal of Agricultural Science*, **13**, 311–320.
- Gregorius, H-R. and Namkoong, G. (1986). Joint analysis of genotypic and environmental effects. *Theoretical and Applied Genetics*, **72**, 413–422.
- Kleinhofs, A., Kilian, A., Maroof, M.S., Biyashev, R., Hayes, P., Chen, F., Lapitan, N., Fenwick, A., Blake, T., Kanazin, V., Ananiev, E., Dahleen, L., Kudrna, D., Bollinger, J., Knapp, S., Liu, B., Sorrells, M., Heun, M., Franckowiak, J., Hoffman, D., Skadse, R., and Steffenson, B. (1993). A molecular, isozyme, and morphological map of the barley (*hordeum vulgare*) genome. *Theoretical and Applied Genetics*, **86**, 705–712.
- Schnabel, S.K. and Eilers, P.H.C. (2009). An analysis of life expectancy and economic production using expectile frontier zones. *Demographic Research*, **21**, 109–134.

Modeling flow in gas transmission networks using shape-constrained expectile regression

Fabian Sobotka¹, Radoslava Mirkov², Benjamin Hofner³, Paul H. C. Eilers⁴, Thomas Kneib⁵

¹ Department of Mathematics, Carl von Ossietzky University Oldenburg, 26111 Oldenburg, Germany, fabian.sobotka@uni-oldenburg.de

² Department of Mathematics, Humboldt-University Berlin, 12489 Berlin, Germany, mirkov@math.hu-berlin.de

³ Institute for Medical Informatics, Biometrics and Epidemiology, Friedrich Alexander University Erlangen-Nürnberg, 91054 Erlangen, Germany, benjamin.hofner@imbe.med.uni-erlangen.de

⁴ Erasmus Medical Center, University Rotterdam, 3015 GE Rotterdam, Netherlands, p.eilers@erasmusmc.nl

⁵ Department of Economics, Georg August University Göttingen, 37073 Göttingen, Germany, tkneib@uni-goettingen.de

Abstract: The flow of natural gas within a gas transmission network is studied with the aim to optimise such networks. The analysis of real data provides a deeper insight into the behavior of gas in- and outflow. A geoaddivitive model for describing the dependence between the maximum daily gas flow and the temperature on network exits is proposed. Semiparametric expectile regression provides the possibility to model the upper tail of the response distribution while accounting for the spatial correlation between different exits. The effect of the temperature is modeled with shape constraints to include knowledge about gas load profiles and to allow for a realistic prediction. Estimates based on least asymmetrically weighted squares (LAWS) and boosting are presented. The forecast of gas loads for very low temperatures is included and the application of the obtained results is discussed.

Keywords: Expectiles; P-splines; Semiparametric Regression; Gas Flow; Boosting.

1 Introduction

Expectiles are a great way to study trends *and* variation, skewness, etc. of observed data. In this paper we will illustrate that by modeling the flow of gas measured at the exits of a gas transmission network. We continue the work of Friedl et al (2011) who used semiparametric mean regression to analyse gas consumption subject to temperature changes for single exits of the gas network. However, the upper tail of the conditional distribution of the gas flow is of more importance. It helps to generate extreme

scenarios of gas consumption that are necessary to examine the capacities of the gas network. Expectile regression offers a flexible and easy way to model the tails of a distribution as it avoids a full specification of the error term distribution. The estimation of expectiles is based on minimising an asymmetrically weighted sum of squared residuals. A semiparametric least squares regression allows for a very flexible regression structure by including nonlinear effects of continuous covariates, random effects or spatial effects. These extensions often rely on penalised least squares or penalised likelihood estimation with quadratic penalties and are therefore natural partners for least squares estimation.

Our geoadaptive model comprises various parametric effects, a nonlinear shape-constrained function of the temperature and a continuous spatial effect based on the longitude and latitude of the exits.

2 Data Description

Data for this study were obtained from measuring stations within the German pipeline network operated by Open Grid Europe GmbH (OGE), one of the leading German gas transmission operators. It contains hourly gas flow for 238 network exits for the period between June 2009 and May 2010. Mean daily temperatures from the corresponding weather stations are also provided. Additionally, we distinguish several exit types. Typical exits in such networks are public utilities, industrial and areal consumers, as well as exits on border and market crossings. Continuous geographic coordinates, i.e., longitude and latitude for every node are also included.

We study the dependence of gas loads on air temperature, exit type and the geographic location of exits within the network, simultaneously on all exits along the pipelines of the gas transmission network. Since we want to maximise the transportation capacity through the pipelines, we concentrate on the daily maximum flows $y_{i,k}^{max}$, $i = 1, \dots, n$ ($n = 365$), at the exits $k = 1, \dots, 238$ in the network. We note here that in this study we observe the so-called H-network, which denotes the network with high Wobbe Index (a measure for the heating value of the gas).

In what follows, we study the daily maximum flows standardised separately for each exit

$$y_{i,k} = \frac{y_{i,k}^{max} - \bar{y}_k}{\hat{\sigma}(y_k)}$$

in order to obtain comparable response values. As we are interested in the upper tail of the conditional distribution of the response, a mean regression is not sufficient. However, a quantile or expectile regression can be a sensible estimate for the quantity of interest.

3 Expectile Regression

The results of an expectile regression can be acquired by computing the least asymmetrically weighted sum of the squared residuals (LAWS) analogue to a quantile regression that minimises the asymmetrically weighted absolute values of the residuals. LAWS minimises

$$S = \sum_{i=1}^n w_\tau(y_i)(y_i - \mu_i(\tau))^2$$

with weights

$$w_\tau(y_i) = \begin{cases} \tau & \text{if } y_i > \mu_i(\tau) \\ 1 - \tau & \text{if } y_i < \mu_i(\tau) \end{cases}$$

where y_i is a continuous response and $\mu_i(\tau)$ is the estimated expectile for different values of the asymmetry parameter $\tau \in (0, 1)$. Hence the computation of expectile regression is very easy, since it avoids the non-differentiable absolute value criterion that is used to estimate quantiles. In further comparison, expectiles lack the intuitive interpretation of quantiles. While the quantile of a random variable Z immediately depicts the amount of probability that lies below it, the τ -expectile $\mu(\tau)$ can only be defined implicitly:

$$\tau = \frac{\int_{-\infty}^{\mu(\tau)} |z - \mu(\tau)| f(z) dz}{\int_{-\infty}^{\infty} |z - \mu(\tau)| f(z) dz} = \frac{G(\mu(\tau)) - \mu(\tau)F(\mu(\tau))}{2(G(\mu(\tau)) - \mu(\tau)F(\mu(\tau))) + (\mu(\tau) - \mu(0.5))}$$

where $G(m) = \int_{-\infty}^m z f(z) dz$ and $G(\infty) = \mu(0.5)$ is the expectation of Z . In addition to the computational advantages of expectiles, one can build additive models that contain different kinds of effects. We portray these effects by design matrices $B^{(j)}$ and assign a vector of regression coefficients β_j to each effect. We can then create the following additive expectile regression model:

$$\mu(\tau) = 1\beta_0 + X\beta_1 + B^{(2)}\beta_2 + \dots + B^{(r)}\beta_r + \varepsilon_\tau.$$

For continuous univariate covariates, smooth expectile curves can be fitted using penalised splines (see Schnabel and Eilers, 2009). Additionally the model can include spatial effects based on either Markov random fields, tensor product splines or Kriging (see Sobotka and Kneib, 2010). The smoothing can be induced by a quadratic penalty on the regression coefficients:

$$pen(\beta_{j,\tau}) = \lambda_j \beta_{j,\tau}^t K_j \beta_{j,\tau}$$

with adaptable smoothing parameter λ and penalty matrix K .

As in the case of modeling gas flow we are interested mainly in extreme expectiles describing the upper tail, we need to clarify the possible interpretations of a single expectile curve. An expectile can be related to the risk measure expected shortfall (ES) as for a random variable Z holds

$$\begin{aligned}
ES_p(t) &= E(Z(t)|Z(t) > \tilde{z}_p(t)) \\
&= \left(1 + \frac{\tau}{(1-2\tau)^p}\right) \mu_\tau(t) - \frac{\tau}{(1-2\tau)^p} \mu_{0.5}(t)
\end{aligned}$$

with the τ -expectile μ_τ and $p = F_t(\mu_\tau(t))$ (see Taylor, 2008). As shown the expected shortfall is the conditional mean above the p -quantile and provides more reliable information about extreme observations than the quantile. This allows us to explicitly compute a risk bound for the maximum daily gas flow.

4 Estimating and Forecasting Gas Flow

The model explaining the standardised maximum gas flows includes parametric effects X indicating a weekend day and the types of the exits. Further a P -spline basis is used to model the effect of the local daily mean temperature \mathbf{t} . Finally, in a spatial effect the longitude \mathbf{u} and latitude \mathbf{v} of each exit are included by a Kriging basis. The knots for this basis are chosen as a subset k_1, \dots, k_K from the covariate observations $(u, v)_1, \dots, (u, v)_{238}$. The basis evaluation is defined by Matérn correlation functions like

$$B_k(r, \phi) = \exp(-|r/\phi|)(1 + |r/\phi|)$$

with the Euclidean distance $r = \|k - x\|$ and a fixed $\phi \propto \max_{i,j}(\|k_i - k_j\|)$. The penalty matrix $K = (B_{k_i}(\|k_i - k_j\|))_{i,j}$ then comprises the evaluated distances between the knots.

The analysed model has the form

$$\mu_\tau(y) = 1\beta_{0,\tau} + X\beta_{1,\tau} + B(\mathbf{t})\beta_{2,\tau} + B(\mathbf{u}, \mathbf{v})\beta_{3,\tau} + \varepsilon_\tau.$$

Due to the physical properties of the gas and the observed behaviour of industries and private households regarding gas consumption, certain restraints can be made to the effect of the temperature on the gas flow. First of all, the demand for gas will generally decrease with higher temperatures. However, even on very warm days there might be a minimum consumption of gas. Even for very low temperatures the loads of gas will not exceed a certain capacity. Hence, we include an additional iteration in the estimation process, restricting the regression coefficients of the P -spline basis to the realistic behaviour. This is achieved similar as in Bollaerts (2006) by an additional iteration within the estimating. First, the shape-constraints are defined as difference penalties on the spline coefficients. After an estimation step those spline basis elements are identified that do not follow the restrictions. Here, the penalty is invoked and the estimation is repeated. Preliminary results are shown in Figure 1. In the results one can easily see the

gain of information through expectile regression. The variance and skewness clearly changes with the temperature. Hence, the information about extreme observations is more accurate than with mean regression. Also the spatial effect is non-informative for the mean while a difference between the north-eastern exits and the south-western exits is clearly shown in the upper tail through the 0.99-expectile.

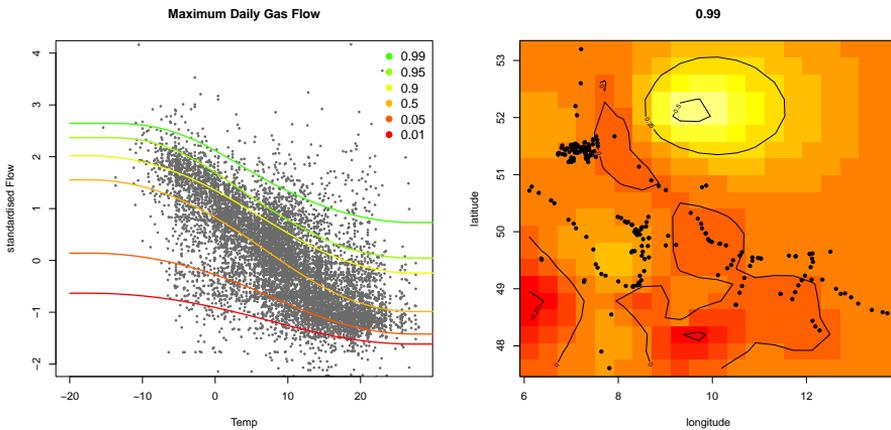


FIGURE 1. Left: Standardised flow for temperature interval. Shape-constrained effects for 6 asymmetries τ . Right: Spatial effect of the 0.99-expectile for 238 exits marked as black points.

The analyses are performed using the R-package “expectreg” (Sobotka, Schnabel, Schulze Waltrup, 2012).

Acknowledgments: The authors are grateful to Prof. Werner Römisch for the possibility to work on this interesting project. Thanks to OGE for supporting the project and providing real data as well as the relevant questions we investigate in this paper. Financial support from the German Research Foundation (DFG) grant KN 922/4-1 is gratefully acknowledged.

References

- Bollaerts, K., P. H. C. Eilers and I. van Mechelen (2006). Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, **59**, 451–469.
- Friedl, H., R. Mirkov, and A. Steinkamp (2011). Modeling and Forecasting Gas Flow on Exits of Gas Transmission Networks. *To appear in: International Statistical Review Special Issue on Energy Statistics*.

- Hofner, B. and J. Müller and T. Hothorn (2011). Monotonicity-Constrained Species Distribution Models. *Ecology*, **92**, 1895-1901.
- Schnabel, S. and P. Eilers (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis* **53**, 4168-4177.
- Sobotka, F. and T. Kneib (2010). Geoadditive expectile regression. *Computational Statistics and Data Analysis*, doi: 10.1016/j.csda.2010.11.015.
- Sobotka, F., S. Schnabel and L. Schulze Waltrup (2012). *expectreg: Expectile and Quantile Regression*. R package version 0.30.
- Taylor, J. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics* **6**(2), 231–252.

Analysing living conditions in Austria by a Bayesian mixed data model

Regina Tüchler¹, Helga Wagner²

¹ Austrian Federal Economic Chamber, Department of Statistics, Vienna, Austria

² Department of Applied Statistics and Econometrics, Johannes Kepler University Linz, Austria

E-mail for correspondence: helga.wagner@jku.at

Abstract: As multidimensional data of different type are nowadays routinely collected in economic and social surveys an increasing need for mixed data model approaches occurs. In this paper we analyse data on living conditions and include two outcomes, the continuous household income and the bivariate material deprivation indicator, in a joint regression model. Variable selection techniques enable us to derive the importance of covariates.

Keywords: Bayesian Mixed Model; Data Augmentation; Variable Selection; Living Conditions.

1 Household income and material deprivation as indicators for living conditions

Indicators on income and living conditions become more and more important for European politics. Nowadays it is commonly agreed that complementary indicators to the GDP should be looked at to measure well-being of societies. Therefore we face an increasing need for analyses of indicators from social statistics, of dependencies between these indicators and of driving factors. Our paper meets these needs and combines the monetary aspect of living conditions via the household income - the money each household has to make its living from, and the material aspect via the so-called material deprivation indicator. A household lives with material deprivation if the members are not capable to meet certain predefined needs like e.g. TV, phone, holiday away from home. We combine the continuous outcome variable household income and the binary outcome variable material deprivation in a mixed data model and analyse the dependence on socio-demographic factors like e.g. the age or activity status of the main-income earner, the household type and migration status. We derive the importance of these explanatory variables by introducing variable selection. The data set comes from the European statistics on income and living conditions

(EU-SILC) 2009 and contains 3 704 Austrian households (BMASK, 2011).

2 Model

Let $\mathbf{y}_i = (y_i^b, y_i^n)$ denote the bivariate response for subject $i = 1, \dots, N$ where y_i^b denotes the binary outcome, material deprivation, and y_i^n the normal outcome, log income. For each outcome component we specify a marginal regression model with linear predictor specified as

$$\eta_i^k = \mathbf{x}_i \boldsymbol{\beta}^k,$$

where \mathbf{x}_i is a design vector of dimension $1 \times d$ and $\boldsymbol{\beta}^k$ is the vector of regression coefficients for component k . For the binary outcome we use the logit link function

$$\mu_i^b = \frac{\exp(\eta_i^b)}{1 + \exp(\eta_i^b)}, \quad (1)$$

whereas for the normal component y_i^n we use the identical link

$$\mu_i^n = \eta_i^n, \quad (2)$$

and assume a constant variance, i.e. $y_i^n \sim \mathcal{N}(\mu_i^n, \sigma^2)$.

To model dependence between binary and normal response we rely on a representation of the logit model as a linear Gaussian model in latent auxiliary variables. The logit regression model results as marginal model from a linear regression model for a latent variable u_i as

$$\begin{aligned} u_i &= \mathbf{x}_i \boldsymbol{\beta}^b + \epsilon_i, \\ y_i^b &= I_{(0, \infty)}(u_i). \end{aligned}$$

Here the error ϵ_i follows the the standard logistic distribution, which can be represented as a scale mixture of 6 normal distributions with fixed scale parameters s_r and component weights w_r , see Frühwirth-Schnatter and Frühwirth (2010). Introducing the mixture component r_i the model for auxiliary variables \tilde{y}_i^b is given as

$$\tilde{y}_i^b = \mathbf{x}_i \boldsymbol{\beta}^b + \tilde{\epsilon}_i^b, \quad \tilde{\epsilon}_i^b \sim \mathcal{N}(0, s_{r_i}^2).$$

We specify a bivariate distribution for \tilde{y}_i^b and the normal component y_i^n as

$$\tilde{\mathbf{y}}_i = \begin{pmatrix} \tilde{y}_i^b \\ y_i^n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i \boldsymbol{\beta}^b \\ \mathbf{x}_i \boldsymbol{\beta}^n \end{pmatrix} + \tilde{\boldsymbol{\epsilon}}_i, \quad \tilde{\boldsymbol{\epsilon}}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \begin{pmatrix} s_{r_i}^2 & s_{r_i} \sigma \rho \\ s_{r_i} \sigma \rho & \sigma^2 \end{pmatrix}$ and ρ is the correlation between the binary and the normal response. A similar approach for modelling data of mixed type

was taken in Wagner and Tüchler (2010) for panel data, where marginal random effects models are joined by correlated random effects.

Bayesian model formulation is completed by specifying a prior distribution for the model parameters. For the full model we consider a prior of the structure $p(\boldsymbol{\beta}, \theta, \rho) = p(\boldsymbol{\beta})p(\theta)p(\rho)$ with $\theta = \ln \sigma$. We use a uniform prior on $[-1, 1]$ for ρ and a normal prior for θ . To incorporate variable selection we specify a spike and slab prior with Dirac spikes and normal slabs for the elements of $\boldsymbol{\beta}$, see e.g. Wagner, H. and C. Duller (2012).

3 MCMC

Let \mathbf{y}_i denote the stacked vector of all \mathbf{y}_i , $\boldsymbol{\beta} = (\boldsymbol{\beta}^b, \boldsymbol{\beta}^n)$ the vector of regression coefficients and $\mathbf{r} = (r_1, \dots, r_n)$ the vector of component indicators. To estimate the model parameters we use an MCMC scheme with the following steps:

- (I) Sample the component indicators \mathbf{r} from $p(\mathbf{r}|\boldsymbol{\beta}, \rho, \theta, \tilde{\mathbf{y}})$.
- (II) Sample (ρ, θ) and the auxiliary variables $\tilde{\mathbf{y}}^b$.
- (III) Perform variable selection and sample the the non-zero elements of $\boldsymbol{\beta}$ from its normal posterior.

Sampling the component indicators. The posterior probabilities $P(r_i = r), r = 1, \dots, 6$ are given as

$$P(r_i = r|\boldsymbol{\beta}, \rho, \theta, \tilde{\mathbf{y}}) \propto \phi\left(\frac{\tilde{y}_i^b - m_{i,r}}{s_{i,r}}\right) \pi_r,$$

where $m_{i,r}$ and $s_{i,r}$ are conditional mean and standard deviation of \tilde{y}_i^b conditional on r_i and the normal response y_i^n .

Sampling the variance and correlation parameter. (ρ, θ) are sampled jointly from the conditional posterior marginalized over the auxiliary variables $\tilde{\mathbf{y}}^b$ given as

$$p(\rho, \theta|\mathbf{y}, \boldsymbol{\beta}) \propto p(\rho)p(\theta) \prod_{i=1}^n p(\mathbf{y}_i|\boldsymbol{\beta}, \rho, \theta).$$

As this posterior is not of closed form we use a MH-step with a tailored proposal to sample (ρ, θ) . The proposal is a bivariate Student t-distribution with 10 degrees of freedom. The mean of this t-distribution is the ML estimate after 10 maximizing iterations of the likelihood $\prod_{i=1}^n p(\mathbf{y}_i|\boldsymbol{\beta}, \theta, \rho)$ and the variance-covariance parameter is the inverse Hessian at this point.

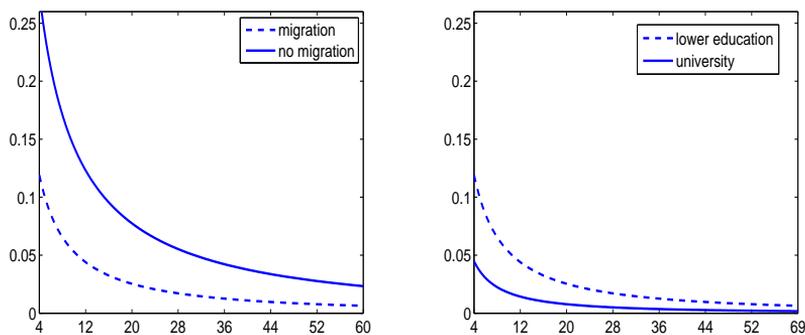


FIGURE 1. Probability of material deprivation conditional on income (in 1000 Euros) for different households.

Sampling the auxiliary variables. The auxiliary variables \tilde{y}_i^b are sampled from $p(\tilde{y}_i^b | r_i, \rho, \theta, \beta, \mathbf{y}_i)$, which is the normal distribution $\mathcal{N}(m_{i,r}, s_{i,r}^2)$ truncated to $(-\infty, 0)$ for $y_i^b = 0$ and to $(0, \infty)$ for $y_i^b = 1$.

4 Results

In Table 1 we give estimates of the mean parameters and the corresponding inclusion probabilities for the two responses. It turned out that the variables activity status, education and migration of the main income earner play an important role in explaining both income as well as material deprivation whereas the inclusion probabilities for the other variables are smaller at least for part of the categories or one response. Households with the main-income earner working only on a part-time basis have less income and higher risk for material deprivation. Naturally effects are even stronger for households with a main-income-earner who is unemployed or out-of-labour force. The higher the level of education the bigger is the estimated effect for the income and the smaller for the material deprivation response. Households with a main-income-earner who currently has or once had a non-EU/EFTA citizenship have less income and a higher risk for material deprivation.

The correlation parameter ρ is equal to -0.25 and indicates that households with higher income are less likely in a situation of material deprivation and vice versa. Joint modelling of both responses allows to determine the conditional probability of material deprivation as a function of the log-income, which is given by

$$p(y_i^b | y_i^n) = \sum \Phi\left(\frac{m_{i,r}}{s_{i,r}}\right) \pi_r.$$

TABLE 1. Estimates for the mean and probabilities to be unrestricted.

variable	material deprivation		log(earnings)	
	$\hat{\beta}$	$\Pr(\delta_j = 1)$	$\hat{\beta}$	$\Pr(\delta_j = 1)$
intercept	-3.43	–	9.82	–
gender (base: male)	0.11	0.34	-0.00	0.05
age (cent. at median 43 y.)	-0.00	0.03	0.01	1.00
age ²	-0.00	0.00	0.00	0.00
activity status (base: full-time)				
part-time	1.02	1.00	-0.26	1.00
unemployed	2.29	1.00	-0.45	1.00
out-of-labour	1.78	1.00	-0.69	1.00
education (base: lower)				
medium	-0.40	0.78	0.15	1.00
higher	-1.55	1.00	0.29	1.00
university	-1.73	1.00	0.45	1.00
migration (base: no migration)	1.40	1.00	-0.28	1.00
type of household (base: single)				
2 adults/no children	-0.17	0.42	0.19	1.00
single-parent	0.19	0.45	-0.02	0.29
2 adults/1 or 2 Children	-0.03	0.18	0.00	0.04
2 adults/+3 children	-0.01	0.21	-0.20	1.00
other	-0.08	0.26	0.11	1.00
type of building (base: single-family)				
2 families	0.02	0.20	-0.00	0.02
3 to 9 families	0.69	0.92	-0.01	0.14
+10 families	0.94	0.99	-0.09	1.00
other	-0.12	0.41	-0.01	0.09
population density (base: high)				
medium	-0.05	0.22	0.00	0.04
low	-0.40	0.75	-0.03	0.54

In Figure 1 we compare households with all covariates but one equal to the baseline value. The left plot compares households with the main-income-earner having no migration background with households where such a mi-

gration background exists. For the median household income of 21 196 Euros the probability for material deprivation is 2.4% for households with no migration background and 7.3% for those with migration background. For households with an income equal to the first decile of 11 513 Euros, the effect of migration is even bigger with a probability of 4.6% without and 12.8% with migration background, respectively. The right plot depicts the probability for material deprivation conditional on fixed incomes for baseline households with lower education and with university degree. For the median income the rate of households in material deprivation equals 0.7% percent if the main-income earner has an university degree and 2.4% percent if he or she has only a lower education.

5 Conclusions

We propose a joint regression model for income and material deprivation which is a bivariate response with components of mixed discrete and continuous type. The mixed data approach yields deeper insights and gives results that could not be obtained by separate modelling of the different data types.

Bayesian estimation by MCMC is straightforward and allows Gibbs sampling for most of the parameters. Additionally Bayesian variable selection and model averaging can easily be incorporated by using spike and slab priors for regression effects. The method applies to a wide range of data from applied sciences, like economics, social science and others.

References

- BMASK (2011). *Armutsgefährdung und Lebensbedingungen in Österreich, Ergebnisse aus EU-SILC 2009. Studie der Statistik Austria im Auftrag des BMASK*. Wien: Sozialpolitische Studienreihe, Bd 5.
- 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.
- Wagner, H. and C. Duller(2012). *Bayesian model selection for logistic regression models with random intercept*. Computational Statistics and Data Analysis **56**, 1256–1274.
- Wagner, H. and R. Tüchler (2010). *Bayesian estimation of random effects models for multivariate responses of mixed data*. Computational Statistics and Data Analysis **54**, 1206–1218.

Comparison of prediction methods for mixed logistic regression

Karin Ayumi Tamura ¹, Viviana Giampaoli¹

¹ Departamento de Estatística, Universidade de São Paulo, São Paulo, Brazil

E-mail for correspondence: karinat@ime.usp.br

Abstract: This article considers a new methodology to predict the random effects for new groups of a mixed logistic regression, by using nonparametric regression. The approach, named as nonparametric prediction method (NPPM), considers the relationship among the random effect and the covariates aggregated in the group level. The aim of NPPM is to predict the response variable of a mixed logistic regression in the observation level. In a context of large data base, in general, the empirical best predictor (EBP) requires a huge computational effort. In our experiments, NPPM drastically reduced the time processing and presented similar results in relation to EBP. Both prediction methods were applied in a data set of a telecommunication company, and were compared with traditional logistic regression.

Keywords: mixed logistic regression; outcome prediction; nonparametric regression; empirical best predictor.

1 Introduction

We address the problem of predicting the outcome of future observations by using mixed logistic regression, which is a particular case of the multi-level model with two levels. Mixed models are adequate for grouped data, providing more reliable estimates by using random coefficients which take into account the variability between groups.

Consider the following application. A mobile company has corporative customers. Associated to each customer, there are employees who use the mobile to work. Suppose that we want to sell a mobile service. The goal is to predict the probability of purchasing by a new customer, based on a mixed logistic regression modeled in a previous period.

This paper presents a novel approach to predict the random effects of future (or new) groups by using nonparametric regression. This method was studied in order to reduce the computational effort, since the EBP method, presented in Tamura and Giampaoli (2011), involved a huge time processing to solve the multi-dimensional integrals.

By assuming independence among observations and ignoring the random effects, the traditional model has been commonly used (palliatively) for

prediction in a future period. In this way, we compared both NPPM and EBP with the traditional logistic regression.

2 Related Works

2.1 Mixed Logistic Regression

Let y_{ij} be the binary outcome corresponding to j -th observation in the i -th cluster, with $i = 1, \dots, q$ and $j = 1, \dots, n_i$. The mixed logistic regression considers that, conditional on α_i , y_{ij} 's are independent Bernoulli, with

$$\text{logit}(P(y_{ij} = 1|\alpha_i)) = \text{logit}(p_{ij}) = \mathbf{x}_{ij}^t \boldsymbol{\beta} + \mathbf{z}_{ij}^t \boldsymbol{\alpha}_i, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects ($p \times 1$) and $\boldsymbol{\alpha}_i$ is a vector of the random effects ($k \times 1$) of the i -th cluster. The vector \mathbf{x}_{ij}^t of known independent variables ($1 \times p$) is associated with $\boldsymbol{\beta}$ and \mathbf{z}_{ij}^t is a vector of known independent variables ($1 \times k$) associated with $\boldsymbol{\alpha}_i$, which are defined by $\mathbf{x}_{ij}^t = (1, x_{1ij}, x_{2ij}, \dots, x_{(p-1)ij})$ and $\mathbf{z}_{ij}^t = (1, z_{1ij}, z_{2ij}, \dots, z_{(k-1)ij})$. This model assumes that $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_q$ 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. This is a special case of generalized linear mixed models, in which the conditional exponential family is Bernoulli, and the link function is logit.

2.2 Empirical Best Predictor

Based on Jiang and Lahiri (2006) and Tamura and Giampaoli (2010), Tamura and Giampaoli (2011) presented a method for the mixed logistic regression to predict the response variable of an observation in a new cluster, considering k random effects.

Lets define $\tilde{\zeta}(\boldsymbol{\beta}, \boldsymbol{\alpha}_i) = \tilde{p}_{ij}$. Since $\tilde{\zeta}$ is usually unknown, we replace $\tilde{\zeta}$ by an estimator $\hat{\zeta}$. Then, for the prediction of future observations, $\hat{\zeta} = \zeta(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}\xi) = \hat{p}_{ij}$ is denominated EBP. Based on (1), EPB is given by

$$\hat{p}_{ij} = \frac{\exp(\mathbf{x}_{ij}^t \hat{\boldsymbol{\beta}}) \int \dots \int_{\xi_1}^{\xi_k} \frac{\exp((y_i + 1)\mathbf{z}_{ij}^t \hat{\boldsymbol{\alpha}})}{1 + \exp(\mathbf{x}_{ij}^t \hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^t \hat{\boldsymbol{\alpha}})} \cdot \prod_{l=1}^{n_i} \frac{1}{1 + \exp(\mathbf{x}_{il}^t \hat{\boldsymbol{\beta}} + \mathbf{z}_{il}^t \hat{\boldsymbol{\alpha}})} f(\xi_1, \dots, \xi_k) d\xi_1 \dots d\xi_k}{\int \dots \int_{\xi_1}^{\xi_k} \exp(y_i \mathbf{z}_{ij}^t \hat{\boldsymbol{\alpha}}) \cdot \prod_{l=1}^{n_i} \frac{1}{1 + \exp(\mathbf{x}_{il}^t \hat{\boldsymbol{\beta}} + \mathbf{z}_{il}^t \hat{\boldsymbol{\alpha}})} f(\xi_1, \dots, \xi_k) d\xi_1 \dots d\xi_k},$$

where $\mathbf{z}_{ij}^t \hat{\boldsymbol{\alpha}} = (\xi_1 \hat{v}_{11} + \dots + \xi_k \hat{v}_{1k}) + \dots + (\xi_1 \hat{v}_{k1} + \dots + \xi_k \hat{v}_{kk}) \mathbf{z}_{(k-1)ij}^t \hat{v}$. are the components of $\hat{\boldsymbol{\Sigma}}$ and $\xi_m \sim \mathcal{N}(0, 1)$, with $m = 1, \dots, k$. Note that, since we do not know y_i , we assumed that $y_i = n_i/2$. More details, see Tamura and Giampaoli (2011).

3 Nonparametric Prediction Method

By the assumption of model (1), $\boldsymbol{\alpha}_i$ follows a multivariate normal distribution. Thus, each marginal of random effects α_{mi} , with $m = 1, \dots, k$ and $i = 1, \dots, q$, follows a univariate normal distribution.

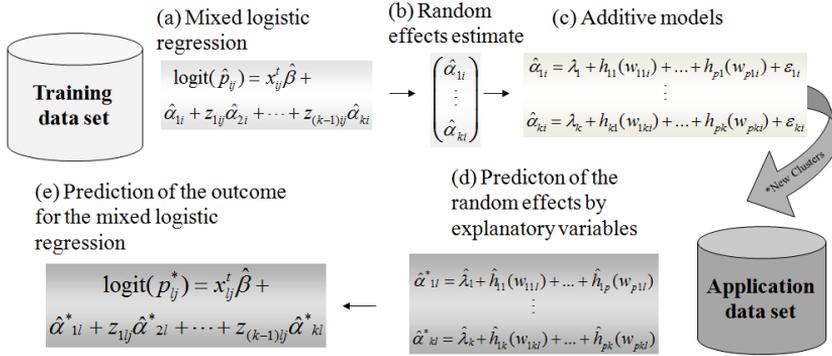


FIGURE 1. A summary of the nonparametric prediction method.

Nonparametric regression can be used as an alternative to linear regression when the empirical data do not present normality. Then, in order to predict the random effects for new groups, we considered each vector of random effect (α_m) as the response variable of a nonparametric model. We worked with additive models which consider an arbitrary univariate function $h(\cdot)$ for each predictor. For more details, see Hastie and Tibshirani (1990).

In this case, we propose the use of spline smoother, which can be easily accomplished to select the smoothing parameters, by using **step.gam** procedure in the R package.

Figure 1 summarizes the steps of NPPM. **Step (a)** In the training data set, fit the mixed logistic model. **Step (b)** Extract the matrix of random effects ($q \times k$) and separate each vector of random effect ($q \times 1$). In this step, aggregate all the covariates in the group level by using some aggregation functions, i.e., $\mathbf{w}_i^t = (\mathbf{x}_i^t, \mathbf{z}_i^t)$. **Step (c)** For the m -th random effect, fit an additive model, where \mathbf{w}_i , with $i = 1, \dots, q$, is the vector of known covariates ($p \times 1$), $h(\cdot)$ is a smooth function. The residuals ε_{mi} 's are independent with $E(\varepsilon_{mi}) = 0$ and $Var(\varepsilon_{mi}) = \sigma_m^2$. **Step (d)** In application data set, predict the m -th new random effect (α_m^*) based on the explanatory variables and the estimate parameters provided by (c). Note that the index for the new groups was replaced by l . **Step (e)** Insert the predicted random effects in the mixed logistic model function, by considering the estimate values of the fixed effects computed in (a). Finally, obtain the outcome prediction for the mixed logistic regression.

Our goal is to model the dependence of the response variable in relation to the independent variables. Thus, based on the covariates, we predicted the random effects for the new groups, which in turn are inserted within the logit function of the mixed model. This function provides the predicted probability of the target event. By using the models described in (c), in Figure 1, the proposed methodology relaxes the assumption of the model (1) (that the average α_{mi} is zero), and we consideres $\alpha_{mi} =$

$\lambda + h_{1m}(w_{1mi}) + \dots + h_{pm}(w_{pmi}) + \varepsilon_{mi}$, with $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_k^2)$. The key idea is that random effects can be explained by including additional information provided by the individual characteristics of each group.

4 Application

We illustrated the usage of the prediction methods by addressing a problem in which a mobile company has corporative customers. Each customer represents the group level and the employee who use the mobile to work represents the unit within the group. The aim is predict the probability of purchasing SMS monthly package for the employees who do not have the service.

The independent variables, in the group and unit levels, were observed from a reference month to back, and the outcome, one month afterwards. The binary outcome of the model was defined as: 1 - the mobile acquired the SMS monthly package and 0 - otherwise.

We considered two reference months. The first period was considered as training data set, with 19,473 mobiles nested in 458 clients, in which we modelled the data. After six months, an application data set was considered for a prediction in a future period, with 21,050 terminals nested in 657 customers.

We considered the traditional (2) and mixed (3) logistic models given by

$$\text{logit}[P(y_{ij} = 1)] = \beta_0 + \mathbf{x}_{ij}^t \boldsymbol{\beta} \quad (2)$$

and

$$\text{logit}[P(y_{ij} = 1) | \boldsymbol{\alpha}_i] = \beta_0 + \mathbf{x}_{ij}^t \boldsymbol{\beta} + \alpha_{1i} + z_{ij} \alpha_{2i}. \quad (3)$$

fitted by using Maximum Likelihood and Laplace Approximation, respectively. Model (3) considered two random effects, one random intercept and one random slope. The covariate z_{ij} , associated to the random slope, corresponds to the percentage of minutes used by the unit inside the group.

The covariates associated to the fixed effects are in Table 1. The same table presents the estimate parameters (est.), standard errors (s.e.) and p-values. The explanatory variables were categorized and the first category of each variable was considered as reference cell. Analyzing the fixed effects, in both models, they were significant (p-value < 0.10), unless the category “minutes of usage roaming > 10”. This dummy was not excluded from both models, because the p-values of the minutes of usage roaming was a significant variable (p-value < 0.10) in the significance factor test. The comparison of nested models used Chi-squared statistics, in which the aim was to test the exclusion of one factor by comparing with the final model (variables in Table 1). All the factors were significant in the traditional and mixed models (p-value < 0.10). In the traditional model, the covariate z_{ij} was tested as fixed effect, but it was not significant (p-value > 0.10).

TABLE 1. Estimate parameters, standard errors and p-values, for the traditional and mixed models.

Explanatory variables	Traditional model			Mixed model		
	est.	s.e.	p-value	est.	s.e.	p-value
(Intercept)	-4.90	0.09	<0.00	-9.31	0.47	<0.00
Qt. of SMS service per day =2	1.26	0.17	<0.00	0.59	0.23	0.01
Qt. of SMS service per day >2	1.68	0.29	<0.00	1.07	0.38	<0.00
Ind. of SMS service monthly in the past	2.34	0.31	<0.00	1.63	0.40	<0.00
Min. of usage roaming =]0,10]	0.70	0.18	<0.00	0.56	0.25	0.02
Min. of usage roaming >10	0.30	0.26	0.25	-0.04	0.35	0.91
Ind. of using e-mail	1.54	0.35	<0.00	-0.71	0.44	0.10
Ind. of using data package	0.79	0.29	0.01	1.97	0.43	<0.00
Random Effects s.d. (σ_1, σ_2)	-			(5.66, 2.41)		
Intraclass correlation	-			(-0.68)		

Abreviations: Qt.=quantity, Ind.=indicator, Min.=minutes and s.d=standard deviation.

We evaluated the predictive performance of the traditional and mixed models in the training and application data sets, by using AUC (Area Under the Curve) and KS (Komolgorov-Smirnov Statistics), presented in Table 2. More details about both measurements, see Fawcett (2005) and Conover (1999), respectively.

In the training data set, analyzing both performance measures, we observed that the mixed model outperformed the traditional model. In the application data set, we analyzed the performance prediction in three different perspectives: only former customers, only new customers and all customers (new and former). For the former customers, we used the estimate random effects available in the training data set. For the new customers, which represent 29% of the application data, we used EBP and NPPM for the mixed model. Both EBP and NPPM presented similar prediction performance and outperformed the traditional model. Finally, joining former and new customers (application data set), the mixed model demonstrated superiority in relation to the traditional model, indicating that the hierarchical structure of the data are suitable for prediction in a future period.

TABLE 2. Performance measures for the traditional and mixed models, in the training and the application data sets.

Performance Measures	Traditional model	Mixed model	EBP	NPPM
Training data set				
AUC	0.713	0.977		
KS	39.037	87.337		
Application Data set				
Former Customers				
AUC	0.737	0.819		
KS	46.467	58.000		
Application Data set				
New Customers				
AUC	0.589		0.687	0.672
KS	17.772		34.767	36.103
Application Data set				
All Customers (new and formers)				
AUC	0.715		0.785	0.779
KS	40.193		52.302	47.228

The computational gain of NPPM in relation to EPB is considerable in the context of huge data sets. For instance, in an ordinary computer, it took around 2 days to predict the outcome by using EBP method for 4,759 new

observations (belonging to 188 new groups), while for NPPM the time was less than 5 minutes.

5 Conclusions

This study has proposed NPPM, which is a novel method for predicting the response variable for new groups, by using logistic mixed models with k random effects. This methodology relaxed the assumption that the average of α_{mi} is zero, by considering a specific average for each group, without requiring a normal distribution for the empirical random effects. As future steps, the same methodology could be tested by using linear models, if the random effects follow a normal distribution. Moreover, the methodology considering regression models to predict the random effects could be applied to other distributions, different from the binary response.

Note that the proposed methodology can greatly reduce the computational effort, while maintaining the same level of prediction in comparison to the EBP method.

Acknowledgments: This work received partial financial support from FAPESP and CNPq.

References

- Conover, W. (1999). *Practical nonparametric statistics*. New York: Wiley.
- Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–74.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.
- Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1–96.
- Tamura, K.A. and Giampaoli, V. (2010). Prediction in multilevel logistic regression. *Communications in Statistics - Simulation and Computation*, **39(6)**, 1063–1076.
- Tamura, K.A. and 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, pp. 593–596.

Predictions from a Markov illness-death model: Application to dementia disease

Célia Touraine^{1,2}, Pierre Joly^{1,2}

¹ INSERM, ISPED, INSERM U897, Bordeaux, F-33000, France

² Univ. Bordeaux, ISPED, INSERM U897, Bordeaux, F-33000, France

E-mail for correspondence: celia.touraine@isped.u-bordeaux2.fr

Abstract: In multi-state models, it is common to model the effects of the covariates on the estimated transition intensities. However, it is sometimes more relevant in a clinical context to predict the health status of a patient at a given time-point. For example, given a set of values for prognostic factors of a patient, it may be useful to estimate the probability of a future event or the expected time without an event. In order to answer these questions we need to estimate quantities such as transition probabilities, cumulative incidence functions and life expectancies. The aim of this paper is to show how to estimate these quantities using the estimated regression parameters and the estimated transition intensities. As an illustration, we give some examples of predictions using data from a large cohort study on cognitive aging.

Keywords: illness-death model; transition probabilities; cumulative incidences; lifetime risks; life expectancies.

Introduction

In longitudinal studies, several events of interest may be recorded on the same individual. One approach to model such data uses multi-state models, which allow subjects to move among a finite number of states over time. We will restrict attention in this paper to the illness-death model (Andersen *et al.*, 1994), also known as disability model, which is widely used in the medical literature. In this model, subjects start out as healthy (state 0). Then, they may become ill (move to state 1) and die (move from state 1 to state 2), or, they may die without first becoming ill (move from state 0 to state 2). We consider here an uni-directional model which means that an ill subject cannot become healthy again (move back to state 0). Figure 1 depicts an example of uni-directional illness-death model.

A multi-state process is fully characterised through transition probabilities or transition intensities. Different assumptions can be made about the relationship between these transition rates and time. It is often assumed that the model is a Markov model where the Markov property implies that the future evolution of a subject depends only on their current state. A

hypothesis of homogeneity can also be made if one assumes that the transition intensities are constant over time. However, in many applications this assumption is too strong. In this work, we consider a non-homogeneous Markov uni-directional illness-death model.

The aim of this paper is to investigate the predictions that can be made using continuous estimations of the transition intensities and to apply these formulae to longitudinal data. We begin by expressing the transition probabilities in terms of the transition intensities. Secondly, we express the cumulative probability of illness, the cumulative probability of death prior illness, the lifetime risk of illness, and life expectancies in terms of the transition probabilities. Finally, we give some examples of predictions using data from a large cohort study on cognitive aging.

1 Predictions

Let s and t be two times. s denotes the time at which a prediction is made for a future time-point t ($t > s$).

1.1 Transition probabilities

Let $\{X(t)\}$ be a process. If $\{X(t)\}$ is Markovian, the future evolution of the process $\{X(t); t \geq s\}$ depends only on the history of the process through the present. Therefore, the transition probability between states i and j (between times s and t) is:

$$p_{ij}(s, t) = \mathbb{P}(X(t) = j | X(s) = i).$$

The transition intensity between states i and j , between times s and t , is by definition given by:

$$\alpha_{ij}(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{ij}(t, t + \Delta t)}{\Delta t}.$$

Let $A_{ij}(s, t)$ be the cumulative transition intensity function: $A_{ij}(s, t) = \int_s^t \alpha_{ij}(u) du$.

This section, based on Putter *et al.* (2007), is devoted to expressing the transition probabilities $p_{ij}(s, t)$ in terms of transition intensities $\alpha_{ij}(s, t)$ and cumulative transition intensities $A_{ij}(s, t)$. The following expressions of $p_{11}(s, t)$, $p_{00}(s, t)$, $p_{01}(s, t)$ can be obtained by solving the Kolmogorov differential equations (see Andersen *et al.*, 1993). We are going to define these expressions before deriving $p_{02}(s, t)$ as the sum of two probabilities depending on whether illness occurs before death or not.

Let us start with probabilities from state 1 *i.e.* probabilities relating to ill subjects at time s . For such subjects, the probability of staying in state 1 (*i.e.* of not dying) until time t is:

$$p_{11}(s, t) = e^{-A_{12}(s, t)}.$$

It follows that the probability of being dead at time t is:

$$p_{12}(s, t) = 1 - p_{11}(s, t).$$

Let's turn now to the probabilities relating to healthy subjects (in state 0) at time s . The probability of still being in state 0 at time t (*i.e.*, of not reaching states 1 or 2 before time t) is:

$$p_{00}(s, t) = e^{-A_{01}(s,t) - A_{02}(s,t)}.$$

The probability of being ill at time t is:

$$p_{01}(s, t) = \int_s^t p_{00}(s, u) \alpha_{01}(u) p_{11}(u, t) du.$$

Indeed, to be ill at time t a subject has to reach state 1 between s and u , where $s < u < t$, and stay in state 1 between u and t .

The probability of being dead at t is the sum of two probabilities since a subject in state 0 at time s can move either directly to state 2 before time t or via the illness state (state 1). Let $p_{02}^0(s, t)$ be the probability of moving directly from state 0 to state 2. We have:

$$p_{02}^0(s, t) = \int_s^t p_{00}(s, u) \alpha_{02}(u) du.$$

Let $p_{02}^1(s, t)$ be the probability of being dead at time t having been ill before. We have:

$$\begin{aligned} p_{02}^1(s, t) &= \int_s^t p_{00}(s, u) \alpha_{01}(u) \left(\int_u^t p_{11}(u, v) \alpha_{12}(v) dv \right) du \\ &= 1 - p_{00}(s, t) - p_{01}(s, t) - p_{02}^0(s, t). \end{aligned}$$

Indeed, the subject has to reach state 1 between s and u , where $s < u < t$, before reaching state 2. Finally, the probability of transition from state 0 to state 2 between s and t is:

$$p_{02}(s, t) = p_{02}^0(s, t) + p_{02}^1(s, t).$$

1.2 Cumulative incidences

We define S to be the exit time of $X(t)$ from state 0, *i.e.*, the time until a subject becomes ill or dies. Let δ be the indicator of cause of exit from state 0 ($\delta = 1$ if illness, $\delta = 2$ if death).

We consider, in the same way as Frydman & Szarek (2010), the functions $F_{01}(s, t)$ and $F_{02}(s, t)$, which are the cumulative probability of illness and the cumulative probability of death before illness:

$$F_{01}(s, t) = \mathbb{P}(S \leq t, \delta = 1 | S > s) \quad ; \quad F_{02}(s, t) = \mathbb{P}(S \leq t, \delta = 2 | S > s).$$

$F_{01}(s, t)$ corresponds to the probability for a healthy subject at time s to become ill at a time u between s and t . Between u and t such a subject may stay in state 1 or may move to state 2. $F_{02}(s, t)$ corresponds to the probability for a healthy subject at time s to die between s and u without going through the illness state before. Therefore, $F_{01}(s, t)$ and $F_{02}(s, t)$ can be expressed in terms of the transition probabilities detailed in subsection 1.1:

$$\begin{aligned} F_{01}(s, t) &= \int_s^t p_{00}(s, u) \alpha_{01}(u) du \\ &= p_{01}(s, t) + p_{02}^1(s, t), \\ F_{02}(s, t) &= p_{02}^0(s, t). \end{aligned}$$

Now $F_{01}(s, \infty) = \lim_{t \rightarrow +\infty} F_{01}(s, t)$ and $F_{02}(s, \infty) = \lim_{t \rightarrow +\infty} F_{02}(s, t)$ can be introduced, which can also be expressed in terms of transition probabilities:

$$F_{01}(s, \infty) = p_{02}^1(s, \infty) \quad ; \quad F_{02}(s, \infty) = p_{02}^0(s, \infty)$$

$F_{01}(s, \infty)$ is the lifetime risk of illness, whereas $F_{02}(s, \infty) (= 1 - F_{01}(s, \infty))$ is the lifetime probability of non-illness.

1.3 Life expectancies

In addition to S the exit time of $X(t)$ from state 0, let T be the entry time of $X(t)$ in state 2; *i.e.*, the death time. We can define three expectations in terms of transition probabilities.

Let's first consider a healthy subject at time s . Their "healthy life" expectancy, *i.e.* the time they can expect to spend without either becoming ill or dying is:

$$\mathbb{E}(S | X(s) = 0) = \int_s^{+\infty} p_{00}(s, u) du.$$

Their life expectancy (*i.e.*, the expected time before death) is:

$$\mathbb{E}(T | X(s) = 0) = \int_s^{+\infty} (p_{00}(s, u) + p_{01}(s, u)) du.$$

Let's now consider an ill subject at time s . Their life expectancy is:

$$\mathbb{E}(T | X(s) = 1) = \int_s^{+\infty} p_{11}(s, u) du.$$

2 Application to a dementia disease study

As an illustration, some examples of predictions are given using data of the Paquid study (Letenneur *et al.*, 1994), a large cohort study of mental and physical aging. The target population consists of subjects aged 65 years and older living in Southwestern France. The sample consists of 3675 initially non-demented subjects followed for dementia and survival. Over a 20-year period, 2937 deaths occurred and 832 incident cases of dementia were observed.

The data have been analysed separately for men and women with and without adjusting for covariates by the illness-death model depicted in Figure 1.

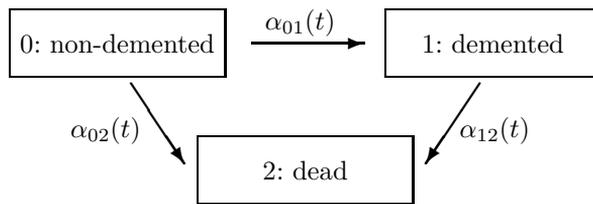


FIGURE 1. Illness-death model

The inclusion of covariates is allowed by a proportional intensity assumption on each $i \rightarrow j$ transition: $\alpha_{ij}(t|Z) = \alpha_{ij,0}(t)e^{\beta_{ij}^T Z}$, where Z is a vector of explanatory variables. The age is taken as basic time-scale. Thus, data are left-truncated. The ages of death are assumed to be known exactly or to be right-censored. The ages of dementia onset are interval-censored or right-censored. Indeed, the illness status is determined only by periodic examinations. When a new case of dementia is detected, we only know that the age of onset of dementia is between the age at the diagnosis visit and the age at the previous visit. Moreover, when a subject diagnosed as non-demented at the last visit dies, they may have become demented between this visit and death. In order to account for interval censoring and to obtain smooth estimations of transition intensities we use a maximum penalized likelihood method where the smoothing parameter is chosen by cross validation. For a detailed explanation of the method, see Joly *et al.* (2002).

Upon obtaining the estimated regression coefficients $\hat{\beta}_{ij}$, the estimated baseline intensities $\hat{\alpha}_{ij,0}$, and the estimated baseline cumulative intensities $\hat{A}_{ij,0}$, we can compute $\hat{\alpha}_{ij}(t|Z) = \hat{\alpha}_{ij,0}(t)e^{\hat{\beta}_{ij}^T Z}$ and $\hat{A}_{ij}(t|Z) = \hat{A}_{ij,0}(t)e^{\hat{\beta}_{ij}^T Z}$ corresponding to a subject with covariate values Z . By replacing α_{ij} by $\hat{\alpha}_{ij}$ in subsection 1.1 we can compute \hat{p}_{ij} , and, by replacing p_{ij} by \hat{p}_{ij} in subsections 1.2 and 1.3 we can compute the other formulae corresponding to

this subject for all s and t . For example, we may calculate lifetime risk over time by education level both for men and women.

Acknowledgments: Special thanks to the Paquid team and specially Mélanie Le Goff for making the Paquid data available to us. We also acknowledge Riccardo Marioni for his helpful comments and the *Région Aquitaine* for financial support.

References

- Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N. (1993). *Statistical models based on counting processes*. Springer: New York.
- Frydman, H., Szarek, M. (2010). Estimation of overall survival in an illness-death model with application to the vertical transmission of HIV-1. *Statistics in Medicine*, **29**, 2045–2054.
- Joly, P., Commenges, D., Helmer, C., 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., Commenges, D., Dartigues, J.F., Barberger-Gâteau, P. (1994). Incidence of dementia and Alzheimer's disease in elderly community residents of south-western France. *International Journal of Epidemiology*, **23**, 1256–1261.
- Putter, H., Fiocco, M., Geskus, R.B. (2007). Tutorial in biostatistics: Competing risks and multi-state models *Statistics in Medicine*, **58**, 2389–2430.

Variable selection for the multinomial logit model

Gerhard Tutz¹, Wolfgang Pöbnecker¹

¹ Department of Statistics, LMU Munich, Germany

E-mail for correspondence: Wolfgang.Poessnecker@stat.uni-muenchen.de

Abstract: Common variable selection for the multinomial logit model is based on forward/backward strategies, which are known to be rather unstable. We propose selection by regularization using an L_1 -type penalization term. The difference to existing methods is that all the parameters that are linked to one variable are penalized simultaneously. The method does not select single contributions of variables but whole variables. An application to data about party choice in Germany demonstrates the advantages of the proposed method.

Keywords: Logistic regression, Multinomial logit model, Variable selection, L_1 -penalty.

1 Introduction

The use of the multinomial logit model is typically restricted to applications with few predictors, because in high-dimensional settings maximum likelihood estimates tend to deteriorate. Therefore variable selection, which reduces the number of parameters to the relevant ones, is important in a parameter intensive model like the multinomial logit model.

The main feature of variable selection in the multinomial logit model is that the effect of one predictor variable is represented by several parameters. Therefore, one has to distinguish between variable selection and parameter selection. The available methods (Krishnapuram et al., 2005; Friedman et al., 2010) that are based on L_1 -type penalties shrink all the parameters simultaneously without using that the parameters are structured in groups with one group collecting all the parameters that refer to one variable.

In the present paper a penalty is proposed which explicitly uses the grouping of parameters that are linked to one predictor. The effect is selection of predictors rather than selection of parameters.

For linear and generalized linear models (GLMs) a variety of penalty approaches for regularized variable selection has been proposed. The most prominent example is the lasso (Tibshirani, 1996) and its extensions to fused lasso (Tibshirani et al., 2005) and grouped lasso (Yuan and Lin, 2006). Alternative regularized estimators that enforce variable selection are the

elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), the Dantzig selector (Candes and Tao, 2007) and boosting approaches (Bühlmann and Yu, 2003; Bühlmann and Hothorn, 2007; Tutz and Binder, 2006).

2 Model and Selection Procedure

2.1 The Multinomial Logit Model

For data $(Y_i, \mathbf{x}_i), i = 1, \dots, n$, with $Y_i \in \{1, \dots, k\}$ denoting the categorical response variable and \mathbf{x}_i the predictor, the multinomial logit model has the form

$$P(Y_i = r | \mathbf{x}_i) = \frac{\exp(\beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r)}{\sum_{s=1}^k \exp(\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s)} = \frac{\exp(\eta_{ir})}{\sum_{s=1}^k \exp(\eta_{is})}, \quad (1)$$

where $\boldsymbol{\beta}_r^T = (\beta_{r1}, \dots, \beta_{rp})$. Since parameters $\beta_{10}, \dots, \beta_{k0}, \boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_k^T$ are not identifiable, additional constraints are needed. Typically one of the response categories is chosen as reference category, for example, by setting $\beta_{k0} = 0, \boldsymbol{\beta}_k = \mathbf{0}$, category k is chosen as the reference category. An extensive discussion of the multinomial logit model as multivariate GLM is given, for example, in Agresti (2002) or Tutz (2012).

2.2 Penalized Estimation

As an alternative to forward/backward procedures we propose a penalized likelihood approach that enforces variable selection. The basic concept is to maximize a penalized log-likelihood

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}),$$

where $l(\boldsymbol{\beta})$ is the usual log-likelihood, λ is a tuning parameter, and $J(\boldsymbol{\beta})$ is a functional that penalizes the size of the parameters. While the tuning parameter determines the strength of the regularization, the functional determines the properties of the penalized estimation.

The most widely used penalty that enforces variable selection is the lasso (Tibshirani, 1996). It has been used in models with unidimensional response models like the classical linear model and univariate generalized linear models (GLMs). For the multinomial logit model direct application of the lasso corresponds to the penalty

$$J(\boldsymbol{\beta}) = \sum_{r=1}^{k-1} \|\boldsymbol{\beta}_r\|_1 = \sum_{r=1}^{k-1} \sum_{j=1}^p |\beta_{rj}|,$$

where $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{k-1})$ collects all the parameters to be estimated. Friedman et al. (2010) used the slightly more general elastic net penalty,

which also has the drawback that selection focuses on parameters but not on variables.

By contrast, the penalty proposed by us penalizes the group of parameters that are linked to one variable. For simplicity let the j -th predictor be metric and the parameters in $\eta_{ir}, r = 1, \dots, k-1$, that are linked to variable j be collected in $\beta_{\bullet j}^T = (\beta_{1j}, \dots, \beta_{k-1,j})$. If no category-specific predictors are included we will use the penalty

$$J(\beta) = \sum_{j=1}^p \|\beta_{\bullet j}\|_2 = \sum_{j=1}^p (\beta_{1j}^2 + \dots + \beta_{k-1,j}^2)^{1/2}. \tag{2}$$

The penalty enforces variable selection, that is, all the parameters in $\beta_{\bullet j}$ are simultaneously shrunk toward zero. It is strongly related to the group lasso (Yuan and Lin, 2006; Meier et al., 2008). However, in the group lasso the grouping refers to the parameters that are linked to a categorical predictor within a univariate regression model whereas in the present model grouping arises from the multivariate response structure.

2.3 Computation of estimates

For maximization of the penalized log-likelihood, we use the general FISTA procedure of Beck and Teboulle (2009). FISTA requires the log-likelihood, the score function and the proximal operator associated with (2). For each actual observation y_i , we define a set of $k - 1$ pseudo-observations y_{ir} :

$$y_{ir} = \begin{cases} 1 & \text{if } y_i = r \\ 0 & \text{otherwise.} \end{cases}$$

With the linear predictors η_{ir} from (1), the log-likelihood can then conveniently be written as

$$l(\beta) = \sum_{i=1}^n \left(\sum_{r=1}^{k-1} y_{ir} \eta_{ir} - \log \left(1 + \sum_{s=1}^{k-1} \exp(\eta_{is}) \right) \right). \tag{3}$$

To be able to give the score function in a concise form, we introduce the following notation: For $r = 1, \dots, k - 1$, let $\mathbf{y}_r = (y_{1r}, \dots, y_{nr})^T$, $\boldsymbol{\pi}_r = (\pi_{1r}, \dots, \pi_{nr})^T$ and $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$. With these definitions, the score function of β_{rj} is given by

$$s(\beta_{rj}) = \frac{\partial l(\beta)}{\partial \beta_{rj}} = \mathbf{x}_j^T (\mathbf{y}_r - \boldsymbol{\pi}_r). \tag{4}$$

For a candidate point $\tilde{\beta}$ (which typically corresponds to some sort of unpenalized estimate), the proximal operator associated with a penalty $\lambda J(\beta)$ is defined as

$$\mathcal{P}_\lambda(\tilde{\beta}) = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2} \|\beta - \tilde{\beta}\|_2^2 + \lambda J(\beta) \right).$$

Because our penalty (2) is block-separable, an analytic form is available for each parameter group:

$$\mathcal{P}_\lambda(\tilde{\beta}_{\bullet j}) = \left(1 - \frac{\lambda}{\|\tilde{\beta}_{\bullet j}\|_2}\right)_+ \tilde{\beta}_{\bullet j}, \quad (5)$$

where $(v)_+ = \max(v, 0)$. Using the equations (3), (4) and (5) as “ingredients”, FISTA is directly applicable. For further details, see Beck and Teboulle (2009).

2.4 Application

We consider the modelling of party preference in Germany. We use data from the German Longitudinal Election Study with the five response categories Christian Democratic Union (CDU: 1), Social Democratic Party (SPD: 2), Green Party (3), Liberal Party (FDP: 4), and Left Party (Die Linke: 5). There are nine potential predictors: age, political interest (1: less interested 0: very interested), religion (1: evangelical, 2: catholic, 3: otherwise), regional provenance (west; 1: former West Germany, 0: otherwise), gender (1: male, 0: female), union (1: member of a union, 0: otherwise), satisfaction with the functioning of democracy (democracy; 1: not satisfied 0: satisfied), unemployment (1: currently unemployed, 0: otherwise), and high school degree (1: yes, 0: no).

Figure 1 shows the buildups of global variables resulting from lasso-type regularization. Only the variables that turned out to be influential are shown. The variables political interest, gender, unemployment, and high school degree were set to zero by the method. The vertical line shows the selected smoothing parameter based on cross-validation.

The grouped selection behavior can clearly be seen from figure 1: the $k - 1$ coefficients that belong to the same predictor always enter or leave the model simultaneously. Thus, in contrast to previous approaches (Krishnapuram, 2005; Friedman, 2010), the method proposed in this paper performs actual variable selection in multinomial logit models.

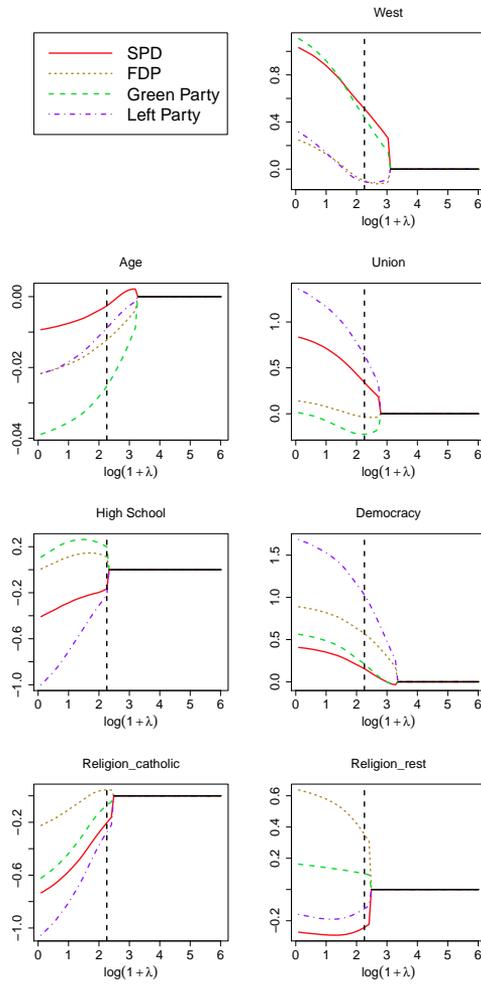


FIGURE 1. Coefficient buildups for selected global variables of party choice data.

References

- Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.
- Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **2**, 183–202.
- Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, **98**, 324–339.
- Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, **22**, 477–505.
- Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, **35**, 2313–2351.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.
- Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.
- Krishnapuram, B., Carin, L., Figueiredo, M. and Hartemink, A. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 957–968.
- Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society B*, **70**, 53–71.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267–288.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Kneight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B*, **67**, 91–108.
- Tutz, G. and Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, **62**, 961–971.
- Tutz, G. (2012). *Regression for Categorical Data*. Cambridge: Cambridge University Press.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, **68**, 49–67.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, **67**, 301–320.

Hierarchical covariance selection models

Insha Ullah¹, Beatrix Jones¹

¹ Institute of Information and Mathematical Sciences, Massey University, New Zealand

E-mail for correspondence: i.ullah@massey.ac.nz

Abstract: We concern ourselves with the multivariate normal hierarchical model where the same set of variables is measured in several different populations. For instance, consider the situation where replicate measurements are made on a single object while at the same time there are many different objects. To capture the multivariate data generating mechanism, reliable estimates are needed for within-group and between-group covariance structures. Recently a penalized likelihood has been proposed to get a reliable estimate in high-dimensional problems for a single population. In this paper the strategy of penalized likelihood has been extended to two-level normal hierarchical models via the EM-algorithm. The performance of the method is illustrated by a number of simulated examples and a real glass chemical composition data set.

Keywords: Hierarchical models; Covariance selection; Lasso; Between-group covariance.

1 Introduction

In many applied problems we come across multivariate data sets that come from multiple groups. One special case is when there are few observations in each group but many more groups. An example of this kind can be found in Aitken and Lucy (2004) where multivariate replicate measurements are taken on the elemental composition of glass from different windows. A data set of similar nature has been collected by Bennett (2002) who made twenty replicate measurements of five elements on each of six different Heineken beer bottles. In both of these cases, since the within-group variation is because of measurement error, our emphasis is on inferring between-group variation while controlling for within-group variation.

A multivariate random-effect model has been used by Aitken and Lucy (2004) to summarize the data. To fit the model one needs reliable estimates of both within-group and between-group covariance components. Traditionally, the estimation of between-group covariance component involves the difference of two mean square matrices: the between-group mean square and the within-group mean square. For a sufficiently large sample size, both mean squares are individually guaranteed to be nonnegative definite,

however, their difference is not, and often produces negative elements on the diagonal. This is pointed out by Hill and Thompson (1978), who have shown that the probability of negative variances increases with increasing the number of variables. Additional problems arise when the sample size n is small relative to the number of variables p : one or both of the mean-square matrices may be ill-conditioned or even singular.

Regularized alternatives have been proposed to deal with the small sample size problem for a single population; see, e.g., Bickel and Levina (2008). Dempster (1972) suggested the ‘‘covariance selection model’’, where the objective is to identify zero elements in the off-diagonal of the inverse covariance matrix. The resulting estimate is interpretable as well as regularized. Zero elements correspond to pairs of variables that are conditionally independent given the others. We find this interpretability appealing and pursue an estimate of this type.

Yuan and Lin (2007) extended the Lasso method of Tibshirani (1996) to the covariance selection setting for estimation of $\Omega = \Sigma^{-1} = (\omega_{ij})_{1 \leq i, j \leq p}$. The following log-likelihood function based on a random sample $X = (X_1, X_2, \dots, X_p)^t$ from a multivariate normal distribution $X \sim N(0, \Sigma)$ is optimized:

$$l(\Omega) = Const - \frac{1}{2} \sum_{i=1}^n \log |\Sigma| - \frac{1}{2} \sum_{i=1}^n X^t \Omega X - \lambda \sum_{j=1}^p \sum_{k=1}^p |\omega_{jk}| \quad (1)$$

where λ is the parameter which shrinks some coefficients towards zero and sets the others as exactly zero. This produces a positive definite solution. Larger values of λ will produce sparser solutions. We follow Gao et al., (2012) and use the Bayesian information criterion (BIC) to select the optimal value of λ .

In this paper, our objective is to summarize a high-dimensional data set in which the same set of variables is measured in many different groups. The method of penalized likelihood is extended to hierarchical covariance selection models via the EM-algorithm in order to get reliable estimates for both within-group and between-group covariance structure. Our primary reason for using EM-algorithm is to avoid the difference of two covariance components while calculating the between-group covariance. We use a positive definite matrix as an initial estimate and update in such a way that the matrices must remain positive definite. The EM-algorithm also allows us to penalize both within-group and between-group covariances separately.

2 The model

We measure p variables from m different groups and there are r measurements from each group. Denote the $n = mr$ observations by $X_{ij} = (X_{ij1}, X_{ij2}, \dots, X_{ijp})^t$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, r$. Let θ be the set of

within-group mean vectors and U be the common within-group covariance matrix. Then, given θ and U , the distribution of X_{ij} is assumed to be normal with $X_{ij} \sim N(\theta_i, U)$. Similarly assume that μ is the between-group mean vector and C is the between-group covariance matrix. Then each θ_i is assumed to be normally distributed with $\theta_i \sim N(\mu, C)$. We write the joint density for this model as:

$$f(X_i, \theta_i) = f(X_i|\theta_i, U)f(\theta_i|\mu, C). \tag{2}$$

The EM-algorithm is designed to compute maximum likelihood estimates in the presence of incomplete data. It iterates between the Expectation-step and the Maximization-step. In the E-step we consider θ as missing and find the conditional expectation of posterior distribution of θ given X , U , and C . Assume that θ is properly centred; the posterior distribution for θ_i given X_i , U , and C is then

$$f(\theta_i|X_i, U, C) \propto \exp\left(-\frac{1}{2}\sum_j (X_{ij} - \theta_i)^t U^{-1}(X_{ij} - \theta_i)\right) \exp\left(-\frac{1}{2}\theta_i^t C^{-1}\theta_i\right) \tag{3}$$

Expanding the exponents and completing the quadratic form for θ_i gives

$$\begin{aligned} f(\theta_i|X, U, C) &\propto \exp\left(-\frac{1}{2}(\theta_i - \theta_0)^t (C^{-1} + rU^{-1})^{-1} (\theta_i - \theta_0)\right) \\ &= N(\theta_0, C^{-1} + rU^{-1}) \end{aligned} \tag{4}$$

where

$$\theta_0 = (C^{-1} + rU^{-1})^{-1} (rU^{-1}\bar{X}). \tag{5}$$

At the M-step we assume that θ is known which allows us to estimate U^{-1} and C^{-1} by independently maximizing $l(U^{-1})$ and $l(C^{-1})$ given by

$$l(U^{-1}) = \sum_{i=1}^m \sum_{j=1}^r \log |U| - \sum_{i=1}^m \sum_{j=1}^r (X_{ij} - \theta_i)^t U^{-1}(X_{ij} - \theta_i) - \rho \sum_{k=1}^p \sum_{l=1}^p |u_{kl}| \tag{6}$$

where u_{kl} is the kl th element of U^{-1} and ρ is the shrinkage parameter; and

$$l(C^{-1}) = \sum_{i=1}^m \log |C| - \sum_{i=1}^m \theta_i^t C^{-1}\theta_i - \lambda \sum_{k=1}^p \sum_{l=1}^p |c_{kl}| \tag{7}$$

where c_{kl} is the kl th element of C^{-1} , and shrinkage parameter λ . The algorithm, for fixed λ and ρ , is then:

1. Initialize all θ_i as \bar{X}_i .
2. Obtain the estimate of U^{-1} by maximizing the expression in (6).

3. Similarly, estimate C^{-1} by maximizing the expression in (7).
4. Update the estimate of each θ_i as $\hat{\theta}_i^1 = \theta_0$ from (5).
5. Repeat step 2-4 until convergence.

This process is repeated for a grid of λ and ρ , and the pair of shrinkage parameters corresponding to a local optimum in BIC is selected.

3 Applications

In this section we assess the performance of hierarchical covariance selection model and report the results of both our simulation study and real data set. In our simulation experiments we first generate U and C using the method of Schafer and Strimmer (2005), which guarantees the generated matrix to be positive definite. The m p -dimensional group means, $\boldsymbol{\theta}$, are drawn from a multivariate normal distribution $\theta_i \sim N(\mu, C)$. The r observations, X_{ij} , for each group are then generated from a multivariate normal distribution i.e. $X_{ij} \sim N(\theta_i, U)$. In all of our simulations we keep the dimension $p = 5$, number of groups $m = 100$, and there are $r = 5$ observations per group.

3.1 Numerical experiments

We consider three types of situations, varying the nature of the within-group and between-group covariance matrices. At first we examine the performance of the hierarchical model in the situation where the between-group variation is dominant and the within-group covariance is roughly diagonal. This is the easiest case in the sense that the EM-algorithm converges quickly and the estimated covariances are relatively close to the true covariances. In the second example we allow a moderate within-group covariance. In third case, we make the within-group variation dominant. This case is comparatively hard and the algorithm needs few more iterations to converge. All the three cases are illustrated in Figure 1(a, c, and e), where we plot the data on the first two principal components with groups represented by numbers 1-100, with points in the same group sharing a common colour. The graphs of corresponding estimated vs true between-group partial correlations are depicted in Figure 1(b, d, and f), and show good agreement.

3.2 Real data example

To illustrate the method we use glass chemical composition data collected by Bennett (2002) (also available in R packages “Hotelling” and “dafs”). The data are the measurements of elemental concentration of five different elements: Manganese, Barium, Strontium, Zirconium, and Titanium.

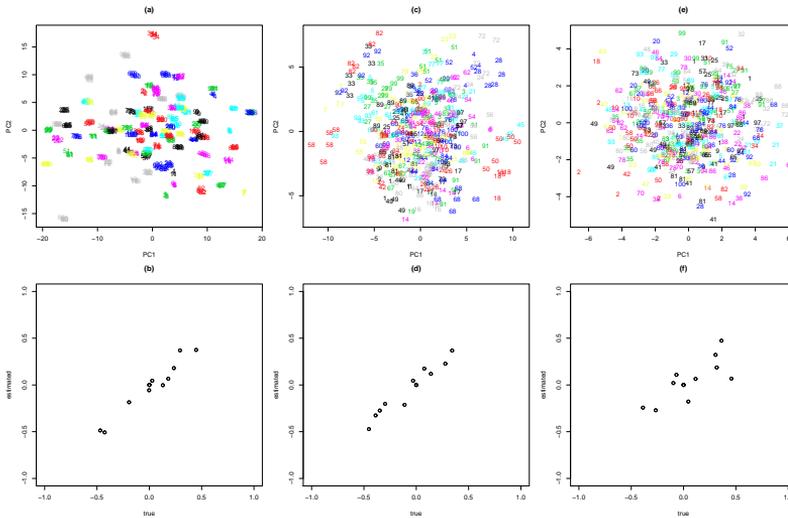


FIGURE 1. Simulated data on first two principle components for 3 cases (a) easiest, (c) moderate, and (e) hard. The true vs estimated between-group partial correlations are in (b), (d), and (f) respectively.

Twenty replicate measurements are taken from six different Heineken beer bottles. Thus, there are 5 variables measured in 6 different groups with 20 replicates in each group. The data is plotted on the first two principal components in Figure 2, with the groups indicated by numbers 1-6. The between-group covariance obtained by the method followed by Aitken and Lucy (2004) is

$$\begin{pmatrix} 26.1 & - & - & - & - \\ 89.1 & 354 & - & - & - \\ 66.5 & 216 & 171 & - & - \\ 59.7 & 183 & 153 & 145 & - \\ 19.1 & 30.2 & 51.1 & 62.7 & 47.7 \end{pmatrix}$$

which is not nonnegative definite. The one obtained by hierarchical covariance selection model is

$$\begin{pmatrix} 18.8 & - & - & - & - \\ 48.3 & 153 & - & - & - \\ 39.4 & 85.3 & 92 & - & - \\ 24.6 & 58.7 & 53.8 & 33 & - \\ 3.09 & -56.5 & 41 & 14.6 & 145 \end{pmatrix}$$

which is positive definite. Thus, for both the real world glass chemical composition data and simulated data, the developed method performs well and produces estimates with the desired property (positive definiteness).

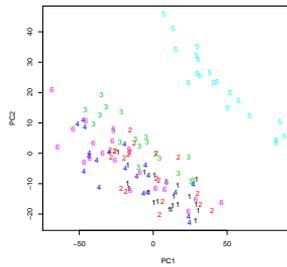


FIGURE 2. First two principle components of glass chemical composition data.

References

- Aitken, C.G.G., and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics, Series C*, **53**, 109–122.
- Bennett, R. L. (2002). *Aspects of the analysis and interpretation of glass trace evidence*. Masters thesis, Department of Chemistry, University of Waikato.
- Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, **36**, 2577–2604.
- Dempster, A. P. (1972). Covariance selection. *Biometrics*, **28**, 157–175.
- Gao, X., Pu, D. Q., Wu, Y., and Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statistica Sinica*, **preprint**, doi:10.5705/ss.2009.210.
- Hill, W. G., and Thompson, R. (1978). Probabilities of non-positive definite between-group or genetic covariance matrices. *Biometrics*, **34**, 429–439.
- Schafer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.
- Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical models. *Biometrika*, **94**, 19–35.

A longitudinal model for latent cognitive function

Ardo van den Hout¹, Jean-Paul Fox², Graciela Muniz³

¹ Department of Statistical Science, University College London, UK

² Twente University, Netherlands

³ MRC Unit for Lifelong Health and Ageing, London, UK

E-mail for correspondence: `ardo.vandenhout@ucl.ac.uk`

Abstract: A mixed-effects regression model with a non-linear predictor is formulated to describe change in latent cognitive function over time. As a latent variable cognitive function is linked to longitudinal questionnaire data by using an item response theory model. Bayesian inference is used, where the Deviance Information Criterion is applied for model comparison. The model is an alternative to a regression model with the manifest sum score as response.

Keywords: Bayesian inference; Change point; Latent variable; Mixed effects.

1 Introduction

Understanding change in cognitive function in the years before death is of interest to researchers in medical science. Potential decline and the possibility of a one-off change in the trend of the decline will be investigated using mixed-effects regression models for longitudinal data with linear and non-linear predictors. The response variable in the regression models is latent cognitive function, which is measured by using a longitudinal item response theory model for observed scores on a questionnaire.

Data stem from the Cambridge City over-75s Cohort Study (CC75C), where cognitive function is measured using a questionnaire with 16 binary items. The discreteness of the sum score and the skewness of its distribution means that a regression model with the manifest sum score as response is problematic when the conditional distribution is assumed to be normal.

As an alternative, we present regression models where the response variable is a latent continuous variable, which explains how well individuals perform in the examination. The link between the latent variable and the longitudinal scores on the individual items in the questionnaire is described by an Item Response Theory (IRT) model. Hence, the latent response is interpreted as the underlying cognitive function which explains observed cognitive performance.

A one-off change of direction in cognitive function can be described by a

change-point model. We will use a smooth version of the broken-stick model called the bent-cable model (Chui et al. 2006).

Bayesian inference will be applied. Markov chain Monte-Carlo (MCMC) techniques will be used for parameter estimation, and the Deviance Information Criterion (DIC) will be used to compare models. Our work is building upon Bayesian inference for cross-sectional IRT models as presented in Johnson and Albert (1999) and Fox (2010).

The combination of an IRT model and a change-point regression model has not been investigated before but seems promising in scope as it can describe cognitive decline using regression models with non-linear predictors and at the same time can take into account the longitudinal question-specific discrete scores.

The time scale that will be used to analyse the CC75C data is rather specific. Almost all (96%) of the participants in CC75C have passed away since the start of the study in 1991. This makes it possible to use years-to-death as the time scale in our models without introducing bias due to selection of the survivors only. The presented methodology, however, is general and can also be used in longitudinal models where age is the time scale.

2 Model

Let the latent variables for individual i be given by $\theta_{i1}, \dots, \theta_{in_i}$ at times t_{i1}, \dots, t_{in_i} , where time of death is $t = 0$ and the times before death are represented by negative values. So $t_{in_i} < 0$ is the last time individual i was observed in the study.

A regression model for θ with a smooth change-point predictor is given by

$$\theta_{ij} = \begin{cases} \eta_{1i} + \eta_{2i}t_{ij} + e_{ij} & t_{ij} \leq \tau_i - \delta \\ \eta_{1i} + \eta_{2i}t_{ij} + \eta_{3i}(t_{ij} - \tau_i + \delta)^2/4\delta + e_{ij} & \tau_i - \delta < t_{ij} \leq \tau_i + \delta \\ \eta_{1i} + (\eta_{2i} + \eta_{3i})t_{ij} - \eta_{3i}\tau_i + e_{ij} & \tau_i + \delta < t_{ij}, \end{cases}$$

$$\begin{aligned} \eta_{1i} &= \beta_1 + b_{1i} & \eta_{3i} &= \beta_3 + b_{3i} \\ \eta_{2i} &= \beta_2 + b_{2i} & \tau_i &= \beta_4 + b_{4i} \\ (b_{1i}, b_{2i}, b_{3i}, b_{4i}) &\sim MVN(0, \Sigma) & e_{ij} &\sim N(0, \sigma^2). \end{aligned}$$

where $\delta > 0$, and τ_i is the random-effect change point. The basic idea in bent-cable regression is that the two linear parts are bridged by a quadratic bend with half-width δ and midway location at τ . In the data analysis, we fix $\delta = 1/2$ year.

Cognitive function is latent since it is not directly observed but measured by a test (a questionnaire). At every observation time, the test consists of K items (questions). For item k with R ordered response categories the data for individual i at time t_{ij} is y_{ij1}, \dots, y_{ijK} . Given response categories 1 up to R (with the latter denoting the best score), the model has $R - 1$ ordered

thresholds parameters d_{k1}, \dots, d_{kR-1} . Together with the bounds $-\infty$ and ∞ and the ordering $-\infty < d_{k1} < \dots < d_{kR-1} < \infty$, these thresholds define segments on the real line. The graded-response model is given by

$$p(y_{ijk} = m | \theta_{ij}, c_k, d_{k1}, \dots, d_{kR-1}) = \Phi(c_k \theta_{ij} - d_{km-1}) - \Phi(c_k \theta_{ij} - d_{km}).$$

For item k , parameter c_k is the discrimination parameter, and d_{k1}, \dots, d_{kR-1} are the difficulty parameters. Given a value of θ_{ij} , these parameters define the probabilities of the answer categories.

In case $R = 2$ with answer categories 1 and 2, the graded response model reduces to $p(y_{ijk} = 2 | \theta_{ijk}, c_k, d_k) = \Phi(c_k \theta_{ijk} - d_k)$.

To identify the model (note that θ does not have a natural scale) restrictions can be imposed either on θ or on both c and d .

3 Markov chain Monte Carlo

In the application, data analysis is with respect to binary items. Using auxiliary variable z allows for a straightforward MCMC implementation for the item parameters. Variable z is a continuous representation of binary data y such that z_{ijk} is normally distributed with mean $c_k \theta_{ijk} - d_k$ and standard deviation 1, and $y_{ijk} = 2$ when $z_{ijk} > 0$, and $y_{ijk} = 1$ when $z_{ijk} \leq 0$.

The conditional distributions in the Gibbs sampler are given by

$$\begin{aligned} p(z|\dots) &= p(z|\theta, c, d, y) & p(\theta|\dots) &\propto p(y|\theta, c, d)p(\theta|\beta, b, \sigma, t) \\ p(\beta|\dots) &\propto p(\theta|\beta, b, \sigma, t)p(\beta) & p(\sigma|\dots) &\propto p(\theta|\beta, b, \sigma, t)p(\sigma) \\ p(b|\dots) &\propto p(\theta|\beta, b, \sigma, t)p(b|\Sigma) & p(\Sigma^{-1}|\dots) &\propto p(b|\Sigma)p(\Sigma^{-1}) \\ p(c|\dots) &\propto p(y|\theta, c, d)p(c) & p(d|\dots) &\propto p(z|\theta, c, d)p(d). \end{aligned}$$

For some of these distributions, e.g., $p(\theta|\dots)$, Metropolis steps are undertaken to sample values. For others, e.g., $p(c|\dots)$ and $p(d|\dots)$, the conditional distribution can be formulated as a normal distribution.

Restrictions are imposed in every run of the Gibbs sampler in the same way as it is done in cross-sectional IRT models, see Fox (2010, Section 4.4.2). If the restriction is with regard to θ , the candidate parameter vector is re-scaled such that mean of θ is 0 and the variance is 1. If the restriction concerns the item parameters, then re-scaling is undertaken such that $\prod_{k=1}^K c_k = 1$ and $\sum_{k=1}^K d_k = 0$.

The above MCMC is extended to take missing data into account when *missing at random*: firstly, missing values are sampled from their conditional distributions given current parameter values, and, secondly, the MCMC steps for complete data are undertaken.

4 Application

4.1 CC75C data

More information on the Cambridge City over-75s Cohort Study (CC75C) can be found at www.cc75c.group.cam.ac.uk. For a preliminary analysis, we take a random subset from CC75C with sample size $N = 200$. In the subset, the frequencies for the number of interviews per individual are 89, 49, 37, 19, 5, and 1, for 1 up to 6 interviews. There are $K = 16$ binary items, and the total number of times a question was asked is $405 \times K = 6480$. There are 354 missing item scores distributed over 50 individuals. The missing item scores are sampled within the MCMC.

Two models will be compared using the DIC, where this criterion is defined using the likelihood for data y given sampled values for θ , c , and d . The first is a model where the change of the latent trend over time is described by a linear predictor, the second is the model with the change-point predictor. Both models are identified by imposing restrictions on the item parameters.

4.2 Model with linear predictor

The model with the linear predictor for latent θ is given by

$$\begin{aligned} \theta_{ij} &= \nu_{1i} + \nu_{2i}t_{ij} + e_{ij} & e_{ij} &\sim N(0, \sigma^2) \\ \nu_i &= (\nu_{1i}, \nu_{2i}) & &\sim MVN((\mu_1, \mu_2), \Sigma). \end{aligned}$$

If θ_{ij} would have been a manifest variable, then the model would be a standard mixed-effects model. The MCMC for the latent-variable model is an adapted version of the MCMC presented in Section 3. Note that once θ has been sampled, the Gibbs sampling for σ , ν_1, \dots, ν_N , μ_1 , μ_2 , and Σ can be undertaken as if θ is manifest, see, e.g., Gelfand et al. (1990) for the conditional distributions.

Vague prior densities are used: $\sigma \sim U(0, 10)$, $\Sigma^{-1} \sim Wishart((\rho R)^{-1}, \rho)$, where $\rho = 2$ and R is the diagonal matrix with diagonal $(1, 1/10)$, and $(\mu_1, \mu_2) \sim N((0, 0), C)$, where $C^{-1} = 0$. For the item parameters, the priors are improper and equal to 1.

The posterior mean of slope parameter μ_2 is -0.105 with 95% credible interval (-0.151, -0.063), which represents a trend of decreasing cognitive function over time. The DIC for this model is 3511.

4.3 Model with change-point predictor

The change-point model for latent θ is specified with a random intercept and a random second slope, and with fixed effects for the first slope and the change point. Table 1 present the posterior means and 95% credible intervals for the regression coefficients and the variance parameters. The

TABLE 1. Posterior means and 95% credible intervals for parameters in change-point regression model for latent θ .

Parameter	Posterior mean	95% credible interval
β_1	1.282	(1.065, 1.611)
β_2	-0.062	(-0.096, -0.027)
β_3	-0.029	(-0.121, 0.043)
β_4	-5.403	(-6.728, -4.428)
Σ_{11}	0.390	(0.258, 0.557)
Σ_{22}	0.048	(0.018, 0.081)
Σ_{12}	0.005	(-0.042, 0.055)
σ	0.274	(0.136, 0.445)

posterior distributions of the item parameters are not reported. The DIC for the change-point model is 3452, which is an improvement over the model with the linear predictor.

The posterior mean of the fixed-effect change point β_4 is -5.403 with 95% credible interval (-6.728, -4.428). This means that if there is a change in the trend of cognitive function, then it occurs on average about five and a half years before death. Figure 1 depicts observed and fitted sum scores for a selection of individuals who had three or more interviews. The fit and the 95% credible band are based on the posterior means of the parameters and their 95% credible intervals. Note that the model is flexible enough to describe trajectories with and without cognitive decline.

References

Chiu, G., Lockhart, R., and Routledge, R. (2006). Bent-cable regression theory and applications, *Journal of the American Statistical Association* **101**, 542–553.

Fox J.-P. (2010). *Bayesian Item Response Modeling*, New York: Springer.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990), Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**, 972–985.

Johnson, V.E., and Albert, J.H. (1999). *Ordinal Data Modeling*. New York: Springer.

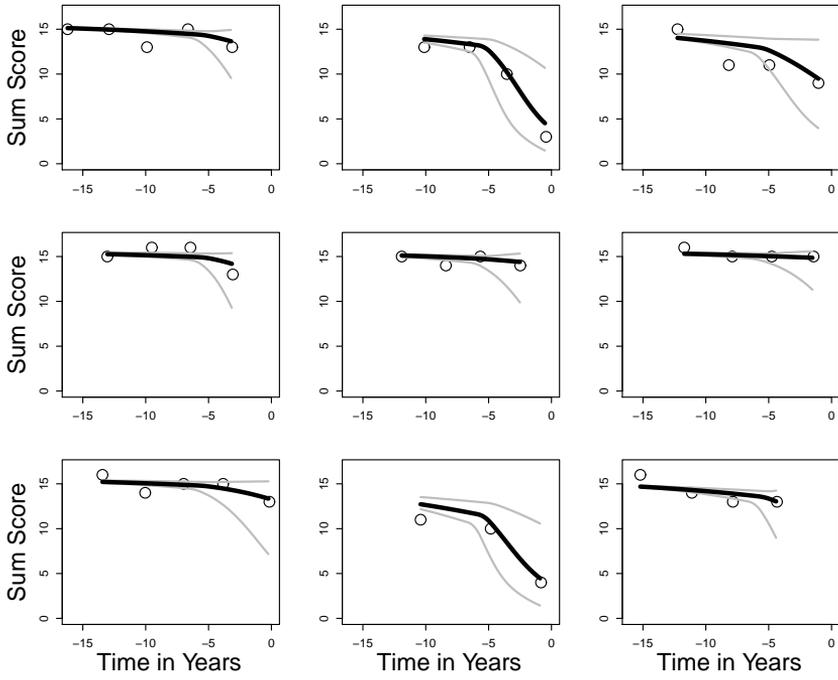


FIGURE 1. Observed and fitted sum score for nine individuals in CC75C.

Marginal beta regression for time series analysis

Cristiano Varin¹, Annamaria Guolo²

¹ DAIS, Università Ca' Foscari, Venezia, Italy

² Dept. of Economics, University of Verona, Italy

E-mail for correspondence: `sammy@unive.it`

Abstract: A marginal beta regression model with autoregressive and moving average errors is developed for the analysis of time series of values in the standard unit interval $(0, 1)$, such as proportions and rates. The dependence structure is conveniently related to the marginal model through a Gaussian copula specification. Likelihood inference, model validation via residual analysis, and prediction are briefly discussed. The methodology is applied to the time series of the rate of hidden unemployment in São Paulo, Brazil.

Keywords: ARMA; Beta Regression; Gaussian Copula; Rate; Time Series.

1 Introduction

Time series of data in the standard unit interval $(0, 1)$ are frequently encountered in economics and other social disciplines. The common approach analyzes the data through standard linear time-series regression after a logit transformation of the observations. As stated by Cribari-Neto and Zeileis (2010), such a strategy can be flawed, since it does not account for heteroskedasticity and asymmetry which often affect the data constrained in the unit interval. Furthermore, the interpretation of the parameters is not straightforward as a consequence of the data transformation. A suitable solution is to exploit the flexibility of the beta distribution (Ferrari and Cribari-Neto, 2004).

In this paper, we specify a convenient class of marginal beta regression models with autoregressive and moving average errors. Our approach is based on Gaussian copula marginal regression following Masarotto and Varin (2011). The methodology is briefly summarized and then illustrated on the time series of the rate of hidden unemployment in São Paulo, Brazil, already analyzed by Rocha and Cribari-Neto (Test, 2009) with an observation-driven model.

2 Marginal beta regression with ARMA errors

Let $Y_t, t = 1, \dots, n$, be a time series of data in the $(0, 1)$ interval and let y_t the corresponding observed values. We assume a beta distribution for Y_t with density function parameterized as in Ferrari and Cribari-Neto (2004),

$$f_t(y_t) = \frac{\Gamma(\kappa_t)}{\Gamma(\mu_t\kappa_t)\Gamma\{(1-\mu_t)\kappa_t\}} y_t^{\mu_t\kappa_t-1} (1-y_t)^{(1-\mu_t)\kappa_t-1}, \quad (1)$$

for $0 < y_t < 1$. The density depends on time-dependent location $0 < \mu_t < 1$ and precision $\kappa_t > 0$ parameters. According to this parameterization $E(Y_t) = \mu_t$ and $\text{Var}(Y_t) = \mu_t(1-\mu_t)/(1+\kappa_t)$. The dependence of the mean of Y_t on covariates x_t , which may include time, can be modeled by assuming

$$g(\mu_t) = x_t^T \beta,$$

where $g(\cdot)$ is a link function from $(0, 1)$ to \mathbb{R} and β is a vector of regression coefficients. A common choice for link $g(\cdot)$ is the logistic function. It is also possible to link the precision parameter to a second set of covariates z_t

$$h(\kappa_t) = z_t^T \gamma,$$

for another vector γ of unknown regression coefficients. A natural choice for the precision link $h(\cdot)$ is the logarithm function.

On the basis of the integral transformation theorem, the above beta regression model can be equivalently expressed in terms of normally distributed errors ε_t , as follows

$$Y_t = F_t^{-1}\{\Phi(\varepsilon_t)\}, \quad t = 1, \dots, n. \quad (2)$$

In the equation above, $F_t(\cdot)$ and $\Phi(\cdot)$ denote the cumulative distribution functions of the beta random variable Y_t and of a standard normal variable, respectively.

Following Masarotto and Varin (2011), serial correlation in time series Y_t can be introduced by assuming that the joint distribution of the errors ε_t is multivariate normal,

$$(\varepsilon_1, \dots, \varepsilon_n)^T \sim \text{MVN}(0, \Omega) \quad (3)$$

where Ω is the correlation matrix of an autoregressive and moving average process of orders p and q , $\text{ARMA}(p, q)$. The nondiagonal terms of Ω are $\text{corr}(\varepsilon_i, \varepsilon_j) = \omega_{|i-j|}$.

If parameters μ_t and κ_t are constant-in-time, then the model is stationary. Otherwise, some forms of nonstationarity can be accounted for by expressing μ_t and κ_t as a function of covariates x_t .

Expressions (2)-(3) identify a Gaussian copula marginal regression model in the terminology of Masarotto and Varin (2011). More generally, this class

of models provides a flexible framework for marginal regression analysis in presence of dependence, whichever type of the response is of interest. Masarotto and Varin (2011) listed three useful properties for model interpretation that relate the correlation structure of the errors to the responses Y_t , conditionally on x_t .

1. Since map (2) is nondecreasing, then the sign of the autocorrelation function of the responses equals that of the normal errors series ε_t .
2. If the errors follow a moving average model of order q , then responses far apart more than q lags are independent.
3. If the errors follow an autoregressive process of order p , then the responses are also a Markovian process of order p .

2.1 Inference, diagnostics, and prediction

Inference

We suggest model fitting through a maximum likelihood approach. Let $\theta = (\psi, \lambda)$ be the parameter vector consisting of the marginal parameter ψ of the beta distribution and of the dependence parameter λ of the ARMA(p, q) correlation matrix. A simple transformation argument yields the following expression for the likelihood function

$$L(\theta) = \phi_n \{ \varepsilon_1(\psi), \dots, \varepsilon_n(\psi); \Omega(\lambda) \} \prod_{t=1}^n \frac{f_t(y_t; \psi)}{\phi \{ \varepsilon_t(\psi) \}},$$

where $\phi_n \{ \cdot; \Omega(\lambda) \}$ is the density function of a multivariate normal distribution with standardized marginals and correlation matrix $\Omega(\lambda)$. The dependence of the normal errors $\varepsilon_t(\psi) = \Phi^{-1} \{ F_t(y_t; \psi) \}$ on the marginal parameter component is stressed out.

The uncertainty of the maximum likelihood estimator is evaluated through the robust sandwich covariance matrix in order to account for possible misspecification of distributional errors assumptions.

Diagnostics

Model validation is based on quantile residuals. Consider that the Rosenblatt (1952) transformations

$$M_t = F_t(Y_t | y_{t-1}, \dots, y_1; \theta)$$

are independent uniform (0, 1) random variables. Hence, model assumptions can be validated through the quantile residuals (Dunn and Smyth, 1996)

$$r_t = \Phi^{-1} \left\{ F_t(y_t | y_{t-1}, \dots, y_1; \hat{\theta}) \right\},$$

where $\hat{\theta}$ is the maximum likelihood estimate. Residuals r_t are realizations of independent standard normal variables under model conditions. Normal probability plots and autocorrelation functions can be used to investigate proper fitting.

Prediction

Prediction is straightforward within the assumed framework. Given the one-to-one relationship between the normal errors and the responses, future values of the errors can be predicted as in standard ARMA models and then transformed back to the response scale through the inverse of map (2).

3 Application to hidden unemployment in São Paulo

We illustrate the methodology on the time series of the rate of hidden unemployment due to substandard work conditions in São Paulo, Brazil. The time series comes from the database of the Applied Economic Research Institute (IPEA) of the Brazilian Federal Government. The observation period is from January 1991 to November 2005, see Figure 1. The data have been previously analyzed by Rocha and Cribari-Neto (2009) by an observation-driven beta autoregressive and moving average model.

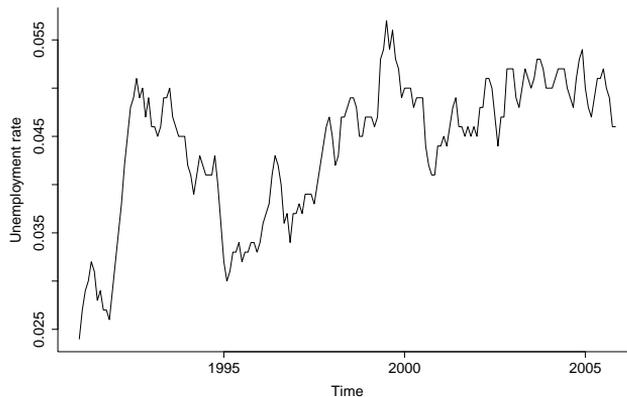


FIGURE 1. Time series of the rate of hidden unemployment in São Paulo, Brazil. Source: Applied Economic Research Institute (IPEA) from the Brazilian Federal Government.

We fit our marginal beta regression model by using the R (R Development Core Team, 2011) package `gcmr` available from the CRAN repository. This package provides a general framework for likelihood inference and model diagnostic for Gaussian copula marginal regression. For illustration, we show the results obtained under the following specification:

1. a linear trend for the mean function, $\text{logit}(\mu_t) = \beta_0 + \beta_1 \text{time}_t$, with time_t denoting the standardized time of the observation y_t ;
2. constant dispersion parameter κ ;
3. ARMA errors with orders p and q in $\{0, 1, 2\}$.

Table 1 reports the values of the Akaike Information Criterion for the models with different choices of the ARMA orders. According to this criterion, the best fitting model has ARMA(2,2) errors.

order p	order q		
	0	1	2
0	-1341.70	-1496.36	-1639.94
1	-1712.80	-1717.57	-1753.45
2	-1721.55	-1721.71	-1761.92

TABLE 1. Values of the Akaike Information Criterion for beta marginal regression models with various choices of the ARMA(p,q) errors.

Table 2 reports estimates and standard errors of all the marginal and the dependence parameters of the selected model. All the estimated parameters are strongly significant, in particular the linear trend and the ARMA coefficients.

parameter	est.	std.err.
intercept	-3.11	0.05
time	0.15	0.06
$\log(\kappa)$	6.96	0.42
ar1	0.53	0.09
ar2	0.29	0.09
ma1	0.80	0.06
ma2	0.78	0.15

TABLE 2. Estimates and standard errors for the marginal beta regression model with ARMA(2,2) errors.

In order to validate the fitted model, we check whether the quantile residuals r_t are realizations of independent standard normal variables. The normal probability plot and the autocorrelation function displayed in Figure 2 support the selected model and provide no evidence of residual serial correlation.

Finally, the one lag ahead predicted value for the percentage of hidden unemployment in December 2005 is 0.0463 with a 95% prediction interval equal to (0.0431, 0.0497).

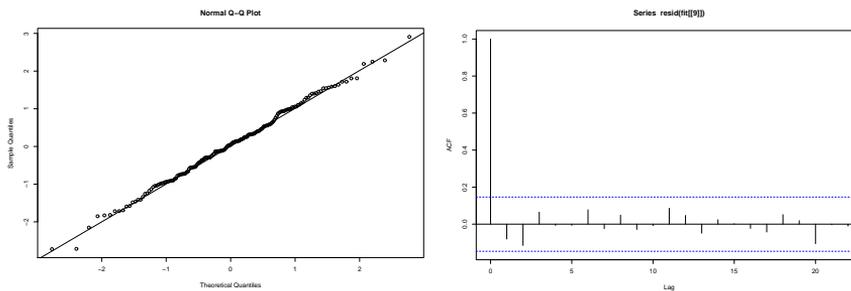


FIGURE 2. Normal probability plot (left panel) and autocorrelation function (right panel) of the quantile residuals for the fitted beta-ARMA(2,2) model.

Acknowledgments: The Authors acknowledge partial financial support from PRIN 2008 MIUR grant.

References

- Cribari-Neto, F. and Zeileis, A. (2010). Beta-regression in R. *Journal of Statistical Software*, **34**(2), 1–24.
- Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.
- Ferrari, S.L.P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799–815.
- Masarotto, G. and Varin, C. (2011). *Gaussian copula marginal regression*. Technical Report Series DAIS, 2011/9. Submitted.
- R Development Core Team (2011). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL www.R-project.org.
- Rocha, V.A. and Cribari-Neto, F. (2009). Beta autoregressive moving average models. *Test*, **18**, 529–545.
- Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, **23**, 470–472.

Variational approximations in Bayesian geoadditive quantile regression

Elisabeth Waldmann ¹, Thomas Kneib ¹

¹ Georg-August-University Göttingen, Germany

E-mail for correspondence: ewaldma@uni-goettingen.de

Abstract: The importance of quantile regression has grown rapidly in the last few years and many different extensions to the original concept of Koenker and Bassett (1978) have been made. In the Bayesian context, the extension of pure linear to more elaborate models has been managed for example by introducing different types of priors on an asymmetric Laplace error distribution (ALD), which was rewritten in order to make it useable for MCMC techniques. In this work we suggest the transfer of this concept to variational approximations (VAs) in order to accelerate inference and avoid problems with convergence monitoring.

Keywords: Bayesian Geoadditive Quantile Regression; Variational Approximations.

1 Variational Aproximations

There is a big variety of approaches in the framework of VAs. In this work, we will present the density approach, which is the most common approach in Bayesian statistics and in the corresponding literature often referred to as variational Bayes. The central idea is to minimize the Kullback-Leibler distance between the posterior $p(\boldsymbol{\theta}|\mathbf{x})$ and a product of densities $q(\boldsymbol{\theta}) = \prod_{i=1}^M q_i(\boldsymbol{\theta}_i)$, with $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M$ being an appropriate decomposition of all parameters. The distance is minimal if $q(\boldsymbol{\theta})$ is proportional to the Quasi-Full-Conditionals (QFC):

$$q(\boldsymbol{\theta}_i) \propto \exp \mathbb{E}_{\boldsymbol{\theta}_{-i}}(\log(q(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}))),$$

for all parameters $\boldsymbol{\theta}_i, i = 1, \dots, M$. This is achieved by an iterative updating approach, which works as follows:

0 Choose a starting value for $\boldsymbol{\theta}$

1 Iterate for $j \in 1, \dots, M$

$$q(\boldsymbol{\theta}_i) \propto \exp \mathbb{E}_{\boldsymbol{\theta}_{-i}}(\log(q(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i})))$$

2 Stop if a certain convergence criterion is reached

The crucial point is finding the appropriate decomposition of the vector of parameters. Not all QFCs appear to be closed form distributions, which arises the necessity of calculating the conditional expectiations numerically.

2 Bayesian Geoaddivitive 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 linear impact of a set of covariates $\mathbf{x}_j, j = 1, \dots, q$ on the τ -quantile of the conditional distribution of $\mathbf{y}|\mathbf{X}$ (where \mathbf{X} is the matrix of all \mathbf{x}_j and $\tau \in (0, 1)$ the quantile), we have to minimize a criterion different to the least squares, which accounts for the asymmetry of the idea of quantiles. This is implied by the asymmetrically weighted absolute deviations (AWAD):

$$\sum_{i=1}^n \rho_{\tau}(y_i - \mathbf{x}'_i \boldsymbol{\beta}_{\tau}) \rightarrow \min_{\boldsymbol{\beta}_{\tau}},$$

where ρ_{τ} stands for the check function:

$$\rho_{\tau}(y_i - \mathbf{x}'_i \boldsymbol{\beta}_{\tau}) = \begin{cases} \tau |y_i - \mathbf{x}'_i \boldsymbol{\beta}_{\tau}| & \text{if } y_i \geq \mathbf{x}'_i \boldsymbol{\beta}_{\tau} \\ (1 - \tau) |y_i - \mathbf{x}'_i \boldsymbol{\beta}_{\tau}| & \text{if } y_i < \mathbf{x}'_i \boldsymbol{\beta}_{\tau}. \end{cases} \quad (1)$$

Minimizing the AWAD requires linear programming techniques, which makes inference for more complex predictor functions difficult.

2.1 Asymmetric Laplace Distribution

In Bayesian inference, we do obviously need an error distribution. We use the ALD, which is defined as follows:

$$f(y|\mu, \delta, \tau) = \tau(1 - \tau)\delta \exp(-\rho_{\tau}(\delta(y - \mu))),$$

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 it is defined above is extremely hard to handle and has to be rewritten in order to be applicable in variational approximation contexts. Using a Gaussian mixture with offsets in variational approximations for the ALD was suggested by Wand et al (2010). Having auxillary parameters $\xi = \frac{1-2\tau}{\tau(1-\tau)}$ and $\sigma^2 = \frac{2}{\tau(1-\tau)}$ and a vector of weights \mathbf{w} with $w_i \stackrel{iid}{\sim} \text{Gamma}(1, \frac{1}{2})$, for $i = 1, \dots, n$, the following equation holds:

$$\mathbf{y}|\mathbf{W}, \xi, \sigma^2 \sim N\left(\boldsymbol{\eta}_{\tau} \frac{1}{2} \xi \frac{\mathbf{w}}{\delta^2}, \frac{\sigma^2}{\delta^2} \mathbf{W}^{-1}\right) \Leftrightarrow \mathbf{y} \sim \text{ALD}(\boldsymbol{\mu}, \delta, \tau),$$

where $\mathbf{W} = \text{diag}(w_i), i = 1, \dots, n$. The independence between δ and the w_i s is important for the decomposition of the posterior distribution.

2.2 Geoaddivitive Modelling

We will now use the geoaddivitive model in its generic notation:

$$\boldsymbol{\eta}_i = \sum_{j=1}^p f(\mathbf{x}_{ij})$$

where $f(\mathbf{x}_{ij})$ can denote the linear, nonlinear, spatial or random effect on individual i . After deriving the corresponding design matrices, we can also write the predictor as a sum over different terms $\mathbf{Z}_j \boldsymbol{\gamma}_j$. A prior choice, which accounts for all the different properties of the coefficients, is a possibly partially improper Gaussian distribution:

$$p(\boldsymbol{\gamma}_j | \mathbf{m}_j, \boldsymbol{\theta}_j, \delta^2) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_j - \mathbf{m}_j)' \mathbf{K}_j(\boldsymbol{\theta}_j)(\boldsymbol{\gamma}_j - \mathbf{m}_j)\right).$$

If we are, for example, looking at linear effects, the $\mathbf{m}_j = 0, \mathbf{K}_j(\boldsymbol{\theta}_j) = \delta^2 \mathbf{M}$ with some large covariance matrix \mathbf{M} . To now develop the exact algorithm, we have to calculate all the QFCs of interest:

- $q(\boldsymbol{\gamma}_j) \propto \exp \mathbb{E}_{-\boldsymbol{\gamma}_j}(\log(p(\boldsymbol{\gamma}_j | \mathbf{m}_j, \mathbf{K}_j(\boldsymbol{\theta}_j), \delta^2, \mathbf{W})))$
- $q(\mathbf{W}) \propto \exp \mathbb{E}_{-\mathbf{W}}(\log(p(\mathbf{W} | \boldsymbol{\gamma}, \mathbf{m}, \mathbf{K}(\boldsymbol{\theta}))))$
- $q(\delta^2) \propto \exp \mathbb{E}_{-\delta^2}(\log(p(\delta^2 | \boldsymbol{\gamma}, \mathbf{m}, \mathbf{K}_j(\boldsymbol{\theta}_j))))$
- $q(\boldsymbol{\theta}_j) \propto \exp \mathbb{E}_{-\boldsymbol{\theta}_j}(\log(p(\boldsymbol{\theta}_j | \boldsymbol{\gamma}, \mathbf{m}, \delta^2, \mathbf{W})))$

The derivation of the QFCs does not differ a lot from calculating classical full conditionals. The main difference is the focus on the expectation and therefore the need to take into account dependencies between the parameters. As an example see the derivation of the QFC for linear effects:

$$q(\boldsymbol{\beta}) \propto \exp\{\mathbb{E}_{-\boldsymbol{\beta}}(\log(p(\boldsymbol{\beta} | \cdot)))\}$$

$$\propto \exp\left\{-\frac{1}{2} \boldsymbol{\beta}^T \underbrace{\left(\delta_{\mathbb{E}}^2 M^{-1} + \frac{\delta_{\mathbb{E}}^2}{\sigma^2} \mathbf{X}^T \mathbf{W}_w^{-1} \mathbf{X}\right)}_{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}} \boldsymbol{\beta} + \boldsymbol{\beta}^T \underbrace{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \frac{\delta_{\mathbb{E}}^2}{\sigma^2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} (\tilde{\mathbf{y}} - M^{-1} \mathbf{m})}_{\boldsymbol{\mu}_{\boldsymbol{\beta}}}\right\}$$

Obviously we get a Gaussian type QFC, which also gives us the knowledge about the covariance of $\boldsymbol{\beta}$ which is needed for the above mentioned cases of dependencies in other QFCs. Also for the weights w_i we derive a closed

form distribution: the inverse Gaussian distribution. The error precision δ^2 however has to be calculated via numerical integration, because there is no closed form for finding its expectation. The same holds for the expectation of the square root δ which is also needed for the algorithm. The result for the linear effect also holds – with little differences in the parameters – for all other effects specified in the above described model type. With this derivation the algorithm is given by:

- 0 Choose a starting value for $\delta_{\mathbb{E}}^2, \boldsymbol{\mu}_{j_{\mathbb{E}}}$ and $\boldsymbol{\Sigma}_{j_{\mathbb{E}}}(j \text{ in } 1, \dots, J)$
- 1 Iterate
 - $\boldsymbol{w}_{\mathbb{E}}^{-1} \mid \delta_{\mathbb{E}}^2, \boldsymbol{\mu}_{j_{\mathbb{E}}}, \boldsymbol{\Sigma}_{j_{\mathbb{E}}}(j \text{ in } 1, \dots, J) \}$ Inverse-Gaussian distribution
 - for $j \text{ in } 1, \dots, J$:
 - $\boldsymbol{\Sigma}_{j_{\mathbb{E}}}^{-1} \mid \delta_{\mathbb{E}}^2, \boldsymbol{w}_{\mathbb{E}}^{-1}, \delta_{\mathbb{E}}$
 - $\boldsymbol{\mu}_{j_{\mathbb{E}}}^{-1} \mid \delta_{\mathbb{E}}^2, \delta_{\mathbb{E}}, \boldsymbol{w}_{\mathbb{E}}^{-1} \}$ Gaussian distribution
 - $\delta_{\mathbb{E}} \mid \boldsymbol{\mu}_{j_{\mathbb{E}}}, \boldsymbol{\Sigma}_{j_{\mathbb{E}}}(j \text{ in } 1, \dots, J), \boldsymbol{w}_{\mathbb{E}}^{-1}$
 - $\delta_{\mathbb{E}}^2 \mid \boldsymbol{\mu}_{j_{\mathbb{E}}}, \boldsymbol{\Sigma}_{j_{\mathbb{E}}}(j \text{ in } 1, \dots, J), \boldsymbol{w}_{\mathbb{E}}^{-1} \}$ Numerical integration
- 2 Stop if relative changes are smaller than a certain stopping criterion.

Further extensions, like regularization techniques could be imposed by putting hyperpriors on $\boldsymbol{\theta}$ and thus extending the algorithm by one step. Simulations using different types of effects have shown the superiority of the VAs over MCMC in terms of speed. The point estimations were of comparable quality, partly even better, while the covariances were consequently underestimated. The latter is typical for the VA approach.

3 Farm Efficiency

As a case study we analyzed a dataset on efficiency of farms in England and Wales. The survey at hand contains the output of farms, observed in the years 2003–2007 and different continuous covariates like for example veterinary costs or working hours, which we incorporated, after some preanalyses, partly as linear and partly as nonlinear effects. Furthermore spatial information is given on county level, which we assigned a Markov-random-field prior. To account for the dependency due to the temporal structure, we also incorporated a random effect. The resulting model hence is:

$$\boldsymbol{\eta}_{\tau} = \mathbf{X}\boldsymbol{\beta}_{\tau} + f_{\tau}(\mathbf{z}_{nl}) + f_{geo_{\tau}}(\mathbf{z}_{sp}) + \mathbf{b}_{\tau},$$

where \mathbf{X} denotes the linear, \mathbf{z}_{nl} the nonlinear and \mathbf{z}_{sp} the spatial covariates; \mathbf{b} is the random effect. As an example for the linear effects see Figure 1, where the linear effect of the veterinary costs on five different quantiles is shown on the right side. While the effect on the lower quantiles is quite high

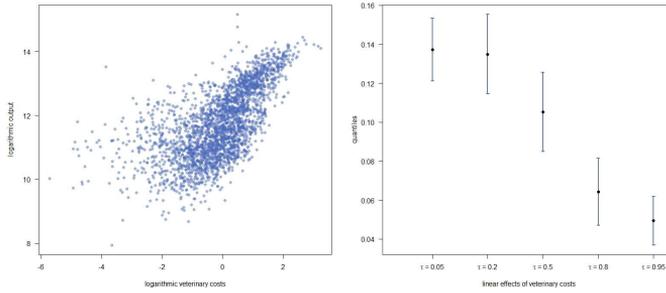


FIGURE 1. Linear effect of the veterinary costs

it decreases for the higher quantiles. This reflects exactly the pattern in the data, where the slope in the lower parts is higher than in the higher parts. Figure 2 shows the nonlinear effect of the logarithmic herdsize. A look at

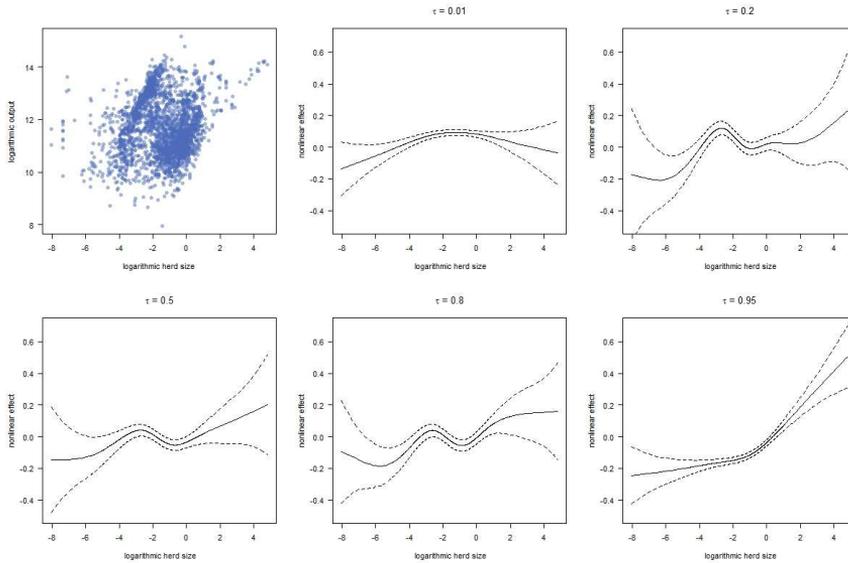


FIGURE 2. Nonlinear effect of the herd size

the scatterplot of the data (upper row, left) shows that the structure is neither linear, nor homoskedastic. This is confirmed by the analysis of the five quantiles, where the behaviour of the curve differs a lot. The last plot (figure 3) presents the spatial effect for $\tau = .05, \tau = 0.50$ and $\tau = 0.95$. Here we see that the spatial effect is higher on the higher quantiles and we get hardly any effect on the low quantiles.

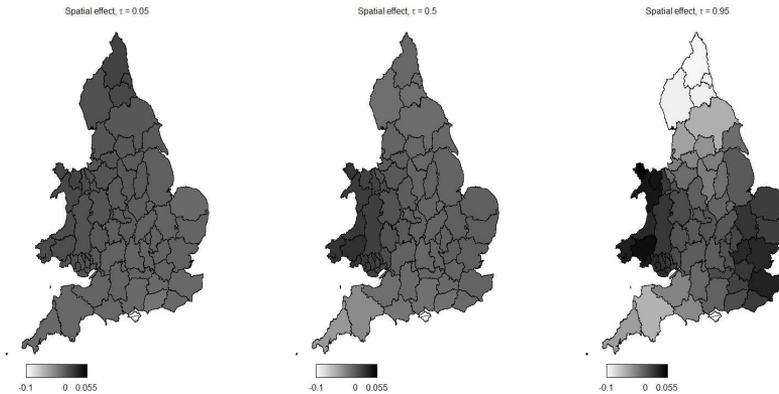


FIGURE 3. Spatial effect of the counties

4 Conclusion

VAs are an interesting alternative to MCMC techniques, in the framework of quantile regression. Especially an astonishing difference in times of speed makes the approach so competitive. In a future step it will be interesting to solve the problems with the overestimation of the precision and then add more complex options to the model like for example variable selection mechanisms like ridge or LASSO estimation.

References

- Koenker, R. and Bassett, G. (1978). Regression Quantiles *Econometrica*, **46**, 33–50.
- Ormerod, J.T. and Wand, M.P. (2010). Explaining variational approximations. *The American Statistician*, **64**, 140–153.
- Wand, M.P, Ormerod, J.T., Padoan, S. A. and Frühwirth, R.(2011). Mean Field Variational Bayes for Elaborate Distributions. *Bayesian Analysis*, **6**, Number 4, 847–900.
- Yue, Y. and Rue, H. (2011). Bayesian Inference for additive mixed quantile regression models. *Computational Statistics and Data Analysis*, **55**, 84–96.

Identifying underlying structure in classification and regression trees using surrogate splits

Deirdre Wall¹, Carl Scarrott², John Newell^{1,5}, Helen Ingoldsby³, Grace Callagy³, Michael J. Kerin⁴

¹ School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway Ireland

² University of Canterbury, Christchurch, New Zealand

³ Discipline of Pathology, NUI Galway, Ireland

⁴ Discipline of Surgery, NUI Galway, Ireland

⁵ HRB Clinical Research Facility, NUI Galway, Ireland

E-mail for correspondence: d.wall3@nuigalway.ie

Abstract: A surrogate split is a split which most accurately predicts the action of s^* , the best split at a node of a tree. In any tree based model there are competing trees with comparable predictive power. Such candidate trees can be found using surrogate splits. In this paper, a Surrogate plot is introduced as a useful visual tool to identify the surrogate splits and their competing trees. An application to Breast Cancer data to predict Oncotype DX classification is used to show the usefulness of surrogate splits.

Keywords: CART; Surrogate Splits.

1 Introduction

Tree based models are useful for classification and prediction problems, where a graphical representation of the underlying structure is presented. Breiman et al (1984) suggests recursive partitioning the data. At any given node the best split s^* is chosen and the data is partitioned into subsets using this criterion. The next split for each of these subsets are found and the data is partitioned again. This continues until some stopping criteria is reached. These trees are usually over-fitted and then pruned back if the extra splits do not improve the misclassification error. Alternatively, trees can be pruned using conditional inference, Hothorn et al (2006), where significance test procedures are used as stopping criteria.

Ensemble methods, such as Random Forests, have been shown to provide improved prediction. Random forests work by growing a large number of trees from resampled data and a prediction is obtained by averaging over all the predicted responses for each tree. The trees grown are not pruned. The

advantage of random forests is a reduction in prediction error however no information relating to the underlying structure of the variables is available. In practice there are often several candidate trees that have comparable prediction power which identify different sets of risk factors (predictors). Surrogate splits are typically used for handling missing data in trees but alternatively, in this paper, they can be used to identify potential candidate trees. An example fitted to breast cancer data where conflicting results in previous research can be consolidated by examining surrogate splits.

2 Surrogate Splits

At any given node in a Classification and Regression Tree (CART), the best split s^* is chosen (i.e. the split which decreases the impurity most). The node impurity is largest when all classes are equally mixed together and the smallest when the node contains only one class, Breiman et al (1984). A surrogate split is a split which most accurately predicts the action of s^* . There are two types of surrogates, primary and secondary. Primary surrogates are the splits with a similar performance in impurity to the best split. Secondary surrogate splits resemble the best split in terms of the number of cases they send the “same way” and are used to handle missing data.

Using the primary surrogate splits, competing trees can be created which may have similar performance or structure to that of the original tree. These trees may identify hidden structure using other key risk factors not identified in the original tree but have comparable prediction power.

3 Case Study: Predicting Oncotype DX Classification

Oncotype DX is an expensive patented test that analyses 16 genes in patients with Oestrogen Receptor positive and Lymph Node negative breast cancer. It assigns each patient with a *Oncotype DX Recurrence Score (RS)*, which are categorized low risk, intermediate risk and high risk. The higher the risk means a higher likelihood of recurrence.

Treatment after surgery is dependent on the Oncotype DX classification. If a patient is categorized as low risk of recurrence only Tamoxifen is required, however if a patient is classified as intermediate or high risk Tamoxifen and Chemotherapy is required.

Many people believe that the results of Oncotype DX can be predicted just as well by routinely (and more cheaply) assessed pathological variables and biomarkers. Previous published studies in this area have used tree based models and each of the papers have identified different set of risk factors, namely

- Grade, progesterone receptor status and Ki67 level, Allison et al (2011)

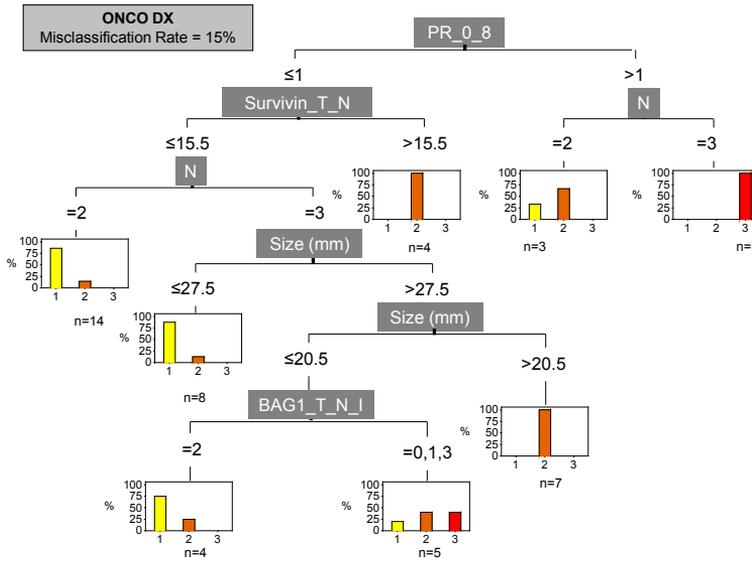


FIGURE 1. Classification tree for the Galway Oncotype DX data.

- Mitotic score (M) greater than 1 combined with a negative progesterone receptor result, Flanagan et al (2008)
- Tubule formation (T), nuclear grade (N), mitotic count (M), oestrogen receptor score, progesterone receptor score and Her2/neu score, Auerbach et al (2010)
- Oestrogen receptor score, progesterone receptor score, Her2/neu score and the 3 components of grade (T, N and M), Geradts et al (2010)

This has lead to considerable debate. The Galway (West of Ireland) dataset contains 52 patients with their Oncotype DX score and 32 useful clinical and pathological variables.

4 Recursive Partitioning Using RPART

The `rpart` library builds classification or regression models of a very general structure using a two stage procedure; the resultant models can be represented as binary trees, Therneau and Atkinson (1997).

The `rpart` tree for the Galway Oncotype DX data given in Figure 1. Some of the variables from previous studies also appear in this tree, such as PR and N. However this tree is a fixed tree and was modeled using a small sample size so it would not be recommended to use this as a predictive model.

There may be more underlying structure if we look at the surrogate splits. Surrogates are useful for identifying variables, which may not appear in the original tree structure, but if the first split is removed from the model, the variable may appear and have a similar tree structure. The idea of an interactive surrogate plot will help identify underlying structure using the surrogate splits.

5 Surrogate Plot

A surrogate plot for the Oncotype DX data is contained in [Figure 2](#). The x-axis is divided into three parts, nodes (the number of times the variable appears as a split in the original tree), primary surrogates and secondary surrogates for the different nodes in the original tree. All the variables in the dataset are given along the y-axis and have been ordered in terms of importance, with the most important risk factors close to the top and the risk factors which do not seem to be as important are closer to the bottom. Five primary surrogates were identified for each of the nodes in the tree. A heat-map is used to compare the predictive power of each surrogate to the best split.

For example, for the first split in the Galway Oncotype DX tree in [Figure 1](#) is $PR(0-8)$, however the surrogates for this split include *Survivin T N*, *Ki67*, *Bcl2*, *CD68* and *N*. If one of these surrogates is selected, lets say *N*, the resultant tree for this surrogate is given in [Figure 3](#). The first split is on the surrogate selected, *N*, then a tree is grown using all the variables for each partition of the parent node. This tree has a misclassification rate of 15%, the same as the original tree, and contains some of the predictors from the paper by Geradts et al (2010). If a tree is created for surrogate *CD68*, similar risk factors to that of Auerbach et al (2010) appear in the tree. These trees have identified more potentially useful variables and biomarkers but they have comparable predictive power to that of the original tree.

This analysis identified risk factors using surrogate splits which consolidated results from previous studies of Oncotype DX classification. Using the tree in [Figure 1](#) as a prediction model on an individual level as not very sensible (due to small sample size). However, these risk factors can now be used on a much larger sample to build a better prediction rule.

6 Conclusions

Surrogate splits may be the equivalent to Best Subsets in regression modeling, as it identifies other trees with comparable prediction error. The surrogates are useful for identifying key risk factors that are missed in the original tree such as *Ki67* and *Survivin T N*. The surrogate plot identifies other comparable and competing trees which in turn identifies other potentially useful variables and underlying structure but with comparable

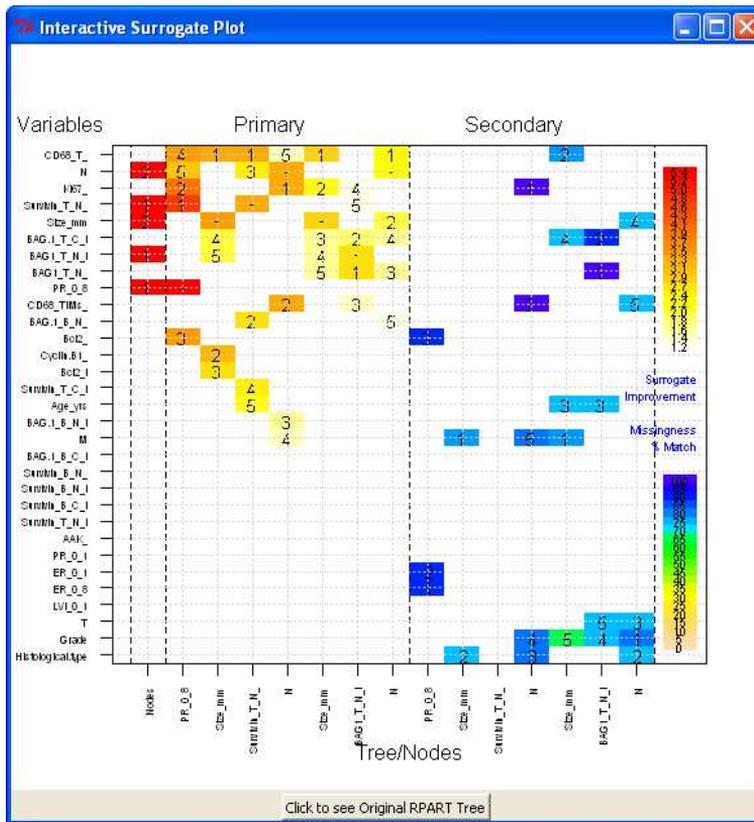


FIGURE 2. Surrogate Plot for the Classification tree of the Galway Oncotype DX data.

predictive power. These surrogates were used to create competing trees which consolidated the results from the previous research of Oncotype DX classification.

Acknowledgments: The first author is grateful to NBCRI (National Breast Cancer Research Institute) for their continued funding of postgraduate students.

References

Allison, K.H. et al(2011). Routine pathologic parameters can predict Oncotype DX recurrence scores in subsets of ER positive patients: who does not always need testing? *Breast Cancer Res Treat.* **131**,413-24.

Auerbach, J. et al(2010). Can features evaluated in the routine Pathologi-

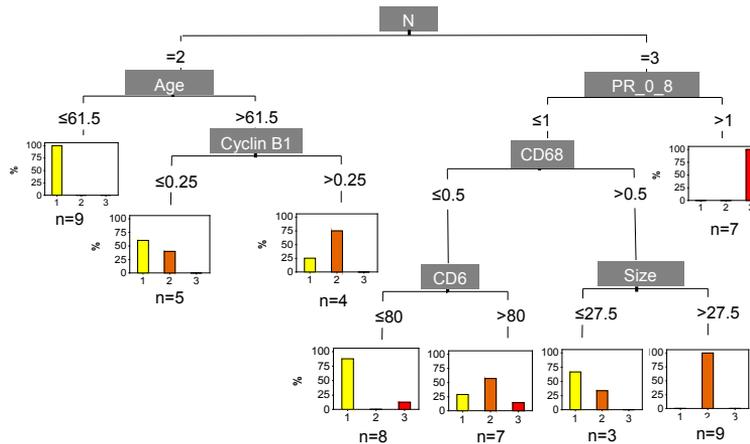


FIGURE 3. Tree for Surrogate N.

cal Assessment of Lymph Node-Negative Estrogen Receptor-Positive Stage I or II Invasive Breast Cancer be used to predict the Oncotype DX Recurrence Score. *Arch Pathol Lab Med.* **122**(4),731-736.

Breiman, L. et al(1984). *Classification and Regression Trees*, Chapman and Hall.

Flanagan, M.B. et al(2008). Histopathologic variables predict Oncotype DX Recurrence Score. *Modern Pathology.* **21**(10), 1255-1261.

Geradts, J. et al(2010). The Oncotype DX Recurrence Score is Correlated with a composite index including routinely reported Pathobiologic Features. *Cancer Invest.* **28**(9),969-977.

Hothorn, T. et al(2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics,* **15**(3), 651-674.

Therneau, T., Atkinson, E.(1997). An Introduction to Recursive Partitioning: Using the RPART Routines. *Technical report, Mayo Clinic Section of Biostatistics.* **No.61.**

Assessing surrogacy of progression free survival for overall survival: A multi-state model approach

Jixian Wang¹

¹ Novartis Pharma AG, Switzerland

E-mail for correspondence: jixian.wang@novartis.com

Abstract: Progression free survival (PFS) is often used as a surrogate endpoint for overall survival (OS) in early line Oncology trials. General approaches to measuring surrogacy such as Kendall's tau and linear model based measures may not be appropriate. We follow the practical approach of using Cox model for analyses of PFS and OS data and propose using concordance between the treatment effect test results for OS and PFS as a measure for surrogacy. We propose a Bayesian bootstrap approach to estimate the measure and apply to a colon cancer dataset as an example. We also examine the properties of the proposed measure using the classical three-state model via simulation.

Keywords: Multi-state models; Overall survival; Progression free survival; Surrogate endpoint

1 Introduction

In Oncology clinical trials, Overall survival (OS), the time from randomization/first treatment to death from any cause, is often used as the primary endpoint for regulatory approval. OS as an endpoint is clinically meaningful and objectively assessed. However, trials using OS endpoints often require large trial size with longer period of follow-up and are often confounded by subsequent therapy use. Surrogate endpoints such as progression-free survival (PFS) can be extremely beneficial if they provides reliable prediction for OS (Burzykowski, Molenberghs and Buyse, 2005). PFS requires less number of patients and shorter follow-up and is in general not impacted by subsequent lines of therapies. Let S_i and T_i be the surrogate and true endpoint variables from patient i , many works on assessing surrogates are based on model:

$$S_i = \mu_s + \beta_s Z_i + e_{si} \quad (1)$$

$$T_i = \mu_t + \beta_t Z_i + e_{ti} \quad (2)$$

where $\mu_{s(t)}$, $\beta_{s(t)}$ and $e_{s(t)i}$ are the intercept, treatment effect and error term for $S_i(T_i)$ and Z_i is a binary treatment indicator. When e_{si} and e_{ti}

are jointly normally distributed, one can predict T_i conditional on S_i with

$$T_i = \gamma_0 + \beta Z_i + \gamma S_i + e_{ti} \quad (3)$$

where β is the treatment effect on T_i after adjusting for S_i . Two measures proposed based on model (2) are Freedman's proportion (of treatment effect) explained: $PE = (\beta_t - \beta)/\beta_t$ and Buyse and Molenberghs' relative effect (RE) β_s/β_t . A general model-independent measure is Kenall's τ

$$\tau = P((T_i - T_j)(S_i - S_j) > 0) - P((T_i - T_j)(S_i - S_j) \leq 0). \quad (4)$$

Model (2) may not be appropriate for survival analysis for time-to-event endpoints such as OS and PFS. PFS is measured by the time to either disease progression (DP) or death, whichever occurs earlier; the relationship between PFS and OS can be described by a model with three states: stable disease, DP and death (Figure 1). However, in practice PFS and OS are often analysed using the Cox model to test the hazard ratio (HR) of a treatment over the control. Therefore, consistency between treatment effect tests on PFS and OS is of great interest to both health authorities as well as pharmaceutical industry. We follow this practical approach and propose a model-independent surrogacy measure using the idea of Kenall's τ for consistency between the PFS and OS tests. Impact of model parameters on this measure is evaluated using simulation from the three-state model. A real-data example is provided using a dataset on colon cancer.

2 Three-state model for PFS and OS

The relationship between PFS and OS can be described by the 3-state model (Figure 1), with transit times D_i , Y_i and U_i . Based on the transit times the PFS time (the shortest time to either DP or death) is $S_i = \min(D_i, Y_i)$ and the OS time is $T_i = \min(D_i + U_i, Y_i)$. For the transit times we consider a multivariate accelerated failure time (AFT) model although other time-to-event models can be used too. Let $\mathbf{R}_i = (\log(D_i), \log(Y_i), \log(U_i))$, the model can be written as

$$\mathbf{R}_i = \mu + \beta Z_i + \mathbf{E}_i \quad (5)$$

where μ are the log-mean transit times, β and \mathbf{E}_i are their treatment and random components. This model is more flexible than the independent exponential model Fleischer et al (2009) used. Dejardin et al (2010) used a 2-state model, assuming no death without DP so that some interesting results can be derived. In general model (5) does not lead to model (2), nor to the Cox model.

3 New surrogacy measure based on Cox regression

Follow the concept of Kendall's τ and the Cox model based analysis, we propose using concordance and discordance between the treatment effect tests on OS and PFS as a surrogacy measure. Specifically, suppose that $z_s(z_t)$ is the test statistic for relative risk reduction in PFS (OS) with rejection region $z_s(z_t) < c_0$, following Kendall's τ we use

$$\phi_c = P((z_s - c_0)(z_t - c_0) > 0) - P((z_s - c_0)(z_t - c_0) \leq 0) \quad (6)$$

where the first term is the probability of concordance and the second one the discordance between the PFS and OS tests. Note that c_0 needs not to be the same for OS and PFS. In a similar way, we can define the false positive (FP) and false negative (FN) rates of predicting OS results by PFS as

$$FP = P(z_s < c_0 \cap z_t \geq c_0 | z_s < c_0) \quad (7)$$

$$FN = P(z_s \geq c_0 \cap z_t < c_0 | z_s \geq c_0). \quad (8)$$

Although FP and FN are not direct measures of surrogacy, it is of interest to explore their relationships with ϕ_c .

In general it is impossible to calculate ϕ_c , FP and FN analytically with given parameters in the three-state models, but simulation for this purpose is simple and straightforward. To this end, we generate data from the three-state model and fit the Cox model repeated for each set of samples. Then ϕ_c can be estimated by the percent of concordance pairs between tests for PFS and OS. In this way one can evaluate the impacts of model parameters to ϕ_c , FP and FN , as shown in the next section.

For a given dataset, we can estimate ϕ_c with Bayesian bootstrap (Rubin, 1981), in which each patient is weighted by a random variable following the standard exponential distribution, then PFS and OS data are fitted to the Cox model separately. These steps repeat many times with a new set of weights applied each time. ϕ_c can be estimated using the percent of concordance pairs of the bootstrapped tests.

4 How model parameters affect surrogacy?

Due to space limit, we only report a small part of results for the impact of parameter changes on ϕ_c , FP and FN . We took model (5) and assumed $\mathbf{E}_i \sim N(0, \Sigma)$ with equal correlation ρ , $\sigma^2 \equiv \text{var}(\log(D_i)) = \text{var}(\log(Y_i))$ but $\text{var}(\log(U_i)) = \sigma^2 C_u^2$ with varying C_u . The means of $\log(D_i)$ and $\log(U_i)$ was set as 0 and -1, respectively, assuming death after DP was much quicker. We assumed no treatment effect on Y_i and equal effects β on U_i and D_i . For a number of combinations of different parameter values we calculated ϕ_c , FP and FN with 5000 simulations each (Table 1). ρ only

TABLE 1. Impact of model parameters on ϕ_c , FN and FP (n is sample size).

n	ρ	σ	β	$E(\log(Y_i))$	C_u	ϕ_c	FP	FN
30	0.5	0.35	0.5	2	2	0.595	0.245	0.027
30	0.6	0.35	0.5	2	2	0.574	0.282	0.007
30	0.8	0.35	0.5	2	2	0.573	0.338	0.000
30	0.5	0.35	0.3	2	2	0.685	0.352	0.021
30	0.5	0.35	0.7	2	2	0.779	0.109	0.155
30	0.5	0.35	0.5	1	2	0.476	0.324	0.010
30	0.5	0.35	0.5	2	3	0.207	0.491	0.007
30	0.5	0.35	0.5	2	1	0.893	0.020	0.191
50	0.5	0.35	0.5	2	1	0.970	0.004	0.257
20	0.5	0.35	0.5	2	1	0.832	0.041	0.162
30	0.5	0.20	0.5	2	2	0.910	0.026	0.782
30	0.5	0.50	0.5	2	2	0.613	0.353	0.025

had a slight impact on ϕ_c , an increase in sample size n or β , or a decrease in σ or C_u (hence σ_u) led to significant increasing ϕ_c , while an increase in $E(\log(Y_i))$ led to a decrease in ϕ_c , all as expected. However, quantitative results for ϕ_c , FP and FN will be useful in trial design when a specific scenario is given.

5 Application to colon cancer data

We consider a dataset on colon cancer patients treated with Lev or Lev+5FU and examine the reliability of treatment comparison based on PFS (for which we treat tumor recurrence as DP) to predict that of OS. The analysis uses Cox models to test the hazard ratio of treatments Lev+5FU to Lev. To calculate ϕ_c , we generated 5000 sets of weights from the standard exponential distribution. For each weight set, the Cox models were fitted to the weighted OS and PFS data and tests for RR was carried. The z-values for the tests were presented in Figure (2). ϕ_c was calculated by the mean of concordance and discordance pairs in the tests and $\phi_c=0.775$ was obtained, suggesting PFS being a good surrogate for OS in this situation. Figure (3) shows the FP and FN rates with different cutoff points for the PFS test.

6 Discussion

In Oncology trials, whether PFS is a valid surrogate for OS remains a very relevant clinical question that have huge impact on trial design and eventually drug approval. Current standard definitions of surrogacy like Presence's are difficult to apply in trials involving PFS and OS because it

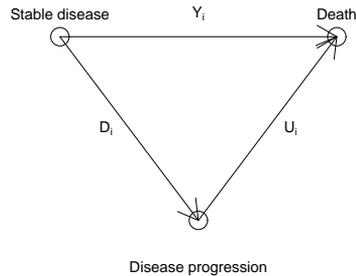


FIGURE 1. A three-state model for stable disease, disease progression and death.

is difficult to model the correlation between these two time-to-event endpoints. In addition, standard definitions may be difficult to interpret. In this paper, we proposed a different definition of surrogacy that is tailored to handle time-to-event endpoints in cancer trials. The definition is based on the concordance and discordance of testing results on PFS and OS effects. This definition can be readily used based on standard Cox regression model. Unlike other definitions, it is also easy to interpret since it answers a very direct clinical question on the concept of surrogacy, that is, how often a PFS testing conclusion will lead to the same conclusion on OS.

References

- Burzykowski T, Molenberghs G, Buyse M. (2005). *The Evaluation of Surrogate Endpoints*. Springer: Heidelberg.
- Dejardin D, Lesaffrea E, Verbeke G. (2010). Joint modeling of progression-free survival and death in advanced cancer clinical trials. *Statist. Med.* **29**, 1724–1734.
- Fleischer F, Gaschler-Markefski B, Bluhmki E. (2009). A statistical model for the dependence between progression-free survival and overall survival. *Statist. Med.* **28**, 2669–2686.
- Rubin D.B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9**, 130–134.

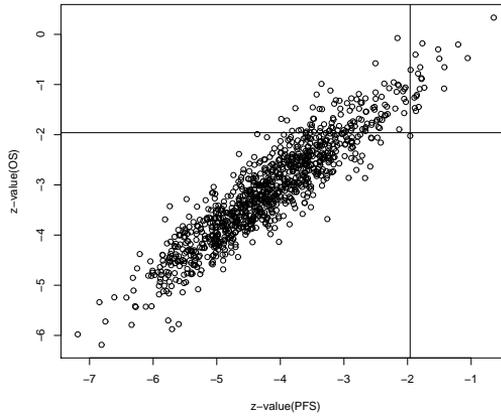


FIGURE 2. Bayesian bootstrapped z-statistics for testing RR of Lev+5FU to Lev, using PFS and OS colon cancer data.

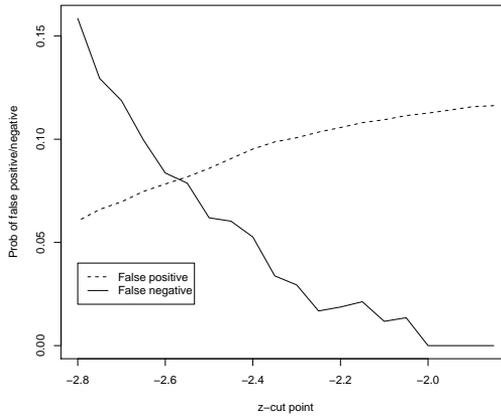


FIGURE 3. False positive and false negative rates as a function of cutoff point for the PFS z-statistic. The cutoff point for OS z-statistic was set at 1.96.

Assessment of model stability for high-dimensional data with applications to complex genomic data

Susan R. Wilson¹

¹ School of Mathematics & Statistics, Faculty of Science, and Prince of Wales Clinical School, Faculty of Medicine, University of New South Wales, Australia

E-mail for correspondence: sue.wilson@anu.edu.au

Abstract: An approach is outlined to model fitting for high-dimensional data, including assessment of the model's stability. An application to a mice obesity data set is given. Some comparisons with alternative approaches to simulations based on real data also are overviewed.

Keywords: Model Stability; High-dimensional Data; Complex Genomic Data; Statistical Modelling.

1 Introduction

One of today's major and outstanding statistical challenges is to determine the quality and robustness of models fitted to high-dimensional data with limited numbers of samples. Collection of high-dimensional data is becoming increasingly easier, and hence more common, even routine in many scientific areas. For example, in molecular biology for each sample one can collect millions of single nucleotide polymorphisms (SNPs), gene expression values and protein concentrations, as well as other (say clinical) covariates. So far, the number of samples is limited to at most a few thousand, but is typically much smaller (hundreds). A wide variety of aims can be the focus of such studies, such as discrimination between diseased and not diseased samples (classification), or prediction of an outcome, say time to some event, based on a selection from the high-dimensional data. Applications to other research areas can be found in Fan & Lv (2010).

Many different approaches have been proposed for fitting multivariate models to such "large p - small n " data. Arguably the most popular are those based on penalisation such as the lasso (Tibshirani, 1996) and its many variations like the hyperlasso (Hoggart et al, 2008), the elastic net (Zhou & Hastie, 2005) and sure independence screening (Fan & Lv, 2010). Although these methods have nice theoretical properties under certain conditions, in practice their performance is less satisfactory. For example, variable selection is highly unstable (Meinhausen & Bühlmann, 2010) and prediction

rules may not be able to be validated on new data (McCarthy et al, 2008). The reasons underpinning these problems include non-homogeneous sample populations, low signal to noise ratios, and unmeasured true covariates or covariates that are only correlated with the true covariates.

It is an open problem regarding the best way to analyse such data. We have introduced a three-step procedure for simultaneous analysis of all covariates, wherein model selection is performed in the first two steps and model stability in the third step. This is outlined in the following section. The following two sections summarise the application of the approach in two different settings.

2 Models and Methods

The proposed three-step procedure involves (i) variable selection in a multivariate setting where there are p covariates typed in n individuals, where $p \gg n$, in such a way as to choose a reasonable number of candidate covariates for step 2; (ii) stepwise regression to select from the covariates found at step 1; (iii) Model stability evaluation using resample model averaging. Since our examples involve continuous phenotypes (response variables), for simplicity an outline for the linear regression modelling framework is given, noting that extensions to other generalised linear models, particularly to logistic regression for a binary (e.g. diseased/non diseased) response, is straightforward.

For the multiple linear regression model

$$y_j = \beta_0 + \sum_{i=1}^p \beta_i x_{ij} + \epsilon_j$$

where y_j is the phenotypic value and x_{ij} is the i th covariate value for the j th individual, ϵ_j the corresponding residual, and the β s are to be estimated. To prevent model over fitting a regularisation method that maximises a penalised form of the likelihood is used. Both the lasso (Tibshirani, 1996) and hyperlasso (Hoggart et al, 2008) were considered; see also Motyer et al (2011a).

Estimates of the regression coefficients, $\hat{\beta}^\ell$, are given by

$$\hat{\beta}^\ell = \arg \max_{\beta} [L(\beta) - f_\ell(\beta; \xi)],$$

where L denotes the log-likelihood, namely

$$L(\beta) = \sum_{j=1}^n \left(y_j - \beta_0 - \sum_{i=1}^p \beta_i x_{ij} \right)^2 + \text{constant},$$

and f_ℓ is a penalty function with a smoothing parameter ξ (giving the maximum likelihood estimates when zero). For the lasso, the penalty function

is given by $f_\ell(\boldsymbol{\beta}; \xi) = \xi \sum_{i=1}^p |\beta_i|$, while for the hyperlasso it is

$$f_\ell(\boldsymbol{\beta}; \lambda, \gamma) = - \sum_{i=1}^p \left(\frac{\beta_i^2}{4\gamma^2} + \log D_{(-2\lambda-1)} \left(\frac{|\beta_i|}{\gamma} \right) \right),$$

where there are now two smoothing parameters λ , γ , and D is the parabolic cylinder function.

From a Bayesian viewpoint, it can be shown that the hyperlasso is a generalisation of the lasso in the prior densities associated with the regression coefficients, and corresponds to a strong prior belief that there are few true causal variants and little prior knowledge of effect size; see Hoggart et al (2008). A disadvantage of the hyperlasso compared to the lasso is that it does not have a closed form solution, and so is very computationally intensive. Values of the relevant smoothing parameters were determined by 10-fold cross validation based on the mean squared prediction error following the first two steps. The hyperlasso favours solutions that are more sparse but having less shrinkage of the variables included in the model compared with the lasso, but at extra computational cost.

In the genomic setting, our approach is only suited to common variants, so SNPs with a low minor allele frequency need to be filtered out.

For the stability analysis, step (iii), the form of resample model averaging known as “bagging”, namely sampling with replacement, is performed. For computational expediency, we suggest 100 resamples, of size equal to the number of individuals in the analysis, with smoothing parameters identical to those found in the initial model selection procedure. The proportion of resamples for which each covariate is selected, namely the resample model inclusion probability (*RMIP*), is plotted. If the covariates selected in the model are those with the highest *RMIP*, then it is indicative that the model is stable under perturbations of the data; pragmatically we have found $RMIP \geq 0.5$ works well.

3 Mice Obesity Data

First, we consider the experimental data that have been analysed previously (Wang et al, 2006, Ghazalpour et al, 2006) where each SNP was evaluated separately. The data are an F2 cross, and we just consider the female mice here, and assume the SNPs follow an additive model. The quantitative trait is abdominal fat mass at 24 weeks. Following quality control there are 151 mice and 1180 SNPs that are reduced to 250 after filtering; see Motyer et al (2011a). Restricting the number of SNPs found at step (i) to 15, the lasso and hyperlasso had 12 in common. If at step (i) we had just considered each SNP separately and then selected the top 15 (as is common practice for such data) then only 6 of these “top” 15 were selected by either lasso or hyperlasso. In the final model for the lasso,

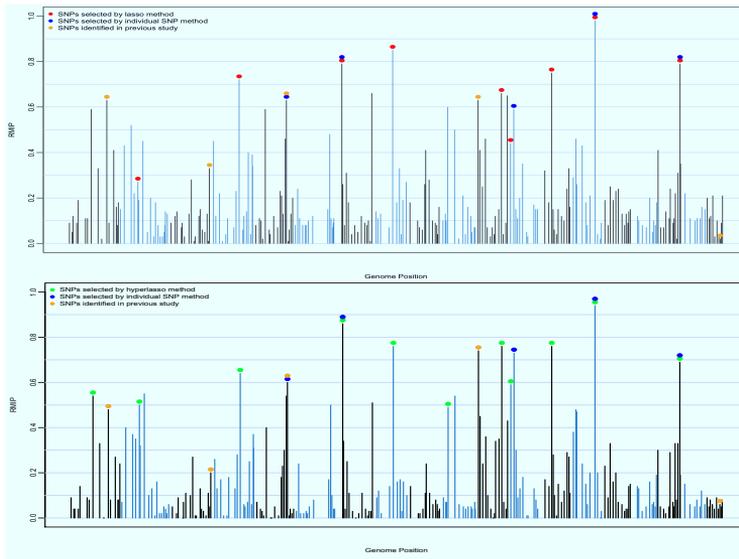


FIGURE 1. $RMIP$ for each SNP from bagging plotted against SNP genome position, with chromosomes plotted in alternative colours. For the lasso (upper figure) SNPs selected are coloured red, for the hyperlasso (lower figure) selected SNPs are coloured green; in both figures SNPs are shown for the individual method (coloured blue) and for the previous study results (coloured yellow).

9 covariates (on chromosomes 2,4,7,8,11,12,13,14,17) were selected and for the hyperlasso 11 covariates selected were selected, the additional 2 being on chromosomes 1 and 10), but only 5 for the single-SNP model (on chromosomes 5,7,12,14,17). Comparing the residual standard error estimates, R -squared values and p -values showed that both the lasso and hyperlasso approach outperform the single-SNP approach, with the hyperlasso having a slight advantage over the lasso. The SNPs selected using the lasso and hyperlasso are not necessarily the SNPs with the strongest association with phenotype; SNPs that should be included in a multi-SNP model may be overlooked if an initial screening of individual SNPs is relied on. Even if we had retained the top 70 SNPs for the stepwise analysis, we would have missed some of the SNPs found by lasso and hyperlasso.

The $RMIP$ for each SNP against genome position is plotted in Figure 1. The plot shows that the majority of the SNPs in the lasso and hyperlasso models have high $RMIP$; note that the 11 SNPs identified using hyperlasso (in green, lower figure) all have $RMIP$ values of around 0.5 or higher, while for the 9 SNPs identified using lasso (in red, upper figure), two (on chromosomes 2 and 12) had values less than 0.5.

4 Simulations based on the 1000 Genomes Project

The biennial Genetic Analysis Workshop (GAW) is a forum for evaluation and comparisons of novel and established statistical modelling methods. The most recent (GAW17) was composed of exome scan data from the 1000 Genomes project, with simulated phenotypes (Ghosh et al, 2011). We analysed the 697 unrelated individuals. There were 24,487 autosomal SNPs in 3,205 genes. Unlike most groups we chose to *not* know the answers before initial analyses. This is in stark contrast to most published simulations that are generally selected in such a way as to show how “well” a new method performs. We subsequently compared our results with the simulating model; see Motyer et al (2011b).

First we used one replicate (of the 200 provided), and one quantitative trait as the response. The non-SNP covariates (Sex, Age and Smoke) were included as unpenalised terms in our model. Filtering out minor alleles left 6,321 SNPs and given this large number, for computational expediency, it was decided to just apply the lasso. We estimated *RMIP* using 100 resamples, and fixed penalty parameter determined from the data. Our initial results performed comparably with those from groups that knew the answers before initial analysis.

In the simulating model only seven SNPs were common variants. We found that the value of the tuning parameter that minimised the CV error worked better than larger values, and AIC arguably better than BIC. Using our approach, we found that only three SNPs had *RMIP* > 0.5. Using all 200 replicates of the simulations we found that after the lasso step, six of the seven SNPs were selected more often than any other SNPs that were not in the simulating model. The three SNPs we consistently found were in the same gene and had the largest effect size in the simulating model. The seventh SNP that was most difficult to identify had a relatively small effect size and the lowest minor allele frequency. Our results reflect the difficulties associated with insufficient sample sizes, and the resultant tradeoff between false negatives and false positives.

Acknowledgments: The research is supported in part by the National Health & Medical Research Council Grant 525453. The figure was prepared by A. Motyer.

References

- Fan, J., and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101–148.
- Ghazalpour, A., et al. (2006). Integrating genetic and network analyses to characterise genes related to mouse weight. *PLoS Genetics*, **2**, e130.

- Ghosh, S., et al. (2011). Identifying rare variants from exome scans: the GAW17 experience. *BMC Proceedings*, **5**, Suppl 9:S1.
- Hoggart, C., et al. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, **4**, e1000130.
- McCarthy, M.I., et al. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews*, **9**, 356–369.
- Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society, Series B*, **72**, 417–473.
- Motyer, A., Galbraith, S. and Wilson, S.R. (2011a). Model selection procedures for high-dimensional genomic data. *ANZIAM Journal*, **52**, C710-726.
- Motyer, A., McKendry, C., Galbraith, S. and Wilson, S.R. (2011b). LASSO model selection with post-processing for a genome-wide association study data set. *BMC Proceedings*, **5**, Suppl 9:S24.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.
- Wang, S., et al. (2006). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, **2**, e15.
- Zhou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.

Author Index

- Šiška, Juraj, 573
- Abad, Ariel Alonso, 395
Adelfio, Giada, 401
Aerts, Marc, 63, 123, 649
Allepuz, Alberto, 717
Altieri, L., 441
Amorim, Leila D. A. F., 407, 495
Andrinopoulou, Eleni-Rosalina, 45
Araújo, Artur Agostinho, 643
Aregay, Mehreteab Fantahun, 99
Arostegui, Inmaculada, 419
- Babanezhad, Manoochehr, 413
Bailey, Trevor C., 435
Barceló, Maria Antònia, 717
Barrio, Irantzu, 419
Bartolucci, Francesco, 51
Bastow, Zachary A., 177
Biggeri, Annibale, 717
Birch, A. Nicholas E., 613
Bloice, Marcus, 303
Borges, Ana, 425
Boscaino, Giovanni, 401
Bowman, Adrian W., 519, 637
Breitner, Susanne, 491
Briz, Teodoro, 501
Burke, Kevin, 431
- Caballero-Águila, R., 513
Cadarso-Suárez, Carmen, 583
Callagy, Grace, 369
Calle, M. Luz, 3
Camarda, Carlo Giovanni, 57
Capursi, Vincenza, 401
Castro-Sánchez, Amparo Yovanna, 63
- Catelan, Dolores, 717
Chen, Ming H., 607
Che Him, Norziha, 435
Christian, Nicholos J., 147
Cobre, Juliana, 69
Cocchi, D., 441
Colosimo, Enrico A., 75, 459, 471, 525
Conde, Susana, 81
Cools, Mario, 263
Cordeiro, Gauss M., 747
Costa, Marco, 447, 531
Crujeiras, Rosa M., 257
Currie, Iain D., 87
Cuzick, Jack, 777
Cyrus, Josef, 491
- Daoudi, Jalila, 455
Deasy, William, 613
Del Fava, Emanuele, 99
Dendale, Paul, 111
de Castro, Mário, 69
de Falguerolles, Antoine, 93
de Kort, Wim, 279
de Oliveira, Maristela Dias, 459
de Sousa, Bruno, 501
Dimitriou-Fakalou, Chrysoula, 465
Domma, Filippo, 105
Duarte, Denise, 471
Durbán, María, 197
- Eckley, Idris A., 141
Efendi, Achmad, 111
Eilers, Paul H. C., 11, 57, 135, 309, 315
Einbeck, Jochen, 117, 595
Elian, Silvia N., 723

- Ender, Manuela, 479
 Ensoy, Chellafe, 123
 Evers, Ludger, 117, 637
 Eze, Jude, 485
- Faes, Christel, 123, 649
 Feist, Michael, 621
 Fenske, Nora, 209
 Fensterer, Veronika, 491
 Fiaccone, Rosemeire, 495
 Filipe, Patrícia A., 501
 Finazzi, Francesco, 129
 Fox, Jean-Paul, 351
 Frasso, Gianluca, 135
 Freitas, Adelaide, 507
 Fulé, Peter Z., 177
- Gámiz-Pérez, M. L., 577
 Gómez, Guadalupe, 297
 Gallastegui, Inmaculada, 655
 Gampe, Jutta, 57
 García-Garrido, Irene, 513
 Ghosh Mukherjee, Kathakali, 519
 Giampaoli, Viviana, 327
 Gilardoni, Gustavo L., 75, 459, 525
 Gillespie, Gordon, 621
 Giordano, Sabrina, 105
 Gomes, Dulce, 501
 Gonçalves, A. Manuela, 447, 531
 Gott, Aimee N., 141
 Grilli, Leonardo, 51
 Grissoto, Laura, 717
 Gu, Jianwei, 491
 Guolo, Annamaria, 357
- Ha, Il Do, 147
 Haines, Linda M., 275
 Hainy, Markus, 537
- Hasso, Sargon, 601
 Heinzl, Felix, 153
 Heller, Gillian Z., 159, 783
 Henderson, Robin, 495
 Hendrych, Radek, 543
 Hens, Niel, 63
 Hofner, Benjamin, 315
 Hothorn, Torsten, 209
 Hsu, Chiu-Hsieh, 549
 Hudecová, Šárka, 555
 Huertas, Jaime-Abel, 297
- Iddi, Samuel, 165
 Ingoldsby, Helen, 369
 Isaac, Benjamin J., 117
- Jacobs, Elizabeth, 549
 Jaeger, Jonathan, 171
 Janssens, Davy, 263
 Jeong, Jong-Hyeon, 147
 Joly, Pierre, 333
 Jones, Beatrix, 345
 Joshi, Chaitanya, 177
- Küchenhoff, Helmut, 491, 567
 Karlis, Dimitris, 263
 Kasim, Adetayo, 673
 Kaspříková, Nikola, 561
 Kato, Bernet S., 673
 Kauermann, Göran, 179, 735
 Kazemi, Iraj, 279
 Kerin, Michael J., 369
 Klupalová, Alena, 573
 Klima, André, 567
 Kneib, Thomas, 23, 315, 363, 735
 Komárek, Arnošt, 33
 Králová, Maria, 573
 Kulich, Michal, 33

- López-Montoya, Antonio Jesús, 577
 López-Ratón, Mónica, 583
 Lachos, Victor H., 607
 Laenen, Annouschka, 589
 Lambert, Philippe, 171, 185
 Lang, Joseph B., 191
 Laughlin, Daniel C., 177
 Lawson, Antony, 595
 Lee, Dae-Jin, 197
 Lee, Duncan, 485
 Lee, Youngjo, 147
 Lesaffre, Emmanuel, 45, 279, 291,
 711
 Letón, Emilio, 583, 627, 759
 Li, Yisheng, 549
 Lie, Bastian, 621
 Lin, Dan, 673
 Linares-Pérez, J., 513
 Long, Qi, 549
 Louzada, Francisco, 69

 Müller, Werner G., 537
 Machado, Luís, 215, 643
 MacKenzie, Gilbert, 81, 431
 Manuguerra, Maurizio, 159
 Martín, Nirian, 203
 Matawie, Kenan M., 601
 Matos, Larissa A., 607
 Maul, Thomas, 621
 Mayr, Andreas, 209
 McLellan, Chris R., 613
 Menezes, Raquel, 425
 Meyer, Renate, 179
 Miklavcic, Stanley, 765
 Miller, Claire, 485, 519
 Mirkov, Radoslava, 315, 621
 Molanes-López, Elisa M., 583, 627,
 633, 759
 Molenberghs, Geert, 99, 111, 165
 Molinari, Daniel Alberto, 637
 Monteiro, Magda, 447
 Moreira, Ana, 215, 643
 Moriña, David, 221
 Muggeo, Vito M. R., 227
 Muniz, Graciela, 351
 Mutambanengwe, Chenjerai Kathy,
 649

 Núñez-Antón, Vicente, 655
 Nagy, Stanislav, 233
 Neubauer, Gerhard, 239
 Newell, John, 369
 Njagi, Edmund, 111
 Novák, Petr, 245
 Nunes, Carla, 501

 O'Hara, Robert B., 37
 Oguiza, Ainhoa, 655
 Ogurtsova, Ekaterina, 251
 Oliveira Perez, Maria, 257
 Oliveira, Maristela D., 75, 525
 Oman, Samuel D., 661
 Omelka, Marek, 33, 667
 Ortega, Edwin M. M., 747
 Ospina, Raydonal, 407
 Otava, Martin, 673

 Pöbnecker, Wolfgang, 339
 Paige, Robert L., 679
 Pardo, Leandro, 203
 Parente, Alessandro, 117
 Patilea, Valentin, 685
 Paula, Gilberto A., 711
 Pešta, Michal, 555
 Perdoná, Gleici, 69
 Peria, Fernanda M., 69

- Perrakis, Konstantinos, 263
Peters, Annette, 491
Pfeifer, Christian, 691
Pieroni, Luca, 51
Pitz, Mike, 491
Pollock, Kevin, 485
Prates, Marcos O., 607
Prates, Weasley, 471
Proctor, Iain, 269
Puig, Pedro, 221
Punt, Leendert, 275

Quintana, Jose María, 419

Ramsey, David, 741
Rikhtehgaran, Reyhaneh, 279
Rizopoulos, Dimitris, 45, 285
Ročková, Veronika, 291
Rocha, Lisandra, 425
Rodríguez-Casal, A., 257
Rodríguez-Álvarez, María Xosé, 419
Rohde, Charles, 697
Romo, Juan, 633
Roque, Sara, 507
Ruli, Erlis, 705
Russo, Cibele M., 711

Sánchez-Sellero, César, 685
Saez, Marc, 717
Santos, Bruno R., 723
Saumard, Matthieu, 685
Scarrott, Carl, 369
Schauberger, Gunther, 729
Schimek, Michael G., 303
Schnabel, Sabine K., 309
Schneider, Alexandra, 491
Schulze Waltrup, Linda, 735
Scott, E. Marian, 269, 441, 485

Serrat, Carles, 297
Sheikhi, Ali, 741
Shkedy, Ziv, 63, 99, 673
Silva, Giovana O., 747
Silva, Marília, 501
Smith, Rognvald I., 269
Sobotka, Fabian, 315, 735
Sousa, Inês, 425
Sparks, Ross, 783
Stefanova, Katia, 753
Stephenson, David B., 435
Stidson, Ruth, 485
Strzalkowska-Kominiak, Ewa, 759

Tüchler, Regina, 321
Takkenberg, Johanna J. M., 45
Tamura, Karin Ayumi, 327
Thiart, Christien, 275
Thomae, Holger, 621
Turner, Paul W., 567
Thut, Gregor, 519
Torokhti, Anatoli, 765
Touraine, Célia, 333
Trindade, A. Alexandre, 679
Tutz, Gerhard, 153, 339, 729

Ullah, Insha, 345

Valero, Jordi, 221
Van Bodegom, Peter M., 177
van den Hout, Ardo, 351
van de Kastelee, Jan, 771
van Eeuwijk, Fred A., 309
van Eijkeren, Jan, 771
Varin, Cristiano, 357
Ventrucci, Massimo, 441
Ventura, Laura, 705
Verbeke, Geert, 279

- Vickerman, Peter, 63
- Wagner, Helga, 321, 537
- Waldmann, Elisabeth, 363
- Wall, Deirdre, 369
- Wallinga, Jacco, 771
- Wang, Jixian, 375
- Wets, Geert, 263
- Wilson, Susan R., 381
- Wood, Graham, 783
- Worton, Bruce J., 613
- Yang, Zihua, 777
- Zamzuri, Zamira, 783
- Zeileis, Achim 691
- Zong, Lu, 479

27th IWSM 2012 Sponsors

We are very grateful to the following organisations for sponsoring 27th IWSM 2012.

- Faculty of Mathematics and Physics, Charles University in Prague
- Faculty of Informatics and Statistics, University of Economics, Prague
- SAS Institute ČR, s.r.o.
- Česká spořitelna, a.s.

**Part 3. Contributed Papers
(Volume II)**

Misspecified random-effects distribution in non-linear mixed models with linear random effects

Ariel Alonso Abad¹

¹ Dept. of methodology and Statistics, Maastricht University, The Netherlands

E-mail for correspondence: ariel.alonso@maastrichtuniversity.nl

Abstract: In non-linear mixed models random effects are a useful device to account for unobserved subject specific characteristics. Typically, it is assumed that these random effects follow a multivariate normal distribution with mean zero and a general covariance matrix. However, the intangible nature of random effects makes this choice rather arbitrary. This raises concerns regarding the robustness of our inferences with respect to this assumption. In the present work we show that, when the random effects enter the model in a linear fashion, the maximum likelihood estimators of the fixed effects and variance components are consistent and asymptotically normal, even if the random-effects distribution is misspecified. Moreover, we show that when the linearity assumption is dropped consistency may not be guaranteed in general and the misspecification can also have a negative impact on the performance of inferential procedures like the Wald test.

Keywords: Loglinear models; Poisson; Misspecification.

1 Introduction

The last twenty years have witnessed the preponderance of non-linear mixed models (NLMM) in a variety of applications. Two standard assumptions made with these models are that (1) conditionally on the random effects \mathbf{b}_i , the outcome variable \mathbf{Y}_i is multivariate normal, and (2) the subject-specific effects \mathbf{b}_i are multivariate normal as well. However, the intangible nature of random effects makes the choice of their distribution rather arbitrary. Actually, their distribution may not be identifiable from the observable data, as pointed out by Alonso *et al.* (2010).

In these circumstances, the robustness of the inferential procedures with respect to the distributional assumptions for the random effects becomes a central issue. Surprisingly, there are not many studies that explore this important matter in the context of NLMMs. For linear mixed models (LMM), Verbeke and Lesaffre (1997) showed that the maximum likelihood estimators (MLE) of the fixed effects and variance components, obtained under

the assumption of normal random effects, are consistent and asymptotically normal even when the random-effects distribution is misspecified. Although LMMs can be seen as a special case of NLMMs, the results obtained in this very specific setting may not carry over to the more general scenario.

Using extensive simulations Hartford and Davidian (1999) found that for NLMMs the point estimates of fixed parameters, given by the maximum likelihood estimators and based on the Laplace approximation or linearization, are fairly robust to mild deviations from normality of the random-effects distribution. However, a different picture emerged for the variance components. Indeed, they reported some bias in the estimation of the variance components and the sampling distribution of the corresponding estimators seems to be markedly skewed. Additionally, these authors found that inference based on these methods can be sensitive to underlying distributional assumptions, and operating characteristics of hypothesis tests can be greatly affected. Finally, they claimed that as these models see even greater widespread use, there is an urgent need for further research in this area.

Along these lines, in the present work we study the impact of misspecifying the random-effects distribution on inferences derived from NLMM. To that effect, we focus on the specific but important special case in which random effects enter the model in a linear fashion. Such models emerge when linear coefficients in a non-linear model are random, or after an originally non-linear mixed model is linearized. Well known examples include the Gompertz and logistic growth curve with a random asymptote. Using theoretical arguments we will show that for these models the MLEs of the fixed parameters and variance components are robust to deviations from normality of the random-effects distribution. We will also show via simulations that consistency cannot be guaranteed if the linearity assumption is dropped.

2 Non-linear mixed models

If $\mathbf{Y}_i = (y_{i1}, y_{i2}, \dots, y_{im})'$ denotes the vector of observations for the i th experimental unit, then the non-linear mixed model can be written as

$$\mathbf{Y}_i = \mathbf{f}(\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{b}_i) + \boldsymbol{\varepsilon}_i \quad (1)$$

where \mathbf{f} is a general nonlinear multivariate function, $\boldsymbol{\beta}$ is a p -dimensional vector of fixed effects, \mathbf{X}_i is a $m \times p$ matrix of covariates, \mathbf{b}_i is a q -dimensional vector of random effects with $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, $\boldsymbol{\varepsilon}_i$ is a vector of error terms with $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Lambda}(\boldsymbol{\beta}, \boldsymbol{\gamma}))$, and \mathbf{b}_i and $\boldsymbol{\varepsilon}_i$ independent. It is further assumed that the matrix $\boldsymbol{\Lambda}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is positive definite for all values of the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. The previous model implies that, conditional on the random effects \mathbf{b}_i , the response vector $\mathbf{Y}_i \sim N(\mathbf{f}(\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{b}_i), \boldsymbol{\Lambda}(\boldsymbol{\beta}, \boldsymbol{\gamma}))$. Marginally, however, the distribution of \mathbf{Y}_i does not need to be normal. In

fact, marginally, \mathbf{Y}_i has density function f_N

$$f_N(\mathbf{Y}_i | \boldsymbol{\Psi}, \mathbf{R}_i) = \int \phi(\mathbf{f}(\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{b}_i), \boldsymbol{\Lambda}(\boldsymbol{\beta}, \boldsymbol{\gamma})) \phi(\mathbf{0}, \mathbf{D}) d\mathbf{b}_i$$

where $\boldsymbol{\Psi} = (\boldsymbol{\beta}', \text{vech}(\mathbf{D})', \boldsymbol{\gamma}')'$ denotes the vector of parameters, $\mathbf{R}_i = \text{vec}(\mathbf{X}_i)$ is the vector of all covariates and ϕ denotes the normal density. For a thorough presentation of this model we remit the reader to Vonesh and Chinchilli (1997) and Davidian and Giltinan (1995).

In what follows, we will focus on the setting where the random effects enter the model linearly, i.e. when $\mathbf{f}(\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{b}_i) = \mathbf{h}(\mathbf{X}_i, \boldsymbol{\beta}) + \mathbf{Z}_i(\boldsymbol{\beta})\mathbf{b}_i$. Model 1 then takes the form

$$\mathbf{Y}_i = \mathbf{h}(\mathbf{X}_i, \boldsymbol{\beta}) + \mathbf{Z}_i(\boldsymbol{\beta})\mathbf{b}_i + \boldsymbol{\varepsilon}_i.$$

In this expression, $\mathbf{h}(\mathbf{X}_i, \boldsymbol{\beta})$ is a general non-linear function and $\mathbf{Z}_i(\boldsymbol{\beta})$ is a $m \times q$ matrix that may depend on covariates and the parametric vector $\boldsymbol{\beta}$.

3 Robustness

Let us consider the non-linear mixed model

$$\mathbf{Y}_i = \mathbf{h}(\mathbf{X}_i, \boldsymbol{\beta}) + \mathbf{Z}_i(\boldsymbol{\beta})\mathbf{b}_i + \boldsymbol{\varepsilon}_i.$$

where $\mathbf{Y}_i = (y_{i1}, y_{i2}, \dots, y_{im})'$ is the vector of responses for unit i with $i = 1, \dots, n$, and $\mathbf{X}_i, \mathbf{Z}_i(\boldsymbol{\beta})$ are matrices of dimension $m \times p$ and $m \times q$ respectively. The vector $\boldsymbol{\beta}$ is p -dimensional and characterizes the population mean and \mathbf{b}_i is a q -dimensional vector of subject-specific effects assumed to follow a multivariate normal distribution with mean zero and covariance matrix \mathbf{D} . Finally, the error term $\boldsymbol{\varepsilon}_i$ is assumed to follow a normal distribution with mean zero and covariance matrix $\boldsymbol{\Lambda}(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Further, let us denote by $\boldsymbol{\Psi} = (\boldsymbol{\beta}', \text{vech}(\mathbf{D})', \boldsymbol{\gamma}')'$ the vector of all parameters. If the true distribution of \mathbf{b}_i is $g(\mathbf{b}_i)$ and $\boldsymbol{\Psi}_0$ denotes the true values of $\boldsymbol{\Psi}$ then, under general regularity conditions the MLE $\hat{\boldsymbol{\Psi}}_n$ of $\boldsymbol{\Psi}$, obtained under the misspecified model, satisfies $\hat{\boldsymbol{\Psi}}_n \xrightarrow{P} \boldsymbol{\Psi}_0$ (strongly).

Notice that if one takes $\mathbf{h}(\mathbf{X}_i, \boldsymbol{\beta}) = \mathbf{X}_i\boldsymbol{\beta}$, $\mathbf{Z}_i(\boldsymbol{\beta}) = \mathbf{Z}_i$ and $\boldsymbol{\Lambda}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sigma_e^2 \mathbf{I}$ then the result presented in Verbeke and Lesaffre (1997) is obtained as a corollary.

4 Simulation studies

In spite of its theoretical implications, the previous result only postulates an asymptotic property. One may wonder if the misspecification may considerably slow down the convergency of the MLEs to the true values of the parameters. Simulations not discuss here clearly showed that the impact

of the misspecification is also negligible in finite samples. Indeed, the rate of convergence of the MLE to the true values under misspecification was almost identical to the rate of convergence in the correctly specified setting. Even when the population was extremely heterogeneous (variance of the random effect $\sigma_{0b}^2 = 30$) the relative bias was smaller than 10% for sample sizes greater than or equal to 100.

The previous theoretical finding and simulation results make one wonder about the importance of the linearity assumption for the random effects. To explore this issue in more detail we carried out a new set of simulations.

4.1 Non-linear random effects

In the second set of simulations the data were generated using model

$$y_{ij} = \frac{\exp(\beta_1 + b_i)}{1 + \exp[-(t - \beta_2)/\beta_3]} + \varepsilon_{ij}, \quad (2)$$

i.e., a logistic growth model with a random asymptote. Notice that the random asymptote was exponentiated to guarantee positiveness. The parameter values were fixed at $\beta_{01} = 1$, $\beta_{02} = 100$, $\beta_{03} = 50$, $\sigma_{0b}^2 = 0.3$, 1 and 3. Further the b_i s were sampled from a general distribution g with mean zero and variance $\sigma_{0b}^2 = 3$, 10 and 30. Three different distributions g for the random effect b_i were included in the study: a normal density, a power function distribution and an asymmetric mixture of two normal densities. If necessary the distributions were shifted over their mean to ensure that $E(b_i) = 0$. For each subject, 4 repeated measurements were generated at times 15, 70, 95 and 140. Four different sample sizes, ranging from small to moderate, were considered: 25, 50, 100 and 200. In each setting, 500 data sets were generated and analyzed under the assumption of normality for the random effect. It is important to point out that the model used in the simulations as well as the values chosen for the parameters were based on the results obtained from a case study not presented here.

The consistency of the MLEs was studied through the evolution of the relative distance between the estimators and the real values, over increasing sample sizes. Let $\boldsymbol{\xi}_0 = (\beta_{01}, \beta_{02}, \beta_{03}, \sigma_{0b}^2, \sigma_{0e}^2)'$ be the vector of true values and $\hat{\boldsymbol{\xi}}_n = (\hat{\beta}_{1n}, \hat{\beta}_{2n}, \hat{\beta}_{3n}, \hat{\sigma}_{bn}^2, \hat{\sigma}_{en}^2)'$ be the corresponding vector of maximum likelihood estimators, then the relative distance between $\boldsymbol{\xi}_0$ and $\hat{\boldsymbol{\xi}}_n$ is defined as

$$d_n(\hat{\boldsymbol{\xi}}_n, \boldsymbol{\xi}_0) = \frac{\|\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0\|}{\|\boldsymbol{\xi}_0\|},$$

where $\|\cdot\|$ denotes the Euclidean norm. If the estimators remain consistent after misspecification of the random-effect distribution, then $d_n(\hat{\boldsymbol{\xi}}_n, \boldsymbol{\xi}_0)$ should converge to zero as the sample size increases.

The results were qualitatively different in this scenario. Indeed, the MLE of σ_b^2 always exhibited a relative bias larger than 55% and in some settings

TABLE 1. Type I error for detecting a significant effect (5% significance level) when $\beta_{01} = 0$ using model 2 assuming normal random effects when the random effects are generated from a normal (No), a power function (PF) and an asymmetric mixture (AM) of two normal distributions with variance σ_{0b}^2 . Values for which the lower bound of the corresponding 95% confidence interval was larger than 0.05 are highlighted.

		$\sigma_{0e}^2 = 0.5$			$\sigma_{0e}^2 = 2$		
	n	$\sigma_{0b}^2 = 0.3$	$\sigma_{0b}^2 = 1$	$\sigma_{0b}^2 = 3$	$\sigma_{0b}^2 = 0.3$	$\sigma_{0b}^2 = 1$	$\sigma_{0b}^2 = 3$
No	25	0.143	0.066	0.032	0.011	0.028	0.048
	50	0.179	0.047	0.051	0.090	0.029	0.036
	100	0.122	0.043	0.055	0.131	0.063	0.032
	200	0.094	0.046	0.033	0.117	0.047	0.050
PF	25	0.146	0.084	0.088	0.026	0.011	0.076
	50	0.177	0.091	0.076	0.072	0.048	0.036
	100	0.126	0.060	0.180	0.106	0.047	0.061
	200	0.064	0.018	0.520	0.110	0.036	0.129
AM	25	0.142	0.131	0.110	0.016	0.019	0.058
	50	0.141	0.077	0.137	0.080	0.059	0.045
	100	0.117	0.036	0.324	0.102	0.046	0.080
	200	0.064	0.008	0.735	0.112	0.021	0.206

it clearly surpassed the 75%. Likewise, a large relative bias was observed for β_1 when $\sigma_{0b}^2 = 1, 3$, even exceeding 45% in some scenarios. Importantly, for the rest of the parameters in the model the relative bias was always smaller than 5%.

Furthermore, we explored the impact of the misspecification on the power and the type I error rate associated with the test of the hypothesis $H_0 : \beta_1 = 0$. To that effect 500 data sets were simulated. In each data set the responses of 25, 50, 100 and 200 experimental units were generated like before, but with $\beta_{01} = 0$. Model 2 was used to analyze these data assuming normality for the random effect. The hypothesis $H_0 : \beta_1 = 0$ was tested using the Wald test. The proportion of cases in which the procedure rejected the null hypothesis (on a 5% significance level) was determined. The results are summarized in Table 1. Clearly, the misspecification can have a substantial impact on the results. For instance, when $\sigma_{0b}^2 = 3$, the type I error rates were considerably larger than the pre-specified 5% significance level, becoming even larger than 50% in some scenarios.

We also evaluated the impact of the misspecification on the power of the Wald test, using similar specifications as before and focusing again on β_1 . The results (not presented) indicate that the misspecification can also have a severe impact on the power as well.

References

- Alonso, A., Litière, S. and Laenen, A. (2010) A note on the indeterminacy of the random-effects distribution in hierarchical models. *The American Statistician* **64** (4), 318-324.
- Davidian, M. and Giltinan, D. (1995). *Nonlinear Models for Repeated Measurement Data*. New York: Chapman & Hall.
- Hartford, A. and Davidian, M. (1999). Consequences of misspecifying assumptions in nonlinear mixed effects models. *Computational Statistics & Data Analysis*. **34**, 139–164.
- Schinckel , A.P., Adeola, O. and Einstein, M. E. (2005). Evaluation of alternative nonlinear mixed effects models of duck growth. *Poult. Sci.* **84**, 256–264 .
- Sheiner, L.B. and Beal, S.L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis–Menten model: routine clinical pharmacokinetic data. *J. Pharmacokin. Biopharm.* **8**, 553–571.
- Tsoularis, A. and Wallace, J. (2002). Analysis of logistic growth models. *Math. Biosci.* **179**, 21-55.
- 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* **53**, 541–556.
- Vonesh, E.F. and Chinchilli, V.G. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. London: Chapman and Hall.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Regression quantiles to assess higher education performance

Giada Adelfio ¹, Giovanni Boscaino ¹, Vincenza Capursi ¹

¹ Dipartimento Scienze Statistiche e Matematiche “S. Vianelli”, Università di Palermo, Italy

E-mail for correspondence: giovanni.boscaino@unipa.it

Abstract: From 2001 Italian university system has adopted credits to measure the workload of the students. The weighted mean of marks with credits as weights is used to measure their performance. In our opinion it does not seem a proper way to measure. We suggest to adopt the median of weighted marks, because we are considering an ordinals variable, with a non-normal empirical distribution, and affected by outliers. Then, instead of using an OLS multiple regression, we investigate the determinants of the performance, measured by the median, using a linear quantile regression for ordinal response. A real dataset concerning 133 students of a degree course of the Faculty of Economics of the University of Palermo is considered. Results show how the quantile regression is more able than the OLS one to catch the effect of each covariate on student performance.

Keywords: measurement of educational path; median of weighted marks; quantile regression

1 Introduction

In the Italian University System (IUS) the degree mark is based on two parts: the first one concerns a measure of the performance of student at the end of educational path (hereafter named MEP) and the second one is the final examination score. Each degree course committee uses to set own criteria to assign a final score. The MEP has changed over the time. Before the 2001 reform it was based just on the examination marks: the mean of the marks was considered as a good indicator of the performance. Since 2001 the European Credit Transfer System has come in force and credits were assigned to every course. The performance indicator became the weighted mean of the marks, with credits as weights. On one hand the IUS can made comparison among European students outcome more easily. On other hand an important matter were not solved. The IUS is affected by high number of students who do not complete their studies in the prescribed time but that are still enrolled, named “fuori corso” (FC). Although the Bologna Process (1999) sets the criteria to recognize the degree among the EU countries, IUS has kept its peculiarity because of the

presence of FC. That is crucial for Italian universities because higher is the number of FC, lower are the public funds. Despite, in the last decade, several IUS reforms tried to reduce the number of the FC, empirical evidence proves that educational careers are getting longer: ministerial data show that between 2005 and 2009 the amount of graduated FC at the first degree course is always over the 50% of the graduated, rising to 63% in 2009. That being so, the aims of this paper are: to highlight some crucial aspects of the current student performance indicator, suggest a new indicator, and investigate the determinants of the student performance via the proposed indicator. The main expected outcome is the attention of the policy makers about those aspects that seem to affect low students performance. Moreover different performances, and determinants, between the FC graduates and the on time ones are expected. The paper is organized as follow. Section 2 concerns the comparison between the current student performance indicator and our proposal. In Section 3 the statistical model used is described. Section 4 is devoted to some results of the application to real data and concluding remarks.

2 The performance indicator

As introduced above, we refer to two student MEPs: before and after 2001. Before 2001 the adopted indicator was the arithmetic mean of marks for each student: $M_i = \sum_{j=1}^J \frac{m_{ij}}{J}$, where m_{ij} is the mark that the i -th student gets for the j -th course. From 2001 the used indicator has been the mean of the weighted marks: $M_i^w = \sum_{j=1}^J m_{ij}^w$, where $m_{ij}^w = \frac{m_{ij}C_j}{\sum_{j=1}^J C_j}$ and C_j is the number of credits of the j -th course. Since the mark variable is measured by an ordinal scale and the empirical distribution of the m_{ij}^w shows a positive skewness affected by several outliers, we suggest the use of the median of the weighted marks as an eligible indicator for the MEP: $Me_i^w = Me_i(m_{ij}^w)$. Moreover, the analysis of three cohorts of students enrolled at three degree courses of different areas (Sciences, Economics, and Arts) shows that the median seems to discriminate better than the mean between the two measurement practice before and after 2001. In fact the Kendall's τ 's reported in table 1 show that there is not a significative difference between the previous practice and the current one, when the arithmetic mean is used; furthermore the median of unweighted marks ($Me_i = Me_i(m_{ij})$) provides different results from the median of weighted ones, and, finally, the comparison between the mean and the median of the m_{ij}^w shows that M_i^w and Me_i^w yield to different conclusions., The non-normality of the distribution of the m_{ij}^w variable, the presence of outliers (Fig.1), and the considerations in the previous paragraph suggest the use of quantile regression approach to investigate the influence of some determinants for the proposed performance indicator. Moreover, although the Ordinary Least Squares regression allows to model the average as a measure of synthesis,

it does not take into account the whole shape of distribution of the outcome variable, as the quantile regression does.

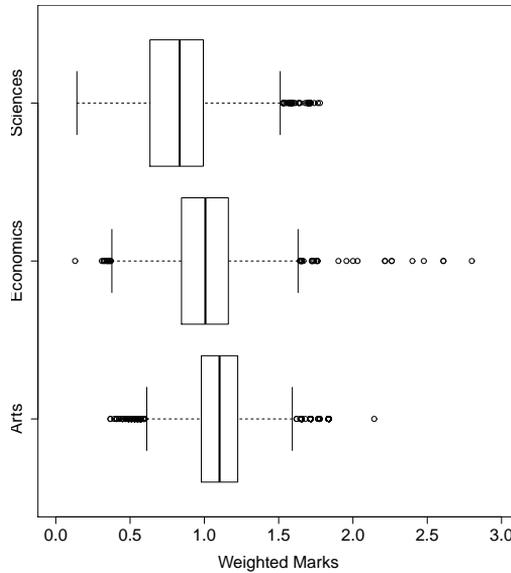


FIGURE 1. Box-plot of m_{ij}^w by faculty.

TABLE 1. Kendall’s τ as a rank-based measure of association between performance indicators : mean of marks and mean of weighted marks (τ_{M_i, M_i^w}), median of marks and median of weighted marks (τ_{Me_i, Me_i^w}), mean of weighted marks and median of weighted marks ($\tau_{M_i^w, Me_i^w}$).

Kendall’s τ	Sciences	Economics	Arts
M_i, M_i^w	0.947	0.959	0.9445
Me_i, Me_i^w	0.487	0.605	0.786
M_i^w, Me_i^w	0.502	0.566	0.740

3 The quantile regression

Quantile regression (QR) (Koenker and Hallock, 2001; Koenker, 2005) deals with the estimation of conditional quantile functions for models in which quantiles of the conditional distribution of the response variable are expressed as functions of observed covariates. Whereas the method of least squares results in estimates that approximate the conditional mean of the response variable, QR aims at estimating either the conditional median or

other quantiles of the response variable; QR also provides more robust estimates against OLS regression. Unlike the ordinary linear regression, the QR parameter measures the change in a specified quantile of the response variable produced by a one unit change in the predictor variable. This allows comparing how some percentiles of the variable of interest may be more affected by certain subject characteristics than other percentiles.

From a more formal point of view we proceed as in a least squares regression. In particular let $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$ be a sample of size n from some unknown population, where $\mathbf{x}_i \in \mathbf{R}^d$. The conditional ϕ th quantile function $f_\phi(x)$ is defined such that $P(Y \leq f_\phi(X)|X = x) = \phi$, for $0 < \phi < 1$.

Therefore, the ϕ th conditional quantile function can be estimated by solving:

$$\min_{f_\phi \in \mathbf{R}} \sum_{i=1}^n \rho_\phi(y_i - f_\phi(\mathbf{x}_i)) \quad (1)$$

where the function $\rho_\phi(\cdot)$ is the tilted absolute value function, that yields the ϕ th sample quantile as its solution and is defined by $\rho_\phi(r) = \phi r$ if $r > 0$, and $-(1 - \phi)r$ otherwise (Koenker and Bassett, 1978).

Setting $f_\phi(\mathbf{x}) = \mathbf{x}^T \beta_\phi$ where $\beta_\phi = (\beta_{\phi,1}, \beta_{\phi,2}, \dots, \beta_{\phi,d})^T$, such that the conditional τ th quantile function $f_\phi(x)$ is a linear function of the parameters β , a linear quantile regression is considered.

In this paper, due to the ordinal nature of the response variable, we consider a transformed QR model based on a jittered response obtained by adding a random noise that smooth the response values, as suggested by Hong and He (2010).

The QR estimates are obtained by the R package `quantreg`, that considers by default the modified version of the Barrrodale and Roberts algorithm described in Koenker and d'Orey (1987, 1994). We have chosen this approach because of the dimension of the analyzed data set, since this approach is quite efficient for problems up to several thousand observations, computing also confidence intervals for the estimated parameters, based on inversion of a rank test (Koenker, 1994).

4 Determinants of Me_i^w in a degree course of Economics

In this section we aim to model the indicator Me_i^w as a function of the following covariates in a QR model: high school diploma type (Lyceum or Other), residence (living in Palermo or not), gender, high school diploma mark, and an indicator variable of being FC or Regular (not-FC) at the moment of the graduation. The dataset concerns the students enrolled in 2002 and graduated from 3 up to 7 years after, in a degree course of the Faculty of Economics of the University of Palermo, Italy.

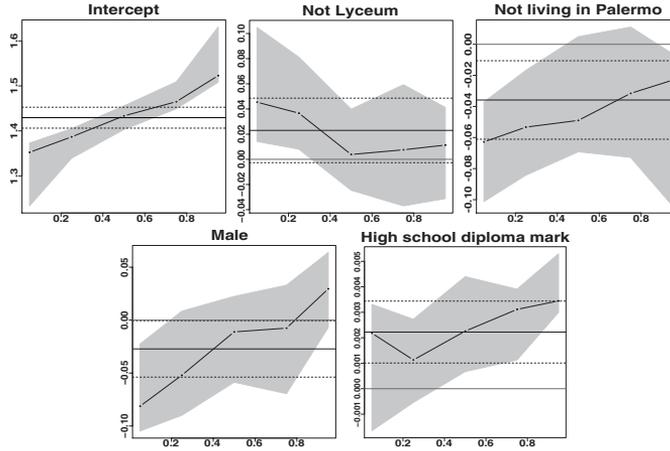


FIGURE 2. Quantile regression estimates for Me_i^w for FC graduates.

The linear QR model fitting suggests that the only covariate with a significant effect (p-value=0.000) is the FC indicator.

Therefore, we consider a linear QR model conditioned to the FC indicator. As it was obviously expected, the results show no significant variable conditioning to the Regular graduates. In our opinion the determinants of the students who get the degree on time do not depend on socio-demographic features but just on student motivation. With respect to the FC graduates, results noticeably change.

In fig. 2 we report a synopsis of QR results for the FC graduates. For each of the estimated coefficients we plot the QR estimates for each value of ϕ ($\phi = 0.05, 0.25, 0.50, 0.75, 0.95$) as 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 OLS estimate of the conditional mean effect is showed by the solid horizontal line, together with its 90 percent confidence intervals (horizontal dashed lines). The grey area represents the 90 percent pointwise confidence band for the QR estimates. The intercept of the model represents the estimated conditional quantile function of Me_i^w distribution of students that are female, with Lyceum diploma, living in Palermo, and with mean diploma mark equal to 89/100. The OLS estimates for this student profile show a mean performance value equal to 1.42, while QR estimates underline the different distribution features conditioned to the considered quantile. The other panels refer to the distribution of the estimated coefficients for different quantiles. The significance of the estimates are highlighted by the lines representing the coefficients when they present a particular slope. In short, we can observe that, although OLS estimates suggest a non-significant effect of “Not Lyceum diploma” with respect to “Lyceum diploma”, QR estimates em-

phasize that for low performer students (first quartile) the diploma has a significant positive effect on performance. In fact, $\hat{M}e_i^w$ of the first quartile of students with “Not Lyceum diploma” is +0.03 points higher than the one of students with “Lyceum diploma”. On the other hand, the effect of “Not living in Palermo” is negative for low-medium performer students (more or less second quartile) and the same happens for the covariate gender for low performer male students (first quartile). Finally, the covariate “High school diploma mark” highlights an expected positive effect on performance for good performer students.

Acknowledgments: The article is the result of the productive collaboration among the authors. In particular, paragraph 1 can be ascribed to Vincenza Capursi, paragraph 2 can be ascribed to Giovanni Boscaino, paragraph 3 can be ascribed to Giada Adelfio and Giovanni Boscaino, and paragraph 4 can be ascribed to Giada Adelfio.

This paper has been supported from Italian Ministerial grant PRIN 2008 “Measures, statistical models and indicators for the assessment of the University System”, n. 2008BWXMLH

References

- Hong, H. G. and He, X. (2010). Prediction of Functional Status for the Elderly Based on a New Ordinal Regression Model. *Journal of the American Statistical Association*, **105**(491), 930-941.
- Koenker, R. W. (1994). Confidence Intervals for regression quantiles. *Asymptotic Statistics*, 349-359, Springer-Verlag, New York.
- Koenker, R. (2005). Quantile Regression, *Cambridge University Press*.
- Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.
- Koenker, R.W. and d’Orey (1987). Computing regression quantiles. *Applied Statistics*, **36**, 383-393.
- Koenker, R.W. and d’Orey (1994). Computing regression quantiles. *Applied Statistics*, **43**, 410-414.
- Koenker, R. and Hallock, K. (2001). Quantile Regression: an introduction. *Journal of Economic Perspectives*, **15**, 143-156

prLogistic: An R package for estimating prevalence ratios using logistic models

Leila D. A. F. Amorim¹, Raydonal Ospina²

¹ Dept. of Statistics, Federal University of Bahia, Brazil

² Dept. of Statistics, Federal University of Pernambuco, Brazil

E-mail for correspondence: leiladen@ufba.br

Abstract: The interpretation of odds ratios (OR) as prevalence ratios (PR) in cross-sectional studies has been criticized since this equivalence is not true unless under specific circumstances. The logistic model is a very well known statistical tool for analysis of binary outcomes and frequently used to obtain adjusted OR. This model can also be used for estimation of PR. Another issue of interest is the estimation of adjusted PR for correlated data. We describe the R package *prLogistic* for estimation of PR using logistic models for analysis of independent and correlated binary data. We show how to use the package, considering two standardization procedures. Delta method and bootstrap are used for obtaining confidence intervals for PR. Our package includes several datasets used to illustrate implementation of these methods.

Keywords: logistic model; delta method; bootstrap.

1 Introduction

The concept of risk is fundamental in several research areas, with the measures of risk being associated to the probability of occurrence of an event. Two commonly used measures related to risk are relative risk (RR) in longitudinal studies and prevalence ratios (PR) in cross-sectional studies. In the simplest scenario, unadjusted measures can be computed easily through analysis of contingency tables. Another measure of association frequently reported by researchers is the odds ratio (OR), which differs mathematically of RR and PR. It is well known that OR overestimates the other two measures when the event is not rare (Greenland, 1987).

Statistical modeling is required when there is interest in estimation of an adjusted risk measure based on covariates. In several applications the outcomes are binary and logistic regression models are widely applied. Using this model, one can easily estimate $OR = \exp(\beta)$, where β is a parameter related to a risk factor of interest. However, interpretation of OR as a risk measure might be misleading. There is some debate in the literature about alternative approaches to obtain adjusted measures of PR. One of

the proposals is to estimate PR using logistic regression (Oliveira, Santana e Lopes, 1997; Localio et al, 2007). A more recent discussion is about how to estimate adjusted PR for correlated data, including analysis of clustered data (Santos et al, 2008).

Implementation of new statistical methods and its availability for applied researchers is another concern among data analysts. Many of the most recently proposed statistical methods can not be applied to data analysis because they are not easily accessible via standard statistical software. Thus, the goal of this paper is to describe the R package *prLogistic* for estimation of PRs using logistic regression models for analysis of both independent and correlated data.

2 Methods

Mathematically OR can be defined by the ratio of two odds, such that $OR = (p_1/(1 - p_1))/(p_0/(1 - p_0))$, where p_1 and p_0 denote, respectively, the prevalence of the event of interest in exposed and non exposed groups. The PR, on the other hand, is defined by the ratio of two proportions given by $PR = p_1/p_0$. Therefore, interpretation of these two measures is not the same, unless for rare events.

Let Y be the binary outcome(0/1) and $\mathbf{X} = \{X_1, X_2, \dots, X_k\}$ a matrix of independent variables. Consider the logistic regression model $E(Y|X) = P(Y = 1|X) = \frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)}$. We are interested in obtaining an expression for estimating PR as a function of β , the model parameters. For instance, suppose that we are evaluating the effect of an exposure X_1 on the occurrence of Y after adjustment by $k - 1$ independent variables (X_2, \dots, X_k). In this case, $PR = \frac{1+\exp\{-\beta_0-\beta_2X_2-\dots-\beta_kX_k\}}{1+\exp\{-\beta_0-\beta_1-\beta_2X_2-\dots-\beta_kX_k\}}$.

2.1 Standardization Procedures

Some standardization procedures for effect measures based on regression models had been proposed in the literature (Wilcosky and Chambless, 1985), being the two most commonly used methods called conditional and marginal standardization. For the conditional standardization procedure, a reference value (for instance, the mean) for each of the variables is defined and, thus, the prevalence for each group is calculated. For the marginal standardization procedure, on the other hand, the prevalence is computed, for each group, using the individual values for the variables and latter getting the mean value among all observations.

As an example consider data on n subjects with a binary exposure X_1 and a continuous variable X_2 . Using the conditional standardization procedure, the adjusted PR is given by $PR = \frac{1+\exp\{-\beta_0-\beta_2\bar{X}_2\}}{1+\exp\{-\beta_0-\beta_1-\beta_2\bar{X}_2\}}$, where \bar{X}_2 denotes the mean of X_2 . For the marginal method the adjusted PR is defined by $PR = \frac{\frac{1}{n} \sum_i (1/\{1+\exp(-(\beta_0+\beta_1+\beta_2X_{2i}))\})}{\frac{1}{n} \sum (1/\{1+\exp(-(\beta_0+\beta_2X_{2i}))\})}$ where the sum is over all n subjects.

2.2 Inference for Prevalence Ratios

We used delta method and bootstrap for obtaining confidence intervals for PR. The delta method is a general technique for asymptotic distribution of random variable functions based on approximation by Taylor series. Using the delta method, the adjusted confidence intervals (CIs) for PR are defined by $\exp(\log(\hat{PR}) \pm z_{\alpha/2} \widehat{se}(\log(\hat{PR})))$, where $\log(\hat{PR})$ is an estimate for adjusted $\log(PR)$, $\widehat{se}(\log(\hat{PR}))$ is an estimate of standard error for $\log(PR)$ and $z_{\alpha/2}$ is the quantile of a standard normal distribution.

The bootstrap approach is based on resampling with replacement. We consider 1,000 bootstrap replicates to produce a bootstrap distribution for PR. The confidence interval based on normal theory is generally approximated for sufficient large samples, considering bootstrap estimates for sample variance. Thus, the bootstrap CI is $\exp(\log(\hat{PR}^*) \pm z_{\alpha/2} \widehat{se}^*(\log(\hat{PR}^*)))$, where $\log(\hat{PR}^*)$ is a bootstrap estimate for adjusted $\log(PR)$ and $\widehat{se}^*(\log(\hat{PR}^*))$ is a bootstrap estimated for standard error of $\log(PR)$. An alternative approach, using percentile interval bootstrap, consider the empirical quantiles of bootstrap estimates for defining the interval. In such situation, the interval limits are given, for instance, by percentiles 2.5 and 97.5 when we are interested in the 95% CI.

2.3 Random Effects Logistic Model

Random effect models are often used for modeling correlated data. These models make adjustment for non observed individual characteristics, which reflect a natural heterogeneity among subjects. Let Y_{ij} be the outcome binary variable at cluster j for subject i , and denote X_{1ij} e X_{2ij} two independent variables. The random effects logistic model can be defined by $\text{logit}[P(Y_{ij}|X_{1ij}, X_{2ij}, u_{oj})] = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + u_{oj}$, where $u_{oj} \sim N(0, \sigma^2)$ represents a cluster specific random effect. Using the estimates for the parameters of this model (β 's), we can obtain PR as defined previously.

3 prLogistic

The R package *prLogistic* (Ospina and Amorim, 2011) contains the following functions:

- `prLogisticDelta`, for estimation of PR and its confidence interval using delta method, considering both marginal and conditional standardization procedures.
- `prLogisticBootCond`, for estimation of PR and its confidence interval using bootstrap for conditional standardization procedures.
- `prLogisticBootMarg`, for estimation of PR and its confidence interval using bootstrap for marginal standardization procedures.

We describe these functions below. All functions allow model fit with standard logistic regression and random-effects logistic model and their outputs present PR and its 95% confidence intervals. Package *prLogistic* also contains several datasets to illustrate the application of these methods. In this article, we describe and analyze two examples.

3.1 Function prLogisticDelta

This function is used as

```
prLogisticDelta(formula, data, pattern = "NULL", cluster="NULL", ...)
```

and takes the following arguments:

formula a symbolic description of the model to be fit. The details of model specification are given below.

data an optional data frame containing the variables named in *formula*. By default the variables are taken from environment(*formula*), typically the environment from which *prLogisticDelta* is called.

pattern optional argument specifying the standardization procedure. The default is set to be the conditional procedure. If *pattern*="marginal" the standardization procedure is set to be the marginal.

cluster optional argument specifying data clustering. The default is *cluster*=FALSE. If data is clustered, it should be set to *cluster*=TRUE.

3.2 Function prLogisticBootCond

This function is used as

```
prLogisticBootCond(object, data, ...)
```

and takes the following arguments:

object any fitted model object from which fixed effects estimates can be extracted. The details of model specification are given below.

data a required data frame containing the variables named in *object*.

3.3 Function prLogisticBootMarg

This function is used as

```
prLogisticBootMarg(object, data, ...)
```

and takes the following arguments:

object any fitted model object from which fixed effects estimates can be extracted. The details of model specification are given below.

data a required data frame containing the variables named in *object*.

4 Examples

4.1 The UMARU Impact Study

Data were from randomized trials related to treatment for drug abuse obtained by the University of Massachusetts Aids Research Unit (UMARU) IMPACT Study (UIS). The aim of the study was to compare treatment programs of different durations in the reduction of drug abuse and in the prevention of high-risk HIV behavior. The variables on the dataset are age at enrollment, back depression score at admission, drug use history at admission, number of prior drug treatments, race, treatment group, treatment site, and patient's status at the end of the treatment program (remained drug free or otherwise) (Hosmer and Lemeshow, 2000).

Analysis was conducted using data from 575 patients, with 26% of them remaining drug free (outcome prevalence). ORs and PRs were estimated using logistic regression model. We observed overestimation of the effects for all risk factors when using ORs. For instance, younger patients (< 32 years-old) were more likely to remain drug free than older patients ($PR = 1.84(95\%CI : 1.38, 2.44)$; $OR = 2.32(95\%CI : 1.54, 3.48)$).

4.2 SCAALA Ecuador-Study

A study was conducted in Ecuador to investigate the impact of long term treatment with the broad-spectrum anthelmintic, ivermectin, used for the control of onchocerciasis, on the prevalence and intensity of soil-transmitted helminth infections in school-age and pre-school children from 31 treated communities and 27 adjacent villages (Moncayo et al, 2008). We analyzed data from a random sample of 2000 children aged 6-16 to evaluate the effect of ivermectina in the occurrence of *Trichuris trichiura*, while adjusting for children's age and gender.

A random effects logistic model was fit. The prevalence of infection was 57.9%. A statistically significant effect of ivermectin on *T.trichiura* was found ($PR=0.33 [95\%CI=0.27; 0.42]$, using conditional standardization with delta method). The bootstrap confidence intervals based on normal theory were narrower than those obtained through delta method.

5 Conclusion

We have shown how logistic regression models can be performed to estimate prevalence ratios and their confidence intervals using our R package *prLogistic*, both in situations where the observations are independent and in clustered studies. The package can accommodate different standardization procedures commonly used as well as distinct approaches for computation of confidence intervals. Our package is easily used and does not involve extensive programming. Our contribution of the package *prLogistic*

will make these methodologies more accessible to applied researchers. In future updates of the package, the functions will also include generalized estimating equations (GEE), a marginal approach for analysis of longitudinal/clustered data.

Acknowledgments: Special Thanks to Professors Mauricio Barreto and Philip Cooper for providing us with data for our second example.

References

- Greenland S. (1987) Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology*, **160**, 301–305.
- Hosmer D.W. and Lemeshow S. (2000) *Applied Logistic Regression* New York: John Wiley & Sons. Second Edition.
- Localio A.R. et al (2007) Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology*, **60**, 874–882.
- Moncayo A.L. et al (2008) Impact of long-term treatment with ivermectin on the prevalence and intensity of soil-transmitted helminth infections. *PLoS Neglected Tropical Disease* **2:e293**.
- Oliveira N., Santana V. e Lopes A. (1997) Razões de proporções e uso do método delta para intervalos de confiança em regressão logística. *Revista de Saúde Pública*, **31**, 90,- 99.
- Ospina R, Amorim L.D. (2011). prLogistic: Estimation of Prevalence Ratios using Logistic Models. R package version 1.1, URL <http://cran.r-project.org/web/packages/prLogistic/index.html>.
- Santos C.A.S.T et al (2008) Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. *BMC Medical Research Methodology*, <http://www.biomedcentral.com/1471-2288/8/80>.
- Wilcosky T.C. and Chambless L.E. (1985) A comparison of direct adjustment and regression adjustment of epidemiologic measures. *J Chron Dis* **34**, 849,- 856.

Matrix correlation and matrix cross spectrum of mismeasured multivariate time series

Manoochehr Babanezhad¹

¹ Department of Statistics, Faculty of Sciences, Golestan University, Gorgan, Golestan, Iran

E-mail for correspondence: m.babanezhad@gu.ac.ir.

Abstract: In many time series analysis one wishes to model a time series data by knowing the behavior of the auto-correlation and spectral density. In this regards, we should expect that the true realizations of considered time series might be mismeasured. For example in bioassay experiments, when time series data represent the values of absorbed water by a plant, the actual value of absorbed water might be mismeasured. Therefore some measurement error may occur over the time on the observed data. In view of this, this study investigates how the random measurement error on a multivariate stationary discrete time series affects its matrix correlation function at the specific Lag h and the cross spectrum in a specific frequency ω . Specifically, it is shown that how the matrix correlation,

$$\left\{ \begin{pmatrix} \rho_{x_t}(h) \\ \rho_{y_t}(h) \end{pmatrix} : h = 0, 1, \dots \right\}$$

and the spectral density,

$$\left\{ \begin{pmatrix} f_{x_t}(\omega) \\ f_{y_t}(\omega) \end{pmatrix} : \omega \in [0, 2\pi] \right\}$$

of a class of linear bivariate time series with zero mean $\left\{ \begin{pmatrix} X_t \\ Y_t \end{pmatrix} : t \in T \right\}$, are affected by the random measurement error.

Keywords: Matrix Correlation; Matrix Cross Spectrum; Bivariate time series; Random measurement error.

1 Introduction

In recent years, there has been considerable progress in the development of inference methods that account for the presence of measurement error in the explanatory variables in regression models (Cook and Stefanski, 1994). However, little works has been done to consider the error contaminated time series data (Fuller, 1995; Tanaka, 2000). In view of this, this study investigates how the random measurement error in a multivariate stationary discrete time series affects its matrix correlation function

and the cross spectrum. In the time series analysis literature, a discrete time series $\{X_t : t \in T\}$ is (weak) stationary (T is an index set), when $E(X_t) = \mu$, $Var(X_t) = \sigma^2$, $\forall t \in T$, and its auto-covariance function at lag h , $\gamma(h) = Cov(X_t, X_{t+h})$ depends only upon the distance h for $h = 0, \pm 1, \pm 2, \dots$ (Brockwell and Davis, 1991; Fuller, 1995; Shumway and Stoffer, 2008). These properties ensure the nature of a stationary time series. One of the main characteristics of a (weak) stationary time series $\{X_t : t \in T\}$ in frequency domain is its spectral density. The Fourier transform of the absolutely summable $\gamma(h)$ function is called the spectral density (Kakizawa, 2006). The spectral density furnishes another important representation of the time series. Spectral density of $\{X_t\}$ in fact describes how the variance of a time series is distributed in different frequencies. Because a (weakly) stationary time series may vary over a continuous range of frequencies, e.g. $[0, 2\pi]$, through its spectral density function $f_{X_t}(w)$. Then $f_{X_t}(w)$ can be computed as follows (Brockwell and Davis, 1991; Shumway and Stoffer, 2008),

$$f_{X_t}(w) = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \gamma_{X_t}(h) e^{-iwh} = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \gamma_{X_t}(h) \cos wh \quad (1)$$

where $w \in [0, 2\pi]$ and $\sum_{h=-\infty}^{+\infty} |\gamma_{X_t}(h)| < \infty$.

2 Mismeasured time series

As stated in time series analysis, we should expect that the true realizations of considered time series might be mismeasured. Now suppose that a discrete time series $\{X_t : t \in T\}$ is measured with error (Fuller, 1995; Tanaka, 2000), that is $\forall t$,

$$Y_t = X_t + u_t \quad (2)$$

where $\{u_t : t \in T\}$ is a time series independent of X_t , $\forall t$. The $\{u_t\}$ is the measurement error or sometimes is called the noise in the system. Time series analysis requires that the pattern of observed time series data to be identified and more or less formally described. Once the pattern is established, we can interpret and integrate it. Here, because $\{X_t\}$ is mismeasured we observe $\{Y_t\}$ instead of $\{X_t\}$ in practice. Although the pattern of the observed time series $\{Y_t\}$ is possible to establish, it follows from (2) that the establishing of time series $\{X_t\}$ depends upon the establishing of the time series $\{u_t\}$. That is $\{X_t\}$ to be stationary if $\{u_t\}$ is stationary. The aim of this work is how the measurement error time series $\{u_t\}$ affects the spectral density of $\{X_t\}$, in case where $\{u_t\}$ is stationary time series or it is possible to make it stationary by removing trend or seasonality variations. Because in practice, when analyzing actual data, we would be faced with the problem of the substantial random fluctuation of the spectral density

of $f_u(w)$. To get through this, we assume that $\{u_t\}$ is an uncorrelated stationary time series with $E(u_t) = 0$, $Var(u_t) = \sigma^2$, and we also assume the covariance functions of X_t and u_t are known. Then in each time t , we obtain

$$\gamma_{Y_t}(h) = Cov(Y_t, Y_{t+h}) = Cov(X_t, X_{t+h}) + Cov(u_t, u_{t+h}) = \gamma_{X_t}(h) + \gamma_{u_t}(h) \tag{3}$$

Therefore it follows from the latter that,

$$\begin{aligned} f_{Y_t}(w) &= \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \gamma_{Y_t}(h)e^{-iwh} = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \gamma_{X_t}(h)\cos wh + \\ \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \gamma_{u_t}(h)\cos wh &+ \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \gamma_{u_t}(h)\cos wh = f_{X_t}(w) + f_{u_t}(w) \end{aligned} \tag{4}$$

where

$$\gamma_{u_t}(h) = Cov(u_t, u_{t+h}) = E\{u_t u_{t+h}\} = \begin{cases} \sigma_u^2 & h = 0 \\ 0 & |h| > 0. \end{cases}$$

and $f_{u_t}(w) = \frac{\sigma_u^2}{2\pi}$. As a result, the spectral density of time series X_t can be obtained as follows,

$$f_{X_t}(w) = f_{Y_t}(w) - \sigma_u^2 = \frac{1}{2\pi} \left\{ 2 \sum_{h=1}^{+\infty} \gamma_{Y_t}(h)\cos(wh) + \sigma_y^2 - \sigma_u^2 \right\} \tag{5}$$

We may suppose that u_t is non-stationary, that is $u_t = g(t)Z_t$ where $g(t)$ is arbitrary function of t (e.g. $g(t) = t$) and Z_t is uncorrelated random variables with mean 0 and variance σ_z^2 for $t = 0, \pm 1, \pm 2, \dots$. Then $E(u_t) = 0$, $\gamma_u(0) = \sigma_u^2 = g(t)^2\sigma_z^2$,

$$\gamma_{u_t}(h) = Cov(u_t, u_{t+h}) = Cov(g(t)Z_t, g(t+h)Z_{t+h}) = \begin{cases} g(t)^2\sigma_z^2 & h = 0 \\ 0 & |h| > 0 \end{cases}$$

and $f_{u_t}(w) = \frac{g(t)^2\sigma_z^2}{2\pi}$. The spectral density of X_t can then be obtained as follow,

$$f_{X_t}(w) = f_{Y_t}(w) - \sigma_u^2 = \frac{1}{2\pi} \left\{ 2 \sum_{h=1}^{+\infty} \gamma_{Y_t}(h)\cos(wh) + \sigma_y^2 - g(t)^2\sigma_z^2 \right\} \tag{6}$$

We will pay the case of non-stationary measurement error in elsewhere.

3 Simulation

To investigate the estimated matrix correlation and matrix cross spectrum of mismeasured in time series in finite samples, we conducted a simulation

experiment. Data were simulated to mimic data on the sediment suspended in the water of the Des Moines River at Boone, Iowa (Fuller, 1995). A portion of the data obtained by daily sampling of the water during 1973. The data are the logarithm of the parts per million of suspended sediment. Since the laboratory determinations are made on a small sample of water collected from the river, the readings can be represented as (2), where Y_t , is the recorded value, X_t is the true average sediment in the river water, and u_t , is the measurement error introduced by sampling and laboratory determination. We estimate auto-correlation function $\rho(h)$ in $h = 1, 2$; and further $f(w)$ in $\omega = \pi/2$ and $\pi/3$ by $\sigma_u^2 = 0.25$. Table 1 summarized the results in two cases where the measurement error is ignored and is corrected. The results show that there are considerable bias when one ignored the measurement error. Table 1 shows that when measurement error is not taken into account, one may wrongly fitted the time series model. Because the results show the auto-correlation tends not to be zero. When the measurement error is corrected the results show the auto-correlation tend to be zero. This issues happens for the spectral density. A rapid deceases is observed in naive estimator. This is not observed for the corrected estimator.

4 Conclusion

While it is possible to use spectral methods to evaluate the analysis of a time series in the frequency domain, we examine the problem of measurement error which is inevitable for many agents of interest. I begin the discussion under the example where the true time series X_t is a stationary, zero mean, first order moving average process, that is, $X_t = e_t + \alpha e_{t-1}$ for $t \in T$ (a special case of the class of linear stationary time series). Assume we are unable to observe X_t , directly. Instead, we observe Y_t where $Y_t = e_t + \alpha e_{t-1} + u_t$, and where u_t is independent of e_j for all t and j . The former equation is called the state equation or the transition equation, and X_t is called the state of the system at time t . The latter equation is called the measurement equation or the observation equation. Then the $f_x(w) = \frac{\sigma_e^2[1+\alpha^2+2\alpha\cos w]+\sigma_u^2}{2\pi}$. To estimate the values of $f_x(w)$, σ_e^2 , σ_u^2 , and α are assumed to be known, or it is important to have an efficient method of computing the estimated values of σ_e^2 and σ_u^2 . These two values can be estimated by methods of simulation-extrapolation (Cook and Stefanski, 1994; Fuller, 1995).

TABLE 1. Auto-covariance function and cross spectrum of mis-measured time series with ignored and corrected measurement error.

Estimators	Auto-correlation	Cross spectrum
Naive estimator	$\begin{pmatrix} 0.52 \\ 0.36 \end{pmatrix}$	$\begin{pmatrix} 4.75/5 \\ 1.75/5 \end{pmatrix}$
Corrected estimator	$\begin{pmatrix} 0.30 \\ 0.03 \end{pmatrix}$	$\begin{pmatrix} 1.00/5 \\ 0.90/5 \end{pmatrix}$

Acknowledgments: I would like to thank Golestan University for supporting my research project financially.

References

- Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods, 2nd Edition*, CRC Press.
- Cook, J. and Stefanski, L. (1994). A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, **89**, 1314–1328.
- Fuller, W.A. (1995). *Introduction to Statistical Time Series, 2nd Edition*, Wiley Series in Probability and Statistics.
- Kakizawa, Y. (2006). Bernstein polynomial estimation of a spectral density. *Journal of Time Series Analysis*, **27**, 253–287.
- Shumway, R. H. and Stoffer, D. S. (2008). *Time Series Analysis and Its Applications With R Examples*, 2nd Edition Springer.
- Tanaka, K. (2000). A unified approach to the measurement error problem in time series models. *Econometric Theory*, **18**, 278–296.

Location of optimal cut-points to categorize continuous variables in clinical studies

Irantzu Barrio^{1,3}, Inmaculada Arostegui¹, María Xosé Rodríguez-Álvarez², Jose María Quintana^{3,4}

¹ Departamento de Matemática Aplicada, Estadística e I.O. Universidad del País Vasco UPV/EHU

² Unidad de Epidemiología Clínica y Biostatística. Complejo Hospitalario Universitario de Santiago de Compostela

³ Unidad de Investigación, Hospital Galdakao-Usansolo. CIBER Epidemiología y Salud Pública (CIBERESP)

⁴ Departamento de Medicina Preventiva y Salud Pública. Universidad del País Vasco UPV/EHU

E-mail for correspondence: irantzubarrio@gmail.com

Keywords: Categorization; Prediction.

1 Introduction

Prediction models are nowadays relevant for decision making in various fields, so they are in medicine. Clinical prediction models may provide the evidence-based input for share decision-making, by providing estimates of the individual probabilities of risk and benefits (Steyerberg, 2006).

An important part when building a prediction model are the covariates that are considered as potential predictors. The association of the predictor with the outcome and the distribution of the predictor determines its strength. The performance of the final model will depend partly on the strength of the covariate. As far as the objective is to build a prediction model to use in the daily practice, we need to use the covariates that are available and can be applicable in the daily practice.

Although there are many scientific texts in which it is recommended not to categorize, but to use the continuous variable instead (Royston et al., 2006), in daily clinical practice physicians encourage statisticians to categorize continuous variables, since at the decision time they systematically think on categories. Even more, medical research studies based on follow-up measurements are often subject to high rates of missingness. In some cases clinical assumptions are made for missing imputation, and this is done into the categorized variable (Quintana et al., 2011).

So far, several approaches to categorize continuous variables have been proposed in the literature based on statistical or clinical criteria (Hin et al,

1999, O'Brien, 2004, Williams et al, 2006, Steyerberg, 2006). However, we are of the opinion that it is necessary to go a step further on to provide categorization methods with the aim of obtaining more accurate predictive models.

The goal of this study is to propose a new approach for the selection of optimal cut-points to categorize continuous variables in order to be incorporated in prediction models to be used in clinical practice, which at the same time are easy to implement and use by clinicians.

2 Methods

Consider we have a dichotomous response variable Y and a continuous covariate X , we want to categorize. Our proposal consists on categorizing X in such a way that we obtain the best predictive logistic model (highest area under the receiver operating characteristic curve - AUC) for Y . To do so, we propose two alternative methods, based on a sequential finding of the cut-points and on the finding of the optimal set of cut-points named *AddFor* and *Genetic* respectively. Given k , the number of cut-points we previously fix to categorize X in $k+1$ intervals, lets denote $\mathbf{v} = (x_1, \dots, x_k)$ the vector of the k cut-points, and X_{cat_k} the categorized variable (taking $k+1$ values, $l = 0, \dots, k$).

With the *AddFor* algorithm we looked for each cut-point at a time. This is, we first look for the x_1 value (in a grid of size M of equally spaced values in the rage of X) in such a way that the AUC of the logistic model

$$P(Y = 1|X_{cat_1}) = \frac{\exp(\beta_0 + \beta_1 1_{\{X_{cat_1}=1\}})}{1 + \exp(\beta_0 + \beta_1 1_{\{X_{cat_1}=1\}})}$$

is maximized. Once we have selected x_1 we fix it and look (in the grid) for x_2 ($x_2 \neq x_1$) so that the AUC for the model $P(Y = 1|X_{cat_2}) = \frac{\exp(\beta_0 + \beta_1 1_{\{X_{cat_2}=1\}} + \beta_2 1_{\{X_{cat_2}=2\}})}{1 + \exp(\beta_0 + \beta_1 1_{\{X_{cat_2}=1\}} + \beta_2 1_{\{X_{cat_2}=2\}})}$ is maximized. We repeat the process until we complete the vector of k cut-points $\mathbf{v} = (x_1, \dots, x_k)$.

For the *Genetic* method, we simultaneously find the vector of k cut-points $\mathbf{v} = (x_1, \dots, x_k)$ which maximizes the AUC of the logistic model

$$P(Y = 1|X_{cat_k}) = \frac{\exp(\beta_0 + \sum_{l=1}^k \beta_l 1_{\{X_{cat_k}=l\}})}{1 + \exp(\beta_0 + \sum_{l=1}^k \beta_l 1_{\{X_{cat_k}=l\}})}$$

We have implemented the required functions for the *AddFor* method in the software R version 2.13.0 and used the `genoud` function of the `rgenoud` package (Walter et al., 2011) for the *Genetic* method. The arguments used in the `genoud` function are the following:

`fn` function to be maximized, this is, AUC.

nvars number of parameters to be estimated k

range the range of the covariate X in which we look for the cut-points.

3 Simulation Study

A simulation study was conducted to examine the behavior of our proposed categorization methods. Given the covariate X defined according to a uniform(0,1), the binary outcome variable Y was generated according to $Y \sim \text{Bernoulli}(p(X))$ where

$$p(X) = \frac{\exp(X-10(X-0.2)^3+110(X-0.6)^3)}{1+\exp(X-10(X-0.2)^3+110(X-0.6)^3)}$$

Simulations were conducted for different number of cut-points to be selected ($k=2, 3$), different grid sizes in which to search for the cut-points in the *AddFor* algorithm ($M=100,1000$), and for different sample sizes ($N=500, 1000$). For each sample size 200 data sets were generated.

For each data set generated, we calculated the AUC value obtained by the proposed categorization, either with the *AddFor* or the *Genetic* method. We then plotted all the result in a box-plot so as to compare the results obtained with each method. Figure 1) shows the results for $N=500$. Finally, to evaluate the prediction accuracy of the proposed categorized variable, we compared the obtained AUCs with the theoretical AUC value for the continuous variable. This value was empirically calculated based on the theoretical probabilities.

In general, we noted that the *Genetic* algorithm provided a better categorization in terms of AUC. Moreover, when the number of cut-points to look for was 3 and the grid for the *AddFor* method was of size 1000, the results were nearly the same and the time required for the *AddFor* method was almost six times lower.

4 Application to the IRYSS-COPD Study

We have applied the methodology proposed in this paper to the IRYSS-COPD study, a prospective cohort of patients with a Chronic Obstructive Pulmonary Disease (COPD) (Quintana et al., 2011). A sample of 2877 with COPD exacerbation patients attending the emergency departments (ED) of 16 participating hospitals in Spain was included in the study. Information was recorded during the time the patient was evaluated in the ED, at the time a decision was made to admit the patient to the hospital or discharge home, and during follow-up after admission or discharge home. The main data collected were those related to the patient's respiratory function (arterial blood gases, respiratory rate (RR), dyspnea) at the arrival to the ED and at the decision time to discharge home or admission to ward.

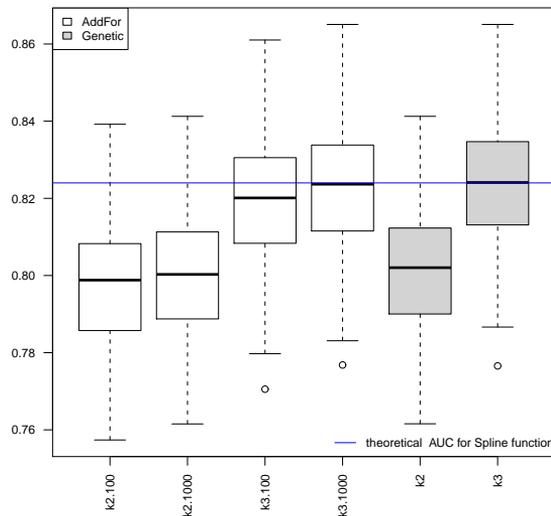


FIGURE 1. Graphical representation of the simulation results obtained with 200 data sets of size 500. km represents the simulations for m cutpoints. km.M represents the simulations for m cutpoints in a grid size of M

The aim was to categorize the covariates arterial blood gas PCO₂ and RR, into 4 and 3 categories respectively, considering their relationship with the outcome. The response variable was poor evolution (death, admission to intensive care unit or intensive care respiratory unit, invasive mechanical ventilation and non-invasive mechanical ventilation). Clinicians decided the number of categories that were clinically significant by clinical expertise and by looking at the graphical relationship between each covariate and poor evolution. Figure 2 shows the estimated relationship between the (continuous) PCO₂ and RR with poor evolution using a logistic generalized additive model (GAM) with P-spline smoothers.

We applied the *AddFor* and *Genetic* methods to this data, and found out that there were not statistically significant differences between the AUCs obtained using both methods not either between the AUC obtained with the categorical variable and the AUC given by the original continuous predictor (using the logistic GAM). More detailed results can be seen in Table 1 and Figure 2, where the cut-points obtained with the two approaches are displayed.

Moreover, the categories obtained matched almost exactly with the categories proposed by clinicians (Quintana et al., 2011), which provides a face-validation for the presented methodology.

TABLE 1. Application of the *AddFor* and *Genetic* methods to PCO2 and RR covariates from the IRYSS-COPD Study

Variable	Method	k	cut-points	AUC	p-value*
PCO2	Continuous	-	-	0.819	-
PCO2	Genetic	3	45.9 ; 55.8 ; 65.9	0.813	0.279
PCO2	AddFor (Grid=100)	3	45.7 ; 54.6 ; 66.5	0.811	0.121
RR	Continuous	-	-	0.673	-
RR	Genetic	2	19.4 ; 24.6	0.672	0.973
RR	AddFor (Grid=100)	2	19.6 ; 24.0	0.672	0.973

* Represents the statistical significance when the null hypothesis considers equal AUC values for the continuous and either the Genetic or AddFor categorization methods.

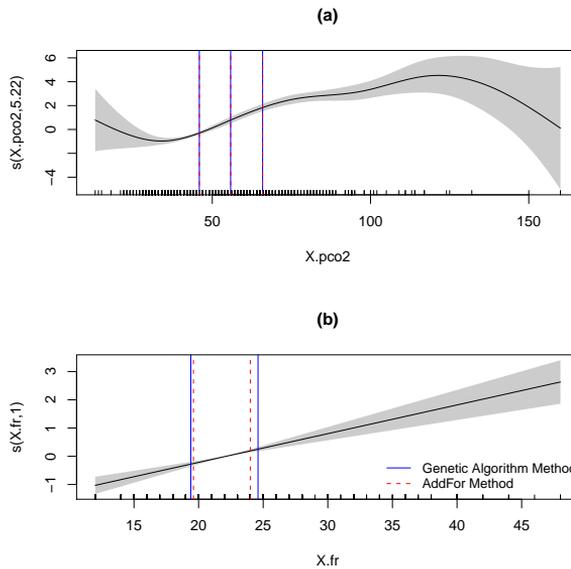


FIGURE 2. Graphical representation of the cut-points obtained in the IRYSS-COPD Study. (a) The relationship between the PCO2 and the poor evolution and (b) the relationship between the respiratory rate and poor evolution

5 Conclusions

When applying to real data, the *Genetic* algorithm provides a higher AUC value than it does the *AddFor*, although there are no meaningful differences between both. However, the *Genetic* method has a higher computational cost than it has the *Addfor* method. This is basically because the *Genetic*

does look for the best vector of cut-points, while the *AddFor* looks sequentially after fixing previous cut-points. Moreover, the simulation study illustrates that this categorization methods, specially the *Genetic* algorithm or the *AddFor* with a grid of size 1000, provide a categorized variable which minimizes the loose of information with respect to the continuous variable. In our opinion, this methodology provides clinicians with an easy to use tool to categorize continuous variables to implement in predictive models to be used in daily clinical practice.

Acknowledgments: This research was supported by grants GIU10/21, UFI11/52, DE2009-0030, MTM2010-14913 and MTM2011-28285-C02-01. Work of MXRA was supported by grant CA09/0053 from the Instituto de Salud Carlos III (Ministerio Español de Ciencia y Tecnología). Work of IB was supported by grant GIU10/21 from the Universidad del País Vasco UPV/EHU. Special thanks to CIBERESP and BIOSTATNET.

References

- Hin L.Y., Lau T.K., Rogers M.S., Chang A.M.Z. (1999). Dichotomization of continuous measurements using generalized additive modelling - application in predicting intrapartum caesarean delivery. *Statistics in Medicine*, **18**,1101–1110.
- O'Brien SM. (2004). Cutpoint selection for categorizing a continuous predictor. *Biometrics*, **60**(2),504–509.
- Royston P, Altman D.G., Sauerbrei W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, **25**(1),127–41.
- Steyerberg, E.W. (2009). *Clinical Prediction Models*. New York: Springer.
- Quintana JM., Esteban C., Barrio I., Garcia S., Gonzalez N., Arostegui I., Lafuente I., Bare M., Blasco JA., Vidal S., The IRYSS-COPD Group (2011). The IRYSS-COPD appropriateness study: objectives, methodology, and description of the prospective cohort. *BMC Health Services Research*, **11**:322 doi:10.1186/1472-6963-11-322.
- Walter R., Jasjeet S.S. (2011). Genetic Optimization Using Derivatives: The rgenoud Package for R. *Journal of Statistical Software*, **42**(11), 1–26.
- Williams BA., Mandrekar JN., Mandrekar SJ., Cha SS., Furth AF.(2006). Finding optimal cutpoints for continuous covariates with binary and time-to-event outcomes. *Technical Report Series*, **79**.

Fixed effects versus random effects in a longitudinal study: A simulation study

Ana Borges¹, Inês Sousa^{2,3}, Lisandra Rocha², Raquel Menezes^{2,3}

¹ Center for Research and Innovation Business Sciences and Information Systems of Polytechnic Institute of Porto, Portugal

² Center of Mathematics of University of Minho, Portugal

³ Department of Mathematics and Applications of University of Minho, Portugal

E-mail for correspondence: ID2704@alunos.uminho.pt

Abstract: In this study we aim to compare parameters estimates obtained from the linear regression model with the ones obtained from the longitudinal linear model with random effects. Making use of the general linear model considering the effect of each individual both as fixed effect and as a random effect. For that we conducted a simulation study, varying the number of individuals, n , and number of observations per individual, m .

Keywords: Longitudinal data; Linear Regression Model; Longitudinal Linear Model with Random Effects.

1 Introduction

In regression analysis, for a variety of outcome measures, the linear regression model (LRM) became very useful as it represents a simple likelihood approach for that purpose. However, when analyzing longitudinal data, where individuals are measured repeatedly over time, it is usual to make use of special statistical methods, since it is necessary to take into account individual heterogeneity. Longitudinal models are also called mixed effects models as they use random effects and possible serial correlation within subjects. The Longitudinal Linear Model with Random Effects (LLM) for longitudinal data can be seen as an extension of linear models and its fundamental feature is the decomposition of variability in between and within subjects.

These random effects describe changes within each individual and flexible representation of the variance-covariance structure. The basic idea underlying a random effects model is that there is natural heterogeneity across individuals in their regression coefficients and that this heterogeneity can be represented by a probability distribution (Diggle et al, 2002). Alternatively to random effects model a linear regression model can be used with

individual effects treated as fixed effects instead. In this case a large number of parameters is estimated, which could mean a greater error in the estimation, compared to the random effects model where only the parameters of the distribution of random effects is estimated (Gardiner et al, 2008). In this study we aim to compare a LLM with the LRM. For this purpose, we make use of the general linear model considering the effect of each individual, i.e. the specificity of each individual, both as random effect and as a fixed effect. We compare parameters estimates for LLM and LRM obtained by a simulated study. The main objective of this work is to see how the two models behave in terms of parameter estimation, and in which conditions we can make use of each model. We conducted a simulation study varying the number of individuals, n , and number of observations per individual, m . We structure the paper as it follows. Firstly, we clarify both LLM and LRM. In the third section we describe the simulation study carried out. In the fourth section we present the main results ending with a conclusion section.

2 Models description

In this section we present the Longitudinal Linear Model with Random Effects and the Linear Regression Model. Let Y_{ij} represents a response variable measured for individual $i = 1, \dots, n$ at time $t_{ij} = 1, \dots, m_i$ and x_{ij} a vector of length p of explanatory variables. For simplicity, we will consider that everyone is measured the same number of the times $m_i = m$. The mean and variance of Y_{ij} are represented by $E[Y_{ij}] = \mu_{ij}$ and $Var(Y_{ij}) = \nu_{ij}$. The set of repeated outcomes for subject i are collected into an m -vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im})$, with mean $E[\mathbf{Y}_i] = \mu_i$ and $m \times m$ covariance matrix $Var(\mathbf{Y}_i) = \mathbf{V}_i$. The response vector for all units are referred to as $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ which is an N -vector with $N = n \times m$.

2.1 Linear Regression Model

The Linear Regression Model (LRM), considered here, assumes the response variable and the explanatory variable related through: $Y_{ij} = \beta_0^* + \beta_1^* t_{ij} + \mathbf{1}\delta_i + \epsilon_{ij}^*$ where β^* is the vector of unknown regression coefficients and δ_i is the fixed effect of the i th individual. That is, the vector parameter δ is associated to a dummy variable \mathbf{Z} at individual level. The ϵ_{ij}^* are independent realizations of a Gaussian random variable with $E[\epsilon_{ij}^*] = 0$ and $Var(\epsilon_{ij}^*) = \tau^{*2}$. The log likelihood for the observed data is

$$\ell(y; \alpha, \tau^{*2}) = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\tau^{*2}) - \frac{1}{2\tau^{*2}} [(Y - X^* \alpha)^T I^{-1} (Y - X^* \alpha)] \quad (1)$$

Equating to zero the derivative of (1) in the order of α and solving it in order for this parameter, we obtain the estimator $\hat{\alpha} = (X^T X)^{-1} X^T y$

where $Var(\hat{\alpha}) = \tau^{*2}(X^T X)^{-1}$. $\hat{\alpha}$ is an unbiased estimator of α , i.e. $E[\hat{\alpha}] = \alpha$ (Gumedze et al,2011). The constant τ^{*2} can be estimated by $\widehat{\tau^{*2}} = \frac{\sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - (\beta_1^* t_{ij} + \delta_i))^2}{(n \times m) - 1}$.

2.2 Longitudinal Linear Model with Random Effects

The Longitudinal Linear Model with Random Effect Model (LLM), considered here, assumes the response variable and the explanatory variable related through: $Y_{ij} = \beta_0 + \beta_1 t_{ij} + U_i + \epsilon_{ij}$ where \mathbf{Y} is a vector $N \times 1$, \mathbf{T} is a matrix $N \times 1$, β is the vector of unknown regression coefficients. The ϵ_{ij} are independent realizations of a Gaussian random variable with $E[\epsilon_{ij}] = 0$ and $Var(\epsilon_{ij}) = \tau^2$. The U_i is a realization of a random effect at individual level, where $U_i \sim Normal(0, \sigma^2)$. Therefore, $E[\mathbf{Y}] = \mathbf{T}\beta_1$ and $Var(\mathbf{Y}) = \mathbf{V}$ where $\mathbf{V} = diag(\mathbf{V}_1, \dots, \mathbf{V}_n)$ is a block diagonal matrix of dimension $N \times N$. One strategy for parameter estimation is to consider simultaneous estimation of the parameters of interest, β , σ^2 and τ^2 (Diggle et al,2002). It is used the maximum likelihood method. We can rewrite $\mathbf{V} = \tau^2 V^*$, where V^* is a block diagonal matrix with blocks V_i^* and only depends on σ^2 . The log likelihood for the observed data is

$$\ell(y; \beta, \tau^2, V^*) = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\tau^2) - \frac{1}{2} n \log |V_i^*| - \frac{1}{2\tau^2} [(Y - X\beta)^T V^{*-1} (Y - X\beta)] \tag{2}$$

For given V^* , equating to zero the derivative of (2) in the order of β and solving it in order for this parameter, we obtain the estimator $\hat{\beta}(V^*) = (X^T V^{*-1} X)^{-1} (X^T V^{*-1} Y)$. Substitution into (2) gives

$$\ell(y; \hat{\beta}(V^*), \tau^2, V^*) = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\tau^2) - \frac{1}{2} n \log |V_i^*| - \frac{1}{2\tau^2} RSS(V^*) \tag{3}$$

where $RSS(V^*) = \{Y - X\hat{\beta}(V^*)\}^T V^{*-1} \{Y - X\hat{\beta}(V^*)\}$. Differentiation of (3) with respect to τ^2 gives the maximum likelihood estimator for τ^2 , for fixed V^* , as $\hat{\tau}^2(V^*) = \frac{RSS(V^*)}{N}$.

2.3 Differences between estimators

In this section we make a brief comparison between both LRM and LLM likelihood functions and explain in which conditions they are mathematically equal. When comparing the likelihood functions we must take into account that in LRM exists, at an individual level, fixed effects so the variability of Y is only explained by the variability of ϵ_{ij} . On the other hand, in LLM the variability of Y is explained, as expected, by the variability of the random effects at individual level and the variability of ϵ_{ij} .

Looking at both equations the differences are in the following terms, related to the variance structure of the models: $\frac{1}{2} n \log |V_i^*|$ and V^{*-1} ; and also in the terms α and β . As Hsiao 2003 refers, both equations became equal when m

is large enough. We can rewrite the matrix V as $V = \tau^2 I_N + \sigma^2 A$, where I is $N \times N$ identity matrix and A is a $N \times N$ block diagonal matrix with blocs J_m . This blocs J_m are $(m \times m)$ matrices of ones. The form of V^{-1} was found to be the following $N \times N$ matrix: $V^{-1} = \frac{1}{\tau^2} (I - \frac{\sigma^2}{\tau^2 + m\sigma^2} A)$. So, in the limit, when m is large ($m \rightarrow \infty$), the term $\frac{\sigma^2}{\tau^2 + m\sigma^2}$ tends to zero and the matrix V^{-1} became $V^{-1} = \frac{1}{\tau^2} I$. In the same way $V^{*-1} = I$. So, substituting V^{*-1} into $\hat{\beta}(V^*)$, the estimator for β becomes $\hat{\beta}(V^*) = (X^T I X)^{-1} (X^T I Y)$ which means that β estimator of LLM becomes equal to the β estimator of LRM when m is large enough.

3 Simulation Study

To compare both LLM and LRM we conducted a simulation study using software *R* (www.R-project.org). Firstly, the data is simulated under the true model LLM with n individuals observed at m times, that is the response variable becomes: $Y_{ij} = \beta_0 + \beta_1 t_{ij} + U_i + \epsilon_{ij}$. For that the data was generated according as: i) $\beta_0 = 1$ and $\beta_1 = 1$; ii) Time $t_{ij} \sim Exp(\lambda_i)$, $j = 1, \dots, m$, where $\lambda_i \sim U[0.1; 0.5]$; iii) $U_i \sim N(0, \sigma^2)$, with $\sigma^2 = 1$; iv) $\epsilon_{ij} \sim N(0, \tau^2)$ with $\tau^2 = 1$. To study the effect of number of individuals and number of observations per individual, we consider all the 36 combinations of $m = 5, 10, 20, 40, 50, 100$ and $n = 20, 40, 50, 100, 200, 400$. For each combination of $n \times m$, the simulation study was performed for 1000 replicates. Each data set was then adjusted to both models, by means of the likelihood method. The estimates obtained for LLM and LRM are: $\hat{\beta}_0$, $SE(\hat{\beta}_0)$, $\hat{\beta}_1$, $SE(\hat{\beta}_1)$, $\hat{\sigma}^2$, $\hat{\tau}^2$, $\hat{\beta}_1^*$, $SE(\hat{\beta}_1^*)$, $\hat{\beta}_1^*$, $SE(\hat{\beta}_1^*)$, $\hat{\delta}_i^2$, $\hat{\tau}^{*2}$ respectively.

4 Results

From the 36 possible combinations, here we show only 6 situations, which are representative of the results in all 36 situations. We decide to show only the results for the situations where the number of total observations is fixed $n \times m = 2000$. In Table 1 we present mean values for replicates of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$, empirical standard deviations for replicates of $\hat{\beta}_1$ ($Sd(\hat{\beta}_1)$) and the mean of the variance of effects (random or fixed) for the 6 cases (A-F). From the values of $\hat{\beta}_1^*$ and $\hat{\beta}_1$ we can observe that the mean values of the estimators of betas for the thousand samples of each combination are very close, meaning that there are no differences in the estimators of betas. On the other hand, we observe that decreasing the number of observations per individual for the same number of individuals, the differences between the values for the standard deviations of the estimated betas $Sd(\hat{\beta}_1)$ increases, and the standard deviations obtained by LRM are smaller. That is, for a smaller number of observations per individual the differences are greater.

TABLE 1. Results for 6 samples simulated.

Case	n	m	LRM				LLM			
			$\hat{\beta}_1$	$Sd(\hat{\beta}_1)$	$SE(\hat{\beta}_1)$	$\overline{Var}(\hat{\delta}_i)$	$\hat{\beta}_1^*$	$Sd(\hat{\beta}_1^*)$	$SE(\hat{\beta}_1^*)$	$\overline{Var}(\hat{U}_i)$
A	20	100	0.99998	0.00221	0.00224	1.09911	0.99998	0.00219	0.00223	0.97613
B	40	50	0.99995	0.00219	0.00225	1.07858	0.99994	0.00218	0.00223	0.98606
C	50	40	0.99998	0.00222	0.00226	1.06234	0.99999	0.00217	0.00223	0.97021
D	100	20	0.99996	0.00222	0.00229	1.07518	0.99994	0.00214	0.00223	0.95514
E	200	10	0.99989	0.00239	0.00235	1.11260	0.99988	0.00230	0.00223	0.90959
F	400	5	1.00003	0.00251	0.00250	1.20669	0.99997	0.00223	0.00223	0.83267

The average standard error is in both cases, LRM and LLM, similar to the standard deviation of punctual estimates obtained from 1000 simulations. As we decrease the number of observations per individual the mean values of the variances of U'_i 's for LRM and LLM get more different. For an even number of observations, with fewer people and a larger number of observations per individual results are similar, but with more people and a smaller number of observations per individual the LLM estimator has less variability. In Table 1 we can also observe that while the results for LLM deviate below the true value of σ^2 (in this case $\sigma^2 = 1$), the results for LRM deviate to values above the real value of σ^2 . The results also show a decrease in variability and the approximation to the true value of σ^2 . For both models, the values for $MSE(\hat{\sigma}^2)$ decreases for larger values of observations per individual (for the same number of individuals). Although, as observed in graphics presented in Figure 1, the values for $MSE(\hat{\sigma}^2)$ are very different between the models for small values of individuals and observations per individuals, this difference decreases considerably for larger number of individuals and observation per individual.

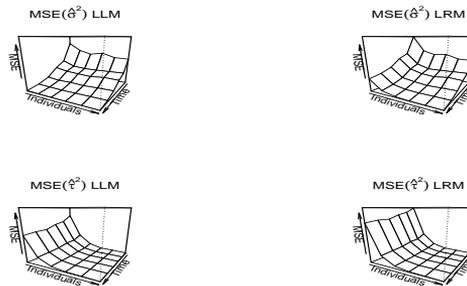


FIGURE 1. Mean Square Error comparison.

For the $MSE(\hat{\tau}^2)$ we observe that there is a decrease for large values of observations per individual. A marked difference between the values, for both models, for small values of individuals and observations per individuals,

that decreases considerably for larger number of individuals and observation per individual. Yet, for a large number of individuals and a small value of observations per individual we also found a noted difference (Figure 1).

5 Conclusions

When comparing the results of our simulation study, for LLM and LRM, they did not show significant differences when the number of observations is very high. Bearing in mind the purpose of a longitudinal analysis and the data set, we should beware of the choice the methods of analysis. In certain conditions it can make sense to choose the Linear Regression Model over the, widely used, Linear Longitudinal Model. Our main objective is to see how the two models behave in terms of parameters estimation, and in which conditions we can make use of the LRM rather than the LLM. In our simulation study we observe that for high values of observations per individual it does not appear to exist significant differences between the values of the parameters estimated, using the maximum likelihood method for both models. Therefore, we conclude that in longitudinal studies where the number of individuals is small and high total number of observations a Linear Regression Model is good enough and sufficient.

Acknowledgments: The authors acknowledge the grant by FCT PTDC/MAT/112338/2009. The authors L. Rocha and A. Borges have a PhD scholarship by FCT SFRH/BD/61368/2009 and SFRH/BD/74166/2010 respectively. 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 a Tecnologia", within the Project Est-C/MAT/UI0013/2011.

References

- Diggle, P.J., Liang, K-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.
- Gardiner, J.C., Luo, Z., and Roman, L.A. (2008). *Fixed effects, random effects and GEE: What are the differences?* *Statistics in Medicine*, 28, pp. 221-239.
- Gumedze, F.N., and Dunne, T.T. (2011). *Parameter estimation and inference in the linear mixed model*. *Linear Algebra and its Applications*, 435, pp. 1920-1944.
- Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge University Press.

Multi-parameter regression survival models

Kevin Burke¹, 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: It is well known that the *proportional hazards* (PH) assumption is a simplifying assumption in survival analysis that may not always be appropriate. However, PH models are routinely fitted and inference is made on the data based on such models. A major flaw here is that if the data are non-PH then we will reach incorrect conclusions by making this assumption. For example we may find a covariate to be statistically insignificant when in fact it is important, but the model fails to pick this up. Even if a PH model *does* pick up the statistical significance of a non-PH 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 PH modelling to try an account for situations such as those described above and offer new, previously unavailable insights, into the data.

Keywords: Crossing hazards, Multi-parameter regression survival models, PH and non-PH models

1 Introduction

Generally, when modelling data the model has multiple parameters. Typically, we choose only to regress one of these parameters to measure the effect of the covariates. For example, in GLMs (McCullagh and Nelder, 1989) we regress the location parameter, $g(\mu) = X\beta$, whilst the dispersion parameter, σ , is often treated as being little more than a nuisance parameter. This view of the role of dispersion is disputed by Pan & MacKenzie (2003) working in the longitudinal data modelling setting. Accordingly, in recent times there has been more work done in regressing dispersion parameters simultaneously with location parameters. However, it is still not common practice among the statistical community.

A similar situation arises in survival analysis. One of the most popular models in this area is the *proportional hazards*, PH, model. Often this model is imposed on data even though it is known that, in reality, the data do not obey the PH assumption. Clearly the more *non*-PH data are, the less appropriate the model will be. The PH model is equivalent to regression of the *scale* parameter, say λ , in a model that has the proportional hazards

property. Our new proposal is to simultaneously regress the *shape* parameter, say γ . 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, can now change with time. The effect that the covariates have on the shape parameter may be of scientific interest in its own right. Here we will focus specifically on the Weibull model, although the methodology can easily be applied to other models.

2 Multi-Parameter Regression Survival Model

In a *parametric* PH model, we regress the scale parameter only. We propose the following multi-parameter regression:

$$g_1(\lambda) = x_1^T \beta, \quad g_2(\gamma) = x_2^T \alpha, \quad (1)$$

where, g_1 and g_2 are appropriate link functions and we have also subscripted the corresponding covariate vectors to highlight the fact that the covariates regressing the scale and shape do not have to be the same.

2.1 MPR Weibull

The 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 $\lambda, \gamma > 0$. The hazard is decreasing for $\gamma < 1$, constant for $\gamma = 1$ and increasing for $\gamma > 1$. The hazard function for the multi-parameter regression Weibull is

$$\lambda(t) = \exp(x_1^T \beta) \exp(x_2^T \alpha) t^{\exp(x_2^T \alpha) - 1}. \quad (2)$$

We have used the log-link for both λ and γ here so that $\lambda, \gamma \in \mathbb{R}_+$. This generalization of the Weibull leads to a time-dependent hazard ratio for variables that are significant in the shape regression.

Moreover it is clear that the addition of a regression for the shape parameter allows the multi-parameter regression Weibull model to deal with crossing hazards data in a natural way, without recourse to the use of the frailty arguments adopted by MacKenzie & Ha (2007). Accordingly, use of this class of models, will render the analysis of crossing hazards data relatively routine.

2.2 Examples

In the presentation we analyse two data sets: 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, and the well-known gastric cancer data set of the Gastrointestinal Tumor Study Group (1982), where we observe the situation of crossing hazards. In both data sets we will illustrate the flexibility of the MPR Weibull model compared with the PH Weibull model.

3 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 ratio tests or AIC values. The extra generality leads of course to additional model selection issues and we describe their solution in the workshop presentation.

Acknowledgments: This work was supported, in part, by the SFI's BIO-SI research programme (www.sfi.ie), 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.
- Gastrointestinal Tumor Study Group. (1982) A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma, *Cancer*, **49**, 1771–1777.
- MacKenzie G, & Ha ID. (2007) Modelling survival data with crossing hazards. *22nd IWSM Proceedings*, Barcelona, Spain, 416–420.
- McCullagh, P. & Nelder, J. (1989) *Generalized Linear Models*, Chapman & Hall/CRC.
- Pan J & MacKenzie G. (2003) On model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, **90**, **1**, 239–244.
- Wilkinson, P. (1995) *Lung Cancer in Northern Ireland 1991–1992*, PhD thesis, Queen's University Belfast.

Climate variability and dengue incidence in Malaysia

Norziha Che Him ¹, Trevor C. Bailey ¹, David B. Stephenson ¹

¹ College of Engineering, Mathematics and Physical Science, University of Exeter, UK

E-mail for correspondence: nc262@exeter.ac.uk

Abstract: We report preliminary results on the use of generalised additive models (GAM) for describing patterns in observed monthly dengue counts in Malaysia and in particular the potential for using climate information as predictors in such models. A GAM is developed to allow for both temporal trend over years and annual seasonal cycle and then to select climate and other covariates which prove significant in prediction of confirmed monthly dengue cases based on data collected across 12 states of Malaysia for the period January 2001 to December 2009. The covariates explored include temperature and precipitation with time lags relevant to dengue transmission, an El Niño sea surface temperature index and other relevant demographic variables. A negative binomial formulation is adopted to allow for overdispersion in the observed dengue counts. We conclude that climate variables could have potential value in helping to predict dengue incidence in Malaysia in both time and space.

Keywords: Generalised Additive Model; Dengue; Climate; Negative Binomial.

1 Introduction

Dengue fever was first reported in Malaysia after outbreaks in Penang during December 1901 (Skae, 1902) and became endemic in Malaysia by the 1960's (Rudnick et al., 1965). In recent years incidence of dengue in Malaysia has steadily increased (Abu Bakar and Shafee, 2002) and has become a significant public health concern with major epidemics in 2002 and 2005. There is some evidence from studies in Singapore, China, Venezuela, Brazil and elsewhere (Hii et al., 2009; Lu et al., 2009; Aura et al., 2010; Lowe et al., 2010) that dengue incidence is influenced by weather and climate variability. For example a recent study by Johansson et al. (2009) in Puerto Rico, investigated biological relationships between temperature, precipitation and dengue transmission, but found empirical evidence of these relationships was inconsistent. However, the same authors did report a strong association between variation in dengue cases and varying climatic conditions in the different regions studied. Relatively little previous work

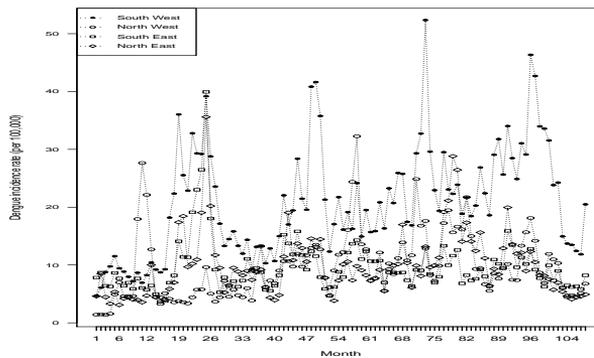


FIGURE 1. Dengue incidence rate per 100,000 population for main regions; North East, South East, North West and South West in Malaysia.

has been done to investigate these relationships in Malaysia, but a surveillance study by Rozilawati et al. (2007) in one district of Malaysia reported a strong correlation between rainfall and egg trap population. Malaysia has a relatively good dengue surveillance system, but successive control programmes have failed to reduce the incidence rates as reported by Lam (1993). If climatic variables can be used to provide early warnings of future dengue outbreaks then this would be of considerable value in allowing public health decision makers to take earlier preventative action. However, so far insufficient modelling studies have been conducted on large enough data sets in Malaysia to establish whether climatic variables can be used to provide early warnings of future dengue outbreaks. In this paper, we develop a negative binomial generalised additive models (GAM) for monthly dengue counts from Malaysia recorded at a state level for the period January 2001 to December 2009. Potential explanatory variables investigated include observed climate data with time lags relevant to dengue transmission along with other demographic covariates.

2 Data

Monthly numbers of confirmed dengue fever cases from twelve of the coterminous states of Malaysia for the nine years from January 2001 to December 2009 were obtained from the Ministry of Health Malaysia. Annual population and population density in each state were obtained from the Malaysian Department of Statistics. In total 309,003 cases were reported across Malaysia during this time period. The monthly dengue incidence rate per 100,000 population over the 108 month of the study period is shown in Figure 1.

Monthly mean temperature, number of rainy days and mean rainfall were

supplied by the Malaysian Meteorology Department and this climate data was supplemented with additional more detailed data on rainfall from the Department of Irrigation and Drainage. Niño4 is an index used to measure the strength of El Niño and La Niña events and is defined as the departure in monthly sea surface temperature (SST) from its long-term mean averaged over the Niño4 region (160 East-150 West, 5 South-5 North) which is most relevant to Malaysia. A time series of Niño4 index was obtained from the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center for the period of the study. In general, the episodes of warm event (El Niño) and cold event (La Niña) in the study area and period were El Niño in 2008 and La Niña for the year of 2002, 2004, 2006 and 2009 with other years being neutral.

3 Model development

Initial exploratory analyses of the data set described in the previous section indicated an increasing trend in dengue incidence rate (DIR) over Malaysia as a whole over the study period. This was particularly marked in those states in the South West of the country where the main urban areas of Malaysia are located. As might be expected DIR is higher in areas where there is a higher population density. This overall increase in dengue superimposed on an annual seasonal cycle which sees DIR peaks in January and July.

Geographical differences in the pattern of DIR were evident at the state level (both in level and to some extent in the annual cycle). This could be explain by fact that Malaysia is characterised by two monsoon regimes, namely, the Southwest Monsoon from late May to September and the Northeast Monsoon from November to March. The Northeast Monsoon brings heavy rainfall, particularly to the east coast states of Malaysia, whereas the Southwest Monsoon normally signifies relatively drier weather. Further investigation indicated that these geographical difference can be adequately captured without significant loss of detail by grouping the 12 states into the four broad regions of ‘North East’, ‘South East’, ‘North West’ and ‘South West’. DIR for those four regions is shown in Figure 1 for the 108 months of the study period and the annual DIR cycle in the same regions averaged over the nine years is shown in Figure 2.

In order to capture the various influences discussed above (global trend, seasonal cycle, regional variations and the impact of population density) whilst at the same time investigating potential association with climate and lagged climate variables, a generalised additive models (GAM) framework was adopted (Hastie and Tibshirani, 1986). The response variable was taken as the monthly number of dengue cases, y_{st} , where s denotes state ($s = 1, \dots, 12$) and t denotes month since the start of the study period ($t = 1, \dots, 108$). Given the high variability in the monthly disease counts and the fact that there are clearly many unmeasured confounding factors in

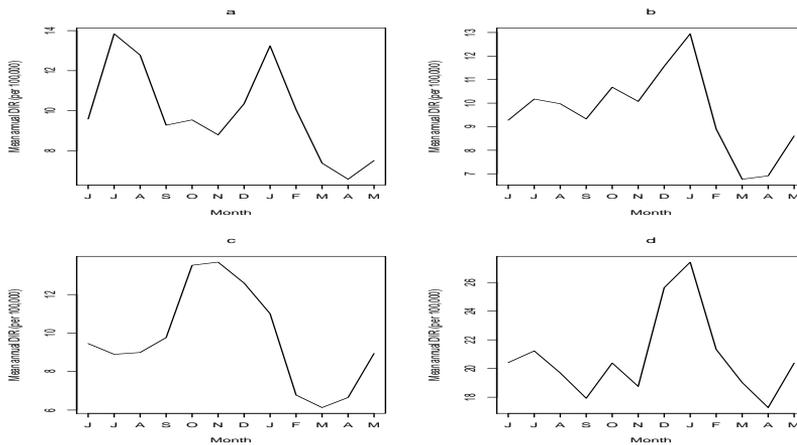


FIGURE 2. The mean annual cycle of dengue incidence rate per 100,000 population for (a)North East, (b)South East, (c)North West and (d)South West in Malaysia.

the data set, y_{st} was assumed to follow a negative binomial distribution to allow for overdispersion.

Population was included as an offset in the model to account for variations in population between state and over time. Smooth functions of year and of calendar month were included to capture global trend and seasonal cycle, while a 4 level factor ('North East', 'South East', 'North West' and 'South West') was incorporated to reflect regional variations. Interactions between the regional factor and the year trend and seasonal cycles were also considered. Additional covariates were then population density and the climate variables: mean monthly rainfall and temperature and number of rainy days, along with the Niño4 SST index. Lagged values of all climate variables (from one to six months previously) were also considered.

The final climate variables to emerge as most significant from the model selection process were rainfall lagged zero and three months, number of rainy days lagged zero and three months, temperature lagged zero month and sea surface temperature lagged six months. A simple linear term was found to be adequate to represent yearly trend, rather than a smooth functions of year, there was no evidence of an interaction between this and the regional factor. The seasonal cycle is represented by a smooth function of calendar month; the interaction between this and region was also significant indicating that the apparent regional differences in seasonal cycle noted earlier do appear to be important even when other variables are allowed for. As expected, population density was highly significant.

So the final model selected was as follows:

$$\begin{aligned}
y_{st} &\sim \text{NegBin}(\mu_{st} = p_{st}\rho_{st}, \theta) \\
\log(\mu_{st}) &= \log(p_{st}) + \log(\rho_{st}) \\
&= \log(p_{st}) + \alpha + \sum_{j=1}^6 \beta_j x_{jst} + \gamma_1 z_{1st} \\
&\quad + \gamma_2 z_{2st} + \delta_{s'(s)} + f_{s'(s)}(z_{3st})
\end{aligned}$$

Here the observed dengue count, y_{st} , for state s ($s = 1, \dots, 12$) and month t ($t = 1, \dots, 108$) is assumed to follow a negative binomial distribution with mean value $\mu_{st} = p_{st}\rho_{st}$ and scale parameter θ , where ρ_{st} is the dengue incidence rate and p_{st} is the known population offset. The climate covariates, x_{jst} ($j = 1, \dots, 6$) are respectively: rainfall with lags zero and three months, number of rainy days lag zero and three months, temperature lag zero month and Niño4 index with a six months lag. The covariate z_{1st} is population density, z_{2st} is year (2001 to 2009) and $\delta_{s'(s)}$ represents a regional effect with $s'(s)$ being an indicator function mapping each state s into one of the four regions ‘North East’, ‘South East’, ‘North West’ or ‘South West’. Finally, $f_{s'(s)}(z_{3st})$ are smooth functions of the calendar month z_{3st} (the latter taking values 1 to 12 where 1 refer to June) in region $s'(s)$. All of the terms in the above model were significant ($p < .05$). Space precludes presentation of detailed model results here, but the broad findings were that mean rainfall three months previously has a positive relationship with DIR, but mean rainfall in the same month has a negative relationship with DIR. This could possibly be because more rainfall earlier in the year could encourage mosquito development, while heavy rainfall in the same month could wash out mosquito breeding places and lower dengue transmission. Number of rainy days both three months previously and in the same month and temperature in the same month all have a positive relationship with DIR. Meanwhile, sea surface temperature (SST) six months previously as defined by Niño4 has a positive relationship with dengue.

4 Conclusions

The preliminary modelling results in this paper indicate that although climate information alone does not account for a large proportion of the overall variation in dengue cases in Malaysia, spatio-temporal climate information does account for some of this variability. Therefore the inclusion of climate information in a dengue epidemic prediction model for Malaysia may well be worth investigating further. The next step would be to refine the preliminary model presented here and to more fully assess its predictive

validity in relation to dengue epidemics. A subsequent step would then be to investigate how the model performs when ‘observed’ climate variables are replaced with retrospective seasonal forecasts made for a historical period. ‘Hindcast’ precipitation, temperature and Niño4 data are available from international forecasting systems which typically produce ensemble predictions with lead times up to 6 months. By replacing ‘observed’ with ‘hindcast’ climate variables in a suitably refined model, dengue predictions for Malaysia could potentially be made several months ahead of the dengue season of interest.

References

- Abu Bakar, S. and Shafee, N.(2002). Outlook of dengue in Malaysia: a century later. *Malaysian Journal Pathology*, **24(1)**, 23–27.
- Aura, D.H.M. and Alfonso, J.R.M.(2010). Potential influence of climate variability on dengue incidence registered in a western pediatric Hospital of Venezuela. *Tropical Biomedicine*, **27(2)**, 280–286.
- Hastie, T. and Tibshirani, R.(1986). Generalized Additive Models. *Statistical Science*, **1(3)**, 297–318.
- Hii, Y.L.(2009). Climate variability and increase in intensity and magnitude of dengue incidence in Singapore. *Global Health Action*.
- Johansson, M.A., Dominici, F., and Glass, G.E.(2009). Local and global effects of climate on dengue transmission in Puerto Rico. *PLoS Negl Trop Dis*, **3(2)**.
- Lam, S.K.(1993). Strategies for dengue control in Malaysia. *Tropical Medicine*, **35(4)**, 303–307.
- Lowe, R.(2011). Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. *Computers and Geosciences*, **37(3)**, 371–381.
- Lu, L., Lin, H., Tian, L., Yang, W., Sun, J., and Liu, Q.(2009). Time series analysis of dengue fever and weather in Guangzhou, China. *BMC Public Health*, **9(395)**.
- Rozilawati, H., Zairi, J., and Adanan, C.R.(2007). Seasonal abundance of *Aedes albopictus* in selected urban and suburban areas in Penang, Malaysia. *Tropical Biomedicine*, **24(1)**, 83–94.
- Rudnick, A., Tan, E.E., Lucas, J.K., and Omar, M.(1965). Mosquito-borne haemorrhagic fever in Malaya. *British Medical Journal*, **1**, 1269–1272.
- Skae, F.M.T.(1902). Dengue fever in Penang. *The British Medical Journal*, 1581–1582.

Modelling urban sprawl patterns in binary raster maps

D. Cocchi¹, Massimo Ventrucci¹, L. Altieri¹, E. Marian Scott²

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

Abstract: In urban studies the situation where an urban area develops in an inefficient manner is generally referred to as urban sprawl. This phenomenon leads to long-term negative effects for the environment, such as soil sealing, pollution and problems related to transport infrastructure. Official land use datasets, such as those from the CORINE Land Cover programme, provide a valuable resource to study urban dynamics. In this work we develop models for representing the urban sprawl phenomenon at a large spatial scale, which involves both estimating the proportion of urbanization over space and finding methods to quantify the urban shape.

Keywords: city boundary; CORINE; P-spline; urban shape; urban sprawl.

1 Introduction

The spatial configuration and the dynamics of urban growth are important topics in the analysis of contemporary urban studies. The situation where an urban area develops in an inefficient manner is described as urban sprawl. This phenomenon is linked to issues such as urban dispersion, low building and population density over rural or semi-rural areas, turning open spaces into built spaces. All these factors lead to long-term negative effects for the environment, above all soil sealing and pollution, and also socio-economic problems especially as regards transport infrastructure.

A clear definition of urban sprawl is missing in the literature. In Tsai (2005) urban sprawl is conceptualized through an intensity-based and a spatial structure-based definition. The former has to do with the proportion of urbanized land in the whole region of interest, while the latter is concerned with the spatial configuration or shape of the urban pattern in that region. Therefore, the intensity of the urban phenomena can vary over space, and as a result of this variation the urban pattern takes a particular shape. As regards the shape Tsai identifies three typical sprawl patterns. The first one is the monocentric case, where urbanization develops around a well compact city core. Second, the polycentric scenario, where urbanization

develops around two or more compact city cores. Third, the decentralized pattern, where urbanization is randomly sparse over space, which represents a case of extremely inefficient urban development. When the interest is in metropolitan areas around a city, normally observed patterns are noisy realizations of either the monocentric or polycentric scenario.

1.1 Motivating example

We aim at studying urban sprawl by using official land use raster data, from the CORINE Land Cover programme (<http://eea.europa.eu>), which basically consist of pixel maps, where each pixel is classified according to a category of land use. At the most detailed level CORINE data provide 44 land use classes available for all European countries, with a pixel size of $100m^2$. Data were extracted for all municipalities included in the Bologna metropolitan area, Italy, and subsequently reduced to binary classes, where 1 is assigned to artificial (i.e. urban) land use pixels and 0 to non artificial (i.e. non urban); see panel a) of FIG. 1.

Several statistical methods have been developed for lattice data, which could also be applied for the analysis of raster maps: for instance, the Moran's spatial association measure I . The use of Moran's I for measuring urban sprawl has been proposed by Tsai (2005). Moran's index varies from -1 to 1 , meaning respectively minimum and maximum spatial correlation between pixel values, while its value is expected to be 0 under spatial independence, i.e. when urban pixels are randomly scattered over the study region. A test to evaluate the null hypothesis of no spatial dependence is easy to compute by taking a standardized version of I , but for the purpose of investigating urban sprawl this test has limited interest. Interesting questions are rather about evaluating whether an observed urban pattern deviates from the monocentric or polycentric compact scenario; or perhaps, which proportion of land in the area under study is taken by a compact urban development and which by a sparse (i.e. sprawled) urbanization; or, producing maps where areas characterized by compact urban development are isolated from those where urbanization develops scarcely and sparsely. Under these circumstances, it appears that simply calculating Moran's I as a measure of global spatial correlation is not informative enough for characterizing urban sprawl. Several sprawl indices based on remote sensing data have been proposed in the geographical and urban literature (Bhatta et al., 2010), but measures able to account for both the proportion of urbanization and the shape of the urban patterns are generally lacking. In the present work we seek to model both components in two separate stages. In Sec. 2.1 we propose a smooth model for the pattern of urbanization over space. In Sec. 2.2 a test to identify what we will call the "city boundary area", i.e. an uncertainty region which lies between the *urban core* and the *suburbs*, is introduced. Also, a method to display the shape of the urban core and highlight possible deviations from a monocentric circular pattern is described.

Some results as regards the urban development of the metropolitan area around Bologna are given in Sec 3. Current and future work is discussed in Sec. 4.

2 Methods

2.1 Modelling the urban proportion over space

For modelling the pattern of urbanization underlying the data we assume response z_i , in each pixel $i = 1, \dots, n$, distributed as $B(1, p_i)$, where p_i is the probability of pixel i being urban. Because the land use phenomenon is a continuous process we can assume that p_i varies smoothly from pixel to pixel. For each pixel centroid location (x_i, y_i) , the model is:

$$g(p_i) = f(x_i, y_i) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \quad (1)$$

where $g(p_i)$ is the logit link function and $f(x_i, y_i)$ is a weighted sum of K cubic b-splines evaluated at each pixel location,

$$f(x_i, y_i) = \sum_{k=1}^K B_k(x_i, y_i) \alpha_k$$

The design matrix B for model (1), of dimension $n \times K$, is the tensor product $B_X \otimes B_Y$, where B_X and B_Y are marginal basis matrices defined over X and Y covariates. We adopt here the P-spline framework to model $f(x_i, y_i)$ in (1), and estimate coefficients α by penalized maximum likelihood (Eilers and Marx, 1996). Computation has been worked out by penalized iterative re-weighted least squares (PIRLS) (Wood, 2006). In the PIRLS solution a penalty term needs to be introduced in order to smooth spline coefficients α over both X and Y covariates, and a parameter needs to be chosen which controls the degree of smoothing.

Estimating α in this case involves a relevant computational challenge because raster data usually contain many pixels. However, the array structure of raster data can be fruitfully exploited. Raster data come in the form of a rectangular matrix of 0's and 1's for pixels located inside the boundary, and undefined value for pixels located outside the boundary; for example, in Fig. 1 the boundary is indicated by the grey line which delimits the metropolitan area around Bologna. Pixels outside the boundary can conveniently be set to 0, i.e. as non-urban, in order to obtain a full data matrix easy to handle for computations. This will not yield any relevant loss except for an estimation bias at the boundary. Therefore, model (1) can be seen as a Generalized Linear Array Model (GLAM) (Currie et al., 2006) and PIRLS be computed with efficient algorithms as those proposed in Eilers et al. (2006) which take advantage of the array structure of the data.

2.2 Modelling the urban shape

Fitted values from model (1) give a probability surface for urbanization. Uncertainty estimates can be provided by computing standard errors. This allows a simple test statistic to be calculated for each pixel:

$$z_i = \frac{\hat{p}_i - 0.5}{s.e.(\hat{p}_i)} \quad (2)$$

This statistic quantifies, in unit of standard error, the distance between the fitted value \hat{p}_i and a cutoff $p_0 = 0.5$ which represents a threshold separating the urban core from the suburban region. The aim of the test is to identify the city boundary area, i.e. a region of pixels with \hat{p}_i not significantly different from p_0 . Due to the asymptotic distribution of estimator \hat{p}_i and the large sample size typical of raster data, we can assume z_i 's distribution to be approximately standard normal. Therefore, pixels with $|z_i| < 1.96$ will be declared as belonging to the city boundary area.

It is of particular interest to find methods to quantify the shape of the urban core of the city, from which urbanization can grow in several directions in suburban areas. One strategy is to compare the observed urban core shape with an appropriate benchmark. We take the circle as benchmark, and refer to it as the *optimal* pattern of urban development, restricting our analysis to the monocentric scenario (i.e. when only one circle is the case).

We present in the following a suitable framework to express both the observed shape and the benchmark shape. First, given the locations of pixels inside the city boundary area, say (x_j, y_j) , and the location of the centre of the urban core, say (x_0, y_0) , we take the Euclidean distance $\hat{r}_j = \sqrt{(x_j - x_0)^2 + (y_j - y_0)^2}$, which is the radius in the case of the circle. Second, for each \hat{r}_j we consider its associated phase information, say $\phi_j \in (0, 2\pi)$ (phase is the angle between the horizontal axis and the line passing through (x_0, y_0) and (x_j, y_j)). An estimate of ϕ_j is calculated by inverting the equation: $x_j = x_0 + \hat{r}_j \cos(\phi_j)$. Because of the properties of the cosine function, and in order for $\hat{\phi}_j$ to span the interval $(0, 2\pi)$, extra care is needed in computation. In particular, pixels located in the north side of the city boundary area ($y_j \geq y_0$) must be handled differently from those located to the south ($y_j < y_0$): in the former case the phase angle estimate is $\hat{\phi}_j = \arccos((x_j - x_0)/\hat{r}_j)$, while in the latter we have $\hat{\phi}_j = 2\pi - \arccos((x_j - x_0)/\hat{r}_j)$.

Finally, we plot each \hat{r}_j as a function of ϕ_j , which returns a curve expressed over radians, say $\hat{r}(\phi)$. If the observed urban core had a perfect circular shape then $\hat{r}(\phi)$ would be a constant straight line. This suggests that visually checking the variability of $\hat{r}(\phi)$ around its mean gives a first insight into the degree of sprawl. Most importantly, new measures of sprawl can be derived from $\hat{r}(\phi)$, which quantify the divergence between the observed urban core pattern and an optimal pattern (the circle) having the same spatial extent.

3 Results

FIG. 1 panel a) shows the observed raster map for the metropolitan area around Bologna, delimited by the grey line. Raster data are arranged in a regular grid of dimension 340×310 ; note, pixels falling outside the metropolitan area are classified as non-urban. Panel b) and c) display a map of fitted values from model (1), i.e. of the smooth probability of urbanization: the intensity of grey colours emphasizes variations over space. The test statistic (2) was calculated in order to identify pixels inside the city boundary area, which is highlighted in red in panel c). Finally, panel d) displays the estimated curve $\hat{r}(\phi)$ (red dots). We see that this curve shows great variability around its mean (red dashed line), as the urban core presents an irregular shape, remarkably far from a circular shape.

4 Current and future work

In this work we presented a model to describe patterns of urban sprawl by using official land use data. P-spline smoothing of large binary raster data was performed and inferential tools were adopted to identify relevant spatial features, such as the city boundary area. The framework presented to express the shape of the city core area as a curve is the major output, which allows for a visual investigation. Future developments will consider more formal sprawl measures of divergence from the scenario of a circular monocentric development.

Further work is needed about the choice of the smoothing parameter. Standard selection techniques such as cross-validation are not appropriate because the raster dataset is too large and pixel values are spatially correlated. Finally, there is a clear multiple testing issue involved in evaluating the test statistics proposed, this will also be the target of future work.

Acknowledgments: Special thanks to Giovanna Pezzi from the Department of Experimental Evolutionistic Biology, University of Bologna, for her support in data preprocessing.

References

- Bhatta, B., Saraswati, S. and Bandyopadhyay, D. (2010). Urban sprawl measurement from remote sensing data. *Applied Geography*, 30, 731-740
- 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.

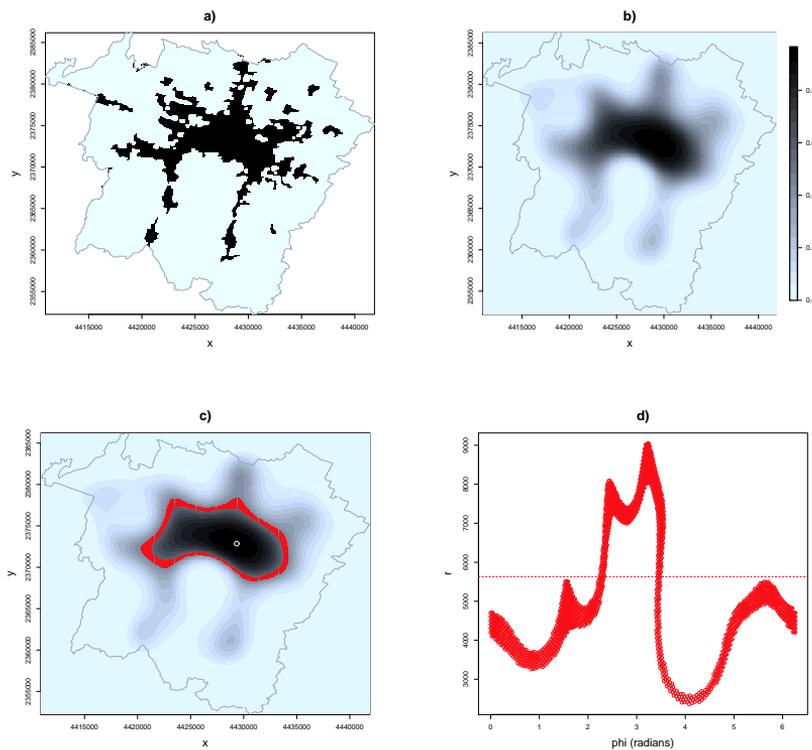


FIGURE 1. Panel a): the observed map of pixels (black=urban, white=non urban) for the Bologna metropolitan area. Panel b): model (1) fitted values which show the smooth pattern of urbanization. Panel c): the city boundary area (red) superimposed to the fitted values; centre (x_0, y_0) is the location of the pixel with maximum \hat{p}_i . Panel d): the curve $\hat{r}(\phi)$ which describes the shape of the urban core area as a function of the phase angle ϕ ; the mean radius (red dashed line) corresponds to a circular urban pattern.

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.

Eilers, P.C.H. and Marx, B.D. (1996). Flexible smoothing with splines and penalties (with discussion). *Statistical Science*, **11**, 735-751.

Tsai, Y.H. (2005). Quantifying urban form: compactness versus 'sprawl'. *Urban Studies*, **42(1)**, 141-161.

Wood, S. (2006). Generalized additive models. An introduction with R. *Chapman & Hall/CRC. Taylor & Francis group*.

Kalman filtering approach in the calibration of radar rainfall data

Marco Costa¹, Magda Monteiro², A. Manuela Gonçalves³

¹ Escola Superior de Tecnologia e Gestão de Águeda - Universidade de Aveiro, Portugal and CMAF-UL

² Escola Superior de Tecnologia e Gestão de Águeda - Universidade de Aveiro, Portugal and CIDMA-UA

³ Departamento de Matemática e Aplicações, Universidade do Minho, Portugal and CM-UM

E-mail for correspondence: marco@ua.pt

Abstract: This work presents a comparative study of some models to estimate radar rainfall in real time using the Kalman filtering approach. This comparison addresses the parameters estimation, the assessment of the accuracy estimates obtained by each model and the impact of the number of rain gauges used in the improvement of radar calibration estimates.

Keywords: Kalman filter, state space model, rainfall estimates, weather radar, calibration

1 Introduction

The weather radar provides precipitation data in a large area, for instance in a radial distance from the radar of 300Km (Figure 1). One of the advantages of radar rainfall over rain gauges is the provision of continuous measurements in real-time, which is unachievable even in a dense telemetered rain gauges network, since there is a large space-time variability of precipitation. However, their estimates have a poor performance, when comparing with gauges estimates, due to errors of either meteorological or instrumental nature which need to be reduced. Having this into account, in the recent years several approaches have been proposed to minimize radar errors, among which is included the combination of radar and gauges measurements, through a state space representation associated to the Kalman filter. This paper aims to discuss and compare different state space formulations through its application to the same data set. This comparison addresses the parameters estimation, the assessment of the accuracy estimates obtained by each model and the impact of the number of rain gauges used in the improvement of radar calibration estimates. It is also important to analyse the behaviour of different state space representations associated to different rain gauges network densities.

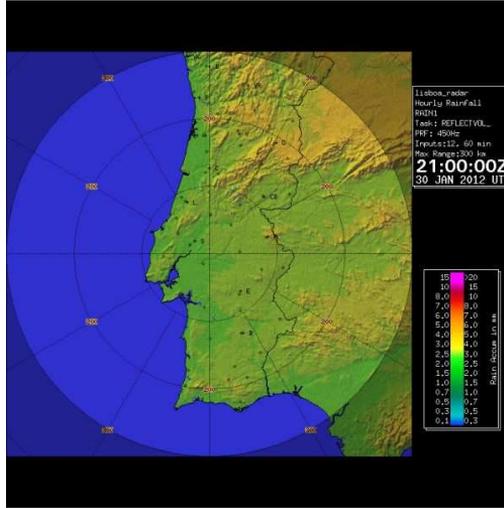


FIGURE 1. Radar umbrella of the weather radar located in Cruz de Leão, Coruche, Portugal

2 State space models and the Kalman filter

The Kalman filter approach provides a real-time scheme to calibrate radar rainfall estimates based on the rain gauge measurements. It is applied to a class of models that admits a state space representation of the form

$$A_t = \beta_t B_t + e_t \quad (1)$$

$$\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t. \quad (2)$$

Equation (1) is the measurement equation and relates the observable variable A_t with the unobservable variable β_t , called the state, while Eq. (2) is the transition or state equation. B_t is a known coefficient and e_t is the measurement error which is a white noise, with variance σ_e^2 . The state β_t is a stationary AR(1) process with mean μ , $|\phi| < 1$, where ε_t is a white noise with variance σ_ε^2 . Furthermore no assumption is made about the distributions of the disturbances e_t and ε_t , only that they are uncorrelated.

Assuming that parameters of the state-space model are known, the Kalman filter is an iterative algorithm that produces, at each time t , an estimator of the state vector β_t , which is the orthogonal projection of the state vector onto the observed variables up to that time. Let $\hat{\beta}_{t|t-1}$ represent the predictor of β_t based on the information up to time $t - 1$ and let $P_{t|t-1}$ be its mean square error (MSE).

The recursive process needs initial values for the state $\beta_{1|0}$ and for its

variance $P_{1|0}$, which in this case, as the state process is assumed to be a stationary AR(1) process it is taken

$\hat{\beta}_{1|0} = \mu$ and $P_{1|0} = \sigma_\beta^2 = \frac{\sigma_\varepsilon^2}{1-\phi^2}$. When the parameters of the model are unknown they have to be estimated and plugged in into the Kalman filter recursive equations. Since precipitation data deviates, in general, from the normal curve it will be considered non parametric methods to estimate the parameters, namely the consistent parameters estimators proposed by Costa and Alpuim (2010). The estimation for the mean of the state process $\{\beta_t\}$, μ , is the average of the ratios A_t/B_t and the remaining parameters of the state process $\{\beta_t\}$ are estimated based on the autocovariance structure of an AR(1) stationary process. The estimator of ϕ is obtained by least square method taking the autocovariances $\hat{\gamma}_k$, with $k = 1, \dots, \ell$, where ℓ is choose having into account the sample dimension. σ_ε^2 is estimated using $\hat{\sigma}_\varepsilon^2 = \frac{1-\hat{\phi}^2}{\hat{\phi}} \hat{\gamma}_1$ and the variance of the measurement equation is done through the relationship $var(\frac{A_t}{B_t}) = \sigma_\beta^2 + B_t^{-2} \sigma_\varepsilon^2$.

3 Models

3.1 Linear calibration (LC)

The linear calibration model was proposed by Alpuim and Barbosa (1999) and Costa and Alpuim (2010) and relates rain gauges and radar measurements through a multiplicative factor of calibration, as follows

$$\begin{aligned} G_t &= \beta_t R_t + e_t \\ \beta_t &= \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t. \end{aligned}$$

G_t is the rain gauge observation in time t , R_t is the radar measurement at the same time and location and β_t is the respective calibration factor. The LC model

does not impose any restrictions to the radar or rain gauges measurements unlike other that will be presented.

3.2 Mean field radar rainfall logarithmic bias modelling (FB)

The mean field radar rainfall logarithm bias model was proposed in Chumchean et al. (2006) and is based on the assumption that there are a consistent bias between radar and rain gauges measurements, that is,

$$Y_t = \frac{1}{k} \sum_{i=1}^k \log_{10} \left(\frac{G_{i,t}}{R_{i,t}} \right)$$

where k is the number of radar-gauge pairs data available in time t , and $G_{i,t}$ and $R_{i,t}$ are the rainfall and unfiltered radar rainfall at time t at location i .

The temporal evolution of the mean field logarithm bias is modeled through the state space model

$$\begin{aligned} Y_t &= \beta_t + e_t \\ \beta_t &= \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t. \end{aligned}$$

3.3 Power law modelling (PL)

Brown et al. (2001) make the assumption that gauge and radar reflectance measurements can be related through a power law, $G_t = bR_t^\alpha$. The authors consider a linearization of the power law where the parameters α and b are not necessarily fixed quantities but may vary stochastically over time. However they concluded that α could be treated as if it is constant, which result in

$$\begin{aligned} Y_t &= \alpha U_t + \beta_t + e_t \\ \beta_t &= \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t, \end{aligned}$$

where $Y_t = \log_{10}(G_t)$, $U_t = \log_{10}(R_t)$ and e_t is a white noise error. α is previously estimated by the method of least squares as the slope of the usual linear regression between Y_t and U_t .

Note that PL and FB models assume that both radar and gauges measurements are nonzero due to the logarithmic function. Another note to point out is that modelling procedure of LC and PL models is based on single-site approach, and it will be necessary to interpolate the predicted calibration factors β_t to other locations where it is intended to correct the radar measurements.

4 COMPARATIVE STUDY

It is used a data set of 17 stratiform storms between September of 1998 and November of 2000 (in a total of 178 hourly precipitation estimates) in a $10 \times 14 \text{ Km}^2$ area, located around 40 Km north of Lisbon at a distance from 31 to 44 Km from the weather radar in *Cruz do Leão*. This area has five rain gauges: Merceana (Mr), Meca (M), Olhalvo (O), Penedos (P) and Abrigada (A) and it has the highest gauge density under the radar umbrella ($\sim 1 \text{ gauge}/28\text{Km}^2$).

The performances of the calibration of the three models are compared in a set of scenarios considering all the combinations using 1, 2, 3 or 4 rain gauges to calibrate the radar estimates in the remaining gauges not used in the parameters estimation procedure in a total of 30 scenarios.

4.1 Models specification and calibration procedure

In order to ensure the independence between parameters estimation and the calibration modelling, three storms occurred in 13 of January, 28 of April and 19 of October of 2000 (62 hours) are used to estimate the models parameters, while the remaining storms are used in the assessment of the performance of the calibration.

The radar calibration procedure focus on the remaining fourteen storms not used in the parameter estimation (116 hourly measurements). The calibration procedure in the scenarios with more than one rain gauges needs interpolating its calibration factors to other locations for models LC and PL. In this case it is considered the inverse square distance method which takes into consideration all available rain gauges to calibrate the radar estimates.

For each scenario it was implemented the Kalman filter equations in order to predict the state β_t at each hour t . As it is considered a real-time procedure, the filtered prediction $\hat{\beta}_{t|t}$ is used to estimate β_t .

When the calibration procedure includes only a single rain gauge, the process to extend the calibration to other location is a straightforward procedure. This remains true even when the model applied is the FB since this model assumes a single mean field bias of calibration. Note that for LC model the radar calibration is obtained by multiplying the radar estimate R_t by the filtered calibration factor $\hat{\beta}_{t|t}^{(LC)}$, while in FB and PL models it is necessary to convert the respective $\hat{\beta}_{t|t}$ into B_t .

4.2 Performance assessment of models

The models performance assessment is done according to the empirical square root of the mean square error of point prediction using the fourteen storms (116 hourly observations) kept for this purpose. It is compared the gauges rainfall estimates G_t with the calibrated radar cell measurement $\hat{R}_t^{(m)}$, with $m = LC, FB$ and PL , at the same location. As it are available five rain gauges in the area under study, it is considered systems with 1, 2, 3 or 4 gauges to the calibration process and for each of these schemes are computed the empirical square root of the mean square error for all combination with each number of gauges.

The empirical square root of the mean square error $RMSE_k^{(m)}$ for the scheme modelled based on k rain gauges with the model m is computed by

$$RMSE_k^{(m)} = \sqrt{\frac{1}{116(5-k)} \sum_i^{5-k} \sum_t^{116} (G_t^i - \hat{R}_t^{(m),i})^2}$$

Table 1 presents the RMSE for the three models considering different numbers of rain gauges in the calibration process. The pre-calibration RMSE

TABLE 1. Square roots of the empirical mean square errors of the three models. In brackets is the % of RMSE reduction comparing to the reference value.

number of rain gauges	Model		
	LC	FB	PL
1	1.38 (-3%)	1.30 (-9%)	1.19 (-16%)
2	1.21 (-16%)	1.23 (-14%)	1.10 (-23%)
3	1.16 (-19%)	1.16 (-19%)	1.11 (-22%)
4	1.09 (-23%)	1.07 (-25%)	1.11 (-22%)
global	1.21 (-15%)	1.19 (-16%)	1.13 (-21%)

of the five rain gauges is taken as reference value to analyse the impact of the calibration procedures. For data set under the calibration procedure (the fourteen events) this value is 1.43. Thereby, it is possible to compare the models performance in view of the percentage of the reference value reduction (indicated in Table 1 in brackets).

It can be state that the models lead to a reduction in the error of the radar rain estimates. Neverytheless, the model PL is less sensitive to the number of rain gauges used in the calibration process. Both RMSE of models LC and FB decrease significantly when it is added more gauges to the calibration process. When the rain gauges density is the lowest (1 gauge per 140Km²) the PL model performed the largest reduction of the RMSE with a strong difference to the other models. On the other hand, when it is considered the highest density (1 gauge per 35Km²), models have similar performances, nevertheless the FB model produces the greatest reduction of the RMSE.

Acknowledgments: Marco Costa was partially supported by Portuguese Foundation for Science and Technology ("FCT-Fundação para a Ciência e a Tecnologia"), PEst OE/MAT/UI0209/2011. Magda Monteiro was partially supported by *FEDER* funds through *COMPETE* – Operational Programme Factors of Competitiveness ("Programa Operacional Factores de Competitividade") and by Portuguese funds through the *Center for Research and Development in Mathematics and Applications* (University of Aveiro) and the FCT, within project PEst-C/MAT/UI4106/2011 with *COMPETE* number FCOMP-01-0124-FEDER-022690. A. Manuela Gonçalves was partially financed by *FEDER* Funds through *COMPETE* and by FCT, within the Project Est-C/MAT/UI0013/2011.

References

- Alpuim T., Barbosa S. (1999). The Kalman filter in the estimation of area precipitation. *Environmetrics*, **10**, 377–394.
- Brown M., Diggle P., Lord M., Young P. (2001). Space-Time Calibration

of Radar Rainfall Data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **50(2)**, 221–241.

- Chumchean S., Seed A., Sharma A. (2006). Correcting of real-time radar rainfall bias using a Kalman filtering approach. *Journal of Hydrology*, **317**, 123–137.
- Costa M., Alpuim T. (2010). Parameter estimation of state space models for univariate observations. *Journal of Statistical Planning and Inference*, **140**, 1889–1902.
- Costa M., Alpuim T. (2011). Adjustment of state space models in view of area rainfall estimation. *Environmetrics*, **22**, 530–540.

Modeling operational risk losses

Jalila Daoudi

¹ Departamento de Estadística. Facultad de Ciencias Sociales y Jurídicas, Universidad Carlos III de Madrid, Spain

E-mail for correspondence: jdaoudi@est-econ.uc3m.es

Abstract: Operational risk has become an area of growing concern for financial institutions. Basel urges these organizations to measure and manage their risk. The observed distributions of operational risk losses often present heavy tails. In fact, many authors suggest use the extreme value theory that was helpful in a first approach to operational risk quantification. Nevertheless, it results in an extremely unrealistic amount of capital. An alternative is to use the g-h distribution or Kernel nonparametric distribution. We propose the DPLN which was proposed in insurance context, is flexible and it allows stable estimations of capital at Risk.

Keywords: Heavy tails; Operacional risk; Double Pareto lognormal distribution

1 Introduction

Under the proposed new accord (Basel II), operational risk has to be treated explicitly. The banks adopt the advanced measurement approach (AMA). This method consist in modeling the aggregate loss distribution which is also known as the loss Distribution Approach(LDA). An aggregate loss over a specified period of time can be expressed as the sum

$$S = \sum_{i=1}^n L_i,$$

where N is a random variable that represents the frequency of losses that occur over the period. L_i is an incident for which an entity suffers damages that can be measured with a monetary value. It is assumed that the L_i are independent and identically distributed, and each L_i is independent from N . The distribution of L_i is called the severity distribution.

The aggregate loss distribution S can be calculated using convolution techniques that are quite laborious and explicit solutions for the cumulative distribution of S can seldom be calculate. In this context, the process of aggregation is carried out by simulation techniques.

2 Modeling the aggregate distribution

In order to estimate the aggregate loss distribution, we first fit the frequency of occurrence and severity and then performs a process of aggregation or convolution of both. A capital at Risk is then calculated as the quantile 99.9% of the capital estimates for this aggregate distribution in holding period. The most frequent probability distributions to estimate the frequency of the events are Poisson distribution and the negative binomial distribution.

The data of operational risk losses presents heavy tails. In literature, many distributions have been proposed to model operational risk severity. The Subexponential family, in particular, the Pareto distribution which has been long used as a model for the tails in several fields such as hydrology, insurance, finance and environmental science, see for example Degen et al. (2007). However the Pareto distribution provides a good model for heavy tails, it is inappropriate for modeling the center of the distribution of many real data sets. Thus, in operational risk results to an inadequate amount of capital with no economic interpretation, infinite expected losses and unstable estimates of capital at Risk.

The medium distributions as the Gamma distribution and exponential distribution provide stability of capital see Degen et al. (2007). Nevertheless, they underestimate the capital at Risk. Therefore alternative models are needed to these distributions. Dutta and Perry (2006) proposed the parametric g-h distribution which is well adopted to operational risk more than the extreme value theory. The g-h distribution allows a reasonable capital with economic sense but it is sensitive to the introduction of new data. In this context, we use the double Pareto lognormal distribution (DPLN) introduced by Reed and Jorgensen (2004),

$$f(x; \alpha, \beta, \tau, \nu) = \frac{\alpha\beta}{\alpha + \beta} \frac{1}{x} \phi\left(\frac{\log x - \nu}{\tau}\right) [R(\alpha\tau - (\log x - \nu)/\tau) + R(\beta\tau + (\log x - \nu)/\tau)],$$

where

$$R(z) = \Phi^c(z)/\phi(z)$$

is the Mill's ratio, where $\Phi^c(z) = 1 - \Phi(z)$ and $\phi(z)$ and $\Phi(z)$ are the standard normal density and cumulative distributions respectively.

The DPLN distribution does not possess a moment generating function in closed form. However, if $n < \alpha$, the moment of order n exists.

Reed and Jorgensen use the maximum likelihood to estimate the parameters and note that this approach suffers from problems of convergence. We use a Bayesian approach proposed by Ramirez et al (2008) to estimate the parameters of DPLN distribution. On the other hand, we use bootstrapping methods to prove that the DPLN yields least varying capital at Risk.

3 Application

The data used for the analysis was collected for several banks participating in the survey to provide individual gross operational losses above a threshold, starting on 2002. This data was grouped in eight standardised business lines and seven event types. The data set used here correspond to the 125 largest losses associated with the business line 'Retail Banking' and the event type 'Internal Fraud'.

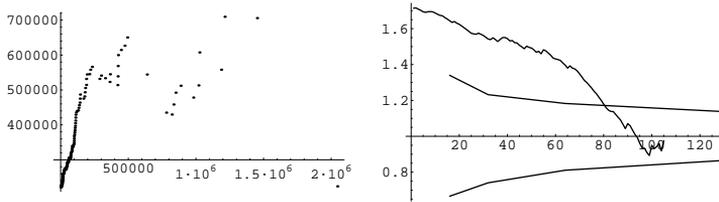


FIGURE 1. In the left, the ME-plot shows increasing line tendence. In the right, the CV-plot is outside the range defined under exponentiality.

The frequencies of events is modeled by Poisson distribution, then we calculate the capital at Risk under the Pareto distribution and the DPLN distribution.

In case of Pareto distribution, although the fluctuation in the estimated parameters appears relatively small, this leads to a significant capital at Risk estimates as we see in Figure 1.

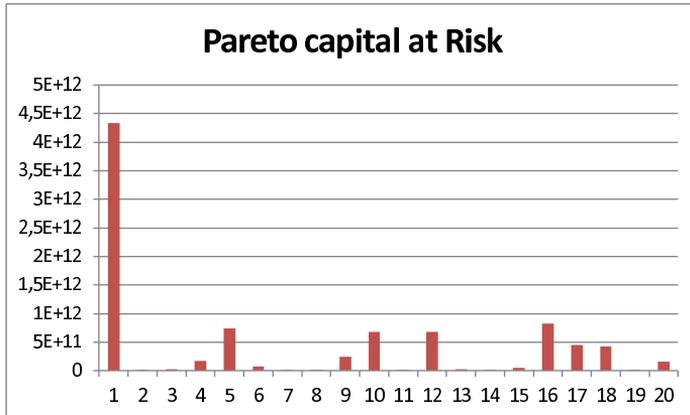


FIGURE 2. Instability of capital at Risk of Pareto using bootstrapping method.

In Figure 2, we can see that The DPLN distribution give the most reasonable with economic sense and a stable capital at Risk estimates.

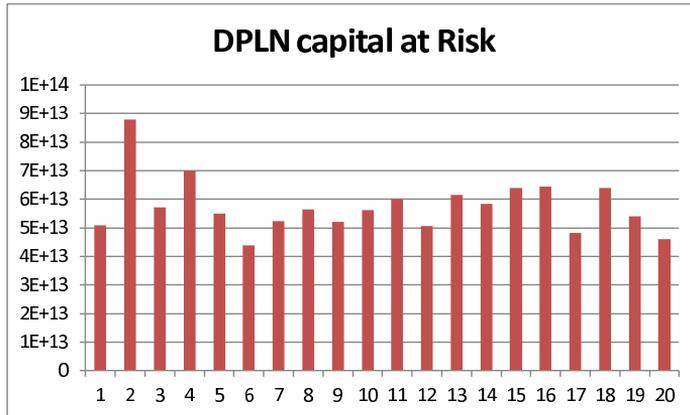


FIGURE 3. Stability of capital at Risk of DPLN using bootstrapping method.

References

- Reed, W. and Jorsensen, M. (2004). The double pareto-lognormal distribution—a new parametric model for size distributions. *Communications in Statistics: Theory and Methods*. **33(8)**, 1733-1753.
- Dutta, K., Perry, J. (2006). A tale of tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational Risk Capital *FRB of Boston Working Paper*, 06-13.
- Degen, M., Embrechts, P., and Lambrigger, D. (2007). The quantitative modeling of operational risk: Between g-and-h and EVT. *ASTIN Bulletin International Actuarial Association- Brussels, Belgium*. **37(2)**, 265-291.
- Ramirez, P., Lillo, R., Wiper, M. and Wilson, S. (2008b). Inference for double pareto lognormal queues with applications. *Statistics and Economics Working papers*, **ws084613**, Universidad Carlos III de Madrid.

Bayesian inference for power law processes with applications in repairable systems

Maristela Dias de Oliveira¹, Enrico A. Colosimo², Gustavo L. Gilardoni³

¹ Federal University of Bahia, Brazil

² Federal University of Minas Gerais, Brazil

³ University of Brasilia, Brazil

E-mail for correspondence: maridias@ufba.br

Abstract: Statistical models for recurrent events are of great interest in repairable systems reliability and maintenance. The adopted model under minimal repair maintenance is frequently a Nonhomogeneous Poisson Process with the Power Law Process (PLP) intensity function. Although inference for the PLP is generally based on maximum likelihood theory, some advantages of the Bayesian approach have been reported in the literature. In this paper it is proposed that the PLP intensity be reparametrized in terms of orthogonal parameters in that the likelihood function is proportional to a product of Gamma densities. Therefore, the family of natural conjugate priors is also a product of Gammas. The idea is extended to the case that several realizations of the same PLP are observed along overlapping periods of time. Some Monte Carlo simulations are provided to study the frequentist behavior of the Bayesian estimates and to compare them with the maximum likelihood estimates. The results are applied to a real problem concerning the determination of the optimal periodicity of preventive maintenance for a set of power transformers. Prior distributions are elicited for the orthogonal parameters based on their operational interpretation and engineering expertise.

Keywords: Conjugate prior; Optimal maintenance; Poisson Process; Posterior distribution; Reference priors.

1 Introduction

Statistical models for recurrent events have been investigated in many papers in the literature. Such models are of great interest to study the reliability and maintenance policies for repairable systems (Ascher and Feingold (1984), Rigdon and Basu (2000), among others). Frequently, the adopted model under minimal repair maintenance is a Nonhomogeneous Poisson Process (NHPP), $\{N(t) : t \geq 0\}$, where $N(t)$ is the number of failures from the beginning of the follow-up until time t (Barlow and Hunter (1960)). A flexible parametric form for the intensity function of the NHPP is the Power Law Process (PLP) (Crow (1974)): $\lambda(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta}\right)^{\beta-1}$, or equivalently

$\Lambda(t) = (t/\theta)^\beta$, where both β and θ are positive.

1.1 Bayesian Inference for PLP

Statistical inference for the PLP is generally based on the maximum likelihood estimator (MLE) and its asymptotic properties (Berman and Turner (1992), Zhao and Xie (1996)). Bayesian approach deals with the uncertainty of the parameters in the model used to describe a recurrent system. A prior distribution is assumed to represent the uncertainty in the model parameters before the current data is observed. Identifying a family of conjugate prior distributions will often result in mathematical and computational simplifications. Based in Huang and Bier (1998), we propose a conjugate prior distribution for the parameters of the PLP looking at the functional form of the likelihood function. Suppose that we observe n events at times $t_1 < \dots < t_n$ until a fixed time T . The likelihood function is

$$L(\beta, \theta) = \left[\prod_{i=1}^n \lambda(t_i) \right] e^{-\Lambda(T)} = \beta^n \theta^{-n\beta} \left[\prod_{i=1}^n t_i \right]^{\beta-1} \exp\{-(T/\theta)^\beta\} \quad (1)$$

(Rigdon and Basu (2000)). We propose here to parametrize the problem in terms of β and $\eta = \Lambda(T) = (T/\theta)^\beta$. On one side, β and η have simple operational definitions which will often make prior elicitation easier. On the other side, in the (β, η) parametrization the likelihood (1) becomes $L(\beta, \eta) = c [\beta^n e^{-n\beta/\hat{\beta}}] [\eta^n e^{-\eta}] \propto \gamma(\beta|n+1, n/\hat{\beta}) \gamma(\eta|n+1, 1)$, where $c = \prod_{j=1}^n t_j^{-1}$, $\hat{\beta} = n / \sum_{j=1}^n \log(T/t_j)$ is the MLE of β and $\gamma(x|a, b) = b^a x^{a-1} e^{-bx} / \Gamma(a)$ ($x, a, b > 0$) is the density of the Gamma distribution with shape and scale parameters equal to a and b , respectively. Notice that β and η are orthogonal and the natural conjugate family has densities of the form

$$\pi(\beta, \eta) = \gamma(\beta|a_\beta, b_\beta) \times \gamma(\eta|a_\eta, b_\eta). \quad (2)$$

The posterior density is $\pi(\beta, \eta|t_1, \dots, t_n, T) \propto \gamma(\beta|a_\beta + n, b_\beta + n/\hat{\beta}) \times \gamma(\eta|a_\eta + n, b_\eta + 1)$, so that both a priori and a posteriori β and η are independent, each following a Gamma distribution. The parametrization (β, η) suggests rather easily how to treat the case when several realizations of the PLP are observed along overlapping time intervals.

1.2 Paper goal and outline

Our interest is in the case of many overlapping realizations. In short, consider a repairable system modeled by a NHPP subject to two types of repairs: either a *minimal repair* after a failure which restores the system (i.e. the intensity) to exactly the same level it was immediately before the failure or a *preventive maintenance* which restores the system to “as good as new” condition. If the preventive maintenances are performed every τ units of

time, the expected cost per unit of time is $H(\tau) = [C_{PM} + C_{MR}EN(\tau)]/\tau = [C_{PM} + C_{MR}\Lambda(\tau)]/\tau$, where C_{MR} and C_{PM} are the expected costs associated to the two types of repair actions. It can be shown (Barlow and Hunter (1960) and Gilardoni and Colosimo (2007)) that the periodicity τ which minimizes $H(\tau)$ satisfies that $\tau\lambda(\tau) - \Lambda(\tau) = C_{PM}/C_{MR}$. In the special case of the PLP, τ becomes $\tau = \theta \left[\frac{C_{PM}}{(\beta-1)C_{MR}} \right]^{1/\beta}$. However, inference about τ only makes sense when $\beta > 1$, i.e., an increasing intensity function.

Besides Section 3 which deals with the many overlapping realizations setting and Monte Carlo simulations, the rest of the paper is organized as follows. In Section 2 we make some additional considerations regarding inference for a single realization of the PLP. The simulation scenarios and informative prior distributions are motivated from the real case discussed in Section 4.

2 A single PLP realization

Suppose a process observed in $(0, T)$. The MLE of β and η are $\hat{\beta}$ and n respectively. From the Fisher information matrix it follows that the asymptotic covariance matrix of $(\hat{\beta}, \hat{\eta})$ is $\text{Var}(\hat{\beta}, \hat{\eta}) \approx n^{-1} \text{Diag}(\beta^2, \eta^2)$. Suppose that the interest is centered in a function $\phi(\beta, \eta)$ such as $\lambda(T) = \beta \eta/T$. Under squared error loss, the Bayes estimate of ϕ is $E[\phi(\beta, \eta)|t_1, \dots, t_n] = \frac{1}{T} \frac{a_\beta+n}{b_\beta+n/\hat{\beta}} \frac{a_\eta+n}{b_\eta+1}$. Credible intervals can use the posterior quantiles of ϕ . The elicitation of proper informative priors in the (β, η) parametrization may be facilitated in view that both β and η have clear operational interpretations. In this sense, $\beta = \frac{d\Lambda(t)/\Lambda(t)}{dt/t}$ is the elasticity of the mean number of events with respect to time. On the other hand, $\eta = EN(T)$ is the expected number of events during the period that the process has been observed.

3 Several overlapping realizations of a PLP

This section considers K independent realizations of the same PLP, say $N_1(t), \dots, N_K(t)$, observed respectively up to times T_1, \dots, T_K . Let t_{ij} be the j -th event time for the i -th realization, $i = 1, \dots, K; j = 1, \dots, n_i$. If we reparametrize the problem in terms of β and $\eta = \sum_{i=1}^K (T_i/\theta)^\beta$, it follows that the likelihood function is $L(\beta, \eta) \propto \gamma(\eta|n+1, 1) \gamma(\beta|n+1, n/\hat{\beta}) e^{nF(\beta)}$ where now $\hat{\beta}$ satisfies $\frac{1}{n} \sum_{i=1}^K \sum_{j=1}^{n_i} \log t_{ij} = \frac{\sum_{i=1}^K T_i^\beta \log T_i}{\sum_{i=1}^K T_i^\beta} - \frac{1}{\hat{\beta}}$ and $F(\beta) = \frac{\sum_{i=1}^K T_i^\beta \log T_i}{\sum_{i=1}^K T_i^\beta} \beta - \log \sum_{i=1}^K T_i^\beta$. Note that β and η are still orthogonal. The Fisher information matrix is $I(\beta, \eta) = n \text{Diag}(\beta^{-2} - F''(\beta), \eta^{-2})$, where $F''(\beta)$ is formally the same as minus the variance of a random variable taking values $\log T_i$ with probabilities proportional to T_i^β ($i = 1, \dots, K$).

3.1 Posterior analysis

Under the prior specification (2), the posterior density becomes

$$\pi(\beta, \eta | D) \propto \gamma(\eta | a_\eta + n, b_\eta + 1) \times \gamma(\beta | a_\beta + n, b_\beta + n/\hat{\beta}) \times e^{nF(\beta)}, \quad (3)$$

where $D = \{t_{ij} : i = 1, \dots, K; j = 1, \dots, n_i\}$. In order to sample from the posterior distribution (3) we use the independence between η and β and obtain first $\eta_1, \dots, \eta_m \stackrel{i.i.d.}{\sim} \text{Gamma}(a_\eta + n, b_\eta + 1)$. Simulation from the posterior distribution of β becomes easy by using, for instance, the rejection sampling algorithms (see Gelman *et al.* (2003)). Since we are interested in estimate τ , the rejection algorithm can also be used when the prior for β is a Gamma distribution truncated to the right of $\beta = 1$. In this case, one just changes the proposal distribution above to be also a truncated Gamma. On other hand if one wants to consider a prior distribution for β which is restricted to have support in $(1, \infty)$ and which is not a truncated Gamma, may be better to consider a shifted Gamma prior, i.e. $\beta - 1 \sim \text{Gamma}(a_\beta, b_\beta)$, and use the Metropolis algorithm to obtain an approximate sample from the posterior of β .

3.2 Monte Carlo Simulation

In this section we describe some Monte Carlo simulations in order to compare MLE and Bayes estimates under different prior specifications in the case of overlapping realizations of a PLP. The quantity of interest in the simulations was τ , defined in Section 1, thus the prior distribution must satisfy $Pr(\beta > 1) = 1$. Different prior distributions for β and η were used in the simulations, according following notations:

* MLE - Maximum likelihood estimate; * BayesE₁ - Bayes estimator by considering a reference prior distribution $\pi(\beta, \eta) \propto \beta^{-2} \eta^{-1}$, truncated at $\beta = 1$; * BayesE₂ - Bayes estimator by considering Jeffrey's prior distribution $\pi(\beta, \eta) \propto (\beta\eta)^{-1}$, truncated at $\beta = 1$; * BayesE₃ - Bayes estimator by considering gamma prior distributions (2) truncated at $\beta = 1$; * BayesE₄ - Bayes estimator by considering a gamma prior distribution shifted to 1 for β and gamma for η and * CP - Interval Coverage Percentage.

Throughout the Monte Carlo study we consider $\tau = 6$. The number of systems K and truncation times T_i 's were set to study three different situations: in Situation 1 $K = 500$ systems all truncated at $T = 100$, Situation 2 considers $K = 50$ systems all truncated at $T = 320$, the third situation has $K = 50$ systems truncated at $T = 30$. The results, based on 3000 replicas, are shown in Table 1. BayesE₄ has the worst interval coverage. In general, all estimates have a small bias, the MLE being the least biased.

TABLE 1. Summary of simulation results

		MLE	BayesE ₁	BayesE ₂	BayesE ₃	BayesE ₄
Situation 1	Mean of $\hat{\tau}$	6.00	6.00	6.00	6.00	6.02
	CP	94.9	94.3	94.4	94.4	95.0
	Mean length	0.475	0.475	0.474	0.474	0.482
Situation 2	Mean of $\hat{\tau}$	6.00	6.00	6.00	6.00	6.03
	CP	94.7	94.5	94.6	94.6	95.0
	Mean length	0.753	0.753	0.752	0.753	0.765
Situation 3	Mean of $\hat{\tau}$	6.11	6.17	6.12	6.13	6.16
	CP	95.4	95.2	95.9	95.1	93.1
	Mean length	2.125	2.149	2.125	2.124	1.971

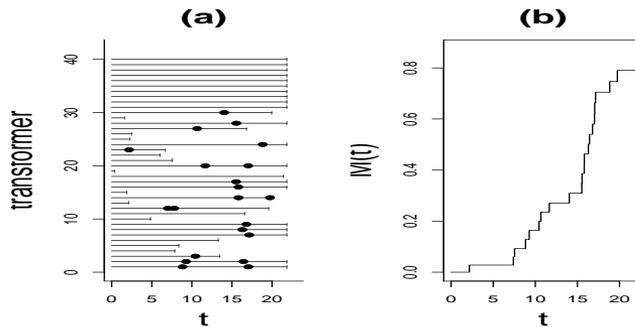


FIGURE 1. (a) Power transformer data. Each horizontal line represents a transformer and dots are observed failure times; (b) Mean Cumulative Failure (MCF) type estimate for Λ .

4 Example: Maintenance of electrical power transformers

Figure 1(a) shows the failure history of 40 electrical power transformers. The usual nonparametric estimate of Λ is the Nelson-Aalen estimate and is shown in Figure 1(b). The convex form of this plot provides some evidence in the sense that the intensity function is increasing. Two informative prior distributions were elicited: BayesI₁, considers $(\beta, \eta) \sim \text{Gamma}(20,10) \times \text{Gamma}(10,0.5)$ and BayesI₂ was set up as $(\beta, \eta) \sim \text{Gamma}(50,25) \times \eta \sim \text{Gamma}(30,1.5)$. These two informative priors are shown in Figure 2. We also consider in the analysis the four noninformative prior distributions discussed previously. The intervals based on the informative priors BayesI₁ and, especially BayesI₂ are the shortest ones, respectively 2.52 and 2.06, nearly ML, 2.83. In other hand, the remaining are bigger than 3.4. Point estimates are in agreement among Bayesian methods taking a value around 6400 hours, although the ML estimate of 6290 hours is slightly smaller.

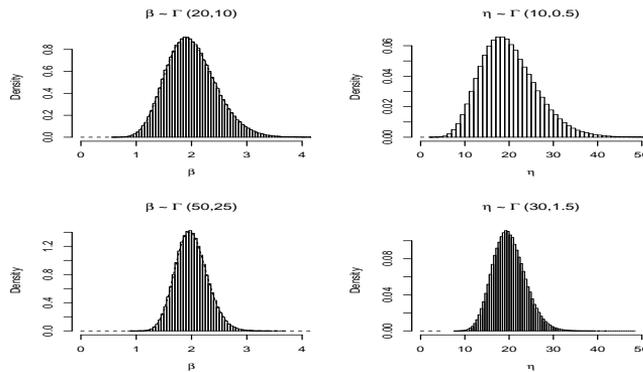


FIGURE 2. Prior distributions: (1) top: least informative: left for β and right for η ; (2) bottom: most informative: left for β and right for η .

References

- Ascher, H. and Feingold, H. W. (1984). *Repairable Systems Reliability: Modeling, Inference, Misconception and their Causes*. New York: Marcel Dekker.
- Barlow, R. E. and Hunter, L. C. (1960). Optimum Preventive Maintenance Policies. *Operations Research* **8**, 90–100.
- Berman, M. and Turner, T. R. (1992). Approximating Point Process Likelihoods with GLIM. *Applied Statistics* **41**, pp. 31–38.
- Crow, L. R. (1974). Reliability Analysis for Complex Systems. *Reliability and Biometry*. F. Proschan and J. Serfling (Eds), pp. 379–410.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman and Hall: New York.
- Gilardoni, G. L. and Colosimo, E. A. (2007) Optimal maintenance time for repairable systems. *Journal of Quality Technology* **39**, pp. 48–53.
- Huang, Y. S. and Bier, V. B. (1998). A Natural Conjugate Prior for the Non-Homogeneous Poisson Process with a Power Law Intensity Function. *Communications in Statistics - Simulation and Computation* **27**, pp. 525-551.
- Rigdon, S. E. and Basu, A. P. (2000). *Statistical Methods for the Reliability of Repairable Systems*. New York: John Wiley.
- Zhao, M. and Xie, M. (1996). On Maximum Likelihood Estimation for a General Non-homogeneous Poisson Process. *Scandinavian Journal of Statistics* **23**, 597-607.

The $(\mathbb{Z}^d \times \mathbb{Z})$ spatial-temporal Auto-Linear-Auto-Regressive model

Chrysoula Dimitriou-Fakalou ¹

¹ School of Mathematics, Statistics and Actuarial Science, University of Kent, United Kingdom

E-mail for correspondence: `C.Dimitriou@kent.ac.uk`

Abstract: The ALAR model is written to express naturally the spatial-temporal second-order dependence in the $(d + 1)$ -dimensional lattice, where d refers to the number of spatial line transects. The new equation extends the spatial Auto-Linear and temporal Auto-Regressive dependence for a parameter driven spatial-temporal process. A suggestion for the estimation of both types of parameters is given and the natural prediction is a direct consequence, such that the new model may be useful for stationary or even non-stationary spatial-temporal series indexed in $(\mathbb{Z}^d \times \mathbb{Z})$.

Keywords: Auto-linear; Auto-regressive; Distribution-free; Weak stationarity

1 Introduction

Following the standard results for stationary time series, the conditional auto-normal models of Besag (1974) and the simultaneous auto-regressive models of Whittle (1954) in the plane were combined for the spatial auto-linear equations of Dimitriou-Fakalou (2010).

Despite the rapid development for the purely spatial or temporal dependencies, until recently the lack of valid nonseparable spatial-temporal covariance functions remained an obstacle for the combined dynamics. Cressie and Huang (1999) followed by Gneiting (2002), Ma (2003) and Stein (2005) have attempted the definition of new classes of nonseparable covariance functions for spatial-temporal processes in $(\mathbb{R}^d \times \mathbb{R})$ and there has been additional work done in $(\mathbb{R}^d \times \mathbb{Z})$.

An auto-linear and auto-regressive type argument will be used here to write a new model, which is appropriate for stationary spatial-temporal processes in $(\mathbb{Z}^d \times \mathbb{Z})$; these are also the main building block for some forms of non-stationary dependence. The parametrization is natural for the prediction over the $(\mathbb{Z}^d \times \mathbb{Z})$ space-time and estimation is straightforward.

2 Natural models

The auto-regressive equation has been the first attempt to express the stationary second-order dependence for a time series $\{X_t, t \in \mathbb{Z}\}$. Then for a finite positive integer p , it holds that

$$X_t - \sum_{i=1}^p \varphi_i X_{t-i} = \varepsilon_t, \tag{1}$$

where $\{\varepsilon_t\}$ are uncorrelated random variables with zero mean and an identical finite variance. The representation (1) must be causal to secure that $\text{Cov}\{\varepsilon_t, X_{t-i}\} = 0, i > 0$.

Next, the auto-linear model expresses naturally the covariance dependence of a stationary spatial process $\{X(\mathbf{s}), \mathbf{s} \in \mathbb{Z}^d\}$ in d dimensions and writes

$$X(\mathbf{s}) - \sum_{n=1}^p \beta_{\mathbf{i}_n} (X(\mathbf{s} - \mathbf{i}_n) + X(\mathbf{s} + \mathbf{i}_n)) = Y(\mathbf{s}); \tag{2}$$

in order to make sure that (2) uses the best linear predictor of $X(\mathbf{s})$ based on all $X(\mathbf{s} - \mathbf{i}), \mathbf{i} \neq \mathbf{0}$ from both the ‘past’ and the ‘future’, Dimitriou-Fakalou (2010) showed that the latent variables $\{Y(\mathbf{s}), \mathbf{s} \in \mathbb{Z}^d\}$ have to be correlated on the ‘lags’ $\pm \mathbf{i}_n \neq \mathbf{0}, n = 1, \dots, p$ with auto-correlations $-\beta_{\mathbf{i}_n}$. Note that for the zero mean variables of interest, the best linear predictor $\hat{X}(\mathbf{s})$ of $X(\mathbf{s})$ based on $X(\mathbf{s} - \mathbf{j}_l), \mathbf{j}_l \neq \mathbf{0}, l = 1, \dots, m$, is such that

$$\text{Cov}\{X(\mathbf{s}) - \hat{X}(\mathbf{s}), X(\mathbf{s} - \mathbf{j}_l)\} \equiv E\{(X(\mathbf{s}) - \hat{X}(\mathbf{s}))X(\mathbf{s} - \mathbf{j}_l)\} \equiv 0$$

holds for all $l = 1, \dots, m$; it is assumed that this predictor is unique and that $\text{Var}(X(\mathbf{s}) - \hat{X}(\mathbf{s}))$ is finite.

For a stationary spatial-temporal process $\{X_t(\mathbf{s}), \mathbf{s} \in \mathbb{Z}^d, t \in \mathbb{Z}\}$, it will be assumed that the natural best linear predictor of $X_t(\mathbf{s})$ based on all the information from the ‘past’ $X_{t-i}(\mathbf{s}^*), i > 0, \mathbf{s}^* \in \mathbb{Z}^d$ as well as the information from the neighbors ‘around’ the location of interest at the same time $X_t(\mathbf{s} - \mathbf{j}), \mathbf{j} \neq \mathbf{0}$ reduces to the finite sum

$$\hat{X}_t(\mathbf{s}) \equiv \sum_{i=1}^{p_t} \sum_{\mathbf{j} \in \mathcal{U}_i} \varphi_{\mathbf{j},i} X_{t-i}(\mathbf{s} - \mathbf{j}) + \sum_{n=1}^{p_s} (\beta_{\mathbf{j}_n} X_t(\mathbf{s} - \mathbf{j}_n) + \beta_{-\mathbf{j}_n} X_t(\mathbf{s} + \mathbf{j}_n)); \tag{3}$$

in (3), p_t and p_s are finite positive integers, $\mathcal{U}_i \subset \mathbb{Z}^d, i = 1, \dots, p_t$ are sets of finite cardinality and the spatial orderings $\mathbf{0} < \mathbf{j}_1 < \dots < \mathbf{j}_{p_s}$ take place for convenience only as, for any $n = 1, \dots, p_s$, both ‘lags’ $\pm \mathbf{j}_n$ are used in the equation. Consequently, it must hold for any $\mathbf{s} \in \mathbb{Z}^d, t \in \mathbb{Z}$ that

$$\text{Cov}\{X_t(\mathbf{s}) - \hat{X}_t(\mathbf{s}), X_{t-i}(\mathbf{s}^*)\} = 0, i > 0, \mathbf{s}^* \in \mathbb{Z}^d \text{ and that} \tag{4}$$

$$\text{Cov}\{X_t(\mathbf{s}) - \hat{X}_t(\mathbf{s}), X_t(\mathbf{s} - \mathbf{j})\} = 0, \mathbf{j} \neq \mathbf{0}. \tag{5}$$

Then $Y_t(\mathbf{s}) \equiv X_t(\mathbf{s}) - \hat{X}_t(\mathbf{s})$, $\mathbf{s} \in \mathbb{Z}^d$, $t \in \mathbb{Z}$ is a best linear prediction error; since $\{Y_t(\mathbf{s}), \mathbf{s} \in \mathbb{Z}^d, t \in \mathbb{Z}\}$ are defined by applying a linear ‘space-time’ invariant filter on a second-order stationary process, then they are also a stationary spatial-temporal process. It holds for all $n = 1, \dots, p_s$, that

$$\begin{aligned} & \text{Cov}\{Y_t(\mathbf{s}), Y_t(\mathbf{s} - \mathbf{j}_n)\} = E\{Y_t(\mathbf{s})Y_t(\mathbf{s} - \mathbf{j}_n)\} = E\{Y_t(\mathbf{s})X_t(\mathbf{s} - \mathbf{j}_n)\} \\ & - \sum_{m=1}^{p_s} [\beta_{\mathbf{j}_m} E\{Y_t(\mathbf{s})X_t(\mathbf{s} - \mathbf{j}_n - \mathbf{j}_m)\} + \beta_{-\mathbf{j}_m} E\{Y_t(\mathbf{s})X_t(\mathbf{s} - \mathbf{j}_n + \mathbf{j}_m)\}] \\ & - \sum_{i=1}^{p_t} \sum_{\mathbf{j} \in \mathcal{U}_i} \varphi_{\mathbf{j},i} E\{Y_t(\mathbf{s})X_{t-i}(\mathbf{s} - \mathbf{j}_n - \mathbf{j})\} = -\beta_{-\mathbf{j}_n} E\{Y_t(\mathbf{s})X_t(\mathbf{s})\}, \end{aligned}$$

where the last equality holds thanks to (4) and (5). Similarly, it can be demonstrated that $\text{Cov}\{Y_t(\mathbf{s}), Y_t(\mathbf{s} + \mathbf{j}_n)\} = -\beta_{\mathbf{j}_n} E\{Y_t(\mathbf{s})X_t(\mathbf{s})\}$ and that it must hold that

$$\beta_{\mathbf{j}_n} = \beta_{-\mathbf{j}_n}, \quad n = 1, \dots, p_s. \tag{6}$$

The fact that $\text{Cov}\{Y_t(\mathbf{s}), Y_t(\mathbf{s} - \mathbf{j}_n)\} = \text{Cov}\{Y_t(\mathbf{s}), Y_t(\mathbf{s} + \mathbf{j}_n)\}$, $n = 1, \dots, p_s$ is true in general, not just for space-time separable auto-covariance functions in $(\mathbb{Z}^d \times \mathbb{Z})$. Thus, the requirement (6) is non-negotiable similarly to the spatial symmetry conditions of the auto-normal models of Besag (1974) or Dimitriou-Fakalou (2010). Hence the definition that follows.

Definition 2.1: The (zero mean) variables $\{X_t(\mathbf{s}), \mathbf{s} \in \mathbb{Z}^d, t \in \mathbb{Z}\}$ form an Auto-Linear-Auto-Regressive process of finite spatial-temporal order (p_s, p_t) if they satisfy the equation

$$X_t(\mathbf{s}) - \sum_{n=1}^{p_s} \beta_{\mathbf{j}_n} (X_t(\mathbf{s} - \mathbf{j}_n) + X_t(\mathbf{s} + \mathbf{j}_n)) - \sum_{i=1}^{p_t} \sum_{\mathbf{j} \in \mathcal{U}_i} \varphi_{\mathbf{j},i} X_{t-i}(\mathbf{s} - \mathbf{j}) = Y_t(\mathbf{s}), \tag{7}$$

where $\mathcal{U}_1, \dots, \mathcal{U}_{p_t} \subset \mathbb{Z}^d$ have finite cardinality and $\{Y_t(\mathbf{s}), \mathbf{s} \in \mathbb{Z}^d, t \in \mathbb{Z}\}$ are zero mean random variables with identical finite variance and auto-correlations

$$\text{Corr}\{Y_t(\mathbf{s}), Y_{t^*}(\mathbf{s}^*)\} = \begin{cases} 1, & t^* = t, \mathbf{s}^* = \mathbf{s} \\ -\beta_{\mathbf{j}_n}, & t^* = t, \mathbf{s}^* = \mathbf{s} \pm \mathbf{j}_n, n = 1, \dots, p_s \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Further, it must hold that $\beta(\mathbf{z}) \equiv 1 - \sum_{n=1}^{p_s} \beta_{\mathbf{j}_n} (\mathbf{z}^{\mathbf{j}_n} + \mathbf{z}^{-\mathbf{j}_n})$, $\mathbf{z} = (z_1, \dots, z_d)$ with $z_k = e^{-i\omega_k}$, $\omega_k \in (-\pi, \pi)$, $k = 1, \dots, d$ and $i = \sqrt{-1}$, generates a positive-definite auto-correlation function in \mathbb{Z}^d and that

$$\left\{ 1 - \sum_{i=1}^{p_t} \beta(\mathbf{z})^{-1} \sum_{\mathbf{j} \in \mathcal{U}_i} \varphi_{\mathbf{j},i} \mathbf{z}^{\mathbf{j}} z_t^i \right\}^{-1} \equiv 1 + \sum_{i=1}^{\infty} \sum_{\mathbf{j}} \Phi_{\mathbf{j},i} \mathbf{z}^{\mathbf{j}} z_t^i$$

has absolutely summable coefficients $\sum_{i=1}^{\infty} \sum_{\mathbf{j}} |\Phi_{\mathbf{j},i}| < \infty$.

3 Initial estimation

For observations $\{X_t(\mathbf{s}), \mathbf{s} \in \mathcal{S}, t = 1, \dots, T\}$ where $\mathcal{S} \subset \mathbb{Z}^d$ is a set of cardinality N , an obvious estimator for the ALAR model is the minimizer $\tilde{\boldsymbol{\theta}}$ of $\sum Y_t(\mathbf{s}, \boldsymbol{\theta})^2$, where the summation can extend over $t = p_t + 1, \dots, T$ and over appropriate values of \mathbf{s} to deal with the ‘edge-effects’ in \mathbb{Z}^d (see Guyon (1982)); note that this is a combination of spatial maximum (Gaussian) pseudo-likelihood and temporal least squares estimation. In order to account for the weak dependence (8), forms of weighted least squares estimation may be considered in the future.

Writing for any $\boldsymbol{\theta}$ (with φ, β) from the allowable parameter values $Y_t(\mathbf{s}, \boldsymbol{\theta}) \equiv Y_t(\mathbf{s}) - \mathbf{X}_t^T(\mathbf{s})(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and $\tilde{Y}_t(\mathbf{s}) \equiv Y_t(\mathbf{s}, \tilde{\boldsymbol{\theta}})$, where $\boldsymbol{\theta}_0$ is the true parameter vector and $\mathbf{X}_t(\mathbf{s})$ is the (column) vector with elements $X_{t-i}(\mathbf{s} - \mathbf{j})$, $i = 1, \dots, p_t$, $\mathbf{j} \in \mathcal{U}_i$ and $X_t(\mathbf{s} - \mathbf{j}_n) + X_t(\mathbf{s} + \mathbf{j}_n)$, $n = 1, \dots, p_s$, it holds that

$$\sum \tilde{Y}_t(\mathbf{s})X_{t-i}(\mathbf{s} - \mathbf{j}) = 0, \quad i = 1, \dots, p_t, \quad \mathbf{j} \in \mathcal{U}_i \quad \text{and that} \quad (9)$$

$$\sum \tilde{Y}_t(\mathbf{s})[X_t(\mathbf{s} - \mathbf{j}_n) + X_t(\mathbf{s} + \mathbf{j}_n)] = 0, \quad n = 1, \dots, p_s. \quad (10)$$

And, of course, (9) and (10) may be summarized in

$$(NT)^{1/2} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left\{ \sum \mathbf{X}_t(\mathbf{s})\mathbf{X}_t^T(\mathbf{s}) / (NT) \right\}^{-1} \{ (NT)^{-1/2} \sum \mathbf{X}_t(\mathbf{s})Y_t(\mathbf{s}) \}. \quad (11)$$

Writing the auto-covariances $c(\mathbf{j}, i) \equiv E\{X_t(\mathbf{s})X_{t-i}(\mathbf{s} - \mathbf{j})\}$ and, in particular, $\gamma(\mathbf{j}) \equiv c(\mathbf{j}, 0)$ for any $\mathbf{j} \in \mathbb{Z}^d$ and $i \in \mathbb{Z}$, it may be written that

$$C \equiv E\{\mathbf{X}_t(\mathbf{s})\mathbf{X}_t^T(\mathbf{s})\} \equiv \begin{pmatrix} C^* & G \\ G^T & \Gamma \end{pmatrix} \quad (12)$$

with elements $c(\mathbf{j} - \mathbf{j}^*, i - i^*)$, $i, i^* = 1, \dots, p_t$, $\mathbf{j} \in \mathcal{U}_i$, $\mathbf{j}^* \in \mathcal{U}_{i^*}$ in C^* , elements $c(\mathbf{j} - \mathbf{j}_n, i) + c(\mathbf{j} + \mathbf{j}_n, i)$, $i = 1, \dots, p_t$, $\mathbf{j} \in \mathcal{U}_i$, $n = 1, \dots, p_s$ in G and, finally, $2(\gamma(\mathbf{j}_n - \mathbf{j}_m) + \gamma(\mathbf{j}_n + \mathbf{j}_m))$, $n, m = 1, \dots, p_s$ in Γ .

On the other hand and for the first part, it is considered that $i, i^* = 1, \dots, p_t$, $\mathbf{j} \in \mathcal{U}_i$, $\mathbf{j}^* \in \mathcal{U}_{i^*}$ and the element

$$\sum_{\mathbf{s} - \mathbf{s}^* \in \mathbb{Z}^d} \sum_{t - t^* \in \mathbb{Z}} E\{Y_t(\mathbf{s})X_{t-i}(\mathbf{s} - \mathbf{j})Y_{t^*}(\mathbf{s}^*)X_{t^*-i^*}(\mathbf{s}^* - \mathbf{j}^*)\}$$

is of interest. For Gaussian random variables, this is equal to

$$\begin{aligned} & \sum_{\mathbf{s} - \mathbf{s}^* \in \mathbb{Z}^d} E\{Y_t(\mathbf{s})Y_{t^*}(\mathbf{s}^*)\}E\{X_{t-i}(\mathbf{s} - \mathbf{j})X_{t^*-i^*}(\mathbf{s}^* - \mathbf{j}^*)\} \\ &= -\nu \sum_{n=1}^{p_s} \beta_{\mathbf{j}_n} [c(\mathbf{j} - \mathbf{j}^* - \mathbf{j}_n, i - i^*) + c(\mathbf{j} - \mathbf{j}^* + \mathbf{j}_n, i - i^*)] \\ &+ \nu \quad c(\mathbf{j} - \mathbf{j}^*, i - i^*), \end{aligned} \quad (13)$$

where it is written $\nu \equiv E\{Y_t(\mathbf{s})^2\} = E\{Y_t(\mathbf{s})X_t(\mathbf{s})\}$. The matrix $B_{(T)}$ with elements (13) for all $i, i^* = 1, \dots, p_t, \mathbf{j} \in \mathcal{U}_i, \mathbf{j}^* \in \mathcal{U}_{i^*}$ is considered. For the second part and $i = 1, \dots, p_t, \mathbf{j} \in \mathcal{U}_i$ and $n = 1, \dots, p_s$, the element

$$\sum_{\mathbf{s}-\mathbf{s}^* \in \mathbb{Z}^d} \sum_{t-t^* \in \mathbb{Z}} E\{Y_t(\mathbf{s})X_{t-i}(\mathbf{s}-\mathbf{j})Y_{t^*}(\mathbf{s}^*)[X_{t^*}(\mathbf{s}^*-\mathbf{j}_n) + X_{t^*}(\mathbf{s}^*+\mathbf{j}_n)]\}$$

will be of interest. For Gaussian random variables, this becomes

$$\begin{aligned} & \sum_{\mathbf{s}-\mathbf{s}^* \in \mathbb{Z}^d} E\{Y_t(\mathbf{s})Y_{t^*}(\mathbf{s}^*)\}E\{X_{t-i}(\mathbf{s}-\mathbf{j})[X_{t^*}(\mathbf{s}^*-\mathbf{j}_n) + X_{t^*}(\mathbf{s}^*+\mathbf{j}_n)]\} \\ &= -\nu \sum_{m=1}^{p_s} \beta_{\mathbf{j}_m} [c(\mathbf{j}-\mathbf{j}_m-\mathbf{j}_n, i) + c(\mathbf{j}-\mathbf{j}_m+\mathbf{j}_n, i) + c(\mathbf{j}+\mathbf{j}_m-\mathbf{j}_n, i) \\ &+ c(\mathbf{j}+\mathbf{j}_m+\mathbf{j}_n, i)] + \nu [c(\mathbf{j}-\mathbf{j}_n, i) + c(\mathbf{j}+\mathbf{j}_n, i)] \end{aligned} \tag{14}$$

and the matrix $B_{(ST)}$ is considered with elements (14) for all $i = 1, \dots, p_t, \mathbf{j} \in \mathcal{U}_i, n = 1, \dots, p_s$. Finally, for $n, m = 1, \dots, p_s$, the element

$$\sum_{\mathbf{s}-\mathbf{s}^* \in \mathbb{Z}^d} \sum_{t-t^* \in \mathbb{Z}} E\{Y_t(\mathbf{s})[X_t(\mathbf{s}-\mathbf{j}_n) + X_t(\mathbf{s}+\mathbf{j}_n)]Y_{t^*}(\mathbf{s}^*)[X_{t^*}(\mathbf{s}^*-\mathbf{j}_m) + X_{t^*}(\mathbf{s}^*+\mathbf{j}_m)]\}$$

is of interest. Using Proposition 7.3.1 of Brockwell and Davis (1991) for Gaussian variables, this is equal to

$$\begin{aligned} & \sum_{\mathbf{s}-\mathbf{s}^* \in \mathbb{Z}^d} E\{Y_t(\mathbf{s})Y_{t^*}(\mathbf{s}^*)\}E\{[X_t(\mathbf{s}-\mathbf{j}_n) + X_t(\mathbf{s}+\mathbf{j}_n)][X_{t^*}(\mathbf{s}^*-\mathbf{j}_m) \\ &+ X_{t^*}(\mathbf{s}^*+\mathbf{j}_m)]\} = -2\nu \sum_{l=1}^{p_s} \beta_{\mathbf{j}_l} [\gamma(\mathbf{j}_n-\mathbf{j}_m-\mathbf{j}_l) + \gamma(\mathbf{j}_n-\mathbf{j}_m+\mathbf{j}_l) \\ &+ \gamma(\mathbf{j}_n+\mathbf{j}_m-\mathbf{j}_l) + \gamma(\mathbf{j}_n+\mathbf{j}_m+\mathbf{j}_l)] + 2\nu [\gamma(\mathbf{j}_n-\mathbf{j}_m) + \gamma(\mathbf{j}_n+\mathbf{j}_m)] \end{aligned}$$

adding an extra $2\nu^2$ if $n = m = 1, \dots, p_s$ only, and the relevant matrix $B_{(S)}$ with these elements $n, m = 1, \dots, p_s$ is considered. Thus, writing

$$B \equiv \begin{pmatrix} B_{(T)} & B_{(ST)} \\ B_{(ST)}^T & B_{(S)} \end{pmatrix}$$

and using (11) and (12), it is possible to prove that

$$(NT)^{1/2} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, C^{-1} B C^{-1}) \text{ as } \min\{N, T\} \rightarrow \infty. \tag{15}$$

4 Concluding remarks

For the sea level pressure data on (5×17) points from the NCER-NCAR that was analyzed by Dimitriou-Fakalou (2010), further information has been made available regularly over time.

The ALAR equation (7) becomes a pure AR equation only if $\beta_{\mathbf{j}_n} = 0$, $n = 1, \dots, p_s$ when $\{Y_t(\mathbf{s})\}$ are allowed to be uncorrelated in \mathbb{Z}^{d+1} . Otherwise, a general auto-regressive equation uses nonzero coefficients on $X_t(\mathbf{s} - \mathbf{j}_n)$ with $\mathbf{j}_n > \mathbf{0}$ only, and it is unnatural over the \mathbb{Z}^d space.

The assumption of Gaussian variables should be possible to relax together with Proposition 7.3.1 of Brockwell and Davis (1991) using a finite fourth moment condition; (15) should still be valid for other distributions and the asymptotic normality result will be distribution-free. Further, the elements of C and B are easy to approximate from the data and $\tilde{\boldsymbol{\theta}}$, and after setting $\tilde{\nu} \equiv \sum \tilde{Y}_t^2(\mathbf{s})/(NT)$ for the statistical inference.

References

- Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion).
Journal of the Royal Statistical Society, Series B, **36**, 192–236.
- Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*.
New York: Springer-Verlag.
- Cressie, N. and Huang, H.-C. (1999). Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions.
Journal of the American Statistical Association, **94**, 1330–1340.
- Dimitriou-Fakalou, C. (2010). Statistical Inference for Spatial Auto-Linear Processes.
Journal of Statistical Theory and Practice, **4**, 345–365.
- Gneiting, T. (2002). Nonseparable, Stationary Covariance Functions for Space-Time Data.
Journal of the American Statistical Association, **97**, 590–600.
- Guyon, X. (1982). Parameter Estimation for a Stationary Process on a d -dimensional Lattice.
Biometrika, **69**, 95–105.
- Ma, C. (2003). Families of spatio-temporal stationary covariance models.
Journal of Statistical Planning and Inference, **116**, 489–501.
- Stein, M. L. (2005). Space-Time covariance functions.
Journal of the American Statistical Association, **100**, 310–321.
- Whittle, P. (1954). On Stationary Processes in the Plane.
Biometrika, **41**, 434–449.

The sample signature of probabilistic context trees with an application to Linguistics

Denise Duarte ¹, Wecsley Prates ², Enrico A. Colosimo ¹

¹ Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

E-mail for correspondence: `denise@est.ufmg.br`

Abstract: We introduce the Sample Signature of a Probabilistic Context Tree (PCT) as a way to distinguish samples coming from different sources. A (PCT) is a class of stochastic chains with memory of variable length which is more flexible than Markov chains. We show that the Sample Signature is capable to distinguish differences between sequences of discrete data even in the case where the estimated PCT generating the data is the same

Keywords: Probabilistic Context Trees , Model selection, penalized likelihood, sample signatures, longitudinal data.

1 Probabilistic Context Trees

Stochastic chains with memory of variable length, also called Probabilistic Context Tree (PCT) models, have the property that, for each string symbols in the past, only a finite suffix (ending string) of the past is enough to predict the next symbol. Following Rissanen (1983) in which these models were introduced, let us call context this relevant part of the past.

Let $p = \{p(\cdot|w) : w \in \tau\}$ be a family of probability measures on A indexed by the elements of τ . The elements of τ will be called contexts and the pair (τ, p) will be called *probabilistic context tree* (PCT).,

A PCT takes values in a finite categorical space A and has a parsimonious structure of transition probabilities, since it is not necessary to consider all $|A|^k(|A| - 1)$ parameters as in a order k Markov chain. The transition probabilities of a PCT must satisfy the following property:

$$\mathbb{P}(X_t = x_t | X_0 = x_0, \dots, X_{t-1} = x_{t-1}) = \mathbb{P}(X_t = x_t | X_{t-l} = x_{t-l}, \dots, X_{t-1} = x_{t-1})$$

where $l = C(x_{-\infty}^{-1})$.

In order to estimate a PCT, named τ , we follow the approach proposed by Csiszar and Talata (2006) who showed that context trees can be consistently estimated in linear time using the Bayesian Information Criteria (Schwarz, 1978).

The BIC estimator with penalizing constant $c > 0$ is defined as

$$\hat{\tau} = \operatorname{argmax}_{\tau \in \mathcal{T}} \{ \log L_{\tau}(X_1^n) - c \log n \}$$

We estimate the tree using the Smallest Maximization Criterion (SMC) algorithm proposed by Galves et al (2011) based on the BIC, but letting the penalizing constant c vary. This variation in the penalizing constant will generate the Sample Signature as we explain in the next section.

2 Sample Signature of a Probabilistic Context Tree

Given a sample X_1, \dots, X_n , the final tree estimated by SMC, $\hat{\tau}_g$ is a function of the sample and the optimal penalizing constant c_{opt} ,

$$\hat{\tau}_g = f(X_1^n, c_{opt}) \quad (1)$$

Where this optimal constant is obtained by changing the values of the penalizing constant in the BIC criterion until we observe a change of regime of the likelihood function, that is when it stops to increase very sharply, indicating that we have reached the true tree. This result is proved in Galves et al (2011). Each value of c stands for the beginning of an interval for which the estimated tree is the same. In this way, before reaching c_{opt} the SMC algorithm generates a non increasing sequence of penalizing constants,

$$c_n > c_{n-1} > \dots > c_{opt} \quad (2)$$

which leads to a non decreasing sequence of candidate trees

$$\tau_n \prec \tau_{n-1} \prec \tau_{n-2} \prec \dots \prec \tau_{opt} \quad (3)$$

We will call the sequence $C_x = (c_{n-1}, c_{n-2}, \dots, c_{opt})$ as the *Sample Signature* of the sample \mathbf{X}

We claim that the Sample Signature identifies the sample in the following sense

- If the samples come from the same tree, the signature of the samples will be very similar.
- If the samples come from different trees, the signature will differ significantly.

3 Simulation results

We have performed 500 monte carlo simulations of size 50000 of the trees A_1 e A_2 and found the Sample Signature for each sample. The results can be viewed in Figure 1. Each interval stands for the 5% and the 95% percentiles

TABLE 1. Transition probabilities of the tree A_1

w	$p(1/w)$	$p(2/w)$	$p(3/w)$
111	0.20	0.24	0.56
211	0.19	0.80	0.01
311	0.10	0.40	0.50
121	0.19	0.39	0.42
221	0.49	0.47	0.04
321	0.48	0.48	0.04
131	0.38	0.38	0.24
231	0.10	0.15	0.75
331	0.50	0.24	0.26
12	0.09	0.09	0.82
22	0.32	0.32	0.36
32	0.40	0.40	0.20
13	0.12	0.12	0.76
23	0.27	0.27	0.46
33	0.49	0.49	0.02

of the value of the penalizing constant for each tree. We observe that the Sample Signature for each tree is completely different. For simulations of the same tree we can observe from the percentiles intervals that the values of the penalizing constant do not differ very much in comparison to the values of the penalizing constants of the other tree.

As an example, for the context 111, the probability of the next symbol to be 1 is 0,2 , to be 2 is 0,24 and to be 3 is 0,56. Transition probabilities of the PCT A_2 are presented in Table 3.

3.1 Linguistic Application

The Corpus of Historical Portuguese Tycho Brahe is an electronic corpus, composed of Portuguese texts written by authors born between 1435 and 1845 and is available to researchers free of charge for academic purposes and teaching through anonymous ftp at <http://www.iel.unicamp.br/tycho>. Sixteen texts of the Portuguese Historical Corpus Tycho Brahe had been modernized, that is, they had been modified so that it was possible to establish a orthographic and grammatical criterion that allowed to codify the words in accordance with the tonality of its syllables. Each syllable of these 16 texts was classified as stressed or non-stressed and if it was the beginning of prosodic word or not. So each syllable received a code: 0 (non-stressed in the middle of the word), 1 (stressed in the middle of the word), 2 (non stressed in the beginning of the word), 3 (stressed in the beginning of the word) and 4 (end point). This code was built by the linguistic team of the project.

TABLE 2. Transition probabilities of the tree A_2

w	$p(1/w)$	$p(2/w)$	$p(3/w)$
111	0.25	0.18	0.57
211	0.20	0.79	0.01
311	0.11	0.40	0.49
121	0.18	0.39	0.43
221	0.40	0.47	0.13
321	0.48	0.38	0.14
131	0.30	0.4	0.20
231	0.15	0.10	0.75
331	0.50	0.25	0.25
12	0.13	0.15	0.72
22	0.32	0.32	0.36
32	0.35	0.45	0.20
13	0.12	0.10	0.78
23	0.25	0.25	0.50
33	0.50	0.35	0.15

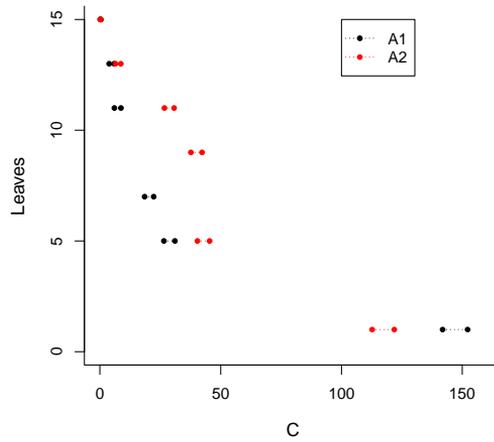


Figure 1: Sample Signature simulations for trees A_1 and A_2 .

Using this code, we estimate the trees of each text, as well as their Sample Signatures. For all texts the estimated tree was the same, with 11 branches and also the transition probabilities.

We could not find any significant difference among the texts neither looking to the structure of the estimated tree behind the data nor by its transition probabilities, since the likelihood ratio test gave no evidence of differences among them.

But the Sample Signatures, that can be seen in Figure 2, suggest that the texts are indeed different. For some of them the value of the penalizing constant for the order zero model is very large, what can be interpreted as for these texts it is much more expensive to accept a independent model than for other texts.

3.2 GEE Approach

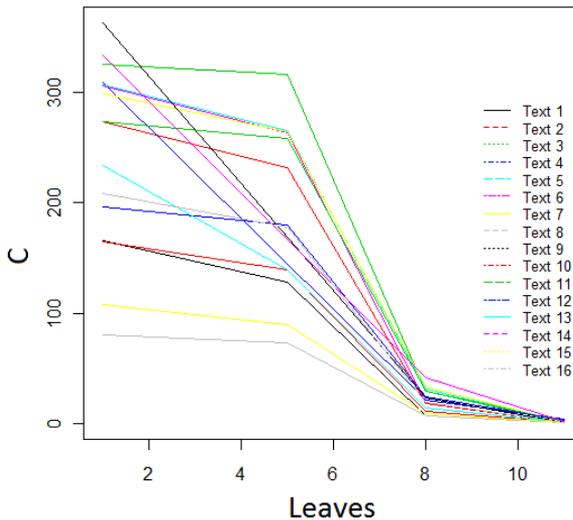


Figure 2: Profiles of the texts listed in Table 6

We will treat the Sample Signatures as longitudinal data, measures repeated in time. The values of the penalization constant will be treated as measures repeated and the number of branches of trees works like the time of realization of the measure. We will model our data through the GEE (Generalized Estimation Equation), Liang and Zeger (1986), so we estimate the parameters considering the regression correlation between individuals without specifying the joint probability function. Our interest is

only to check for differences between groups of texts over the years. We have proposed the following model:

$$y_i = \beta_0 + \beta_1 Leaves_5 + \beta_2 Leaves_8 + \beta_3 Leaves_{11} + \beta_4 Group_2 + \beta_5 (Folha_5 * Group_2) + \beta_6 (Leaves_8 * Group_2) + \beta_7 (Leaves_{11} * Group_2) + e_i$$

It was found that there was a change only when we divide the time between the groups until 1675 and after 1702, when all other combinations did not present any significant difference, using the Model Marginal GEE variance and covariance structure AR(1).

There was no significant interaction between leaves and group thus we remove them from the initial model. The results of the final model are summarized in Table 3. It can be observed that all coefficients are significant and the coefficient of the groups is significant at 3% level.

Table 3 - Estimates of Marginal Model GEE- AR (1)

Coefficients	Estimates	S.E.Naive	S.E.Robust	Z-Robust	P-Valor
Intercept	269.34	17.41	17.23	15.62	0.0000
Leaves 5	-46.65	13.26	9.39	-4.96	0.0000
Leaves 8	-233.36	14.99	19.19	-12.15	0.0000
Leaves 11	-250.92	16.85	20.65	-12.14	0.0000
Group 2	-36.03	20.08	16.94	-2.12	0.0334

4 Conclusions

We conclude that the Sample Signature was able to find differences between groups of texts even in the case where the estimated tree for all samples was the same.

References

- Csiszár, I., Talata, Z. (2006) *Context tree estimation for not necessarily nite memory processes, via BIC and MDL*. IEEE Trans. Inform. Theory, 52(3)
- Diggle, P.J., Liang, K-Y, Zeager, S.L. (2002). *Analysis of Longitudinal Data*. 1st ed. Oxford
- Galves, A., Galves, C, Garcia, J., Garcia, N., Leonardi, F.(2011) *Context tree selection and linguistic rhythm retrieval from written texts*. ArXiv: 0902.3619v2.

- Rissanen, J.(1983) *A universal data compression system*. IEEE Trans. Inform.Theory, 29(5).
- Schwarz, G. (1978) *Estimating the dimension of a model*. Annals of Statistics, v.6, p.461-464.
- Verbeke,G.; Molenberghs,G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Analysis of temperature-based weather derivatives in mainland China: Pricing and simulation

Manuela Ender¹, Lu Zong¹

¹ Department of Mathematical Sciences, Xian-Jiaotong Liverpool University, Suzhou, Jiangsu Province, China

E-mail for correspondence: manuela.ender@xjtlu.edu.cn

Abstract: In this paper, we present an analysis of weather derivatives in consideration of the climatic regionalization of mainland China. We apply Alaton et al's (2002) temperature model to twelve cities in China. Given the estimated parameters we provide values of certain weather derivatives for all climatic zones in China.

Keywords: weather derivatives; temperature modelling; climatic regionalization; pricing; simulation.

1 Introduction

Weather derivatives, especially temperature-based derivatives, have become a flourishing financial product worldwide in terms of risk hedging. Even though weather derivatives have not yet been traded in China, we are convinced that these financial products are essential, or at least contributing factors to those industries exposed to weather risks of the country. But challenges still exist. Apart from the difficulty in pricing a non-tradable underlying (i.e. weather indexes), working out a multi-regional joint model could also be regarded as an important step.

In this paper, we present the results of tests whether the architectural climatic zones of China are appropriate for a weather derivatives trading system. Twelve cities scattered throughout mainland China will be used applying Alaton et al's (2002) model for this purpose. Finally, we are going to price weather options based on both, the approximation formula of Alaton et al (2002) and Monte Carlo simulation.

2 Standard of climatic regionalization and datasets

The Standard of climatic regionalization is a division method used in architecture to partition building standards in areas with different climate

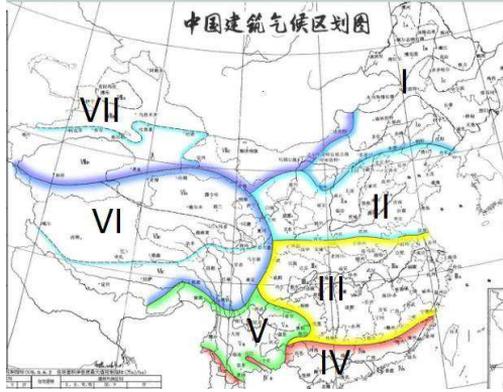


FIGURE 1. Map of climatic regionalization in China

characteristics. It divides the main land of China into seven climatic zones which are normally represented by Roman numerals. Figure 1 gives a detailed map of partition.

In our analysis, twelve cities, Haerbing and Changchun (I), Beijing and Tianjin (II), Shanghai, Hangzhou, Nanjing (III), Guangzhou and Hainan (IV), Kunming (V), Lhasa (VI), Urumchi (VII), were selected according to this partition method. Apart from Shanghai, the duration of the temperature dataset is thirty years, covering January 1981 to December 2010. However, the duration of the dataset for Shanghai is twenty years. The truncation of dataset is caused by a change of the meteo-station.

3 Model selection and parameter estimation

Even though the Black-Scholes model is no longer valid in the study of weather derivatives, there are still some alternatives to replace it. For this paper, Alaton et al's (2002) model was selected with the following formula for the temperature model:

$$T_t = (T_s - T_s^m)e^{-\alpha(t-s)} + T_s^m + \int_s^t e^{-\alpha(t-\tau)} \sigma_\tau dW_\tau,$$

where $T_t^m = A + Bt + C \sin(\omega t + \varphi)$.

At the beginning of Table 1, an overview of the parameters and their corresponding properties of the temperature dataset is given.

In order to get the monthly standard deviation of temperature σ , we employ the mean value of two statistic estimators (Alaton et al (2002)). The first one is the quadratic variation of temperature at time t (Basawa et al (2008)). The second one is derived from a discretized equation using the knowledge on regressive process.

Table 1 gives our results of A , B , C , φ and α of the twelve cities.

TABLE 1. Estimated values of A , B , C , φ and α of twelve cities in mainland China

	A	B	C	φ	α
	Mean temp.	Warming pace	Range of temp.	Translation of sine function	Mean-reverting
Haerbing	3.986	$1.776 * 10^{-4}$	20.165	-1.851	0.3121
Changchun	5.527	$1.25 * 10^{-4}$	18.876	-1.863	0.28
Beijing	12.406	$1.022 * 10^{-4}$	15.017	-1.852	0.2821
Tianjin	12.814	$0.278 * 10^{-4}$	15.205	-1.874	0.3091
Shanghai	16.29	$2.212 * 10^{-4}$	11.703	-2.071	0.3008
Hangzhou	15.951	$1.937 * 10^{-4}$	11.82	-2.015	0.2636
Nanjing	14.999	$1.75 * 10^{-4}$	12.561	-1.974	0.2515
Guangzhou	21.819	$1.123 * 10^{-4}$	7.649	-2.021	0.2279
Hainan	24.916	$0.673 * 10^{-4}$	5.237	-1.86	0.1954
Kunming	14.505	$1.872 * 10^{-4}$	5.898	-1.739	0.2911
Lahsa	8.060	$1.441 * 10^{-4}$	8.496	-1.84	0.2616
Urumchi	5.942	$3.241 * 10^{-4}$	17.81	-1.839	0.1977

4 Underlying

The two basic underlyings for weather derivatives are CDD (cooling-degree day) and HDD (heating-degree day). A degree day measure is the accumulation of degrees that deviates from a proxy temperature (normally 18°C or 65°F) during a specified period of time. The equation of HDD and CDD can be defined as follows:

$$HDD_{t_1, t_2} = \int_{t_1}^{t_2} \max(18 - T_t, 0) dt, \quad CDD_{t_1, t_2} = \int_{t_1}^{t_2} \min(T_t - 18, 0) dt,$$

where T_t denotes the daily observed temperature.

Basically, a HDD contract is traded during cold seasons, and a CDD contract is traded during hot seasons. Hence, determining the cold/warm seasons for each city is also an essential step for a trading system.

5 Temperature joint modelling

In order to find a multi-regional joint model to reduce dimensions in a trading system, in Ender and Zong (2012) the noise distribution, the parameters, and the cold/warm seasons were tested whether they can be assumed to be the same within the same climatic zones or not. The findings of these tests are listed below:

Proposition 2.1: *The noise distribution is the same within the same climatic zones.*

Proposition 2.2: *The values of A and B vary among different cities within the same climatic zones.*

Proposition 2.3: *The values of C , φ , α , and σ stay constant within the same climatic zones.*

Proposition 2.4: *The cold/warm season division is the same within the same climatic zones.*

6 Alaton et al's pricing formula (2002)

According to Alaton et al (2002), an approximation of the price of weather derivatives could be derived by manipulating the martingale measure \mathbb{Q} . In order to simplify the deriving process, we let the market price of risk (MPR) be equal to a constant λ . The converting ratio (a.k.a. principal nominal) is one unit of currency pays for one degree Celsius. The pricing method is based on the assumption that the probability of the daily temperature in the HDD contract period being larger than 18°C is zero. Hence, the pay-off of such a degree day based contract is normally distributed under measure \mathbb{Q} with the mean μ_n and standard deviation σ_n . In other words, we have the pay-off of a HDD contract given by:

$$H_n = \sum_{t=1}^n \max(T_t - 18, 0) = \sum_{t=1}^n T_t - 18n.$$

Consequently, the following equation gives the pricing formula of a HDD call option:

$$\begin{aligned} C(t_0) &= e^{-\gamma(t_n-t_0)} \int_K^\infty (x - K) f_{X_n}(x) dx \\ &= e^{-r(t_n-t_0)} \left[(\mu_n - K) \Phi\left(\frac{K - \mu_n}{\sigma_n}\right) + \frac{\sigma_n}{\sqrt{2\pi}} e^{-\frac{(K - \mu_n)^2}{2\sigma_n^2}} \right], \end{aligned}$$

where r stands for the risk free rate, t_0 stands for the first day of the contract, t_n stands for the last day of the contract, K stands for the strike price, and $\Phi(\cdot)$ is the c.d.f of the standard normal distribution.

7 Monte Carlo simulation

In order to have a second pricing method for comparison, we additionally use a Monte Carlo simulation. The biggest advantage of a Monte Carlo simulation is that it skips Alaton et al's (2002) assumption on temperature distribution. To simulate the daily temperature, it follows:

$$T_t = T_t^m + (T_0 - T_t^m) \exp(-\alpha t) + \int_0^t \exp(-\alpha(t-s)) \sigma dW_s.$$

In the performed simulation, a sample size $N = 100,000$ is used. From the comparison of the observed temperature and the simulated values we can conclude that the estimation errors are larger in winters and smaller in summers.

8 Numerical results for temperature-based options

In a weather derivative option contract, the information involved is similar to regular options. Basically, we need to specify the strike level, the expiry date, the duration, and the location in the contract. Hence, the option holder can get a payoff or nothing according to the strike level. For example, the call option holder will get a payoff if the cumulative HDD or CDD is greater than the strike level.

Table 2 and Table 3 represent numerical results of pricing some specified HDD and CDD call option contracts. Since there is no existing market for weather derivatives in China, the estimation of the market price of weather risk is not possible. In this case, the market price of risk is assumed to be zero. The risk free rate equal to 6.5% is used. The duration of the HDD and CDD contracts is respectively January and July of 2010.

In Table 2 and Table 3, the difference in price between the approximation formula and the Monte Carlo simulation is relatively small, but should not be neglected. Since there is no existing contract to refer to, an conclusion which method produces more accurate prices on temperature-based weather options is not possible at the moment.

9 Conclusion

In this paper, we continued to propose the idea of a weather derivative trading system using the climatic regionalization in China. This is the first paper on temperature modelling and temperature-based option pricing using such a large number of cities in China. We see this paper as an additional step on the way to introduce a systematic weather derivatives trading system in China.

References

- Alaton, P., Djehiche, B., and Stillberger, D. (2002). On Modeling and Pricing Weather Derivatives. *Applied Mathematical Finance*, **9**.
- Basawa, I.V., Rao, P., and B, L.S. (2008). *Statistical Inference for Stochastic Processes*. Academic Press.

TABLE 2. HDD call options pricing (Contract Period: January 2010)

Climatic zone	City	Strike price (in RMB)	Option price	Option price
			(in RMB) MC simulation	(in RMB) Alaton(2002)
I	Haerbing	750	43.01	42.85
	Changchun		35.08	34.75
II	Beijing	500	24.42	24.42
	Tianjin		27.45	27.44
III	Shanghai	200	25.25	25.15
	Hangzhou		23.59	23.70
	Nanjing		32.75	32.74
IV	Guangzhou	50	7.73	7.53
	Hainan		0	0
V	Kunming	100	15.75	15.74
VI	Lahsa	400	21.66	21.52
VII	Urumchi	600	35.12	35.50

TABLE 3. CDD call options pricing (Contract Period: July 2010)

Climatic zone	City	Strike price (in RMB)	Option price	Option price
			(in RMB) MC simulation	(in RMB) Alaton(2002)
I	Haerbing	100	11.92	11.79
	Changchun		10.40	10.65
II	Beijing	150	23.07	23.06
	Tianjin		20.95	20.81
III	Shanghai	200	18.48	18.39
	Hangzhou		18.60	18.54
	Nanjing		16.61	16.7
IV	Guangzhou	250	15.48	15.37
	Hainan		17.46	17.44
V	Kunming	50	8.12	8.08
VI	Lahsa	20	0.99	0.770
VII	Urumchi	100	13.36	13.73

Ender, M. and Zong, L. (2012). Analysis of temperature-based Weather Derivatives in Mainland China: Temperature Joint Modelling *Working paper*.

Heimfarth, L. and Musshof, O. (2011). Weather index-based Insurances for Farmers in the North China Plain: An Analysis of Risk Reduction Potential and basis Risk. *Agricultural Finance Review*, **71**.

Modelling the association between bathing water quality and gastrointestinal illness in South West Scotland

Jude Eze¹, E. Marian Scott¹, Kevin Pollock², Ruth Stidson³,
Claire Miller¹, Duncan Lee¹

¹ University of Glasgow, School of Mathematics and Statistics, G12 8QW, UK

² Health Protection Scotland, 4th Floor, 5 Cadogan Street, Glasgow, G2 6QE

³ Scottish Environment Protection Agency, Edinburgh, EH14 4AP

E-mail for correspondence: jude.eze@glasgow.ac.uk

Abstract:

The relationship between bathing water quality and health outcomes is investigated, using measured data on bathing water quality from 22 bathing waters and monthly counts of laboratory confirmed cases of gastroenteritis reported in two Scottish health boards. We modelled gastroenteritis counts as a smooth spline function of log faecal coliform (FC) and streptococci (FS) respectively, adjusting for the effects of season, long term trend and the confounding effect of temperature. Results indicate significant association between levels of the faecal indicator organisms and the incidence of different gastro-intestinal pathogens in the two health boards. An increase in FC by 10cfu/100 ml was associated with 12.86 (12.08, 13.68) increase in average monthly counts of viral gastroenteritis.

Keywords: Bathing water; faecal indicators, gastroenteritis; cubic splines.

1 Introduction

The association between health outcomes and recreational water quality has been studied by researchers for more than six decades (Stevenson 1953; Moore 1959). Over the years, many studies have reported increased risk of illness resulting from exposure to pollution at bathing waters (Prüss 1998; Wade et al. 2006).

Given the health risk posed by exposure to bathing waters, regulatory authorities have enacted various legislations which set quality standards for recreational waters. In Europe, the revised European Union (EU) Bathing Water Directive (2006/7/EC) replaces Directive 76/160/EEC and makes specific provisions for monitoring, classification and management of bathing water quality and provision of information to the public. The overall objective of the 2006 directive is to preserve, protect and improve the environment and for protection of public health. This has to be achieved

across member states by 2015. The directive specifies tighter microbiological standards that use more reliable indicator parameters for predicting microbiological health risks associated with bathing in marine and fresh waters (Prüss 1998; WHO 2003).

In this paper, the association between bathing water quality and the incidence of gastro-intestinal illnesses in two Scottish Health Boards is studied, after adjusting for the confounding effect of temperature.

2 Data

Routine data on bathing water quality from 22 bathing waters located within two NHS health board areas (Ayrshire and Arran (AA), and Glasgow and Clyde (GC)) were supplied by the Scottish Environment Protection Agency (SEPA). Bathing waters were sampled between 1998 and 2010 at least twice monthly during the bathing season (May to September) and monthly during the off season. Bathing water quality is measured by the faecal indicator organisms (FIOs) faecal coliform (FC) and faecal streptococci (FS) as colony forming units per 100 ml (cfu/100ml). Weather information was sourced from Weather Underground (www.wunderground.com). Faecal indicator enumerations were generally reported as actual counts. However, occasionally, censored data values were reported. Roughly 18% and 14% of the data series for FS and FC respectively, were recorded as censored below three limits (2, 10 and 100). Imputation for these censored values was conducted using the Regression on Order Statistics (ROS) method (Lee and Helsel 2005).

In addition, laboratory-confirmed cases of viral and non-viral groups of infectious gastrointestinal pathogens (between 1988 and 2008 for the two NHS Health Boards) were supplied by the Health Protection Scotland (HPS).

3 Methods

Based on a preliminary site-specific analysis, we observed that the variability, seasonality and trend across the 22 bathing sites were similar. Consequently, monthly averages across the 22 sites were obtained to produce a single monthly series for each of the water quality measures. We then used this series to separately model the relationship between FIOs and gastroenteritis groups of pathogens in the two health board areas. The use of the single series for both health boards is informed by the understanding that individuals residing in the health boards could visit any of the bathing water locations. We have used the monthly counts of the two groups of gastro-pathogens for Glasgow and Clyde and Ayrshire and Arran health board areas. These counts are assumed to be Poisson distributed and log base 10 transformations of the bathing water quality indicators (FC and FS) were obtained and used in the model. The following generalized additive models

(GAM) were fitted in R using the `mgcv` package (Wood 2006). Smoothing parameters were selected using the unbiased risk estimator (UBRE). Given that $Y_{it} \sim \text{Poi}(\lambda_{it})$, then

$$\text{Log}(\lambda_{it}) = \mu + s(\text{month}_i) + s(\text{year}_t) + s(\text{log}(F)_{it}) \quad (1)$$

Adjusting for the effect of temperature

$$\text{Log}(\lambda_{it}) = \mu + s(\text{month}_i) + s(\text{year}_t) + s(\text{temp}_{it}) + s(\text{log}(F)_{it}) \quad (2)$$

F = FS or FC. Y_{it} is the count of gastrointestinal infections in month i and year t . λ_{it} is the expected value of the counts. The intercept μ is the average log count; s is a nonparametric smooth spline function which represents the flexible functional form of the association between each covariate and expected log count of gastrointestinal infections. The penalized cubic regression spline was used. The smooth term for month was included to account for seasonal variations not explained by the effects of temperature and $\text{log}(F)$. The smooth effect of year represents the long term trend effect. Model (3) allows the seasonal pattern to vary with year.

$$\text{Log}(\lambda_{it}) = \mu + s(\text{month}_i, \text{year}_t) + s(\text{temp}_{it}) + s(\text{log}(F)_{it}) \quad (3)$$

4 Results

The seasonal behaviours of the FIOs are similar to the seasonal patterns exhibited by the two gastro-pathogens. Figure 1 shows the seasonal plots of the non viral group of gastroenteritis pathogens (dotted lines) with that of FC and FS (solid lines) in the Glasgow and Clyde (GC) health board, respectively. Peak periods of infections closely match the peak of the faecal indicator organisms and this coincides with the bathing season.

Estimates of seasonal, trend and log FS effects on the number of viral associated gastroenteritis cases in Glasgow and Clyde from Model (1) are plotted in Figure 2. All three smooth model terms were statistically significant. Peak viral infection is in May and the trend over the years is increasing. An increasing number of viral infections was associated with increasing levels of log FS. Figure 3 shows results from model (2) fitted to viral gastro-intestinal pathogens with faecal streptococci and temperature in GC. Results from model (2) indicate that faecal streptococci (FS) and faecal coliform (FC) have a statistically significant positive association with the number of viral gastroenteritis cases in Glasgow and Clyde (GC) area after accounting for the effect of temperature. Interestingly, the addition of temperature into the model slightly improved the explained deviance but did not affect the significance of the faecal indicator organisms. Temperature has a negative effect on the number of viral-associated gastroenteritis cases in Glasgow and Clyde.

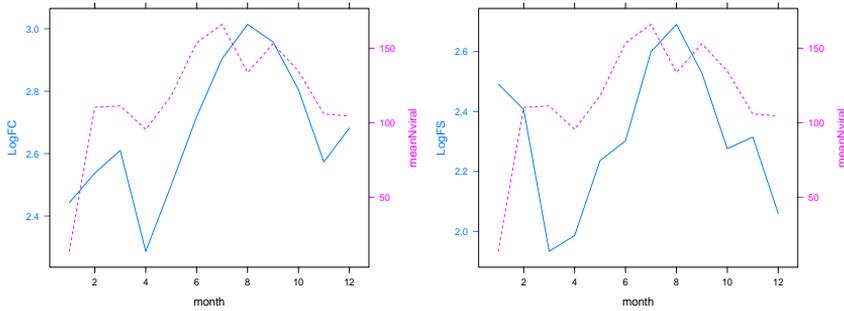


FIGURE 1. Seasonal patterns of FC, FS and non-viral gastroenteritis. The lines are the monthly averages of FC and FS and the dotted lines are that of non viral gastroenteritis in GC.

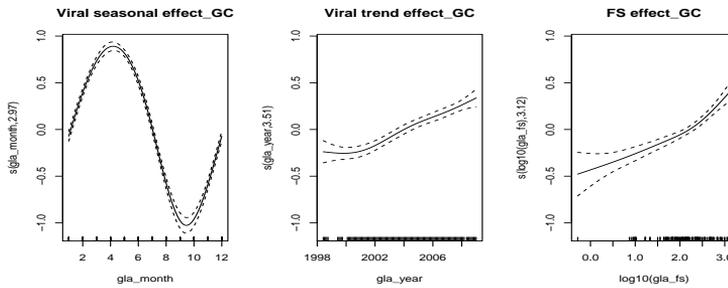


FIGURE 2. : Estimates from model 1 showing seasonal, trend and FS effects on number of cases of viral gastroenteritis in Glasgow and Clyde

TABLE 1. Results from model 3 showing p-values of smooth model terms for gastro-pathogens in Glasgow and Clyde.

Pathogen	Month:Year	Temp	FS
Viral	< 0.001	< 0.001	< 0.001
Non Viral	< 0.001	< 0.001	0.009

By comparing the AICs, UBRE score and deviance explained for each of the three models, it appears that model (3), which allows the seasonal pattern to vary with year is the most appropriate of these models, describing 72.9% of the deviance. Table 1 shows the p-value for each of the smooth terms model (3) for the two gastrointestinal pathogens in GC.

Figure 3 suggests that the relationship between FS and viral gastro-pathogens can be approximated by a straight line. Therefore, fitting a semi parametric

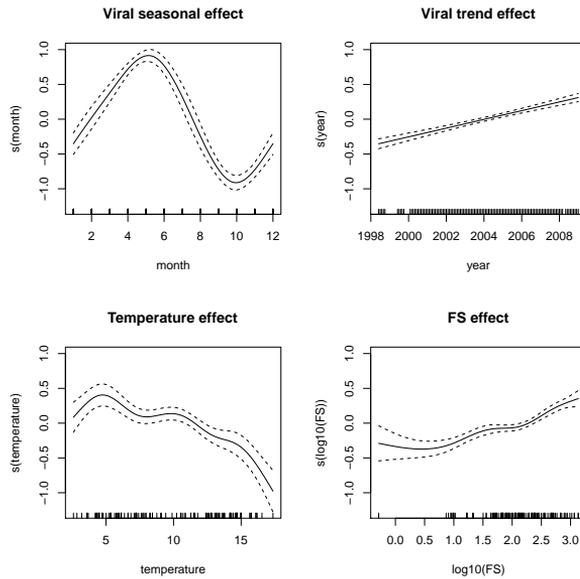


FIGURE 3. Estimates from model 2 showing seasonal, trend, temperature and FS effects on counts of viral gastroenteritis in Glasgow and Clyde

model to viral gastroenteritis in Glasgow and Clyde with smooth functions for season, trend and temperature and a linear function for FS, results indicate that an increase in FC by 10cfu/100 ml increased average monthly counts of viral gastroenteritis by 12.86 (12.08, 13.68).

5 Conclusion

This study investigates the association between bathing water quality and cases of laboratory-confirmed gastroenteritis in two Scottish health boards. Results suggest positive association between the two infectious intestinal pathogens and faecal streptococci in Glasgow and Clyde (GC). Temperature is also statistically significant but the direction of the effect depends on the type of pathogen. For the viral gastro-pathogen, increasing temperature is inversely associated with the number of viral intestinal infections. However, the number of non viral gastroenteritis is positively associated with temperature. Our results are consistent with other findings discussed in Kovats et al. (2004) and D'Souza et al. (2007). Adjusting for temperature slightly improved the explained deviance but did not affect the significance or otherwise of both bathing water indicators.

Acknowledgments: Special thanks to NERC for funding this project (GE/9001170/1) and to the Scottish Environment Protection Agency (SEPA) and Health Protection Scotland for providing the data.

The views and opinions expressed in this publication are those of the contributors and are not necessarily the views and opinions of the agencies involved.

References

- D'Souza RM, Hall G, and Becker NG. (2008). Climate factors associated with hospitalizations for rotavirus diarrhoea in Children under 5 years of age. *Epidemiology and Infection* **136**, 56–64.
- Kovats RS, Edwards S, Hajat S, Armstrong B, Ebi KL, Menne B and C.G. (2004). The effect of temperature on food poisoning: time series analysis in 10 European countries *Epidemiology and Infection* , **132**, 443–453.
- Lee K. and Helsel D. (2005). Statistical analysis of environmental data containing multiple detection limits: S-language software for regression on order statistics. *Computers in Geoscience* , **31**, 1241–1248.
- Moore B. (1959). Sewage contamination of coastal bathing waters in England and Wales: A bacteriological and epidemiological study. *Br J Hygiene*, **57**, 435–472.
- Pruss A. (1998). Review of epidemiological studies on health effects from exposure to recreational water. *Int J Epidemiol*, **27**(1), 1–9.
- The Council of the European Communities (2006). Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 and repealing Directive 76/160/EEC
- Stevenson A.H. (1953). Studies of bathing water quality and health. *Am J Public Health Nations Health*, **43**, 529–538.
- Wade, T.J., Calderon, R.L., Sams, E., Beach, M., Brenner, K.P., et al. (2006). Rapidly measured indicators of recreational water quality are predictive of swimming-associated gastrointestinal illness. *Environmental Health Perspectives*, **114**(1), 24–28.
- WHO (2003). Guidelines for safe recreational water environments. Vol 1: coastal and fresh waters. *World Health Organisation, Geneva*, 219 pp.
- Wood, S.N. (2006). Generalized Additive Models: An Introduction with R. *Chapman and Hall/CRC*.

Measurement error in the personal exposure to air pollution

Veronika Fensterer¹, Helmut Küchenhoff¹, Josef Cyrys^{2,3},
Susanne Breitner², Alexandra Schneider², Mike Pitz^{2,3}, Jianwei
Gu^{2,3}, Annette Peters²

¹ Statistical Consulting Unit, Department of Statistics, Ludwig-Maximilians-Universität (LMU), Munich, Germany

² Helmholtz Zentrum München - German Research Center for Environmental Health (HMGU), Institute of Epidemiology II, Neuherberg, Germany

³ University of Augsburg, Environment Science Centre (WZU), Augsburg, Germany

E-mail for correspondence: Veronika.Fensterer@stat.uni-muenchen.de

Abstract: Data collection of the individual exposure to air pollution is concerned with measurement error from various aspects. Whereas mobile devices exhibit classical measurement error, measurements from central stations have Berkson error. The combination of the two data sources (mobile and stationary measurements) results in a mixture of the two error types. Furthermore, the examination of longitudinal data from the “Augsburger Umweltstudie” showed person-specific structures within the errors. We extended the bias analysis of Wang et al. (1998) to random effects in the error structure, to autocorrelated errors and to the mixture of classical and Berkson error. Based on these results we combined central and personal measurements and corrected the estimations of the Augsburg Umweltstudie with the method of moments.

Keywords: measurement error; method of moments; air pollution.

1 The Study: Augsburg Umweltstudie

The association between human health and air pollution is a wide area of common research. A big part of the research work refers to the examination of a long-term impact of air pollution e.g. on mortality; however, also short-term effects of air pollution e.g. on changes in the respiratory system or in the cardiovascular system are under consideration. The investigation of short-term effects involves a preferably accurate measurement of the personal exposure to air pollution. Central measurement stations permanently record air quality at several places, but only insufficiently reflect the personal exposure. To overcome this drawback, data on personal exposure was collected in the context of the Augsburg Umweltstudie: The study participants carried devices recording their current exposure to ultrafine particles

and simultaneously wore devices collecting their ECG-parameters in everyday situations. Two difficulties in the particle measurements were faced: Firstly, breakdowns and incorrect appliance of the mobile devices resulted in about 23 % of missing data; this proportion will even increase if lagged effects of personal measurements are investigated. Imputation of data from the central measurement station induced partly individual-specific Berkson error (Carroll et al., 2006) in the data. Secondly, the mobile devices exhibit an uncertainty of $\pm 20\%$ entailing partly individual-specific classical measurement error.

2 Statistical model

A linear mixed model for an health outcome Y_{it} for person i , $i = 1, \dots, n$, at time t , $t = 1, \dots, T$ depending on the personal exposure X_{it} is considered as main outcome model: $Y_{it} = \beta_0 + \beta_1 X_{it} + \tau_i + \varepsilon_{it}$. The mixture of classical and Berkson error in the particle measurements is formalized as follows:

$$X_{it}^{*M} = \begin{cases} X_{it}^{*B} & \text{for } p \cdot 100\% \text{ of the measurements} \\ X_{it}^{*C} & \text{for } (1 - p) \cdot 100\% \text{ of the measurements} \end{cases}$$

The true personal exposure X_{it} is composed of the Berkson error-prone measurement X_{it}^{*B} , a random person effect ν_i^B and a random error U_{it}^B : $X_{it} = X_{it}^{*B} + \nu_i^B + U_{it}^B$. The mobile devices, measuring the classical measurement error-prone exposure X_{it}^{*C} , overlay the true exposure X_{it} with a random person effect ν_i^C and a random error U_{it}^C : $X_{it}^{*C} = X_{it} + \nu_i^C + U_{it}^C$. All errors and random effects are assumed to be independent from each other and normally distributed: $\tau_i \sim N(0, \sigma_\tau^2)$, $\nu_i^B \sim N(0, \sigma_{\nu^B}^2)$, $\nu_i^C \sim N(0, \sigma_{\nu^C}^2)$, $\varepsilon_{it} \sim N(0, \sigma_\varepsilon)$, $U_{it}^B \sim N(0, \sigma_{U^B})$ and $U_{it}^C \sim N(0, \sigma_{U^C})$. The error terms ε_i , U_i^B and U_i^C are supposed to follow a first-order autoregressive process with autocorrelation coefficients ρ , ρ^B and ρ^C .

3 Bias analysis

3.1 Berkson error

In simple linear regression, a covariate measured with Berkson error yields unbiased effects estimates, but with less power (e.g. Carroll et al., 2006). The extension of the arguments to random effects in the Berkson error and to AR(1)-errors in a mixed model context is straight forward. The random effect variance of the Berkson error is assigned to the random effect variance of the main model; respectively, the variance of the random part of the Berkson error is added to the variance of the error term in the main model.

3.2 Classical measurement error

Classical covariate measurement error causes an attenuation of the effect in simple linear regression. The attenuation factor is determined by the ratio between the covariance of true and error-prone data and the variance of the error-prone data: $\text{Cov}(X, X^{*C})/\text{Var}(X^{*C})$. Besides other works on measurement error in models with random effects (e.g. Tosteson et al., 1998; Buonaccorsi et al., 2000), Wang et al. (1998) analyzed the bias of the effect estimate in mixed models with classical measurement error in heterogeneous covariates. Their derivation of the attenuation factor can be transferred to a heterogeneous structure of the classical measurement error. If the number of observations for each person goes to infinity, the attenuation factor does not depend on the variance of the random effect $\sigma_{\nu^C}^2$. Moreover, we show to what extent the presence of autocorrelation influences the attenuation factor.

3.3 Mixture measurement error

Similar calculations as for the classical measurement error reveal that the random effects of the errors, ν_i^B and ν_i^C , intensify the attenuation of the effect estimate. In the case with independence of the errors of the measurements for each person and independent normally distributed X_{it}^B , with $X_{it}^{*B} \sim N(0, \sigma_{X^{*B}}^2)$ and $T \rightarrow \infty$, the attenuation factor λ^M is

$$\lambda^M = \frac{\sigma_{X^{*B}}^2 + (1-p)\sigma_{U^B}^2}{\sigma_{X^{*B}}^2 + (1-p)(\sigma_{U^B}^2 + \sigma_{U^C}^2) + p(1-p)(\sigma_{\nu^B}^2 + \sigma_{\nu^C}^2)}.$$

3.4 Simulations

We illustrated the theoretical findings with a simulations. The design and parameter definition were guided by the empirical estimations in the Augsburg-Umweltstudie and an attendant validation study. The influence of the number of observations T for each person and the size of autocorrelation were examined.

4 Application to the ‘‘Augsburger Umweltstudie’’

Data from an external validation study and from comparison measurements were used to estimate the size of the different measurement errors and other parameters. Using the method of moments (Carroll et al., 2006), the attenuation factor has been calculated with these estimates. We corrected the effect estimations of the Augsburg-Umweltstudie with the corresponding attenuation factor in two ways: firstly, only the personal exposure data was analyzed with the correction methods for classical measurement error; secondly, the missing values for the personal exposure were replaced by

the measured values from the central measurement station and corrected with the estimation of the attenuation factor for mixture measurement error. Also the bias of the confidence intervals was corrected. Furthermore, the lagged effects of particles on the ECG-parameters were evaluated. The advantage of the imputation of values from the central measurement station is that also missing values resulting from considering lagged measurements, which are abundant in the longitudinal data setting, can be replaced and further analyzed.

References

- Buonaccorsi, J.P., Dimidenko, E., Tosteson, T., (2000). Estimation in longitudinal random effects models with measurement error. *Statistica Sinica*, **10**, 885–903.
- Carroll, R.J., Ruppert, D. Stefanski, L.A., Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models - A Modern Perspective*. Boca Raton: Chapman & Hall/CRC.
- Tosteson, T., Buonaccorsi, J., Dimidenko, E. (1998). Covariate measurement error and the estimation of random effect parameters in a mixed model for longitudinal data. *Statistics in Medicine*, **8**, 1069–1082.
- Wang, N., Lin, X., Gutierrez, R.B. Carroll, R.J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, **93**, 249–261.

Correlated frailty model for multivariate longitudinal count data

Rosemeire Fiaccone¹, Robin Henderson², Leila D. A. F. Amorim¹

¹ Statistics Department, Federal University of Bahia, Brazil

² School of Mathematics and Statistics, Newcastle University, UK

E-mail for correspondence: fiaccone@ufba.br

Abstract: A joint model of a sequence of different longitudinal counts using a correlated Poisson-gamma model was extended to allow frailties to be marginally gamma distributed with different variance parameters. In this case the frailty factor is split into the sum of two components: one is shared by entire group and the other is an individual term. That is, a time varying frailty term is included in the model to handle both between heterogeneity and within-subject serial correlation. Because of the intractability to obtain the full likelihood, we use the composite likelihood procedure based on all bivariate contributions of a multivariate longitudinal vector. We illustrate the methodology by modeling the number of mild and moderate/severe episodes of diarrhoea occurred between two following treatments in a placebo-controlled trial conducted in Brazil. The proposed model can be reduced to that proposed by Henderson and Shimakura (2003) when there is evidence to support the hypothesis of frailty variances being equal. A limitation of this model is that it assumes that there is always a positive correlation between frailties.

Keywords: Correlated data; Frailty; Count; Longitudinal Data Analysis.

1 Introduction

Many sets of data collected in human and biological science have a multi-level or hierarchical structure. Hierarchy in this context means that units at a certain level are grouped into, or nested within, higher level units. Formally, multilevel data structures can be regarded as encompassing nearly all types of correlated data, including multivariate data, clustered data, longitudinal data, spatial data and genuine multilevel data. This correlation can be explained by heterogeneity across individuals or event dependence. Heterogeneity can reflect unmeasured covariates or susceptibility of the individual to develop some disease, for example. Over the last twenty years, there has been increasing interest in developing suitable techniques for the statistical modelling and analysis of correlated data.

Although each subject in longitudinal studies often yields a single series on a variable of primary interest, it also often happens that each subject can yield two or more series on the primary variable. For example, in ophthalmologic studies in which visual acuity is measured on both eyes at a common set of times, yielding two longitudinal series per subject, termed "repeated-series" by Heitjan and Sharma (1997). Much of literature assumes that the data are multivariate normal with a variance-covariance matrix structure which takes into account the serial dependence, heterogeneity and dependence between response variables. However, very few tools exist in literature when the responses are not normal or/and when their marginals distributions are not in the same family.

This work is motivated by a double-blind, placebo-controlled trial carried out from December 1990 to December 1991 in Serrinha (Northeast region, Brazil). The study was longitudinal, with a fixed cohort of 1,240 children aged 6-48 months old at baseline. One of main purpose of this trial was to assess the effect of administering large, four-monthly doses of vitamin A on diarrhoea and acute lower respiratory infection. Full details of study design and data collection are given in Barreto et al. (1994). The aim in this work is to obtain a joint model for a sequence of different longitudinal counts using a correlated Poisson-gamma model. The methodology used is based on a foundational assumption of Poisson regression with the possibility of a latent process (or frailty) to take into account the overdispersion and correlation structure. For our application the idea is to model the number of mild and moderate/severe diarrhoea episodes that occurred between two following treatments.

2 Model

Frailty models were originally developed to allow for unobserved heterogeneity in survival analysis. Later they have been extended to model patterns of dependence. When frailties are common among groups of individuals but are randomly distributed across groups they are called shared frailty models. A shared frailty model can be considered as a random effects model with two sources of variation: group variation and individual random variation. The dependence structure is conceptually similar to compound symmetry models for multivariate data.

When the frailty factor is split into the sum of two components: one is shared by entire group and the other is an individual term, then the model is named a correlated frailty model. This model is based on the idea of correlated rather than shared frailties because it may be unrealistic to assume that the same random effect is present for all responses within a cluster. The specific feature of clusters, represented by association, is captured by the correlation between the individual frailties. Therefore, this

model measures the effect of the association and the heterogeneity parts of the frailty separately. Thus, it is necessary to define the joint distribution for these frailty terms.

Henderson and Shimakura (2003) proposed an extension of time constant shared frailty approach using a time-varying serially correlated gamma frailty in order to get a more realistic correlation structure in longitudinal count data. That is, a time varying frailty term is included in the model to handle both between heterogeneity and within-subject serial correlation. This model has different parameters to denote the heterogeneity (ξ) and the association (ρ). To illustrate this methodology, consider a longitudinal data set consisting of a count response variable N_{ij} , with the corresponding frailties Z_{ij} , observed at time $j = 1, 2, 3, \dots, t$, for independent subjects $i = 1, 2, \dots, m$ and a $p \times 1$ vector x_{ij} of covariates. Suppose that all subjects are observed over a common period. For simplicity the subject subscript i will be dropped for the moment. Then, N_1, N_2, \dots, N_t are the event counts and Z_1, Z_2, \dots, Z_t are the frailties respectively. To specify a time-varying frailty model the following criteria are used:

- the event counts are conditionally independent Poisson variables;
- gamma frailty Z_j for time j and subject i with mean one and variance ξ , written $\Gamma(1/\xi, 1/\xi)$;
- within-subject correlation $Corr(Z_j, Z_k) = \rho^{|j-k|}$.

According to Henderson and Shimakura (2003), the first two criteria allow a closed form for the marginal distribution of the event counts, which is negative binomial. The third admits a correlation structure which depends on the separation. Let x denote fixed and time-constant covariates and $u_j = \exp(\alpha_j + \beta x)$, where α_j determines the interval-specific baseline rate. The authors expressed a multivariate gamma distribution for $Z = (Z_1, Z_2, \dots, Z_t)'$, without a closed form, through the following Laplace transform:

$$\begin{aligned} L(u) &= E[\exp\{-u'Z\}] \\ &= |I + \xi C \text{diag}(u)|^{-1/\xi} \end{aligned} \quad (1)$$

where $\xi > 0$, $u = (u_1, \dots, u_t)$ and C is some positive definite correlation matrix which represents the association among gamma variables. Different correlation structures can be used allowing a valid distribution in $L(u)$. Inference can be based in part on the marginal distribution of the count data after integrating out the frailty terms.

Many models with components of repeatedly observed multivariate outcomes are analysed separately. Motivated by the Serrinha study, the idea is to allow a joint fit of the number of mild and moderate or severe diarrhoea

episodes which occurred per child. We aim to deduce a joint model for a sequence of different longitudinal measurements using a serially correlated Poisson-gamma model. The idea is to include a time varying frailty to account for heterogeneity between and within subjects in multivariate count data. In this case the heterogeneity and dependence between response variables should be taken into account in the model.

Consider M clusters of size t , each cluster representing a sequence of t observations on a q -vector N on subject i , $i = 1, 2, \dots, M$. Let $N_{ij}^{(l)}$ be the type (l) count at time j on subject i with $Z_{ij}^{(l)}$ the corresponding frailty and suppose there is also a $p \times 1$ vector x_{ij} of covariates, $j = 1, \dots, t$ and $l = 1, 2, \dots, q$. For our data let $q=2$. In a compact notation, this model can be represented by $N_i = (N_i^{(1)}, N_i^{(2)})'$ a $2t \times 1$ count vector with $Z_i = (Z_i^{(1)}, Z_i^{(2)})'$ a $2t \times 1$ frailty vector. As a result, more parameters in the correlation structure are needed. One of them expresses cross-component correlation representing the association between $Z_i^{(1)}$ and $Z_i^{(2)}$ and the other to represent association within frailty vectors of type 1 and type 2. It is important to highlight that in the absence of cross-component correlation, the joint model reduces to two separate models presented by Henderson and Shimakura (2003) and under perfect correlation this model reduces to the shared frailty model. We assume a multivariate gamma distribution where all components have gamma distribution with variance parameters ξ_1 and ξ_2 . Also, it is assumed that the event counts, $N_{ij}^{(l)}$ given frailties, are conditionally independent Poisson variables with mean equal to $Z_{ij}^{(l)} \exp(\alpha_j^{(l)} + \beta x_{ij})$ where $\alpha_j^{(l)}$ determines the interval-specific baseline rate for each type set of events $l = 1, 2$.

Because of the intractability in obtain the full likelihood, we use the composite likelihood procedure (Lindsay, 1988) based on all bivariate contributions of a multivariate longitudinal vector. Composite likelihood refers to a pooling of log likelihood contributions in an additive fashion which provide consistent parameter estimates.

To illustrate, fix $q = 2$ and let $N_i^{(1)} = (N_{i1}^{(1)}, \dots, N_{it}^{(1)})'$ refers to vector of type 1 counts with $Z_i^{(1)} = (Z_{i1}^{(1)}, \dots, Z_{it}^{(1)})'$ the corresponding vector frailties and vector type 2 in a similar way. In a compact notation, this model can be represented by $N_i = (N_i^{(1)}, N_i^{(2)})'$ a $2t \times 1$ count vector with $Z_i = (Z_i^{(1)}, Z_i^{(2)})'$ a $2t \times 1$ frailty vector. In the application to follow, $N_i^{(1)}$ will be the number of mild diarrhoea episodes for child i and $N_i^{(2)}$ will be the number of moderate or severe episodes.

In this approach there are two sets of correlated frailties, $Z_i^{(1)}$ and $Z_i^{(2)}$. Consider the general form of this correlation structure represented by a matrix R partitioned into 2^2 blocks, where each block has dimension $t \times t$:

$$R = \begin{vmatrix} \Sigma_1 & \Omega_{12} \\ \Omega'_{12} & \Sigma_2 \end{vmatrix}$$

The upper left (Σ_1) and lower right (Σ_2) $t \times t$ sub matrices are component-specific longitudinal correlation matrices representing association within frailty vectors of type 1 and type 2, respectively. The off-diagonal sub matrix (Ω_{12}) expresses cross-component correlation representing the association between $Z_i^{(1)}$ and $Z_i^{(2)}$.

As a prelude to the composite likelihood technique consider a bivariate distribution with Laplace transform:

$$L(u_1, u_2) = \left(\frac{1}{1 + \xi_1 u_1} \right)^{\frac{1}{\xi_1}} \left(\frac{1}{1 + \xi_2 u_2} \right)^{\frac{1}{\xi_2}} \left(\frac{1}{1 - \frac{\rho \xi_1 \xi_2 u_1 u_2}{1 + \xi_1 u_1 + \xi_2 u_2}} \right)^{\frac{1}{\sqrt{\xi_1 \xi_2}}}$$

In this case $Z_i \sim \Gamma(\frac{1}{\xi_i}, \frac{1}{\xi_i})$ and $Corr(Z_i, Z_j) = \rho$. By differentiation, it is possible to find the approximate marginal bivariate distribution of N_1, N_2 . The composite log-likelihood $l_i(\theta)$ contribution for subject i is

$$\sum \frac{1}{(2t - 1)} \log pr(N_{ij}^{(l)} = n_{ij}^{(l)}, N_{ik}^{(m)} = n_{ik}^{(m)})$$

where the sum is over j, k, l, m giving all contributions of two distinct counts with $\theta = (\{\alpha_j\}, \beta, \xi_s, \rho_s)$ representing the vector of unknown parameters. Maximum composite likelihood estimates of (ξ_1, ξ_2) and (ρ_1, ρ_2, ρ_3) are found numerically through a non-linear search routine following a Newton-type algorithm.

3 Results

The results suggested the presence of overdispersion in our data. It was assumed an exchangeable structure for component-specific longitudinal correlation matrix, represented by ρ_1 , and for off-diagonal submatrix we had the cross-correlation between frailties represented by ρ_3 . To fit multivariate correlated gamma frailty model we performed a two-stage modelling procedure. In the first stage, the effect of the covariates were modeled by GEE method. In the second stage a four-parameter search method is used for correlation and frailty parameters estimation.

Results from covariate effects (fixed) of the first stage are interpreted as: there is a decrease in the number of moderate/severe episodes when children received vitamin A, but this effect is not significant; the absence of a toilet at home is more strongly associated with the occurrence of moderate/severe episodes; as well as the absence of some treatment of drinking

water. Furthermore, the results from fitting correlated model in the second stage showed strong evidence to support the presence of correlation between frailties and correlation within frailty type 1.

4 Discussion

Originally, correlated frailty models were developed for the analysis of bivariate failure time data, in which two associated random variables are used to characterize the frailty effect for each pair. Through the Laplace transform, a correlated gamma frailty model was extended to allow frailties to be marginally gamma distributed with different variance parameters. However, this model can be reduced to the model proposed by Henderson and Shimakura (2003) when there is evidence to support the hypothesis of frailty variances being equal. Further, it is shown that in the absence of cross-correlation between frailties, we can fit two models separately. A limitation of this model is that it assumes that there is always a positive correlation between frailties. Also the correlation structure used to fit the multivariate count vector was exchangeable. However, the proposed procedure works for other correlation structures.

References

- BARRETO, M.L. et al (1994). Effect of vitamin A supplementation on diarrhoea and acute lower respiration infection. *Lancet*, **344**, 228-231.
- HEITJAN, D.F., SHARMA, D. (1997). Modelling repeated-series longitudinal data. *Statistics in Medicine*, **16**, 347-355.
- HENDERSON, R., SHIMAKURA, S (2003). A serially gamma frailty model for longitudinal count data. *Biometrika*, **90**, 355-366.
- LINDSAY, B. G. (1988). Composite likelihood methods. *Contemp. Math*, **80**, 221-239.

Delay in diagnosis of pulmonary tuberculosis in Portugal

Patrícia A. Filipe¹, Dulce Gomes¹, Carla Nunes^{2,3}, Marília Silva², Bruno de Sousa³, Teodoro Briz^{2,3}

¹ Escola de Ciência e Tecnologia, Universidade de Évora, CIMA/UE, Portugal

² Escola Nacional de Saúde Pública, Universidade Nova de Lisboa, Portugal

³ CMDT, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Portugal

E-mail for correspondence: pasf@uevora.pt

Abstract: Our main purpose is to achieve a better understanding of the several factors that are explicitly related with the time between the onset of the first symptoms and the diagnosis of Pulmonary Tuberculosis (“delay” in diagnosis) and its possible role on the level of the disease incidence. The length of such delay under scope is extremely important in Tuberculosis dissemination, since the bacilliferous patient circulates freely, being a source of infection and constituting a danger to the health of susceptible people. The period 2000–2009 was studied and the official Tuberculosis Surveillance database was used. Survival Analysis methodologies were applied for the characterization of the delay in diagnosis of Pulmonary Tuberculosis in which the event of interest is the diagnosis of the disease. Some explanatory variables for the delay were also considered: region, age, sex and some risk factors for Tuberculosis (e.g., drugs consumption, HIV infection), among others. At this stage factors connected to the health services system that may explain the delay were not considered. The delay in diagnosis looks, in general, excessive and much variable for a satisfactory control, and presents some differences between the categories for sex, age groups, drugs consumption, new case/relapse and HIV status. Only between critical (high risk) and non critical geographical areas the delay in diagnosis seems identical. As a global conclusion this study indicates a possibly higher probability of (earlier) diagnosis when patients are male, young or HIV positive. This evidence suggests that the delay in diagnosis is reflecting some differences according to risk factors, which may be substantially improved by both customized services and patients’ awareness.

Keywords: Pulmonary Tuberculosis Control; Delay in Diagnosis; Survival Analysis

1 Introduction

In Portugal, Pulmonary Tuberculosis (PTB) is frequently mentioned with concern by several authors and health authorities. On a global scale the endemic level is medium-low, with a slowly decreasing trend, but still at the

more unfavorable level of Western Europe. The current availability of effective means for the disease control, such as a sound intervention program with a disease surveillance system and a very reasonable overall performance, sustains the expectation of a greater influence over the endemic. Difficulty of control can be, partially, due to HIV coinfection in at least 15% of new cases of Tuberculosis and to some lack of coordination between services. Thus, there is still important progress to be promoted, particularly in the regions with a higher risk of disease, since the factors that mostly perpetuate the disease in populations are mainly related to socioeconomic, cultural, behavioral and organizational contexts that maintain the transmission of the bacillus, the non-detection of new cases or situations of latent disease, and the non compliance with therapy in specific groups (Briz *et al.*, 2009).

Nunes (2007), Nunes and Gomes (2009), Nunes *et al.* (2011) developed some early work in Spatial Epidemiology, and in particular spatiotemporal clustering processes, where they characterize the notified incidence rates distribution in Portugal. From this work, the existence of a pronounced geographical heterogeneity in the disease incidence notified rate was confirmed and significant space-time clusters were identified and described in detail.

In the current study, taking into account the dynamics of the endemic, we intended to explore several factors that may explain the time between the onset of the first symptoms and the diagnosis of Pulmonary Tuberculosis ("delay" in diagnosis). The analysis of this delay is extremely important for future implementation of disease control measures, since it is when the baciliferous patient circulates freely, being a source of infection of tuberculosis for other susceptible people. Relevant differences regarding "delay" between risk groups were looked for, by applying Survival Analysis techniques, as they may focus the period of time between the appearance of the first symptoms of PTB and the diagnosis of the disease. We included in the analysis variables that are known to be related to the dissemination of the disease, such as: sex, age group, drugs consumption, new case/relapse and HIV infection. The previously known geographical areas of higher/lower PTB risk were also considered (Nunes *et al.*, 2011).

2 Material and Methods

We worked with data concerning the period 2000 to 2009, per municipality, provided by two official sources: National Program for Tuberculosis Control (data on Pulmonary Tuberculosis notified cases) and National Statistical Institute (population data). Constant detection rates among municipalities and in time were assumed (according to the *WHO 2011 Global Tuberculosis Control Report*, in 2010 the Portuguese detection rate estimate was 87%.) The geographical areas of high/low incidence of PTB, already identified using the most recent data, are shown in Nunes *et al.* (2011) and Couceiro

et al. (2011). In order to identify critical incidence areas, spatiotemporal clustering analysis, based on the space-time scan statistic (Kulldorff, 1997), were applied then.

Survival Analysis methods allow us to describe and model time elapsed since a defined starting point until the occurrence of a certain event, usually called time-to-event. In this study, the diagnosis is the event of interest and the time-to-event is defined as the time between the first symptoms and the diagnosis of PTB. Thus, "time-to diagnosis" allows us to characterize the delay in diagnosis. The accuracy of the calculated delay depends on the accuracy of the two dates involved, date of the symptoms onset and date of the establishment of the diagnosis. Conclusions must take this into account. *Kaplan-Meier* survival curve was used to estimate the distribution of delay in diagnosis. The variables sex, age group, drugs consumption, new case/relapse, HIV infected/not infected and geographical areas of high/low risk were considered as explanatory factors. These variables were also used as predictors (covariates) when applying *Cox Regression* (Cox (1972), Cox and Oakes (1984)) to model the data. Cox regression analysis is used to model the *hazard function* $h(t)$, risk of occurring the event after time t . The model can be written as follows:

$$h(t|X) = h_0(t) \exp(B_1X_1 + B_2X_2 + \dots + B_kX_k) = h_0(t) \exp(BX)$$

where $h_0(t)$ represents the baseline hazard, $X = (X_1, X_2, \dots, X_k)$ is the vector of covariates and $B = (B_1, B_2, \dots, B_k)$ is the parameters vector. The *hazard ratio* is given by $\exp(B)$.

3 Main Results

From all notified cases we have considered diagnosed cases from 2000 until 2009, with a delay in diagnosis between 0 and 365 days. Our database has 35711 notified cases of PTB where 67.4% are male and 32.6% are female. The age groups considered (and respective frequencies) are: 0–4 (1.1%), 5–14 (2.4%), 15–24 (11%), 25–34 (23.1%), 35–44 (22.6%), 45–54 (14.9%), 55–64 (9.1%), 65–74 (8.4%) and > 74 (7.3%). From the 23049 cases to which HIV tests results are available, 20% are HIV positive. A similar approach indicates that 9% of the 31841 individuals that explicitly answered the question on drugs consumption (yes/no) were drug users. From the total of the notified cases selected for this study, about 90% corresponded to new cases of PTB. In terms of high PTB incidence rates level, based on critical areas identified in Nunes *et al.* (2011) - Oporto and Lisbon Metropolitan Areas - 15% belong to high risk areas. Based on selected cases, the delay in diagnosis presents a minimum of 1 day, a maximum of 365 days, a mean of 73.66 days, a median of 55 days and a very high standard-deviation of 61.44 days. We have applied Kaplan-Meier analysis to this delay using each variable described above as factors. A short summary of the Kaplan-Meier analysis is shown in Table 1, where we can see the median values of

the delay in diagnosis and the p-value for the Log-rank (Mantel-Cox) test. Considering sex as a factor and stratifying by each of the previous variables, we did not observed any changes in the results presented in Table 1, with the exception for HIV positive, individuals where we did not obtained a significant difference between sexes (p-value=0.069).

TABLE 1. Median delays according to explanatory variables. Log-rank test results.

Variable	p-value	categories	Median
sex	< 0.001	F	58
		M	54
age	< 0.001	0-4	33
		5-14	29
		15-24	48
		25-34	52
		35-44	55
		45-54	58
		55-64	64
		65-74	65
HIV	< 0.001	Positive	51
		Negative	59
drugs	< 0.001	Yes	53
		No	56
new case	0.002	Yes	55
		No	57
high risk area	0.704	Yes	57
		No	55

We have used a Cox Regression model to explain the probability of occurring the diagnosis of the disease after a certain number of days, considering the variables mentioned above as predictors. The results are shown in Table 2. Based on Wald test ($H_0 : B = 0$), only age group, sex and HIV were identified as significant predictors. The value of $\exp(B)$ for sex reveals that men have an increased probability of earlier diagnosis of 11%, as compared with women. For the age group variable, only the 5-14 group has an increased probability of 30% as compared with the reference group (0-4). For all the other age groups it is observed a decreased probability pattern of diagnosis varying from about 40% (55-64) to 20% (15-24). A similar behavior to sex is observed for the variable HIV, revealing that HIV positive have an increased probability of earlier diagnosis of 13%, as compared with HIV negative. The assumptions and suitability of the model were confirmed.

TABLE 2. Cox Regression results for delay, as a function of sex, age group and HIV infection.

Variable	B	p-value	Exp(B)	CI 95% Exp(B)
sex (F*)	0.105	< 0.001	1.111	1.078-1.144
age (0-4*)				
5-14	0.267	0.007	1.307	1.075-1.587
15-24	-0.214	0.012	0.807	0.683-0.954
25-34	-0.329	< 0.001	0.720	0.610-0.848
35-44	-0.380	< 0.001	0.684	0.580-0.806
45-54	-0.455	< 0.001	0.635	0.538-0.749
55-64	-0.515	< 0.001	0.597	0.505-0.707
65-74	-0.506	< 0.001	0.603	0.509-0.714
> 74	-0.499	< 0.001	0.607	0.511-0.720
HIV (neg*)	0.122	< 0.001	1.130	1.089-1.172

* Reference class.

4 Conclusions

These results may strongly reflect probable accuracies of the original data in PTB database; therefore conclusions have to be cautious and essentially indicative. In particular, it is important to highlight the following: (1) “delay” values are extremely variable, showing a very wide range; (2) both “date of disease onset” and “date of diagnosis”, used to calculate “delay”, are much dependent on patient’s behavior and memory, and on health services efficiency, thus impacting measurement precision; (3) no possible time trends in diagnosis practices of services, strongly influencing the second date, could be taken into account - the 10-year period observed was considered as homogeneous in this regard. The delay in diagnosis looks, in general, rather excessive and does not seem to have the same pattern between individuals of the different categories for sex, age group, drugs consumption, new case/relapse and HIV status. However, the delay in diagnosis distribution is identical in critical (high risk) and non critical geographical areas. Only age group, sex and HIV status were identified as significant covariates in the Cox Regression model used to assess the probability of occurring the diagnosis of the disease. Being male, young (<15) and HIV positive, as opposed to being female, in older age groups and HIV negative, respectively, indicate a possibly higher probability of (earlier) diagnosis. This slightly earlier diagnosis in some groups may suggest a “positive discrimination” attitude by clinicians. In Public Health terms, this study highlights some important facts that should be considered in future decisions: the magnitude of diagnosis delay has a huge impact in the population transmission rate. Therefore, all the possible reasons for this delay must be better understood (*e.g.*, validity and precision of relevant dates in the database, reasons regarding patients, services response, diagnosis procedures). A not so ex-

pected result in this study is that critical and non-critical areas seem to have the same delay pattern. Local Public Health decision makers have to adapt their actions considering whether they are or not in a critical area; this is why the focus of next research is being driven to deeper investigate the patterns of possible explanatory dimensions among critical and non-critical municipalities.

Acknowledgments: This work, within the research project PTDC/SAU-SAP/116950/2010, was financed by FCT/MCTES.

References

- Briz, T., Nunes, C., Alves, J. and Santos, O. (2009). O Controlo da Tuberculose em Portugal: uma apreciação crítica epidemiológica global. *Revista Portuguesa de Saúde Pública*, **1**, 19–54.
- Couceiro, L., Santana, P. and Nunes, C. (2011). Pulmonary tuberculosis and risk factors in Portugal: a spatial analysis. *The International Journal of Tuberculosis and Lung Disease*, **15**(11), 1445–1454.
- Cox, D. R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.
- Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.
- Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: theory and methods*, **26**(6), 1481–1496.
- Nunes, C. (2007). Tuberculosis incidence in Portugal: spatiotemporal clustering. *International Journal of Health Geographics*, 6–30.
- Nunes, C., Briz, T., Gomes, D. and Filipe, P.A. (2011). Pulmonary Tuberculosis and HIV/AIDS: joint space-time clustering under an epidemiological perspective. In: *Proceedings of the Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition*. Barbara Cafarelli (Ed.), Foggia e Gargano.
- Nunes, C. and Gomes, D. (2009). Processo de detecção de aglomerações espaço-temporais: alguns condicionantes. In: *Estatística. Arte de Explorar o Acaso. Actas do XVI Congresso da Sociedade Portuguesa de Estatística*. (Oliveira, I., Correia, E., Ferreira, F., Dias, S. and Braumann, C., eds.), Edições SPE, 477–488.
- WHO (2011). *Global tuberculosis control 2011: WHO Report*. Geneva.

A comparison of several background correction and normalization methods on microarray data

Adelaide Freitas¹, Sara Roque¹

¹ University of Aveiro, Portugal

E-mail for correspondence: adelaide@ua.pt

Abstract: We investigate the effect of 36 preprocessing strategies resultant of combination of one background correction method with one normalization method on three published microarray data sets which microarrays are already classified by cancer type. For each data set, we analyze the ability for detecting differentially expressed genes provided by the popular algorithm SAM when applied on each of the 36 preprocessed microarray data matrices and measured in terms of the false discovery rate. Furthermore, based on the estimates of the classification error rates of two distinct types of classifiers induced from each of the 36 preprocessed microarray data matrices determined in other paper, we propose the visualization of effects of both background correction and normalization methods using biplots. Our study evidences that the rate of differentially expressed genes incorrectly detecting by SAM and the predictive performance of the two classifiers induced from the data depends strongly on the data set and the selected preprocessing method. There is no preprocessing strategy that outperforms all others in all circumstances.

Keywords: Background correction; Normalization; False discovery rate; SAM; Biplot.

1 Introduction

Microarray technology is aimed at monitoring thousands of genes simultaneously in one single experiment. Microarray experiments involve many steps and each step can introduce experimental bias and random fluctuations in the measured intensities that can affect the quality of raw data. Background correction (BC) and normalization (NM) are preprocessing techniques aimed at correcting the raw data at the undesirable fluctuations arising from technical factors such that intrinsic biological variations are still retained. Based on three published cDNA microarray data sets (Lymphoma, Lung and Liver available in <http://genome-www5.stanford.edu/>) which microarrays are already classified by cancer type, Freitas et al. (2009) investigated the predictive performance on cancer classification of two clas-

TABLE 1. Estimates of the classification error rates (%) of kNN classifiers induced from preprocessed microarray data matrices obtained from Lymphoma and Lung data sets.

Data set	Methods	none	sub	half	min	edw	nexp
Lymphoma	NN	26.85	21.20	23.14	21.29	25.92	19.44
	IG	12.96	16.66	20.37	16.66	22.22	11.1
	IL	12.03	15.74	17.59	15.74	21.29	11.1
	SI	16.66	18.51	17.59	18.51	20.37	12.96
	IG-SL	12.03	16.6	20.37	15.74	21.29	9.25
	IL-SL	9.25	16.6	17.59	16.6	21.29	11.11
Lung	NN	35.38	26.15	36.92	36.92	35.38	35.38
	IG	23.07	29.23	41.53	41.53	41.53	32.3
	IL	20	27.69	38.46	38.46	41.53	29.23
	SI	32.3	27.69	29.23	29.23	36.92	30.76
	IG-SL	24.61	30.76	35.38	36.92	44.61	29.23
	IL-SL	21.53	24.61	41.53	41.53	35.38	27.69

sifiers (k-nearest neighbor -kNN- and support vector machine with linear kernel -SVM-) induced from microarray data where a particular BC method was applied, individually and in combination with a single-bias or double-bias-removal NM method. They considered six BC methods available in the R package *limma* from Bioconductor (<http://www.bioconductor.org>) through the function *backgroundCorrect* (**none**: no BC, **sub**: subtraction, **half**: half, **min**: minimum, **edw**: edwards, **nexp**: normexp) and six NM methods available in the R package *marray* from Bioconductor through the functions *maNorm* and *maNormMain* (NN: no NM, IG: Intensity Global loess, IL: Intensity Local loess, SL: Spatial Local loess, IG-SL: Intensity Global loess followed by Spatial Local loess, IL-SL: Intensity Local loess followed by Spatial Local loess) (for more details, see Freitas et al (2009)). The estimates of the classification error rates of kNN classifier for both Lymphoma and Lung data sets are depicted in Table 1. Due to limited space, results for the other data set and for the SVM classifier will be omitted here.

Assuming a two way additive model for the classification error e_{ij} , with $i \in \{\text{none, sub, half, min, edw, nexp}\}$ e $j \in \{\text{NN, IG, IL, SL, IG-SL, IL-SL}\}$, Freitas et al evaluated the mean effect of the application of BC methods on the performance of classifiers induced from each preprocessed data matrix, for each of three data sets. While the exploratory analysis suggested that **sub** and **nexp** may be the best BC methods, there was no possible to guarantee significant gains on the predictive performance of both kNN and SVM classifiers.

In order to explore the dependence on data sets, we now investigate effects produced by those 36 preprocessing strategies, resultant of combinations of

the six BC and the six NM methods, by two others different approaches. In Section 2 we visualize the effect on the performance of both kNN and SVM classifiers induced from preprocessed microarray data sets displaying correspondence analysis (CA) biplots as discussed in Park et al (2008). In Section 3 we analyze the effect on the correct identification of differentially expressed genes given by the popular algorithm SAM (Tusher et al (2001)) when applied on each preprocessed microarray data set and measured in terms of the false discovery rate.

2 Biplots to visualize effects

In Park et al (2008), CA biplots showed to be a useful exploratory tool to visualize relations between multiple levels of two factors based on a dependent variable observed on each level. Taking, for each data set, the matrix of the estimates of classification error rates, we explore relations between column and row factors (BC and NM methods) applying CA biplots. For the two data matrices depicted in Table 1, we have the CA biplots displayed in Figure 1. For Lymphoma data set (graph on the left side), we observe:

- different effects when IG, IL, IG-SL, IL-SL and NN and SL are used (blue points on different positions on the plane: left versus right side);
- sub and min BC methods tend to produce similar profile (the nearest red points), and so the effect of a combination of these BC techniques with any NM strategy is similar;
- none as BC method and IL-SL as NM method are negatively related, and so IL-SL is opposed to the effect yielded when no BC is applied (blue and red points in opposed positions).

From CA biplot for Lung data set (graph on the right side in Figure 1), we clearly check that the latter two remarks above are not detected.

3 Assessing differential expression

Now we investigate the effect of preprocessing methods on the ability of the SAM algorithm for correctly detecting differentially expressed genes. The SAM algorithm was applied on each of the 36 preprocessed microarray data matrices for Lymphoma, Lung and Liver data sets using the *samr* software (<http://www-stat.stanford.edu/~tibs/SAM>). Comparing the rate of false discoveries versus the number of differentially expressed genes selected by SAM from each of the preprocessed data matrices, there were no preprocessing strategy that outperformed all others in all circumstances. Figure 2 shows the rate of false discoveries versus the number of differentially expressed genes selected by SAM from preprocessed data matrices

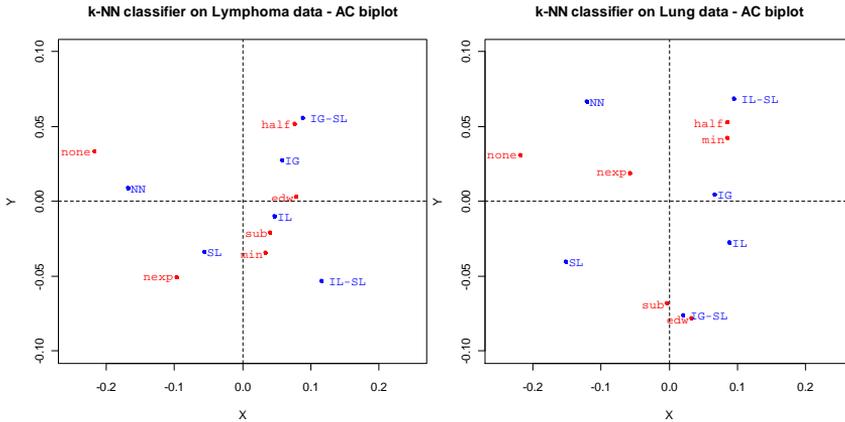


FIGURE 1. AC biplots obtained from the two data matrices depicted in Table 1.

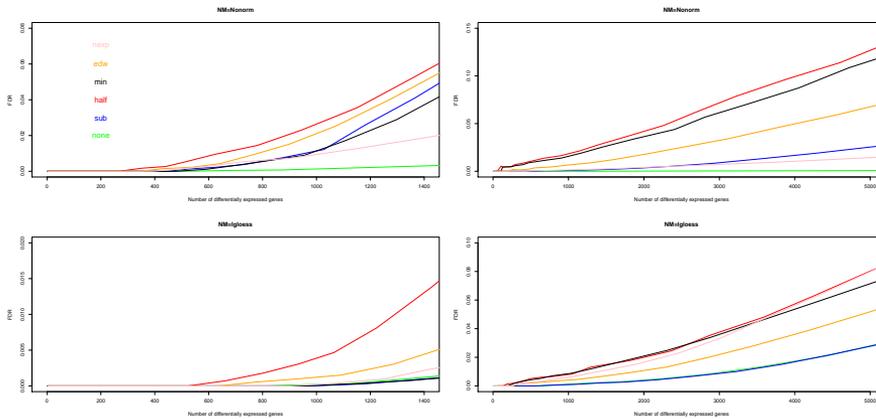


FIGURE 2. False discovery rates (FDR) obtained by the SAM algorithm when applied on preprocessed data microarrays from Lymphoma (on the left) and Lung (on the right) data sets where one of six BC methods is combined with the NM methods NN (on the top) and IG (on the bottom).

from Lymphoma and Lung datasets when each BC method is combined with no normalization and IG normalization method. The curves show that no BC method does better. However, there is a clearly differentiable behavior among the cases reported in Figure 2 for the sub BC method, which is the standard and most common method of background correction of microarray data.

4 Conclusion

Application of preprocessing methods on data resultants from microarray experiments is an important issue to take into account before to make any statistical analysis of gene expression levels. Preprocessing techniques are aimed at deleting experimental bias and random fluctuations that can affect the quality of raw data. Since there is no preprocessing strategy that outperforms others in all circumstances, it is highly recommended to applied more than one preprocessing strategy on microarray data and compare results before any conclusion and subsequent analysis from preprocessed data.

Acknowledgments: This research was supported by the Foundation for Science and Technology (FCT-Portugal) through Center of Research and Development in Mathematics and Applications of University of Aveiro. The preprocessed data sets was obtained during the project PTDC/MAT/72974/2006 (FCT-Portugal).

References

- Freitas, A., Castillo, G., So Marco, A. (2009) Effect of Background Correction on Cancer Classification with Gene Expression Data. In: *Proceedings of AIME 2009, Lecture Notes in Artificial Intelligence*, Springer Verlag. pp. 416–420.
- Park, M., Lee, J., Lee J.B., Song, S.H. (2008) Several biplot methods applied to gene expression data. *Journal of Statistical Planning and Inference*, **138**, 500–515.
- Tusher, V., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. In *Proceedings of the National Academy of Sciences*, **98**, 9, pp. 5116-5121

Recursive linear estimation from multi-sensor observations with correlated uncertainties

Irene García-Garrido¹, J. Linares-Pérez¹, R. Caballero-Águila²

¹ Dpto. Estadística e I.O., Universidad de Granada, Spain

² Dpto. Estadística e I.O., Universidad de Jaén, Spain

E-mail for correspondence: irenegarciag@ugr.es

Abstract: The least-squares linear estimation problem in linear systems with uncertain observations coming from multiple sensors with different uncertainty characteristics is addressed. Such systems are characterized by including in the observation equation not only additive noise, but also a multiplicative noise described by a sequence of Bernoulli random variables whose values -one or zero- indicate the presence or absence of the state in the corresponding observation. Recursive filtering and fixed-point smoothing algorithms are obtained assuming that the Bernoulli variables are correlated at instants that differ m units of time and using an innovation approach. The performance of the proposed estimators is illustrated by a simulation example.

Keywords: Least-squares estimation; Uncertain observations; Multiple sensors.

1 Introduction

The least-squares (LS) linear estimation problem in systems with uncertain observations has been widely studied (see e.g. Sahebsara et al. (2007), Hermoso-Carazo et al. (2008) and references therein). This is mainly due to the applicability of such systems to many practical situations where the signal to be estimated can be randomly missing in the observations (for example, situations with intermittent failures in the observation mechanism, accidental loss of some measurements or data inaccessibility during certain times). These systems are characterized by the fact that the observation equation contains, besides additive noise, a multiplicative noise component defined by a sequence of Bernoulli random variables, which models the presence or absence of the state in the observations.

On the other hand, over the past few years, the estimation problem of random signals from noisy measurements coming from different multiple sensors is gradually becoming an active research topic due to its importance in many engineering application fields, where sensor networks are used to obtain all the available information on the system state, and its estimation must be carried out from the observations provided by all the sensors.

Most papers assume that the observations come from multiple sensors with identical uncertainty characteristics; nevertheless, in the last years, this situation has been generalized by several authors considering uncertain observations whose statistical properties are assumed not to be the same for all the sensors (see e.g. Hounkpevi & Yaz (2007) for the case when the uncertainty is modelled by independent random variables, or Caballero-Águila et al. (2011) for the case when such variables are correlated at consecutive sampling times).

In this paper, recursive algorithms are proposed for the LS linear filtering and fixed-point smoothing problems of discrete-time signals from uncertain observations coming from multiple sensors, under the assumption that the variables describing the uncertainty in the observations are correlated in instants that differ m units of time. This correlation model for the uncertainty, more general than that considered in Caballero-Águila et al. (2011), allows us to model situations where the signal cannot be absent in $m + 1$ consecutive observations.

2 Model Description

Consider an n -dimensional stochastic process $\{x_k; k \geq 0\}$ representing the state of a discrete-time linear stochastic system with scalar uncertain observations $\{y_k^i; k \geq 1\}$, $i = 1, \dots, r$, coming from r sensors and perturbed by an additive noise and a multiplicative noise describing the uncertainty:

$$\begin{aligned} x_k &= F_{k-1}x_{k-1} + w_{k-1}, \quad k \geq 1, \\ y_k^i &= \theta_k^i H_k^i x_k + v_k^i, \quad k \geq 1, \quad i = 1, \dots, r, \end{aligned}$$

where $\{w_k; k \geq 0\}$ and $\{v_k^i; k \geq 1\}$ are zero-mean white noise processes with $Cov[w_k] = Q_k$ and $Cov[v_k^i] = R_k^i$, respectively. For $i = 1, \dots, r$, $\{\theta_k^i; k \geq 1\}$ is a sequence of Bernoulli random variables with $P[\theta_k^i = 1] = \bar{\theta}_k^i$. For $i, j = 1, \dots, r$, the variables θ_k^i and θ_s^j are independent for $|k - s| \neq 0, m$ and $Cov[\theta_k^i, \theta_s^j]$ are known for $|k - s| = 0, m$. Defining $\theta_k = (\theta_k^1, \dots, \theta_k^r)^T$, the covariance matrices of θ_k and θ_s will be denoted by $K_{k,s}^\theta$. The initial state x_0 is a random vector with $E[x_0] = \bar{x}_0$ and $Cov[x_0] = P_0$. Finally, we assume that the initial state x_0 and the noise processes $\{w_k; k \geq 0\}$, $\{v_k^i; k \geq 1\}$ and $\{\theta_k^i; k \geq 1\}$ are mutually independent.

Note that, when $\theta_k^i = 1$, which occurs with known probability $\bar{\theta}_k^i$, the state x_k is present in the observation y_k^i coming from the i -th sensor at time k , whereas if $\theta_k^i = 0$ such observation only contains additive noise v_k^i with probability $1 - \bar{\theta}_k^i$.

Denoting $y_k = (y_k^1, \dots, y_k^r)^T$, $v_k = (v_k^1, \dots, v_k^r)^T$, $H_k = (H_k^{1T}, \dots, H_k^{rT})^T$ and $\Theta_k = \text{Diag}(\theta_k^1, \dots, \theta_k^r)$, the observation equation is equivalent to the following stacked observation equation

$$y_k = \Theta_k H_k x_k + v_k, \quad k \geq 1. \quad (1)$$

3 LS linear estimation problem

Our aim in this paper is to derive recursive algorithms for the LS linear filtering and fixed-point smoothing problems using uncertain observations (1). Due to the correlation assumed in the uncertainty, this model covers, for example, situations where the signal cannot be absent in $m + 1$ consecutive observations. Specifically, the problem is to obtain the LS linear estimator, $\hat{x}_{k/L}$, of the signal x_k based on the observations $\{y_1, \dots, y_L\}$, with $L \geq k$, by recursive formulas. For this purpose an innovation approach will be used, consisting of obtaining the estimators as a linear combination of the innovations, which are defined as $\nu_k = y_k - \hat{y}_{k/k-1}$, where $\hat{y}_{k/k-1}$ is the one-stage LS linear predictor of y_k . Then, the LS linear estimator is expressed as

$$\hat{x}_{k/L} = \sum_{i=1}^L S_{k,i} \Pi_i^{-1} \nu_i, \quad k \geq 1,$$

where $S_{k,i} = E[x_k \nu_i^T]$ and $\Pi_i = E[\nu_i \nu_i^T]$ is the covariance of ν_i .

3.1 Linear filtering algorithm

The linear filter, $\hat{x}_{k/k}$, of the state x_k is obtained as

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + S_{k,k} \Pi_k^{-1} \nu_k, \quad k \geq 1; \quad \hat{x}_{0/0} = \bar{x}_0,$$

where the state predictor, $\hat{x}_{k/k-1}$, is given by

$$\hat{x}_{k/k-1} = F_{k-1} \hat{x}_{k-1/k-1}, \quad k \geq 1.$$

The innovation process verifies

$$\begin{aligned} \nu_k &= y_k - \bar{\Theta}_k H_k \hat{x}_{k/k-1}, \quad k \leq m, \\ \nu_k &= y_k - \bar{\Theta}_k H_k \hat{x}_{k/k-1} \\ &\quad + \Psi_{k,k-m} \left[\nu_{k-m} - \sum_{i=1}^{m-1} T_{k-i,k-m}^T \Pi_{k-i}^{-1} \nu_{k-i} \right], \quad k \geq m + 1, \end{aligned}$$

where $\bar{\Theta}_k = E[\Theta_k]$ and $\Psi_{k,k-m} = K_{k,k-m}^\theta \circ (H_k \mathbb{F}_{k,k-m} D_{k-m} H_{k-m}^T \Pi_{k-m}^{-1})$, with \circ the Hadamard product, $\mathbb{F}_{k,i} = F_{k-1} \cdots F_i$ and $D_k = E[x_k x_k^T]$ recursively obtained by

$$D_k = F_{k-1} D_{k-1} F_{k-1}^T + Q_{k-1}, \quad k \geq 1; \quad D_0 = P_0 + \bar{x}_0 \bar{x}_0^T.$$

The matrices $T_{k,k-i}$ are given by

$$\begin{aligned} T_{k,k-i} &= \bar{\Theta}_k H_k \mathbb{F}_{k,k-i} S_{k-i,k-i}, \quad 2 \leq k \leq m, \quad 1 \leq i \leq k-1, \\ T_{k,k-i} &= \bar{\Theta}_k H_k \mathbb{F}_{k,k-i} S_{k-i,k-i} \\ &\quad - \Psi_{k,k-m} T_{k-i,k-m}^T, \quad k \geq m+1, \quad 1 \leq i \leq m-1. \end{aligned}$$

The covariance matrix of the innovation, $\Pi_k = E[\nu_k \nu_k^T]$, satisfy

$$\begin{aligned} \Pi_k &= K_{k,k}^\theta \circ (H_k D_k H_k^T) + R_k + \bar{\Theta}_k H_k S_{k,k}, \quad k \leq m, \\ \Pi_k &= K_{k,k}^\theta \circ (H_k D_k H_k^T) + R_k \\ &\quad - \Psi_{k,k-m} \left(\Pi_{k-m} + \sum_{i=1}^{m-1} T_{k-i,k-m}^T \Pi_{k-i}^{-1} T_{k-i,k-m} \right) \Psi_{k,k-m}^T \\ &\quad + \bar{\Theta}_k H_k S_{k,k} + S_{k,k}^T H_k^T \bar{\Theta}_k - \bar{\Theta}_k H_k P_{k/k-1} H_k^T \bar{\Theta}_k, \quad k \geq m+1. \end{aligned}$$

The matrix $S_{k,k}$ is determined by the following expression

$$\begin{aligned} S_{k,k} &= P_{k/k-1} H_k^T \bar{\Theta}_k, \quad k \leq m, \\ S_{k,k} &= P_{k/k-1} H_k^T \bar{\Theta}_k - (S_{k,k-m} \\ &\quad - \sum_{i=1}^{m-1} S_{k,k-i} \Pi_{k-i}^{-1} T_{k-i,k-m}^T) \Psi_{k,k-m}^T, \quad k \geq m+1, \end{aligned}$$

where $P_{k/k-1}$, the prediction error covariance matrix, is obtained by

$$P_{k/k-1} = F_{k-1} P_{k-1/k-1} F_{k-1}^T + Q_{k-1}, \quad k \geq 1,$$

with $P_{k/k}$, the filtering error covariance matrix, verifying

$$P_{k/k} = P_{k/k-1} - S_{k,k} \Pi_k^{-1} S_{k,k}^T, \quad k \geq 1; \quad P_{0/0} = P_0.$$

3.2 Fixed-point smoothing algorithm

For each fixed $k \geq 1$, the linear smoothers, $\hat{x}_{k/k+N}$, $N \geq 1$, of the state x_k , are calculated as

$$\hat{x}_{k/k+N} = \hat{x}_{k/k+N-1} + S_{k,k+N} \Pi_{k+N}^{-1} \nu_{k+N}, \quad N \geq 1,$$

whose initial condition is the filter, $\hat{x}_{k/k}$, and

$$\begin{aligned} S_{k,k+N} &= (D_k \mathbb{F}_{k+N,k}^T - M_{k,k+N-1} F_{k+N-1}^T) H_{k+N}^T \bar{\Theta}_{k+N}, \quad k \leq m, \quad N \geq 1, \\ S_{k,k+N} &= (D_k \mathbb{F}_{k+N,k}^T - M_{k,k+N-1} F_{k+N-1}^T) H_{k+N}^T \bar{\Theta}_{k+N} \\ &\quad - \left(S_{k,k+N-m} - \sum_{i=1}^{m-1} S_{k,k+N-i} \Pi_{k+N-i}^{-1} T_{k+N-i,k+N-m}^T \right) \\ &\quad \times \Psi_{k+N,k+N-m}^T, \quad k \geq m+1, \quad N \geq 1, \end{aligned}$$

where the matrices $M_{k,k+N}$ satisfy the following recursive formula

$$\begin{aligned} M_{k,k+N} &= M_{k,k+N-1} F_{k+N-1}^T + S_{k,k+N} \Pi_{k+N}^{-1} S_{k+N,k+N}^T, \quad N \geq 1, \\ M_{k,k} &= D_k - P_{k/k}. \end{aligned}$$

The innovations ν_{k+N} , their covariance matrices Π_{k+N} and the matrices $\Psi_{k+N,k+N-m}$ are given in the filtering algorithm. Finally, the estimation error covariance matrices, $P_{k/k+N}$, are given by

$$P_{k/k+N} = P_{k/k+N-1} - S_{k,k+N} \Pi_{k+N}^{-1} S_{k,k+N}^T, \quad N \geq 1,$$

with initial condition given by the filtering error covariance matrix.

4 Computer-simulation example

To illustrate the performance of the proposed estimators, consider a scalar first-order autoregressive model, $x_k = 0.95x_{k-1} + w_{k-1}$, $k \geq 1$, where x_0 is a zero-mean Gaussian variable with variance $P_0 = 1$ and $\{w_k; k \geq 0\}$ is a zero-mean white Gaussian noise with $Q_k = 0.1, \forall k \geq 0$. The uncertain observations coming from two sensors are specified by

$$y_k^i = \theta_k^i x_k + v_k^i, \quad k \geq 1, \quad i = 1, 2,$$

where the noises $\{v_k^i; k \geq 1\}, i = 1, 2$, are independent zero-mean white processes with constants variances $R_k^1 = 0.5$ and $R_k^2 = 0.9, \forall k \geq 1$, respectively. The variables modelling the uncertainty of each sensor, θ_k^i , are defined from two independent sequences of independent Bernoulli random variables, $\{\gamma_k^i; k \geq 0\}, i = 1, 2$, with $P[\gamma_k^i = 1] = \gamma_i$, as follows

$$\theta_k^i = 1 - \gamma_{k+m}^i (1 - \gamma_k^i), \quad i = 1, 2.$$

Since the variables γ_k^i and γ_s^i are independent, θ_k^i and θ_s^i are also independent for $|k - s| \neq 0, m$. The mean of these variables is given by $\bar{\theta}^i = 1 - \gamma_i(1 - \gamma_i)$, thus being a decreasing function of γ_i .

Under these hypotheses, the estimation error variances have been calculated for different values of the probabilities γ_1 and γ_2 , which provide different values of $\bar{\theta}_1$ and $\bar{\theta}_2$ (since the values of $\bar{\theta}^i$ are the same if $1 - \gamma_i$ is used instead of γ_i , only the case $\gamma_i \leq 0.5$ has been considered). For instance, Figure 1 displays the filtering and the fixed-point smoothing error variances at $k = 50$ versus γ_1 (for constant values of γ_2). This figure shows that the smoothing error variances are less than the filtering ones and also that, as γ_2 decreases (thus increasing the probability that the signal is present in the observations), the error variances are smaller in both cases and, hence, better estimations are obtained.

5 Conclusion

The state estimation problem for discrete-time linear systems with uncertain observations has been addressed when the uncertainty at any sampling time k depends only on the uncertainty at the previous time $k - m$. Applying an innovation technique, recursive algorithms for the filter and fixed-point smoother are derived.

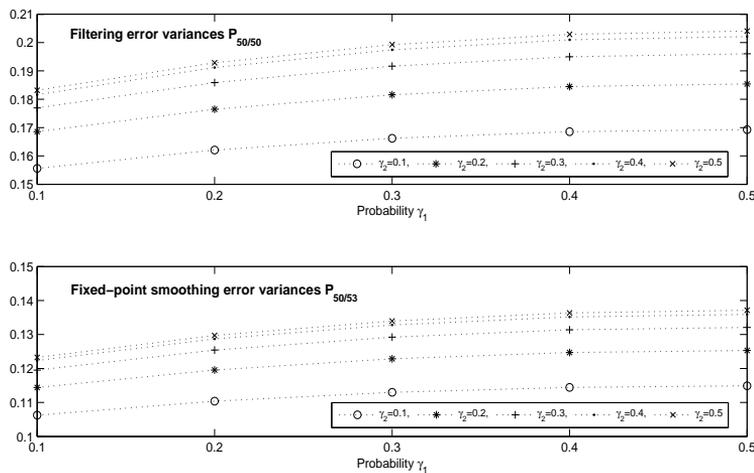


FIGURE 1. Filtering and fixed-point smoothing error variances at $k = 50$ versus γ_1 with γ_2 varying from 0.1 to 0.5, for $m = 3$.

Acknowledgments: This research is supported by Ministerio de Educación y Ciencia (Programa FPU and grant No. MTM2011-24718) and Junta de Andalucía (grant No. P07-FQM-02701).

References

- Caballero-Águila, R., Hermoso-Carazo, A., and Linares-Pérez, J. (2011). Linear and quadratic estimation using uncertain observations from multiple sensors with correlated uncertainty. *Signal Processing*, **91**, 330–337.
- Hermoso-Carazo, A., Linares-Pérez, J., Jiménez-López, J.D., Caballero-Águila, R., and Nakamori, S. (2008). Recursive fixed-point smoothing algorithm from covariances based on uncertain observations with correlation in the uncertainty. *Applied Mathematics and Computation*, **203**, 243–251.
- Houkpevi, F.O., and Yaz, E.E. (2007). Minimum variance linear state estimators for multiple sensors with different failure rates. *Automatica*, **43**, 1274–1280.
- Sahebsara, M, Chen, T., and Shah, S.L. (2007). Optimal \mathcal{H}_2 filtering with random sensor delay, multiple packet dropout and uncertain observations. *International Journal of Control*, **80(2)**, 292–301.

A flexible regression framework for TMS-EEG signals

Kathakali Ghosh Mukherjee¹, Claire Miller¹, Adrian W. Bowman¹, Gregor Thut²

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.ghosh.1@research.gla.ac.uk

Abstract: Electroencephalogram (EEG) recordings of transcranial magnetic stimulation (TMS) experiments present noisy time series data. There is a lot of interest in eliminating known artefacts and modelling the remaining variation in such datasets. This paper investigates the components of TMS-EEG signals by using an additive model framework to estimate artefacts and to distinguish brain signal of interest at the single replicate level. In addition, the time series of the signal after TMS are investigated for evidence of entrainment to an alpha-band frequency and variation in the signal across replicates and between different subjects is studied.

Keywords: Additive Model; TMS; EEG; Entrainment.

1 Introduction

Transcranial Magnetic Stimulation (TMS), developed by Cracco et.al.(1989), is used to study the functions of the brain by exposing it to controlled non-invasive magnetic stimuli. Recently it has been possible to obtain electroencephalography (EEG) recordings of TMS leading to experimentation in TMS-EEG. This generates functional data which provides information on how TMS affects brain function. A set of time series, which are spatially correlated, are simultaneously recorded from a number of channels or electrodes. EEG recordings of TMS induced brain activity are contaminated by the magnetic field induced by the magnetic nature of the stimulus (Thut *et. al.*(2011)). TMS-EEG recordings are characterized by strong pulses, as a result of the magnetic stimulation forcing the magnetic field to change instantaneously, heavily masking any brain signals that are registered. The task then becomes retrieval of the brain signals, after eliminating the strong artefact.

The first goal is to sufficiently model a set of noisy time series simultaneously recorded from a number of channels or electrodes, accounting for spatiotemporal patterns in the signal. This entails extraction of the signal

of interest in a single replicate, sufficiently eliminating the noise and artefacts. Secondly, any evidence of change and/or entrainment in the brain signals during or after the stimulus (thought to be a direct effect of TMS) is of interest. Thirdly, the variation in the traces of brain signals over time - between different channels in a single replicate, between replicates within a subject exposed to different conditions, and between subjects is of interest. A wide range of parametric and semi-parametric modelling strategies have been proposed for pre-processing, estimation and inference of EEG data. Most of these models, however, are targeted either at the spatial or the temporal domain. This paper applies an additive model (AM) described by Hastie and Tibshirani (1986) and Wood (2006) for EEG signal responses in a TMS setting. This class of flexible models provides a method for data to be modelled simultaneously in space and time. It allows for the data to guide the model in estimating the functional dependence of the mean response on the predictors for a single replicate. In this paper, the EEG signals are assumed to be a sum of different components - mean response, the artefact components and the signal of interest.

1.1 Data

The data consist of EEG recordings from 8 subjects in a TMS-EEG cognitive experiment with 4 conditions, each having 54 trials (Thut *et. al.*(2011)). Data are recorded at $S = 64$ channels connected to an EEG cap and $T = 5501$ time points spanning a time of 1.1 seconds (-0.1 to 1.0 s with 0 s being the time of stimulation). Each trial consists of pre-stimulus recordings of the EEG signal (-0.1 to 0 s), followed by five TMS pulses (approximately 0 to 0.4 s) and the signal post-stimulus (0.4 to 1s). This experiment studies the effect of TMS pulses on specific brain activation as recorded by EEG. The TMS pulses are administered within the alpha frequency band (8-12 Hz) with the exact frequency determined previously for each subject. A characteristic of the EEG signal is the presence of an underlying mains current of 50 Hz due to the electrical apparatus used. The magnitude of this 'cyclic' component varies depending on the position and orientation of the equipment, its distance from the electrodes, etc. Initially we consider a single replicate spatiotemporally for one subject to fit the model. We then consider variation across several replicates for this subject.

2 Methods

We propose an additive model for this data. The underlying mains current can be estimated either (i) by approximating it with a 50 Hz cosine function or (ii) as one of the temporally smoothed components in an additive model with a periodicity of 0.02 s. In addition, a smooth component designed to track the spike effect in the data for 10 milliseconds after each TMS pulse

is included. The remainder of the trace for this component is set to 0. A spatiotemporally smooth function estimates the remaining signal.

2.1 Additive Model

For a single channel j the following model is proposed:

$$y_j(t) = \mu + m_1(t) + m_2(t) + m_3(t) + \varepsilon \quad \forall t = 1, \dots, T; j = 1, \dots, n \quad (1)$$

where t is the time duration of the recordings and n is the total number of channels (or electrodes). Here m_1 represents the smooth trace of the TMS pulses, m_2 is the smooth signal of interest and m_3 is a smooth function defining the underlying cyclic 50 Hz. component. These three terms are all estimable because they operate on different timescales. m_1 is active only over a 10 ms period after each pulse, m_3 is cyclical with period of 20 ms and m_2 is unrestricted. The error ε is assumed to be independent and normally distributed. More generally, the spatiotemporal model may be written as:

$$y(s, t) = \mu + m_1(s, t) + m_2(s, t) + m_3(s, t) + \varepsilon \quad \forall t = 1, \dots, T \quad (2)$$

The components corresponding to the TMS pulses and mains current (m_1 and m_3) are smoothed over time only, as they are assumed to be spatially unrelated between the channels. Local mean smoothers are employed for temporal smoothing to estimate the spikes $m_1(s, t)$ as well as for spatiotemporal smoothing to estimate the signal of interest $m_2(s, t)$. A cyclic smoother is applied to estimate the temporally smoothed cyclic mains current with a frequency of 50 Hz. The smoothing parameters are predetermined and the components estimated using the backfitting algorithm. The basic matrix of weights for temporal smoothing of $m_1(s, t)$ is constructed to ensure that each spike is only estimated for a short window (10 ms) in the data (Bowman and Azzalini, 1997, 2003, Ventrucchi *et. al.*(2011)).

3 Results

After applying (2) to the data from a single replicate in a subject the three components shown in Figure 1 were obtained.

It is of interest to investigate the remaining estimated smooth signal once the effects of the TMS pulses and the mains current have been eliminated. For each channel, particularly those that are in closest proximity to the TMS equipment (viz. P4, CP4 and P2), the signal of interest (Figure 2) reveals evidence of entrainment to the alpha band frequency (8-12 Hz), especially after the third TMS pulse.

In a spatiotemporal map of the brain (Figure 3), considerable variability is evident across several time instances between channels in a single replicate as well as between replicates within a subject. Figure 3 shows the spatial variation of the signal of interest in three replicates of the same subject. Further, the range of amplitude of the signal also varies from replicate to replicate (as seen in the figure) and between subjects.

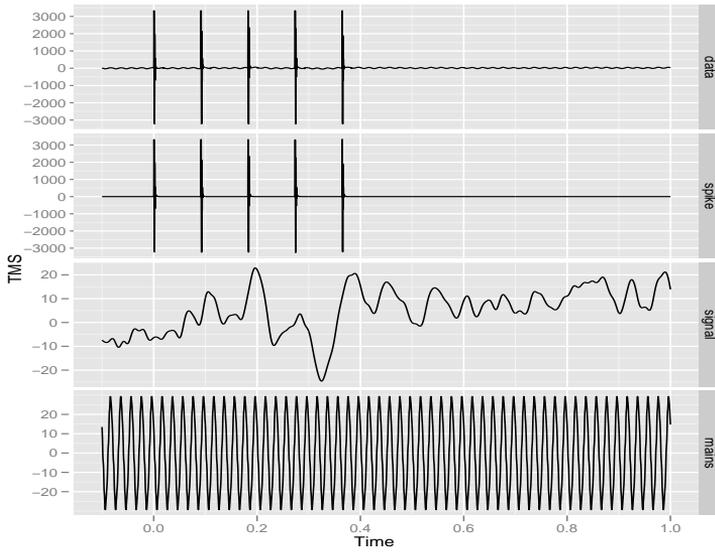


FIGURE 1. Additive components for channel CP4 in a single replicate of a subject: trace of data across the time span (topmost) and estimated spike, signal and mains current components from the additive model.

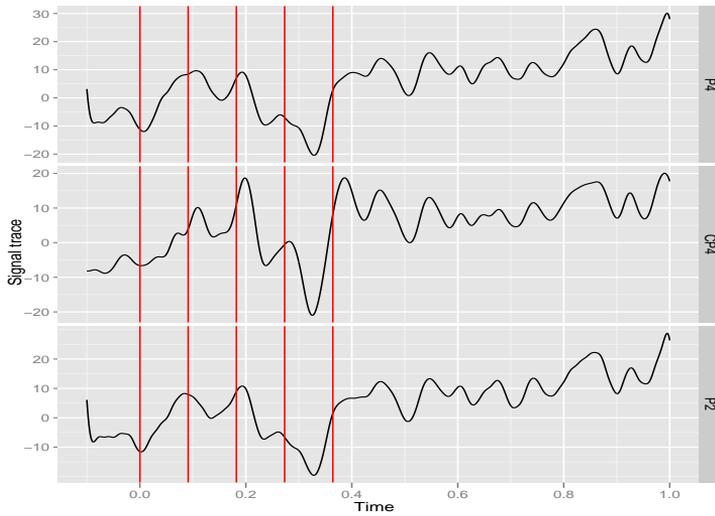


FIGURE 2. Signal of interest for channels P4, CP4 and P2 in a single replicate of a subject. The TMS pulses shown as vertical lines on the time axis are administered at intervals of approximately 0.09 s to this subject.

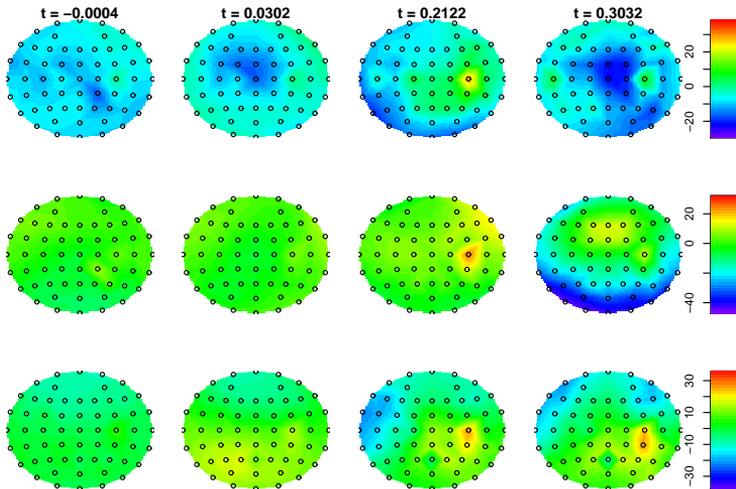


FIGURE 3. The spatial distribution of the signal of interest for three replicates of the same subject. Rows 1-3 show spatial brain maps for time $t = -0.0004$ (at an instant before the first TMS pulse), $t = 0.0302$ (30 milliseconds after the first pulse), $t = 0.2122$ (30 milliseconds after the third pulse) and $t = 0.3032$ (30 milliseconds after the fourth pulse) for three replicates of the same subject.

4 Discussion and Future Work

The decomposition of the EEG signal into smooth additive components captures the spike effect and mains current reasonably well. Entrainment of the remaining ‘brain’ signal to the alpha band frequencies is evident across channels especially after the third spike has been administered, once the data have been corrected for the artefacts. Future work will involve characterising the smooth ‘brain’ signal using a small number of parameters such as phase, amplitude and frequency. The exploratory maps of the brain show variation across several replicates of a subject. This will enable an investigation of the variability across channels, replicates and subjects with the aim of developing a spatiotemporal random effects model.

References

- Bowman, A.W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the Kernel approach with S-plus illustrations*. Oxford University Press.
- Bowman, A.W. and Azzalini, A. (2003). Computational aspects of non-parametric smoothing with illustrations from the sm library. *Computational Statistics and Data Analysis*, **42**, 545–560.
- Cracco, R.Q., Amassian, V.E., Maccabee, P.J. and Cracco, J.B. (1989). Comparison of human transcallosal responses evoked by magnetic coil and electrical stimulation. *Electroencephalogr Clin Neurophysiol*, **74**, 417–424.
- Hastie, T.J and Tibshirani, R.J (1990). *Generalized Additive Models*. Chapman & Hall/CRC.
- 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.
- Ventrucci, M., Miller, C., Gross, J., Schoffelen, J.M. and Bowman, A.W. (2011). Spatiotemporal smoothing of single trial MEG data, *Journal of Neuroscience Methods*, **200**:2 219-228.
- Wood, S.N. (2006). *Generalized Additive Models, An Introduction with R*. Chapman & Hall/CRC.

Bootstrap confidence intervals for the optimal maintenance time of a repairable system

Gustavo L. Gilardoni¹, Maristela D. Oliveira², Enrico A. Colosimo³

¹ Universidade de Brasilia, Brazil,

² Universidade Federal da Bahia, Salvador, Brazil

³ Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

E-mail for correspondence: gilardon@unb.br

Abstract: This paper discuss how to construct bootstrap confidence intervals for the optimal maintenance periodicity of a repairable system operating under a maintenance strategy that calls for complete preventive repair actions at pre-scheduled times and minimal repair actions whenever a failure occurs. The paper departs from the usual power-law-process parametric approach by using the constrained nonparametric maximum likelihood estimate of the intensity function to estimate the optimum preventive maintenance policy. The methodology is applied to a real data set concerning failure histories of a set of power transformers.

Keywords: Constrained maximum likelihood estimation; greatest convex minorant; minimal repair; Poisson process; total time on test.

1 Introduction

Consider a nonhomogeneous Poisson process (NHPP) having an increasing intensity $\lambda(t)$ and hence a convex mean function $\Lambda(t) = \int_0^t \lambda(u) du$, and let $N_i(t)$ be K independent realizations observed along possibly overlapping time intervals $0 \leq t \leq T_i$ ($1 \leq i \leq K$). This paper deals with the construction of bootstrap confidence intervals (CIs) for a functional $\tau = \tau[\lambda(\cdot)]$ based on the constrained nonparametric maximum likelihood estimate (NPMLE) of $\lambda(t)$. Its motivation lies on the problem of estimating the optimal periodicity τ of perfect preventive maintenance (PM) for a repairable system that is subject to minimal repair (MR) after each failure (cf. Barlow and Hunter, 1960 or Gilardoni and Colosimo, 2007). More precisely, suppose that under MR the system failures are modeled as an NHPP with an increasing intensity $\lambda(t)$ and let C_{PM} and C_{MR} be the fixed costs of the PM and MR actions. In order to minimize the system operational cost per unit of time, PMs should be performed at every τ units of time, where τ is the solution of

$$D(\tau) = \tau\lambda(\tau) - \Lambda(\tau) = C_{PM}/C_{MR}. \quad (1)$$

Since the intensity $\lambda(t)$ is typically unknown, in practice one obtains an estimate $\hat{\lambda}(t)$ based on the failure histories of one or more realizations of the system under consideration and use it to compute an estimate of τ by solving (1) with $\hat{\lambda}$ instead of λ . Gilardoni and Colosimo (2007) and Oliveira et al. (2011) consider respectively frequentist and bayesian estimates of τ using a power law process (PLP) parametric intensity $\lambda(t) = (\beta/\theta) (t/\theta)^{\beta-1}$. On the other hand, Gilardoni and Colosimo (2011) discuss a nonparametric approach using kernel estimates of $\lambda(t)$, although their approach has the drawback that it does not take into account the monotonicity restriction on $\lambda(t)$, without which (1) may not have a solution.

Suppose first that only one system is observed up to time T and let $N(t) = \sum_{j=1}^n I(t \geq t_j)$ be the number of failures up to time t , where $0 < t_1 < \dots < t_n < T$ are the observed failure times. Following Rigdon and Basu (2000), the log-likelihood is

$$\ell(\lambda) = \sum_{i=1}^n \log \lambda(t_i) - \int_0^T \lambda(u) du. \tag{2}$$

The unconstrained NPMLE of $\Lambda(t)$ is the step function $\tilde{\Lambda}(t) = N(t)$. Since its derivative is almost everywhere null, the unconstrained NPMLE of $\lambda(t)$ does not exist. However, depending on the type of restriction imposed to $\lambda(t)$, the constrained NPMLE may exist. Indeed, Boswell (1966) showed that the maximum of (2) among increasing intensities is attained for a right continuous step function $\hat{\lambda}(t)$ which has jumps at a subset of the t_i 's and such that $\hat{\lambda}(0) = 0$ and

$$\hat{\lambda}(t_j) = \max_{1 \leq h \leq j} \min_{j \leq k \leq n+1} \frac{k - h}{t_k - t_h}, \tag{3}$$

where we define $t_0 = 0$ and $t_{n+1} = T$. An alternative representation of the constrained NPMLE is that $\hat{\lambda}(t) = \hat{\Lambda}'(t+0)$, where $\hat{\Lambda}(t) = \sup\{f(t) : f \text{ is convex and } f(u) \leq \tilde{\Lambda}(u) \text{ for all } u\}$ is the GCM of $\tilde{\Lambda}(t)$ (see Figure 1a).

Suppose now that we observe K independent realizations of the same NHPP truncated at possibly different times T_1, \dots, T_K . Let the i -th realization be $N_i(t) = \sum_{j=1}^{n_i} I(t \geq t_{ij})$, where t_{ij} is the time of the j -th failure of the i -th system, and define the TTT transformation $R(t) = \sum_{i=1}^K \min(t, T_i)$ and its generalized inverse $R^{-1}(s) = \inf\{t : R(t) \geq s\}$. Gilardoni and Colosimo (2011) showed that in this case the log-likelihood can be written as $\ell(\lambda) = c + \sum_{i=1}^K \sum_{j=1}^{n_i} \log \lambda_S(s_{ij}) - \int_0^S \lambda_S(s) ds$, where $s_{ij} = R(t_{ij})$, $\lambda_S(s) = \lambda[R^{-1}(s)]$ and $S = R(\max\{T_1, \dots, T_K\}) = \sum_{i=1}^K T_i$. Comparing with (2) and noting that $\lambda(t)$ is increasing if and only if $\lambda_S(s)$ is too, one concludes that computation of the constrained NPMLE in the many realizations set up can be reduced to the single realization one discussed before. Essentially, one defines the superimposed process $N_S(s) =$

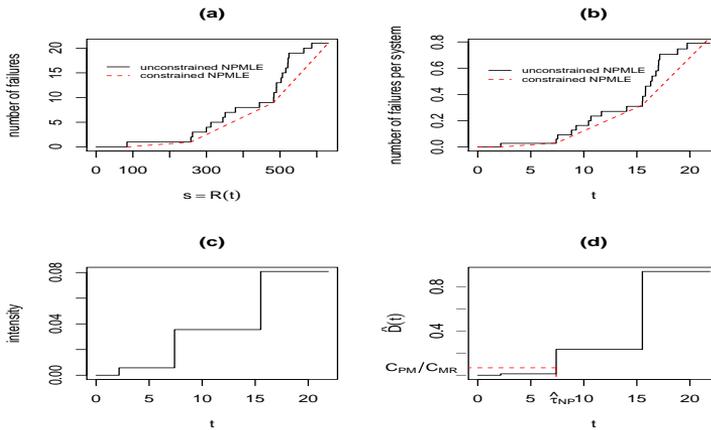


FIGURE 1. NPMLEs of (a) the mean function $\Lambda_S(s)$ for the single realization of the superimposed process in the TTT scale; (b) the mean function $\Lambda(t)$ of the original process; (c) the intensity $\lambda(t)$ in the original time scale and (d) $D(t)$ (cf. equation (1)) and optimal PM time.

$\sum_{i=1}^K N_i[R^{-1}(s)]$ and its intensity $\lambda_S(s) = \lambda[R^{-1}(s)]$, computes the constrained NPMLE $\hat{\lambda}_S(s)$ using (3) and then goes back to the original time scale by letting $\hat{\lambda}(t) = \hat{\lambda}_S[R(t)]$. This process is shown in figures 1a-1c for the power transformers data set.

The rest of the paper is organized as follows. Section 2 describes in detail several strategies used to generate bootstrap samples of τ . Section 3 shows an application of the methodology to a real data set consisting of the failure histories of 40 electrical power transformers (cf. Gilardoni and Colosimo, 2007, 2011 and Tsai et al., 2011). Some final remarks are given in Section 4.

2 Bootstrap Methods

We consider five different strategies to bootstrap the estimates of the optimal periodicity. The first one, denoted below by **boot.sys**, bootstraps the systems or realizations in the original time scale as if they were clustered data (cf. Field and Welsh, 2007). More precisely, we first sample with replacement K integers between 1 and K , say i_1^*, \dots, i_K^* . Based on the set of failure times $t_{i_h^*,j}$ and truncations $T_{i_h^*}$, we compute the TTT transform and the estimates $\hat{\lambda}_S^*(s)$ and $\hat{\lambda}^*(t)$ of the intensities in the TTT and original time scale and, finally, we use $\hat{\lambda}^*(t)$ to compute a bootstrap estimate $\hat{\tau}^*$ of the optimal periodicity.

All other strategies resample failure times from the superimposed process in the TTT time scale. The second one, denoted by **boot+n+t**, generates

TTT failure times $s_1^*, \dots, s_{n^*}^*$ by simulating an NHPP whose intensity function is the constrained NPMLE $\hat{\lambda}_S(s)$ computed with the original data set. We remember that, to generate an NHPP with intensity $\hat{\lambda}_S(s)$ and mean function $\hat{\Lambda}_S(s)$, one first generates $n^* \sim \text{Poisson}(\hat{\Lambda}_S(S))$ and then obtains the TTT failure times as the order statistics of a size n^* iid random sample from the cdf $\hat{\Lambda}_S(s)/\hat{\Lambda}_S(S)$, $0 < s < S$, where $S = \sum_{i=1}^K T_i$ is the truncation time of the superimposed process. Once this failure times are obtained, we use (3) to compute $\hat{\lambda}_S^*(s)$, then the inverse TTT transform to get $\hat{\lambda}^*(t)$ and finally (1) to get a resample $\hat{\tau}^*$ of the estimate of the optimal periodicity. The last three strategies are variants of the one just described. More precisely, the third one, **boot-n+t**, proceeds exactly as **boot+n+t** but for the fact that n^* is kept fixed equal to n , the number of failures in the original data set. The fourth one, **boot+n-t**, differs from **boot+n+t** only for the fact that, instead of resampling the TTT failure times $s_1^*, \dots, s_{n^*}^*$ from the cdf $\hat{\Lambda}(s)/\hat{\Lambda}(S)$, they are the order statistics of a size n^* random sample with replacement from the set of original TTT failure times $\{s_{ij} : 1 \leq i \leq K, 1 \leq j \leq n_i\}$. Finally, the last strategy **boot-n-t** both kept $n^* = n$ fixed and resamples from the original TTT failure times.

A Monte Carlo simulation was run to compare the performance of these strategies. The main finding of the simulation was that the two strategies which actually resample the failure times from the cdf $\hat{\Lambda}(s)/\hat{\Lambda}(S)$ ($0 < s < S$), namely **boot+n+t** and **boot-n+t**, behaved similarly and were far superior than the other three strategies.

3 Application

Here we consider the data set consisting of $n = 21$ failure times for $K = 40$ power transformers introduced in Gilardoni and Colosimo (2007). The cost ratio was assumed to be $C_{PM}/C_{RM} = 1/15$. NPMLEs of the mean function $\Lambda(t)$ and the intensity $\lambda(t)$ are shown in Figures 1b and 1c. The corresponding estimates $\hat{D}(t)$ and $\hat{\tau}_{NP} = 7.396$ are shown in Figure 1d. Besides the PLP model we also adjusted a loglinear intensity $\lambda(t) = \exp\{\alpha + \beta t\}$ and a bounded intensity $\lambda(t) = \alpha[1 - 1/\sqrt{1 + t\beta}]$ by numerically maximizing the corresponding likelihood functions. The MLEs were $\hat{\theta} = 24.366$, $\hat{\beta} = 1.995$ and $\hat{\tau}_{PLP} = 6.286$ for the PLP, $\hat{\alpha} = -4.505$, $\hat{\beta} = 0.094$ and $\hat{\tau}_{loglin} = 8.586$ for the loglinear and $\hat{\alpha} = 0.561$, $\hat{\beta} = 73.138$ and $\hat{\tau}_{bounded} = 6.140$ for the bounded intensity model. Figure 2a shows the unconstrained NPMLE estimate of $\Lambda(t)$ along with the three parametric estimates and suggests that the PLP and bounded intensity models seem to adjust better to the data than the loglinear model. Figure (2)b shows the parametric estimates of $\Lambda(t)$ for the observed failure times for the PLP and bounded intensity models. Since the points lie almost exactly on the line $y = x$, it suggests that for the range of values considered in this data set the two models are to a large extent indistinguishable.

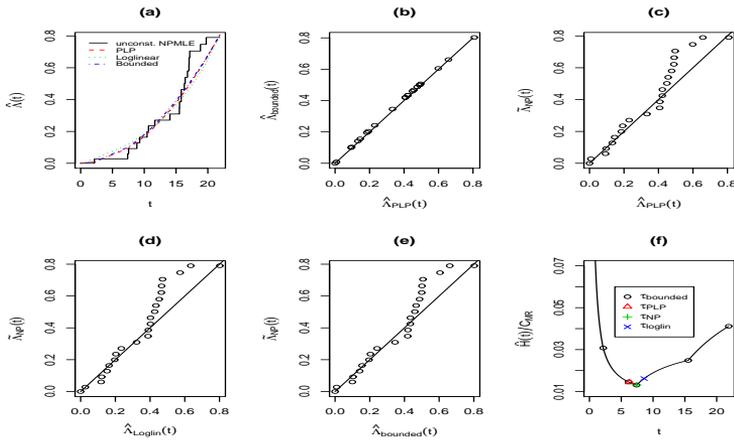


FIGURE 2. (a) to (e): Estimates of $\Lambda(t)$ for observed failure times: (a) NPMLE, PLP, loglinear and bounded model estimates against time, (b) PLP against bounded model estimates, (c)-(e) parametric against NPMLE estimates; (f) non-parametric estimate of $H(t)/C_{MR}$, the mean cost per unit of time measured in C_{MR} monetary units.

The maximized log-likelihoods were respectively $\hat{\ell}_{PLP} = -87.671$, $\hat{\ell}_{loglin} = -88.913$ and $\hat{\ell}_{bounded} = -87.639$. Since all three models have exactly two parameters, the bayesian information criterion (BIC) to compare them differs from the log-likelihoods by the same constant. Hence, assuming equal prior probabilities for the three models and diffuse priors for the parameters, we can interpret the likelihoods to be approximately proportional to the posterior probabilities of the models. These would give probabilities 0.431, 0.124 and 0.445 respectively for the PLP, loglinear and bounded models. Hence, even if the PLP and bounded intensity models adjust better, it is not possible to totally dismiss the loglinear model. Moreover, the diagnostic plot of the parametric estimates $\hat{\Lambda}_{PLP}(t)$ against the unconstrained NPMLE estimate, shown in Figures 2c-2e, leave some doubts about the quality of the three parametric fits.

The parametric 90% CI based on the Delta Method was $5.203 < \tau < 7.593$, which is based on an estimated error equal to 0.115 for $\log \hat{\tau}_{PLP}$. On the other hand, the 90% CI using the **boot-n+t** was $5.235 < \tau < 10.894$. We note that the difference between the two CIs seems to be almost exclusively along the upper tail of the sampling distributions, in the sense that the two lower limits are quite similar. Given the small number of failures and the previous observations about the fit of the three parametric models, we feel that the parametric CI may be misleading in the sense that it may fail to capture all the uncertainty in the data.

4 Final Remarks

This paper presents a bootstrap nonparametric approach to construct CIs for functionals of the intensity of an NHPP subject to a monotonicity constraint. When several realizations of the NHPP are observed, our results suggest that one should first pool or superimpose the realizations using the TTT time transform and then bootstrap the unique realization of this superimposed process. In a sense, this is consistent with the fact that, in the context of kernel estimation of the intensity function, there is substantial evidence that one should pool first and estimate second and not *viceversa* (see, for instance, Chiang et al., 2005 or Gilardoni and Colosimo, 2011).

Acknowledgments: GLG was financed by UnB/DPP and Capes grants.

References

- Barlow, R. and L. Hunter (1960). Optimum preventive maintenance policies. *Operations Research*, **8**, 90–100.
- Boswell, M. T. (1966). Estimating and testing trend in a stochastic process of Poisson type. *The Annals of Mathematical Statistics*, **37**, 1564–1573.
- Chiang, C. T., M. C. Wang, and C. Y. Huang (2005). Kernel estimation of rate function for recurrent event data. *Scandinavian Journal of Statistics*, **32**, 77–91.
- Field, C. A. and A. H. Welsh (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society B*, **69**, 369–390.
- 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.
- Oliveira, M. D. de, E. A. Colosimo and G. L. Gilardoni (2012). Bayesian inference for power law processes with applications in repairable systems. *Journal of Statistical Planning and Inference*, **142**, 1151–1160.
- Rigdon, S. E. and A. P. Basu (2000). *Statistical Methods For the Reliability of Repairable Systems*. New York: John Wiley.
- Tsai, T. R., P. H. Liu, and Y. L. Lio (2011). Optimal maintenance time for imperfect maintenance actions on repairable product. *Computers & Industrial Engineering*, **60**, 744–749.

Water monitoring sites discrimination using clustering water variables time series data and main latent factors identification

A. Manuela Gonçalves¹, Marco Costa²

¹ Department of Mathematics and Applications, University of Minho
CMAT-Center of Mathematics of Minho University, Portugal
mneves@math.uminho.pt

² Higher School of Technology and Management of Águeda-University of Aveiro
CMAF-UL, Portugal
marco@ua.pt

E-mail for correspondence: mneves@math.uminho.pt

Abstract: Application of multivariate methodologies for the evaluation and interpretation of space-time variations in an environmental monitoring data set from an hydrological basin is presented in this study. The results obtained allowed detecting natural clusters of monitoring sites with similar water quality type and identifying important discriminant variables for each cluster. Furthermore, this type of analysis allows reducing the number of models in the variables future modelling process.

Keywords: surface water quality; discrimination; cluster analysis; principal components analysis; latent factors identification.

1 Introduction

Both anthropogenic pressures and natural processes account for degradation in surface water quality. Surface water quality monitoring has as its main objective the characterization of water resources, as well as the monitoring of its space-time evolution in order to achieve an appropriate administration. Multivariate statistical analyses have become widely applied in water quality assessment and sources apportionment of water over the last years (Shrestha and Kazama, 2007). In this study, multivariate statistical methods as Cluster Analysis (CA) and Principal Component Analysis (PCA) will be applied with the aim of evaluating and interpreting the space-time variations of surface water quality of any given hydrological basin. These methodologies allowed a reduction in the dimensionality of the large data set. In particular, CA has allowed the identification of homogeneous regions (i.e., groups of monitoring sites with similar characteristics in terms of quality variables), thus reducing the large number of

monitoring sites into a small number of homogeneous groups, and PCA delineated a few latent factors indicating the variables responsible for large variations in water quality. This global reduction will be very useful in the future modelling process (Costa and Gonçalves, 2012).

2 Methods

2.1 Study area and data description

The Ave hydrological basin is located in Northwest Portugal and its main adjacent streams are the rivers Este, Selho, and Vizela. The surface water of River Ave has high pollution levels and the water quality measurements failed to comply with the objectives of minimum quality for surface waters. Since 1988 the Portuguese government has been monitoring the surface water quality monthly along the river Ave and its main adjacent streams by means of 20 monitoring sites. It was analysed a data set of 11 quality variables (ten physicochemical ones and one microbiological) relevant to the evaluation of surface water quality of rivers subjected to discharges of industrial effluents: pH, dissolved oxygen (DO), chemical oxygen demand (COD), biochemical oxygen demand (BOD5), total suspended solids (TSS), oxygen demand (OD), ammonical nitrogen (NH4-N), nitrate nitrogen (NO3-N), conductivity (COND), water temperature (WT), and faecal coliforms (FC). This data set consists of monthly values of water quality variables measured between 1988 and 2006.

2.2 Cluster analysis

We performed cluster analysis for the grouping of monitoring sites with similar water quality characteristics. In this study, hierarchical agglomerative CA was performed on the normalized data set. For the hierarchical agglomerative CA procedure purposes it will be considered the measure of dissimilarity proposed in Gonçalves and Alpuim (2011). The main problem is that for all locations and variables there are not observations for all months under study. Therefore, let us consider x_{ikt} the value of the quality variable k , measured at location i , in time t . Let P_t be the set of all quality variables measured at the same time t , in both sites i and j . The Euclidean distance at this time instant between locations i and j is given by the expression

$$dist_{ij}(t) = \left[\sum_{k \in P_t} (x_{ikt} - x_{jkt})^2 \right]^{\frac{1}{2}} .$$

This dissimilarity measure corresponds to the average of this distance over all months t , where there is at least one quality variable with measurements in the two sites, that is

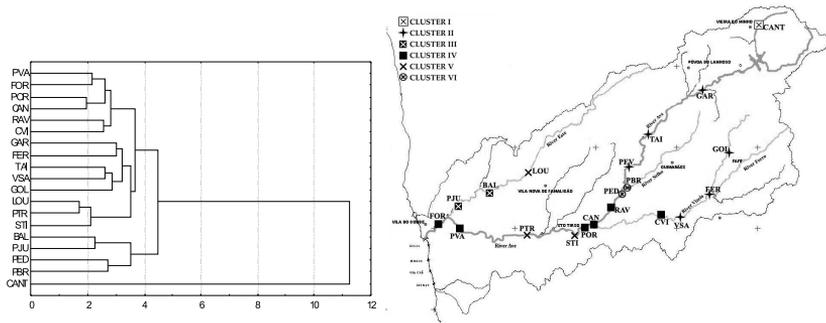


FIGURE 1. Dendrogram showing clustering of sampling sites according to surface water quality variables without PEV, and also the spatial representation of River Ave’s clusters.

$$d_{ij} = \frac{1}{\#M_{ij}} \sum_{t \in M_{ij}} \left[\sum_{t \in P_t} (x_{ikt} - x_{jkt})^2 \right]^{\frac{1}{2}}, \quad i, j = 1, \dots, 20,$$

where M_{ij} is the set of all months with at least one variable measured in both sites i and j .

The final result of the obtained groups was discussed according to the complete linkage method. This method was used because, in this case, it rendered well-defined clusters and according to the reality of the river basin. According to the dendrogram analysis six well-differenced clusters are obtained. The obtained dendrogram of the monitoring sites and the clusters geographical representation are shown in Figure 1. Taking into consideration the quality variables averages within each cluster, they are classified into five categories according to their pollution levels established by the National Department for Pollution Control (NDPC): "Without Pollution (WP)", "Moderately Polluted (MP)", "Polluted (P)", "Very Polluted (VP)" and "Extremely Polluted (EP)". Each cluster is classified based on the NDPC criteria, which are determined according to the worst value of a given variable observed in the cluster. The resulting classifications of the six clusters confirm the previous knowledge that effluent discharge varies according to natural and geographical/economical reasons. Cluster I (consisting of just one site-CANT) may be characterized as Without Pollution and corresponds to the source of River Ave. Then there is a set of locations which can be defined as Moderately Polluted (Cluster II, composed by GAR, TAI, PEV, GOL, FER and VSA), including 6 sites in both adjacent streams Este and Vizela. These stations receive pollution mostly from domestic wastewater and from agricultural and manure discharges. Cluster III, classified as Polluted 1 (P1), is composed by BAL and PJU located in River Este, where the quantity of nitrate-nitrogen has been relatively high.

In Cluster IV, six of the monitoring sites (RAV, CAN, POR, PVA, FOR, and CVI) are situated in the River Ave and only one, CVI, is located at the most downstream site of River Vizela. This is a densely populated region, with high industrial productivity, and here the River Ave receives similarly polluted waters (Polluted 2 (P2)) from its adjacent rivers. In Cluster V, with 3 sites, LOU (in river Este), STI and PTR (located near the most polluted area of the basin), there is a growing urban population and a high concentration of industrial activity. Cluster V was classified as Very Polluted. Finally, the most polluted cluster, Cluster VI (EP), consists of two stations, PBR and PED, located near the mouth of the Selho tributary and represents a highly polluted area. These sites receive pollution from domestic wastewater and industrial effluents located in city areas.

2.3 Principal component analysis

PCA is designed to transform the original variables into new, uncorrelated variables, called the principal components (PCs), which are linear combinations of the original variables (see Barnett, 198). PCA allows us to explain and evaluate the correlation structure between observed variables in water quality sampling stations and to identify relevant factors. The PCA technique is separately applied to the homogeneous groups of water monitoring sites (6 clusters), as obtained in the first clustering procedure, by taking into account all 11 water quality variables. The Kaiser-Meyer-Olkin (KMO) statistics and Bartlett's test were performed in order to examine the data suitability for PCA (KMO= 0.85). Spearman rank-order correlations were used to study the correlation structure between variables in order to account for non-normal distribution of water quality variables. A modified PCA was separately performed on the raw data sets (11 variables) for the six different regions (clusters) WP, MP, P1, P2, VP and EP, as delineated by CA techniques, in order to compare the compositional pattern among the analysed water monitoring sites and to identify the factors influencing each one. To further reduce the contribution of variables with minor significance, the PCs were subjected to varimax rotation (raw) generating varifactors (VFs). The PCA of the six data sets yielded four PCs for the WP and MP sites, three PCs for the P1, P2, VP and EP sites with eigenvalues > 1, explaining 69.86, 63.52, 69.91, 69.31, 65.66 and 73.12 of the total variance in the respective water quality data sets. Corresponding VFs, variable loadings and explaining variance are obtained for 6 clusters. Here we will only present the results for Cluster VI (EP), in Table 1, and then their respective interpretation.

Concerning the data set pertaining to EP sites, PBR and PED, among the three varifactors kept in the application of ACP, Varifactor 1 explains 43.66 percent of the total variance, has strong negative loadings on the variables DO and nitrate-nitrogen and also strong positive loadings on ammonical nitrogen and temperature variables. This varifactor contains the variables

TABLE 1. Loadings of experimental variables (11) on the first three rotated PCs for EP sites data set. Bold values indicate strong loadings.

Variables	Varifactor 1	Varifactor 2	Varifactor 3
BOD5	0.235	0.778	0.222
COD	0.251	0.864	0.063
TSS	-0.303	0.751	-0.260
DO	-0.885	-0.262	-0.093
OD	0.279	0.905	0.117
NH4-N	0.806	0.086	0.092
NO3-N	-0.776	-0.171	0.184
FC	0.379	0.273	0.301
COND	0.560	0.588	0.290
WT	0.885	0.019	-0.079
pH	-0.121	0.064	0.913
Eigenvalue	4.803	2.133	1.108
% Variance explained	43.663	19.387	10.070
Cumulative % variance	43.663	63.050	73.119

most related to pollution of anthropogenic origin. The same happens with Varifactor 2 (an organic factor), which explains 19.38 % of the total variance and presents strong positive loadings on BOD5, COD, TSS, OD and COND (representing influences from domestic wastewater and industrial effluents). Varifactor 3 explains 10.07 % of variance, with strong positive loadings on pH, and presents the influence of this variable on the chemical processes in EP waters. The obtained latent multifactors, with hydrochemical meaning, indicate that the responsible variables for the variation of the basin's water quality are mainly related with effluent discharges of anthropogenic origin (agricultural and industrial origin) along the River Ave and its tributary streams. Only in areas WP or MP do latent factors represent the variability inherent to the natural climatic seasonality and the variability associated to the basin's geomorphological characteristics, both of which naturally influence the hydrochemistry of the river's surface water. The obtained varifactors indicate that the quality variables responsible for water quality variations are mainly related to discharge and temperature (natural origin), nutrients and organic pollution in relatively less polluted areas, pollution by organic matter and nutrients from anthropogenic sources (mainly as discharges of industrial and municipal wastewater), and manure affecting the quality and hydrochemistry of river water in highly polluted areas around the basin.

3 Conclusions

Hierarchical CA grouped 20 monitoring sites into six clusters of similar water quality characteristics and, based on the obtained information, it is possible to design a future, optimal spatial sampling strategy which could reduce the number of sampling monitoring sites and associated costs. The results of CA confirm the expected behaviour of the temporal/spatial dynamics of pollutants concentration (along the river and its main streams), thus allowing to reduce the large number of monitoring sites into a small number of homogeneous groups.

The ACP analysis indicates that clusters have distinct factors/sources responsible for variations in River Ave's water quality (identifying environmental, social and industrial aspects which influence water quality variations) and identifies important latent variables for each cluster. A statistical modelling procedure will be applied to a set of water monitoring sites grouped in homogeneous clusters (like in Costa and Gonçalves, 2011).

Acknowledgments: This research was partially financed by FEDER Funds through "Programa Operacional Factores de Competitividade - COMPETE" and by Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia", within the Project Est-C/MAT/UI0013/2011. Marco Costa was partially supported by Fundação para a Ciência e a Tecnologia, PEst OE/MAT/UI0209/2011.

References

- Barnett, V. (1981). *Interpreting Multivariate Data*. Sheffield: John Wiley & Sons.
- 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 .
- Costa, M. and Gonçalves, A.M. (2012). Combining Statistical Methodologies in Water Quality Monitoring in a Hydrological Basin - Space and Time Approaches. Chapter in: *Water Quality Monitoring and Assessment*, ISBN 978-953-51-0486-5, Intech Publisher.
- Gonçalves, A.M. and Alpuim, T. (2011). Water quality monitoring using cluster analysis and linear models. *Environmetrics*, **22**, 933–945 .
- Shrestha, S. and Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling and Software*, **22**, 464–475 .

Simulation-based D_B -optimal designs: Conception and implementation issues

Markus Hainy¹, Werner G. Müller¹, Helga Wagner¹

¹ Department of Applied Statistics, Johannes Kepler University Linz, Austria

E-mail for correspondence: `werner.mueller@jku.at`

Abstract: We demonstrate which issues arise when a simulation-based design strategy, where the interest is in maximizing an integral expression representing expected utility with respect to the design variables, is implemented in the framework of Bayesian D -optimal design. We provide suggestions for solutions and exemplary computations for a second order polynomial regression model.

Keywords: MCMC; D -optimality; Bayesian design.

1 Concept

1.1 Model: We apply the standard linear regression model, so we assume that

$$p(y|\theta, X) \sim \mathcal{N}(X\theta, \sigma^2 I_N).$$

That is, the expected value of the dependent variable is a linear combination of the parameter values $\theta \in \Theta \subseteq \mathbb{R}^k$ and depends on the design through the design matrix $X = (f(x_1), \dots, f(x_N))^T$, where $f(\cdot)$ is a k -dimensional function of the design variables $x_i \in [-1, 1]$. The N observations are assumed to be normally distributed, independent, and homoscedastic with known variance σ^2 . Furthermore, the parameters θ follow the prior normal distribution

$$p(\theta) \sim \mathcal{N}(\theta_0, \sigma^2 R^{-1}).$$

Define $M = X^T X$. Then the posterior distribution of the parameters θ is

$$p(\theta|y, X) \sim \mathcal{N}((M + R)^{-1}(X^T y + R\theta_0), \sigma^2(M + R)^{-1}).$$

We take $u(y, \theta, X) = \log\left(\frac{p(\theta|y, X)}{p(\theta)}\right)$ as our utility function, so that the expected utility for a specific design X is the expected gain in Shannon information (see Chaloner and Verdinelli, 1995):

$$\begin{aligned} U(X) &= \int u(y, \theta, X) p(y, \theta|X) d\theta dy \\ &= \int \log\left(\frac{p(\theta|y, X)}{p(\theta)}\right) p(y|\theta, X) p(\theta) d\theta dy. \end{aligned}$$

Since $\int \log(p(\theta))p(y|\theta, X)p(\theta)d\theta dy = \int \log(p(\theta))p(\theta)(\int p(y|\theta, X)dy)d\theta$ does not depend on X , it is sufficient to compute

$$U^*(X) = \int u^*(y, \theta, X)p(y, \theta|X)d\theta dy = \int \log(p(\theta|y, X))p(y|\theta, X)p(\theta)d\theta dy.$$

For our particular model, the integral can be computed analytically. It is

$$U^*(X) = -\frac{k}{2} \log(2\pi) - \frac{k}{2} + \frac{1}{2} \log \det(\sigma^{-2}(M + R)),$$

which has the same maximum as the criterion for D_B optimality, $\Psi(X) = \det(M + R)$ (cf. Atkinson et al., 2007). Note that the D_B -optimal design does neither depend on σ^2 nor on the prior mean θ_0 .

1.2 Simulation method: We use the simple Metropolis Hastings (MH) Markov chain Monte Carlo (MCMC) scheme proposed by Müller (1999) to create a Markov chain for the stationary distribution

$$h(y, \theta, X) \propto u(y, \theta, X)p(y, \theta|X) = u(y, \theta, X)p(y|\theta, X)p(\theta).$$

The optimal design can be read off as the mode of this sampled marginal distribution of X . However, the usual aim of MCMC methods is to generate a sample from some distribution and not to find an optimum. Therefore, this scheme will produce proper draws from the distribution $h(y, \theta, X)$, but it may be very hard to identify the maximum of the pdf from these draws, especially in multi-dimensional problems. In order to alleviate this problem, Müller (1999) proposes an extension similar to simulated annealing. The MCMC scheme is implemented for the stationary distribution

$$h_J(y_1, \dots, y_J, \theta_1, \dots, \theta_J, X) \propto \prod_{j=1}^J u(y_j, \theta_j, X)p(y_j, \theta_j|X),$$

which leads to the expected utility

$$U^J(X) = \int \prod_{j=1}^J u(y_j, \theta_j, X)p(y_j, \theta_j|X)d\theta_1 \dots d\theta_J dy_1 \dots dy_J,$$

the J th power of the expected utility $U(X)$. Thus, the marginal distribution of X becomes more and more peaked as J increases and turns into a point mass at the optimal design for $J \rightarrow \infty$. J can be regarded as the inverse of the annealing temperature. It is also possible to construct an inhomogeneous Markov chain where J is increased according to some annealing schedule until all draws of X are within a certain prespecified range, see Müller et al. (2004). They suggest a logarithmic annealing schedule, whereas Amzal et al. (2006) favor a linear schedule. Amzal et al. (2006)

incorporate the idea of simulated annealing into a system of interacting particles. Their approach provides a way to concentrate the draws around the modes of the distribution even tighter. Another improvement involving several chains with different annealing temperatures is put forward by Ruiz-Cárdenas et al. (2011). Their evolutionary MCMC algorithm uses the genetic operators *crossover* and *exchange* to swap elements between the chains. This procedure prevents chains with low annealing temperatures to get trapped in local modes, which is also a relevant issue in our setting.

2 Implementation issues

2.1 Utility function restrictions: The utility function $u(d, \theta, y)$ has to be nonnegative and bounded. This precludes the application of this algorithm to many problems, among them the computation of Bayesian D -optimal designs. Solution options:

- Add a positive constant if there is a lower bound to the utility values.
- Use the transformed utility $u_2(y, \theta, X) = -1/u^*(y, \theta, X)$. This can only work if all utility values are negative. If $u^*(y, \theta, X) = \log p(\theta|y, X) < 0 \forall y, \theta, X$ (i.e. $p(\theta|y, X) < 1$), then the transformed utility $u_2(y, \theta, X) = -1/(\log p(\theta|y, X)) \geq 0 \forall y, \theta, X$. It can be shown that in our setting this condition is met if $\sigma^2 > (\det(M + R))^{1/k}/(2\pi)$. The optimal design does not depend on σ^2 , so we might just choose σ^2 large enough to avoid any problems. The transformed utility function has the same peaks as the original utility function. However, its shape may be quite distorted, and so the expected utility integral

$$U_2(X) = \int u_2(y, \theta, X)p(y, \theta|X)d\theta dy$$

may lead to a completely different optimal solution for X .

- Use $u(y, \theta, X) = \log(p(\theta|y, X)) - \log(p(\theta))$ as utility function. This should be positive for most proposed values. One option is to set the utility to $\max(\log(p(\theta|y, X)) - \log(p(\theta)), 0)$, or one might repeat drawing from the proposal distribution until the utility value is positive. If there are not too many draws with negative utility, both strategies bring about only minor distortions of the expected utility function.

2.2 MH proposal distribution: Choose a suitable proposal density to account for restrictions on the design variables: $x_i \in [-1, 1]$. Suggestions:

- Uniform distribution or Dirichlet distribution;
- Random walk using a truncated distribution for the increments (e.g. a truncated normal distribution);

- Discretize design set and use a discrete proposal distribution, e.g. a discrete uniform distribution;
- Adjust scaling of the proposal distribution if Markov chain moves close to the boundaries;
- Just set x_i to -1 or 1 if the proposed draw is out of the boundaries. This leads to discontinuities at the boundaries, therefore it may be difficult to derive the correct proposal distribution for the MH acceptance probability formula. An advantage would be that if the boundaries are indeed optimal, then many draws are made at the exact optimum values and not just close to them. This can also be achieved with a discrete design set including the boundaries (see above).

2.3 Mode detection and runtime: It is very difficult to detect the peak(s) of $h(y, \theta, X)$ for a multi-dimensional design space. $h(y, \theta, X)$ is usually rather flat and there are only marginally more draws near the optimal design than at the rest of the design space. The simulated-annealing-like extension discussed earlier can alleviate that problem to a great extent, but it also has some drawbacks. One needs to take J draws of y and θ in each simulation step, which amounts to a considerable increase in computing time. If the utility function is too peaked, it is possible to get trapped in one mode (if there are several).

2.4 Identifiability: If the N units are symmetric, it may happen that they “switch” positions (e.g. first x_1 is at a high level and x_2 is at a low level, then they switch). This makes it also very hard to deduce optimal designs because there are usually multiple equivalent modes. Due to the curse of dimensionality there may not be enough draws to clearly identify any of these multiple modes.

3 Example

We examine the performance of the basic simulation-based optimal design algorithm on our toy example. This little exercise is meant to assess the usefulness and limitations of the algorithm. It is not difficult to obtain the exact solution for our example, making it easy to check the MCMC-results. The following setting was used: the predictor is a polynomial of order two in one factor, i.e.

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{pmatrix}.$$

The continuous optimal design for this problem puts equal weights of $1/3$ on the three design points -1 , 0 , and 1 , see Atkinson et al. (2007). Likewise, if the number of trials of an exact design is divisible by three, then

at the optimal design 1/3 of the trials are set to -1 , 0 , and 1 , respectively. Therefore, if the prior information matrix R is chosen to represent prior information equivalent to one trial taken at the design point -1 , i.e. $R = f(-1)f^T(-1) = (1, -1, 1)^T(1, -1, 1)$, and we have $N = 2$ trials, it is optimal to set one trial to 0 and the other trial to 1 .

As proposal distribution for the Metropolis-Hastings sampler we used a uniform distribution over $[-1, 1]^2$. This sampler is very easy to implement. The uniform proposal distribution visits every region of the design space with equal probability irrespective of the current position of the chain, which prevents the sampler from getting stuck in local modes. On the downside, if the utility function is very peaked, rejection rates will be high. Nevertheless, for our purposes this simple proposal distribution is sufficient. The plots of the results for $J = 1$ (i.e., the simple simulation-based optimal design algorithm) and for $J = 100$ can be found below. This example is meant to illustrate the effect of tightening the expected utility function. We took a sample of 500000 draws for $J = 1$ and a sample of 50000 draws for $J = 100$. In the case of $J = 1$, the plot showing all draws is omitted, because it is almost impossible to detect any clusters due to the rather flat utility function: the whole square is filled up almost uniformly. We therefore plot only the highest one percent of the draws with respect to $u(y, \theta, X)p(y, \theta|X)$. These draws are supposed to be in regions where the expected utility integral is also high. The left plot in Figure 1 shows these draws of x_1 and x_2 . One can see that they cluster around the true optimum values $(x_1, x_2) = (0, 1)$ and $(x_1, x_2) = (1, 0)$ (bimodal). If $J = 100$, the draws cluster much more tightly around the optimum values, as can be seen from the right plot in Figure 1.

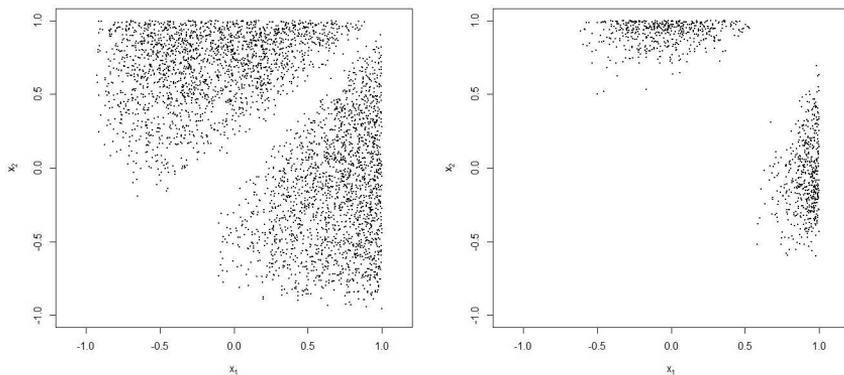


FIGURE 1. Left: $J = 1$, highest 1 % of draws for $u(y, \theta, X)p(y, \theta|X)$. Right: $J = 100$, all draws.

4 Outlook

It is evident that our rather cumbersome procedure is not suitable for simple D_B -optimal designs in linear regression, for which much more efficient methods exist (see for example Atkinson et al., 2007). However it will certainly be useful in more complex situations, such as nonlinear regression with little prior information on the parameters, compound design criteria, models with correlated errors or adaptive designs. Also, as Solonen et al. (2011) point out, as a by-product the MCMC approach delivers a kind of map of designs with similarly high efficiencies. In any case, knowing the performance of the investigated procedure for the simple D_B setting will always serve as a useful benchmark.

References

- Amzal, B., Bois, F.Y., Parent, E., and Robert, C.P. (2006). Bayesian-Optimal Design via Interacting Particle Systems. *Journal of the American Statistical Association*, **101**, pp. 773–785.
- Atkinson, A., Donev, A., and Tobias, R. (2007). *Optimum Experimental Designs, with SAS*. New York: Oxford University Press.
- Chaloner, K. and Verdinelli, I. (1995). Bayesian Experimental Design: A Review. *Statistical Science*, **10**, pp. 273–304.
- Müller, P. (1999). Simulation-Based Optimal Design. In: *Bayesian Statistics 6*, Bernardo, J.M., Berger, J.O., Dawid, P., and Smith, A.F.M. (Eds.), New York: Oxford University Press, pp. 459–474.
- Müller, P., Sansó, B., and De Iorio, M. (2004). Optimal Bayesian Design by Inhomogeneous Markov Chain Simulation. *Journal of the American Statistical Association*, **99**, pp. 788–798.
- Ruiz-Cárdenas, R., Ferreira, M.A.R., and Schmidt, A.M. (2011). Evolutionary Markov Chain Monte Carlo Algorithms for Optimal Monitoring Network Designs. *Statistical Methodology*, **9**, pp. 185–194.
- Solonen, A., Haario, H., and Laine, M. (2012). Simulation-Based Optimal Design Using a Response Variance Criterion. *Journal Of Computational And Graphical Statistics*, **21**, pp. 234–252.

Conditional correlation modelling: Simulation study

Radek Hendrych¹

¹ Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

E-mail for correspondence: hendrych@karlin.mff.cuni.cz

Abstract: This contribution deals with a simulation analysis of time varying correlations. Several different approaches are assumed: the exponential smoothing, the moving averages, the technique based on the Cholesky decomposition with various types of estimators and Engle's dynamic conditional correlations models.

Keywords: conditional correlations; DCC models; simulations.

1 Introduction

Consider a stochastic vector process $\{\mathbf{X}_t\}_{t \in \mathbb{Z}}$ of the dimension $(n \times 1)$. Denote \mathcal{F}_{t-1} the information set (σ -algebra) generated by observed multivariate time series $\{\mathbf{X}_t\}$ up to and including time $t - 1$.

Assume the following model of the conditional covariance structure:

$$\mathbf{X}_t = \mathbf{H}_t^{1/2} \cdot \boldsymbol{\varepsilon}_t \quad (1)$$

given the information set \mathcal{F}_{t-1} , where $\mathbf{H}_t^{1/2}$ is a $(n \times n)$ positive definite matrix of (unknown) parameters and $\{\boldsymbol{\varepsilon}_t\}$ is an i.i.d. vector process independent of $\{\mathbf{X}_t\}$ such that $\mathbf{E}\boldsymbol{\varepsilon}_t = \mathbf{0}$ and $\mathbf{E}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' = \mathbf{I}_{n \times n}$. It can be easily viewed that

$$\begin{aligned} \mathbf{E}(\mathbf{X}_t | \mathcal{F}_{t-1}) &= \mathbf{H}_t^{1/2} \mathbf{E}(\boldsymbol{\varepsilon}_t | \mathcal{F}_{t-1}) = \mathbf{0}, \\ \text{var}(\mathbf{X}_t | \mathcal{F}_{t-1}) &= \mathbf{H}_t^{1/2} \mathbf{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' | \mathcal{F}_{t-1}) (\mathbf{H}_t^{1/2})' = \mathbf{H}_t. \end{aligned}$$

Thus, it is evident that $\mathbf{H}_t^{1/2}$ is any $(n \times n)$ positive definite parameter matrix such that \mathbf{H}_t is the conditional covariance matrix of \mathbf{X}_t , e.g. $\mathbf{H}_t^{1/2}$ may be obtained by the Cholesky decomposition of \mathbf{H}_t . From the theoretical point of view, the conditional correlation matrix \mathbf{R}_t can be acquired by obvious adjustment of the conditional covariance matrix, i.e. $\mathbf{R}_t = \text{diag}\{\mathbf{H}_t\}^{-1/2} \mathbf{H}_t \text{diag}\{\mathbf{H}_t\}^{-1/2}$. To catch up the structure of the (individual) diagonal elements of \mathbf{H}_t , i.e. the conditional variance of the elements \mathbf{X}_t , a wide range of possibilities exists, e.g. the univariate ARCH

and GARCH models, several filtering techniques or methods based on non-parametric approaches, see Engle (2009) for more details. For this reason, in the case of this contribution, the simplified situation when the matrix \mathbf{H}_t has unit diagonal elements is considered, i.e. $\mathbf{H}_t = \mathbf{R}_t$.

2 Different Ways of Conditional Correlation Modelling

The following part of the paper reviews several techniques which can be used in modelling of the conditional covariance (correlation) structures. These methods are compared in a simulation analysis in the next section. Note that the whole study is reduced to the situation of the (2×2) matrix \mathbf{H}_t with unit diagonal elements, i.e.

$$\mathbf{H}_t = \begin{pmatrix} 1, & \rho_t \\ \rho_t, & 1 \end{pmatrix} \quad [= \mathbf{R}_t]. \quad (2)$$

2.1 Simple Models

The (simple) moving average in its general form:

$$\hat{\mathbf{H}}_t^M = \frac{1}{M} \sum_{s=t-M}^{t-1} \mathbf{X}_s \mathbf{X}_s', \quad M \in \{2, 3, \dots\}. \quad (3)$$

The exponential smoothing usually used by Riskmetrics (Engle, 2009):

$$\hat{\mathbf{H}}_t^{EXP} = (1 - \lambda) \mathbf{X}_{t-1} \mathbf{X}_{t-1}' + \lambda \hat{\mathbf{H}}_{t-1}^{EXP}, \quad \lambda \in (0, 1). \quad (4)$$

Both previous estimators of the conditional covariance matrix \mathbf{H}_t should be normalized to match the specification (2).

2.2 Cholesky Decomposition

Let \mathbf{H}_t , a positive definite matrix, has the reparametrization in the form

$$\mathbf{H}_t = \mathbf{L}_t \mathbf{G}_t \mathbf{L}_t', \quad (5)$$

where \mathbf{L}_t is a lower triangular matrix with unit diagonal elements and \mathbf{G}_t is a diagonal matrix with positive elements on its diagonal. The Cholesky decomposition (5), in its general form, has one great advantage in estimation as it requires no parameter constraints for the positive definiteness of \mathbf{H}_t . Taking (2) and (5), it is clear

$$\mathbf{H}_t = \begin{pmatrix} 1, & \rho_t \\ \rho_t, & 1 \end{pmatrix} = \begin{pmatrix} g_{11,t}, & \ell_{21,t} g_{11,t} \\ \ell_{21,t} g_{11,t}, & g_{22,t} + \ell_{21,t}^2 g_{11,t} \end{pmatrix}.$$

Comparing both sides of the last equality, one can obtain: $g_{11,t} = 1, g_{22,t} = 1 - \rho_t^2$ and $\ell_{21,t} = \rho_t$. The estimator of ρ_t in each time given \mathcal{F}_{t-1} is then obviously reached as the estimator of the parameter β in the linear regression model $X_{2,k} = \beta_t X_{1,k} + Y_{2,k}$, where $Y_{2,k}$ is an error term and $k = t, t-1, \dots$. Taking into account the previous assumptions, the following statements can be viewed:

$$\beta_t = \frac{\text{cov}(X_{1,t}, X_{2,t} | \mathcal{F}_{t-1})}{\text{var}(X_{1,t} | \mathcal{F}_{t-1})} = \ell_{21,t} (= \rho_t), \quad \text{var}(Y_{2,t} | \mathcal{F}_{t-1}) = 1 - \rho_t^2$$

$$\text{and } \text{cov}(X_{1,t}, Y_{2,t} | \mathcal{F}_{t-1}) = 0.$$

The lower triangular matrix with unit diagonal elements, \mathbf{L}_t , provides an orthogonal transformation of the vector \mathbf{X}_t such that

$$\mathbf{Y}_t = \mathbf{L}_t^{-1} \mathbf{X}_t, \quad \text{cov}(\mathbf{Y}_t | \mathcal{F}_{t-1}) = \mathbf{G}_t. \tag{6}$$

The straightforward recursive generalization of this method for $n > 2$ and the non-normalized diagonal of \mathbf{H}_t can be found in Tsay (2005).

Four different choices of estimation techniques are considered: the standard OLS method, its exponentially weighted version, the $L1$ estimator and its exponentially weighted version. The proper weights are selected which minimize an adequate sum of calculated errors.

2.3 Dynamic Conditional Correlation Models

Suppose three different types of the *dynamic conditional correlation models* (DCC models) which have been introduced by Engle (2002):

$$\mathbf{Q}_t = (1 - \lambda) \mathbf{X}_{t-1} \mathbf{X}'_{t-1} + \lambda \mathbf{Q}_{t-1}, \tag{7}$$

$$\mathbf{Q}_t = \mathbf{\Omega} + \alpha \mathbf{X}_{t-1} \mathbf{X}'_{t-1} + \beta \mathbf{Q}_{t-1}, \tag{8}$$

$$\mathbf{Q}_t = \mathbf{\Omega} + \alpha \mathbf{X}_{t-1} \mathbf{X}'_{t-1} + \gamma \boldsymbol{\nu}_{t-1} \boldsymbol{\nu}'_{t-1} + \beta \mathbf{Q}_{t-1}, \tag{9}$$

where $\boldsymbol{\nu}_t = \min\{\boldsymbol{\varepsilon}_t, \mathbf{0}\}$ and the intercept matrix $\mathbf{\Omega}$, both in (8) and (9), is supposed to be symmetrical and positive (semi-)definite with a normalized diagonal with respect to the unique parametrization. In this case, the matrix \mathbf{Q}_t can be looked upon as an approximation to the conditional covariance (correlation) matrix \mathbf{H}_t as $\mathbf{H}_t = \text{diag}\{\mathbf{Q}_t\}^{-1/2} \mathbf{Q}_t \text{diag}\{\mathbf{Q}_t\}^{-1/2}$. The *integrated* DCC model (7) is a direct analogue of exponential smoothing. There is only one unknown parameter, $\lambda \in (0, 1)$, and it is used in each equation of the system. The process for \mathbf{Q}_t has a unit root and it has no tendency that its covariances revert to a constant value. This model can be easily rewritten as an integrated moving average model without intercept, put $e_{ij,t} = X_{i,t} X_{j,t}$:

$$(e_{ij,t} - e_{ij,t-1}) = -\lambda(e_{ij,t-1} - q_{ij,t-1}) + (e_{ij,t} - q_{ij,t}), \quad i, j = 1, \dots, n. \tag{10}$$

The *mean-reverting* DCC model (8) is an analogue of the scalar diagonal GARCH model (in terms of the volatility adjusted data). This process has two unknown dynamic parameters and the intercept matrix with $\frac{1}{2}n(n-1)$ unknown parameters. However, Engle (2009) presents so-called correlation targeting, a simple estimator of the intercept matrix, which allows to work only with two remaining unknown parameters α and β . There one also shows sufficient conditions which guarantee the matrices \mathbf{Q}_t to be positive definite.

The *asymmetric* DCC model (9) respects the fact that dynamic adjustment of correlations may be different for positive/negative variables. Estimation of the intercept matrix is a little more complicated than it is in the case of the intercept matrix in (8), see Engle (2009) for more details.

Estimation of the DCC models can be formulated as a maximum-likelihood problem once a specific distributional assumption is made for the data. Obviously, it is supposed that the data are multivariate normal with the given mean and covariance structure. Fortunately, the considered estimator is a quasi-maximum-likelihood, in the sense that it will be consistent but inefficient, if the mean and covariance assumptions are correctly specified even if other distributional assumptions are incorrect. See Engle (2002 and 2009) for more information and other references.

3 Simulations

In this section, the introduced models are compared by simulations.

Following Engle (2002), various types of correlation functions ρ_t were generated for $t = 1, \dots, 1000$: (i) Three constant correlations: $\rho_t^1 = 0.05$, $\rho_t^2 = 0.5$ and $\rho_t^3 = 0.95$. (ii) Two periodical ones: $\rho_t^4 = 0.25 + 0.7 \cos(\pi t/125)$ and $\rho_t^5 = 0.25 + 0.7 \cos(\pi t/20)$. (iii) Three jumping correlations: $\rho_t^6 = 0.5 + 0.4I_{[t>500]}$, $\rho_t^7 = -0.25 + 0.3I_{[t>500]} + 0.5I_{[t>750]} + 0.1I_{[t>875]} + 0.3I_{[t>937]}$ and $\rho_t^8 = \text{mod}(t/200)/200$. (iv) The ninth correlation function is a realization of the ARIMA(1, 1, 1) process with innovations distributed as $N(0, 1)$. These processes were chosen because they exhibit jump changes, gradual changes and periods without changes.

The moving average (3) with $M = 50, 100, 150, 200$ (MA50, MA100, MA150 and MA200), the exponential smoothing (4) with the standardly used choice $\lambda = 0.94$ (EX0.94), the Cholesky decomposition technique with four mentioned estimation procedures of ρ_t (CHOLols, CHOLwls, CHOLl1 and CHOLw1l), the DCC models (7)-(9) (DCCin, DCCmr and DCCa) and the estimation alternative of (7), the model (10) (DCCima), are evaluated.

The capabilities of considered models are measured in three ways. First, there is a simple comparison of the estimated correlations with their true counterparts, i.e. the mean absolute error which is defined as $MAE = \frac{1}{T} \sum_{t=1}^T |\rho_t - \hat{\rho}_t|$. The second and third measure, respectively, is the Ljung-Box test for existence of correlations in the series of the calculated and

squared calculated (second) residuals, respectively. In particular, the number of rejections using the 5% critical value is a measure of the performance of the estimator. The more rejections mean the more evidence that the calculated residuals have remaining time varying structures.

The process \mathbf{X}_t defined by (1) was simulated 1000 times with multivariate normal and t_4 distributed errors, respectively. The results were quite similar. Therefore, only the case with the t_4 distributed errors is presented. The graphical results of the simulations are shown in Figures 1 and 2. In both figures, each of the 13 stacks represents one of the introduced techniques for the conditional correlation modelling. Within each stack, the models ρ^1, \dots, ρ^9 are then confronted with each other with regard to given criteria. It is clear that the class of the DCC models overcomes the others (in the sense of the defined criteria). Especially, the mean-reverting (8) and asymmetric (9) DCC methods are successful. The Cholesky decomposition method and its related different estimators are not very suitable from this point of view (see the Figure 2). Even, the classical and frequently used models, i.e. the exponential smoothing and moving averages, lose compared with Engle's DCC models.

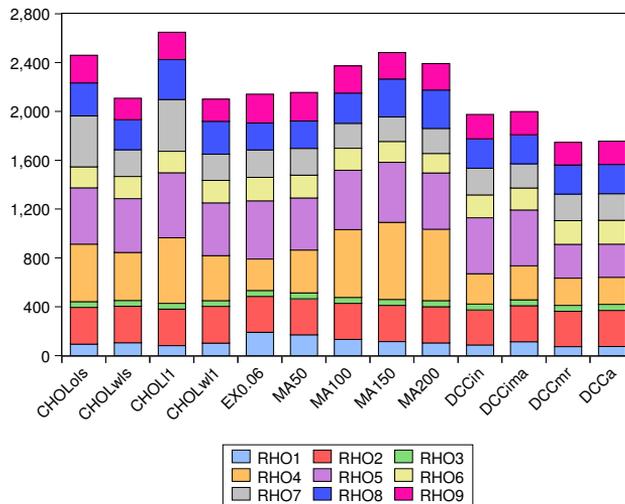


FIGURE 1. The comparison of the total sums of the mean absolute errors.

4 Conclusion

In the present contribution, several methods which estimate time varying correlations were introduced and compared in the simulation framework.

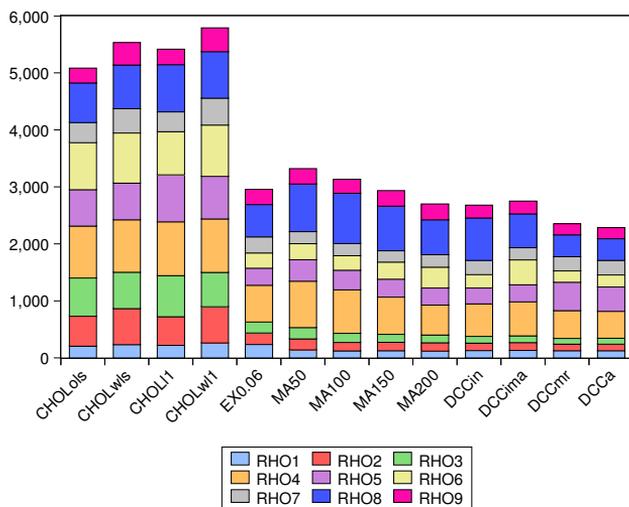


FIGURE 2. The comparison of the total number of the test rejections in the series of the calculated or squared calculated (second) residuals. The Ljung-Box test with the 5% critical value for existence of correlations is used.

It was clearly shown that Engle's dynamic correlation techniques outperform the usual moving averages models, the exponential smoothing procedure or the method based on the Cholesky decomposition. In the class of the dynamic correlations, two models dominate, i.e. the mean-reverting and the asymmetric version. Therefore, it is really reasonable to use them to analyze time varying correlation structures.

Acknowledgments: This work was supported by SVV 265315/2012. The author also wishes to express his thanks to Prof. RNDr. Tomáš Cipra, DrSc. for several helpful comments.

References

- Engle, R. (2009). *Anticipating Correlations: A New Paradigm for Risk Management*. New York: Princeton University Press.
- Engle, R. (2002). Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroscedasticity Models. *Journal of Business and Economic Statistics*, **20**, 339–350.
- Tsay, R.S. (2005). *Analysis of Financial Time Series*. New York: Wiley.

A robust nearest neighbor-based multiple imputation approach for data with missing covariate values

Chiu-Hsieh Hsu¹, Qi Long², Yisheng Li³, Elizabeth Jacobs¹

¹ University of Arizona, USA

² Emory University, USA

³ MD Anderson Cancer Center, USA

E-mail for correspondence: phsu@azcc.arizona.edu

Abstract: In regression analysis, often there are some subjects with missing covariate information. Ignoring these subjects may result in a loss of efficiency or bias in the estimation of the parameters of interest. Here, a nearest neighbor-based multiple imputation approach is proposed to recover missing covariate information. To conduct the imputation, two working regression models are fitted to define an imputing set. One is a regression model for predicting the missing values. The other is a regression model for predicting the probabilities of missingness. This imputation approach is robust to misspecifications of either of the two working models and the underlying distribution of the data. We study the finite sample performance of the proposed approach through simulation and apply the method to estimate the association between the occurrence of any large colorectal adenomas and the serum vitamin D level using the data from a colorectal adenoma intervention trial. The results indicate that the proposed imputation approach can improve efficiency and reduce bias in estimating the association.

Keywords: missing at random, multiple imputation, nearest neighbor.

1 Introduction

There is an extensive body of literature on statistical methods that use covariates to predict either missing observations or the probabilities of missingness (Little, 1992). Most of these methods predict either the missing observations or the probabilities of missingness and only a few predict both simultaneously. Also, most of them rely on strong distributional assumptions or specific missing data mechanisms for predicting missing observations or probabilities of missingness. Here, we propose a direct approach, multiple imputation (Rubin, 1987), for handling missing observations that employs covariates to predict both missing observations and probabilities of missingness. Specifically, for each missing observation, we will use two working regression models to define a set of similar observations called the

imputing set. One model is for predicting the missing values. The other is for predicting the probabilities of missingness. This approach relies on weaker assumptions because parametric models containing the covariates are not directly used to predict the missing observations or the probabilities of missingness. We also expect this approach will induce a double robustness property under a missing at random (MAR) mechanism and This paper is organized as follows. In the Method Section, we describe the imputation procedures. In the Simulation Section, we study the performance of the proposed imputation approach in finite sample. In the Application Section, we demonstrate the imputation approach using baseline data from an ursodeoxycholic acid (UDCA) colorectal polyp prevention study in which the serum vitamin level was only measured on a fraction of participants who had an observed clinical endpoint. We conclude with a discussion about the performance and potential generalizations and limitations of the proposed imputation approach.

2 Method

2.1 Notation

For simplicity, we consider a situation that only one covariate has missing values. Let Y denote the outcome variable, X denote the covariate with missing observations, M denote the missingness indicator, and Z denotes the fully observed covariates that are predictive of either X or M . We briefly describe the multiple imputation procedures below.

2.2 Imputation procedures

For each missing observation, we seek an imputing set consisting of observations from participants without missing data who are similar to the participant with a missing observation through a four step procedures.

Step 1 Calculating Scores: We propose to fit a working regression model for cases with no missing values to derive the predictive scores for both the non-missing and missing cases. The score provides a profile of an individual's X . We also propose to fit a working logistic regression model to derive a missingness score. This strategy summarizes the multi-dimensional structure of the variables into a two-dimensional summary score. If one of these two working models is correctly specified, conditional on these two scores, the covariate with missing values is independent of the missing status. Based on this property, we expect the combination of these two scores will have a double robustness property (Robins, 2000) in a regression setting with a missing covariate under an MAR mechanism.

Step 2 Defining the Imputing Set: We propose to calculate a weighted Euclidean distance to define similarity between subjects using information from the two predictive scores, where the weights are used to control the

effects of the two scores on the distance. For each subject with missing X , this distance is then employed to define a set of nearest neighbors. This neighborhood consists of subjects who have the smallest distances from subject with missing X .

Step 3 Imputation Schemes: For a subject with missing X , after the imputing risk set is defined, an observation is drawn equally likely from the imputing set. The procedure will be independently repeated K times (5 in this paper) to obtain multiple imputed datasets for use in estimation.

Step 4 Analyzing imputed datasets: The methods for analyzing multiply imputed data sets have been well established in Rubin (1987). The MI procedure by itself does not incorporate the full uncertainty in the imputed values, because it does not include a first stage of an initial parameter draw. This can be improved by conducting MI on bootstrap samples (Heitjan and Little, 1991; Rubin and Schenker, 1991). The above nearest neighbor-based MI approach conducted on the bootstrap sample is denoted as $BNNMI(NN, w_1, w_2)$, where NN is the size of the nearest neighbor, w_1 is the weight for the missing value predictive score and w_2 is the weight for the missingness predictive score.

3 Simulation

We performed a simulation study to investigate the finite sample properties of the BNNMI method in a generalized linear model (GLM) setting with a Poisson outcome (Y), a continuous covariate (X) subject to missing and one fully observed continuous covariate (Z). We mainly focused on comparing the estimate of the regression coefficient of the missing covariate (X) in a GLM for Y with X and Z as the covariates between fully observed (FO), which was treated as the gold standard since all X were fully observed, complete case (CC) and BNNMI methods. In addition, we were also interested in exploring the effects of NN , w_1 , w_2 and misspecification of the underlying distribution of X on the BNNMI method.

For each of 500 independent simulated datasets, X was generated from $N(0, 0.5)$, Z was generated from $Exponential(1/e^{-1-2X})$, Y was generated from a $Poisson(e^{1+2X-Z})$ distribution and M was generated from $P(M = 1) = 1/1 + e^{1+5Z-0.5Y}$. Those parameters were chosen to control the missing rate at approximately 35%. A sample size of 200 was considered in this paper. For the FO method, a Poisson regression model with X and Z as the covariates was fitted to the data (Y) before the missing indicator was applied to the data. For the CC method, a Poisson regression model was fitted using the completed cases only. For the BNNMI method, a working linear regression model for X based on the completed cases (M_1) and a logistic regression model for M (M_2) were fitted to the observed data to derive two predictive scores to select an imputing set for each missing observation. Four scenarios of the two working models were considered, including both models with only Z as the covariate, i.e. both mis-specified,

($BNNMI_{11}$), M_1 with Z as the covariate and M_2 with Y and Z as the covariates, i.e. M_1 mis-specified and M_2 correctly specified, ($BNNMI_{12}$), M_1 with Y and Z as the covariates and M_2 with Z as the covariates, i.e. both mis-specified, ($BNNMI_{21}$), and both models with both Y and Z as the covariates, i.e. M_1 mis-specified and M_2 correctly specified, ($BNNMI_{22}$). Note: M_1 with Y and Z as the covariates was considered as mis-specified because X conditional on Y and Z does not follow a normal distribution. The results are provided in Table 1. The CC method had the largest bias in estimating the regression coefficient compared to the other methods for both X from a normal distribution and an exponential distribution. The bias was because the missing mechanism was MAR and the CC method did not take that into account while estimating the regression coefficient. For the BNNMI method, if one of the working models was correctly, as the weight on the correct working model increased, the bias decreased. The bias also decreased as the size of nearest neighborhood decreased and as sample size increased (not shown here). Under $BNNMI_{21}$ that both models were mis-specified), the bias was smaller than CC, as well. This indicates the BNNMI method is robust to the mis-specification of the working models.

4 Application

The UDCA data consist of 1,192 patients, who underwent removal of colorectal adenomas between January 1996 and January 2000, from a colorectal adenoma prevention trial conducted at the Arizona Cancer Center (Alberts et al., 2005). Due to a limited budget, of the 1,192 participants, only 598 (50.2%) participants were selected to perform an assay to measure the serum vitamin D level. For those participants who were not selected for the assay, their serum vitamin D levels were regarded as missing data. We demonstrated the proposed MI approach on estimating the association between any large baseline colorectal adenomas and serum vitamin D. The covariates significantly associated with the observed serum vitamin D level were used to derive a predictive score of serum vitamin D levels through a linear regression model. The covariates significantly associated with the missing probabilities were used to derive a predictive score of missingness through a logistic regression model. These two scores were then used to select an imputing set for each missing observation.

The CC analysis showed no statistically significant association between any large baseline colorectal adenomas and the serum vitamin D level with an odds ratio estimate of 0.84 (95% CI: 0.70, 1.01), similar to what was reported for this population previously (Jacobs et al., 2007). Based on the findings in simulation and a weak MAR mechanism for the data, we only demonstrated the BNNMI approach with the entire weight on the score derived from the working linear regression model. The BNNMI method indicated that there was a statistically significant association between any

large baseline colorectal adenomas and the serum vitamin D level. Specifically, BNNMI produced an estimate of odds ratio of 0.79 with a 95% CI of (0.66, 0.95). In summary, the BNNMI method using the predictive covariates in the estimation had potential to improve efficiency and reduce bias in the estimate of the association between any large baseline colorectal adenomas and the serum vitamin D level.

5 Discussion

This paper describes a simple nonparametric multiple imputation procedure which uses predictive variables to recover information for missing observations. This multiple imputation method indirectly incorporates the information from the predictive covariates into estimation of the association. In this sense the properties of the estimation are derived mainly from the data, rather than from the assumptions in the working models, and, therefore, the proposed approach is robust to misspecification of the underlying distribution of the covariate with missing observations. In contrast, most of the methods in the literature directly incorporate the information from the predictive covariates into estimation of the association and, therefore, their performance highly depends on the assumptions in the models. Our data analysis and simulation study show that the use of this multiple imputation method has potential to lead to improved performance.

The performance of the proposed imputation method in improving efficiency and reducing bias depends on how predictive the variables are for both the missing values and missing probabilities. In the data analysis, the magnitude of the missing at random mechanism was weak and, therefore, the use of the scores based on the model predicting missing probabilities provided little contribution in estimation. Thus, it is more important to seek good models for predicting missing values than to find reasonable working models for both missing values and the probabilities of missingness. Also, the adequacy of the imputation procedures will depend on the "nearness" of the imputing set. When the nearest neighborhood contains some observations that are not close enough to the missing observation, some remnant of the missing at random mechanism remains within the neighborhood, which could contribute bias into estimation. The "nearness" of the imputing set will depend on the quality of the parameter estimates from the two working models, especially the parameters from the working regression models, and the size of the nearest neighborhood. In this paper, we simply fitted a linear regression model to predict the covariate with missing observations. Potentially, when the covariate is not normal, a generalized linear model may be fit to predict the values of the missing covariate. As for the selection of the weights, a small weight (e.g. 0.2) for the predictive score derived from the missing probability model is usually sufficient even under a MAR mechanism.

TABLE 1. Monte Carlo Simulation Results

Method	Estimate	Bias	SD	SE	CR
FO	1.999	0.001	0.0954	0.0933	0.95
CC	1.829	-0.171	0.2511	0.2580	0.90
$BNNMI_{11}(3, 0.8, 0.2)$	0.431	-1.568	0.2245	0.3899	0.15
$BNNMI_{12}(3, 0.8, 0.2)$	1.880	-0.119	0.4022	0.4708	0.96
$BNNMI_{21}(3, 0.8, 0.2)$	2.052	0.053	0.4198	0.4546	0.93
$BNNMI_{22}(3, 0.8, 0.2)$	2.048	0.049	0.4187	0.4557	0.93
$BNNMI_{11}(3, 0.2, 0.8)$	0.430	-1.569	0.2238	0.3946	0.15
$BNNMI_{12}(3, 0.2, 0.8)$	1.930	-0.069	0.4238	0.4957	0.96
$BNNMI_{21}(3, 0.2, 0.8)$	2.015	0.016	0.4059	0.4469	0.95
$BNNMI_{22}(3, 0.2, 0.8)$	2.029	0.030	0.4282	0.4640	0.94
$BNNMI_{11}(5, 0.8, 0.2)$	0.418	-1.581	0.2184	0.3880	0.11
$BNNMI_{12}(5, 0.8, 0.2)$	1.690	-0.309	0.3800	0.4580	0.93
$BNNMI_{21}(5, 0.8, 0.2)$	1.897	-0.102	0.3956	0.4362	0.93
$BNNMI_{22}(5, 0.8, 0.2)$	1.909	-0.090	0.4041	0.4359	0.94
$BNNMI_{11}(5, 0.2, 0.8)$	0.412	-1.587	0.2076	0.3837	0.09
$BNNMI_{12}(5, 0.2, 0.8)$	1.746	-0.253	0.4032	0.4798	0.93
$BNNMI_{21}(5, 0.2, 0.8)$	1.878	-0.121	0.3884	0.4274	0.94
$BNNMI_{22}(5, 0.2, 0.8)$	1.870	-0.129	0.4046	0.4552	0.95

References

- Alberts, D.S., Martinez, M.E., Hess, L.M., et al (2005) Phase III trial of ursodeoxycholic acid to prevent colorectal adenoma recurrence. *Journal of the National Cancer Institute*, **97**, 846–853.
- Heitjan, D.F., Little R.J.A. (1991) Multiple imputation for the fatal accident reporting system. *Applied Statistics*, **40**, 13–29.
- Jacobs, E.T., Alberts, D.S., Benuzillo, J., et al. (2007) Serum 25(OH)D levels, dietary intake of vitamin D, and colorectal adenoma recurrence. *Journal of Steroid Biochemistry & Molecular Biology*, **103**, 752–56.
- Little, R.J.A. (1992) Regression with missing X's: a review. *Journal of the American Statistical Association*, **87**, 1227–1237.
- Robins, J.M., Rotnitzky, A., van der Laan, M. (2000) Comment on ‘On profile likelihood’. *Journal of the American Statistical Association*, **95**, 477–482.
- Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*. Wiley.
- Rubin, D.B., Schenker N. (1991) Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine*, **10**, 585–598.

Generalized estimating equations in claims reserving

Šárka Hudecová¹, Michal Pešta¹

¹ Charles University in Prague, Faculty of Mathematics and Physics, Czech Republic

E-mail for correspondence: `hudecova@karlin.mff.cuni.cz`

Abstract: Some commonly used generalized linear models, namely the over-dispersed Poisson model, gamma model, and Tweedie's compound Poisson model, are considered for a standard actuarial data set on claims amounts. The parameters of the model are fitted using the standard generalized linear models techniques as well as using the generalized estimating equations. The estimated outstanding claims reserves are computed and compared.

Keywords: Claims reserving; GEE; Tweedie distribution.

1 Introduction

Claims reserving is a classical problem in general insurance. The aim is to quantify the outstanding reserves for insurance claims which have been incurred but have not yet been settled. A number of different methods to this problem has been invented, see, e.g., Wüthrich and Merz (2008) for an overview.

England and Verrall (2002) give an overview on a wide range of stochastic reserving models in general insurance. These methods are then illustrated and compared on a standard actuarial data set. In this paper, we apply some recently proposed techniques on the same data set in order to extend this comparison. Namely the Tweedie's compound Poisson model from Wüthrich (2003) is fitted. The estimated outstanding claims reserves are computed and compared with the results obtained by some standard approaches, namely the over-dispersed Poisson model and gamma model. Moreover, parameters of the three models are estimated by standard GLM techniques as well as using generalized estimating equations (GEE). The observed discrepancies are discussed.

The paper is organized as follows. The next section introduces notation and an analyzed data set. The applied methods are briefly described in Section 3. Section 4 provides results from the analysis. Discussion is given in Section 5.

TABLE 1. Run-off triangle for incremental claims $X_{i,j}$.

Accident year i	Development year j				
	1	2	\dots	$n-1$	n
1	$X_{1,1}$	$X_{1,2}$	\dots	$X_{1,n-1}$	$X_{1,n}$
2	$X_{2,1}$	$X_{2,2}$	\dots	$X_{2,n-1}$	
\vdots	\vdots	\vdots	\ddots		
			$X_{i,n+1-i}$		
$n-1$	$X_{n-1,1}$	$X_{n-1,2}$			
n	$X_{n,1}$				

2 Notation and Data

We introduce the classical claims reserving notation and terminology. Outstanding loss liabilities are structured in so-called claims development triangles. Let us denote $X_{i,j}$ all the claim amounts in development year j with accident year i . Therefore, $X_{i,j}$ stands for the *incremental claims* in accident year i made in accounting year $i+j$. The current year is n , which corresponds to the most recent accident year and development period as well. That is, our data history consists of observations $X_{i,j}$, where $i = 1, \dots, n$ and $j = 1, \dots, n+1-i$ (run-off-triangle), see Table 1. The aim is to estimate the outstanding claims reserves $R_i = \sum_{j=n+2-i}^n X_{ij}$ for all $i = 2, \dots, n$.

In this contribution, we analyze a standard actuarial data set previously studied in England and Verrall (2002). The data come from Automatic Facultative General (AFG) Liability (excluding Asbestos and Environmental) from Historical loss development study published in 1991 by Reinsurance Association of America. The incremental claims amounts are listed in Table 2.

Notice that there is one negative incremental value in cell (2,7). This causes a problem in the analysis, because solely positive values are required by some of the applied methods. For this reason, the negative increment has been weighted out of the whole analysis, similarly as in England and Verrall (2002, Section 7.7).

3 Methodology

Recall that X_{ij} stands for incremental claims in accident year i and development period j , and denote $\mu_{ij} = EX_{ij}$. Following the notation from England and Verrall (2002), we assume the multiplicative model

$$\mu_{ij} = x_i y_j,$$

TABLE 2. Incremental claims of Automatic Facultative General Liability data.

$i \setminus j$	1	2	3	4	5	6	7	8	9	10
1	5012	3257	2638	898	1734	2642	1828	599	54	172
2	106	4179	1111	5270	3116	1817	-103	673	535	
3	3410	5582	4881	2268	2594	3479	649	603		
4	5655	5900	4211	5500	2159	2658	984			
5	1092	8473	6271	6333	3786	225				
6	1513	4932	5257	1233	2917					
7	557	3463	6926	1368						
8	1351	5596	6165							
9	3133	2262								
10	2063									

where $\sum_{j=1}^n y_j = 1$. Here x_i can be interpreted as expected ultimate claims in accident year i , and y_j is the proportion paid in development year j . For the estimation purposes, the model is reparametrized by taking the logarithm and one gets an additive model

$$\log(\mu_{ij}) = c + \alpha_i + \beta_j \tag{1}$$

with constraints $\alpha_1 = \beta_1 = 0$.

In order to estimate the parameters of the model (and thus the outstanding reserves), one needs to make some assumptions about the distribution of X_{ij} . Commonly, the gamma distribution or the over-dispersed Poisson distribution are used for the incremental claims. Recently, the Tweedie's compound Poisson distribution has been found as suitable for modeling incremental claims as well, see Wüthrich (2003) and Peters et al. (2009), providing more flexibility in modeling. This distribution (also referred to as Poisson-gamma or compound gamma) arises if the number of claims in year i and development period j is a Poisson random variable, and the sizes of individual payments are independent gamma distributed variables with common shape parameter, see, e.g., Jørgensen and Souza (1994). All the three mentioned distributions belong to the exponential family and, therefore, the parameters of model (1) can be estimated by standard GLM methods. For the Tweedie's distribution, this is preceded by the estimation of an additional parameter p by a profile likelihood plot.

However, the framework of GLM requires the necessary assumption that the incremental claims X_{ij} are independent variables for all $i, j = 1, \dots, n$. If this assumption is violated, then the predicted reserves might be misleading. One possible way to deal with correlated data in GLM is based on GEE, see, e.g., Ziegler (2011). The unknown correlation structure is modeled using so called working correlation matrices. The advantage is that estimates of the parameters are consistent under mild regularity conditions even if the variance structure is not correctly specified.

All the considered models were fitted in program R. The over-dispersed

Poisson model and gamma model were fitted by `glm` function. For estimation of the Tweedie's compound Poisson model, packages `statmod` and `tweedie` are required. The GEE estimates can be obtained by `gee` function from `gee` library. The exchangeable working correlating matrix was used for the Poisson and gamma models (the program was not able to handle autoregressive and m -dependent structures). The function `gee` does not allow using the Tweedie's distribution and, thus, the GEE estimation for this model is based on code from Swan (2006). Here, the autoregressive autocorrelation structure AR(1) was used. Nevertheless, consistent estimates are obtained from the GEE approach even if the correlating matrix is misspecified.

4 Results

The estimates of the parameters from model (1) are listed in Table 3. Note

TABLE 3. Estimated parameters of model (1) for AFG data.

	Poisson		Gamma		Tweedie	
	GLM	GEE	GLM	GEE	GLM	GEE
c	7.6446	7.6462	7.6954	7.6954	7.6926	7.4376
$\widehat{\alpha}_2$	-0.0547	-0.0546	0.0428	0.0428	-0.0559	0.0419
$\widehat{\alpha}_3$	0.2454	0.2452	0.1427	0.1427	0.1798	0.3839
$\widehat{\alpha}_4$	0.4204	0.4201	0.3566	0.3566	0.3728	0.5305
$\widehat{\alpha}_5$	0.4396	0.4393	0.2444	0.2444	0.3273	0.6801
$\widehat{\alpha}_6$	0.0453	0.0453	-0.0898	-0.0898	-0.0358	0.2269
$\widehat{\alpha}_7$	-0.0488	-0.0488	-0.2790	-0.2789	-0.1663	0.1733
$\widehat{\alpha}_8$	0.2537	0.2535	0.0689	0.0689	0.1560	0.4526
$\widehat{\alpha}_9$	-0.1498	-0.1497	-0.0361	-0.0361	-0.1101	-0.1912
$\widehat{\alpha}_{10}$	-0.0127	-0.0127	-0.0635	-0.0635	-0.0607	0.1943
$\widehat{\beta}_2$	0.6928	0.6924	0.7141	0.7141	0.6924	0.7785
$\widehat{\beta}_3$	0.6260	0.6256	0.7289	0.7289	0.6567	0.6752
$\widehat{\beta}_4$	0.2769	0.2767	0.2562	0.2561	0.2678	0.3780
$\widehat{\beta}_5$	0.0606	0.0605	0.1041	0.1040	0.0810	0.0973
$\widehat{\beta}_6$	-0.1958	-0.1957	-0.1526	-0.1526	-0.1687	-0.2489
$\widehat{\beta}_7$	-0.8304	-0.8297	-0.7615	-0.7615	-0.8030	-0.8353
$\widehat{\beta}_8$	-1.2791	-1.2777	-1.3172	-1.3173	-1.2943	-1.1929
$\widehat{\beta}_9$	-1.9323	-1.9296	-2.0489	-2.0490	-1.9679	-1.7570
$\widehat{\beta}_{10}$	-2.4971	-2.4930	-2.5479	-2.5479	-2.5451	-2.2901

that our results for the over-dispersed Poisson model and gamma model slightly differ from those presented in England and Verrall (2002). This is probably due to omitting the one negative observation. Some differences in the parameter estimates can be observed comparing the three different

models. In addition, the two sets of estimates obtained by the two estimation methods (GLM and GEE) slightly differ as well. The differences are more noticeable from the estimated outstanding reserves given in Table 4.

TABLE 4. Estimated outstanding reserves for AFG data. Model (P) stands for the Poisson model, (G) for gamma model, and (T) for Tweedie’s compound Poisson model.

Model	Estimation of reserves for year i								
	2	3	4	5	6	7	8	9	10
(P)	163	607	1608	3052	3855	5623	11161	10820	16534
(P) GEE	164	609	1614	3061	3865	5635	11181	10839	16561
(G)	180	525	1492	2644	3618	4840	9896	13305	17159
(G) GEE	180	525	1492	2644	3618	4840	9896	13304	17158
(T)	163	573	1567	2859	3775	5327	10701	11988	16718
(T) GEE	179	683	1666	3390	3816	5843	11624	8862	17525

One can see that the estimated reserves obtained by GLM and GEE are practically the same for the gamma model (the differences are erased by rounding). For the Poisson model, the two methods give only negligible differences. The total estimated reserves are listed in Table 5.

TABLE 5. Total estimated reserves.

Poisson		Gamma		Tweedie	
GLM	GEE	GLM	GEE	GLM	GEE
53422	53528	53657	53656	53670	53589

5 Conclusions and Discussion

Stochastic methods handling the claims reserving problem usually assume independent claim amounts in the development years. If this assumption is violated, then the classical techniques provide incorrect prediction of the claims reserves. Hence, methods that enable modeling the dependencies in the development years are needed. In this article, we show the application of generalized estimating equations (GEE) for estimation of claims reserves. The GEE approach is compared with the GLM one on two commonly used models, i.e., over-dispersed Poisson model and gamma model. The Tweedie’s compound Poisson model is also added into the comparison for its gaining popularity.

It is not possible to say that one of the models would lead generally to larger (or smaller) estimates than the others. Generally, the GEE method gives substantially higher estimates than the GLM method, especially in the case of Tweedie’s compound Poisson model. The probable reason is

that the classical GLM assume independent development in claims. The only exception is in the estimated reserves for the 9th accident year in the case of Tweedie's compound Poisson model. Possible explanation lies in the sensitivity of the multiplicative model for the last accident and development years. Moreover, the uncertainty is largest in the two recent years 9 and 10, and thus there is highest variability in the reserves estimation. Nevertheless, the total estimated reserves are quite comparable in all the models.

Acknowledgments: This paper was written with the support of project "DYME Dynamic Models in Economics" No. P402/12/G097 of the Czech Science Foundation.

References

- England, P. and Verrall, R. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, **8**, 443–544.
- Jørgensen, B. and Souza, M.C.P.D. (1994). Fitting Tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, **1**, 69–93.
- Peters, G.W., Schevchenko, P.V., and Wütrich, M.V. (2009). Model uncertainty in claims reserving within Tweedie's compound Poisson models. *ASTIN Bulletin*, **39**(1), 1–33.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Swan, T. (2006). Generalized estimating equations when the response variable has a Tweedie distribution: An application for multi-site rainfall modelling. Master's thesis, The University of Southern Queensland, Toowoomba, QLD.
- Wüthrich, M. and Merz, M. (2008). *Stochastic claims reserving methods in insurance*. Wiley finance series. John Wiley & Sons.
- Wütrich, M.V. (2003). Claims reserving using Tweedie's compound Poisson model. *ASTIN Bulletin*, **33**(2), 331–346.

Application of text mining for media analysis

Nikola Kaspříková¹

¹ University of Economics in Prague, Czech Republic

E-mail for correspondence: data@tulipany.cz

Abstract: Media analysis using basic text mining techniques is reported. Analysis was based on 52 news articles on one of the U.S. presidential elections candidate, documents coming from three sources (U.S. based broadcaster, U.K. based broadcaster and documents referenced from the candidate's website). Cluster analysis was performed to tell if there is some classification structure in the data set and the results suggested partitioning the documents into two clusters, location (U.S. vs. U.K.) of the source being more important factor for cluster identification than expected independence of the source (broadcasters vs. candidate's website).

Keywords: Media analysis; News; Text mining; Clustering, U.S. presidential elections.

1 Introduction

Many news articles and blog posts are being published online and these articles may reach and influence a lot of readers. Documents published online are suitable for automated data retrieval, processing and analysis using text mining techniques. Common text mining tasks in analysis of news include topic identification, opinion extraction and sentiment analysis (see e.g. Junqué de Fortuny et al. (2012)). Models for demand prediction or price forecast based on published news articles are being developed in economic applications of text mining (Yu et al. (2007) and Mittermayer and Knolmayer (2006)). Methods used for analysis include techniques for supervised and unsupervised classification and latent variables construction. Various special tools for data preparation such as part-of-speech tagging or lexicons of words with positive or negative polarity are used in many text mining analyses.

This paper reports on media analysis using basic text mining techniques. Analysis was based on a collection of news articles on one of the U. S. presidential elections candidate. Documents were coming from three sources: U. S. based broadcaster, U. K. based broadcaster and finally documents referenced from the candidate's website. The issue addressed is to tell if there is some classification structure in this collection of documents and how it corresponds with the source of the documents.

2 Material and methods

Documents included in analysis came from three sources. All 52 documents were published within the same period, so it can be expected that the documents could have reported on the same set of events and all the documents included phrase "Romney" in the headline. 15 documents were published by a British public service broadcaster, 15 documents were published by a U.S. cable news channel and 22 documents were referenced on <http://www.mittromney.com> website. There was no duplicity in the collection of the documents.

Data preparation included standard document processing steps, i. e. eliminating punctuation and extra whitespace, converting to lower case, stemming and removing stop words. Porter's word stemming algorithm implemented in R package Rstem (Lang (2011)) was used for stemming. After constructing document-term matrix, which was constructed using term frequency (TF) weighting, sparse terms were removed. Using another weighting scheme (TF-IDF) has also been tried and it has given similar results in analysis of this collection of documents.

We use simple count-based methods, as results of these methods can usually be easily interpreted. R software (R Development Core Team, 2012) and text mining algorithms available in R package tm (Feinerer, Hornik and Meyer (2008)) have been used for data analysis. Hierarchical clustering with Ward's method was used for unsupervised classification of documents and classification tree as implemented in rpart R package (Therneau and Atkinson, 2012) was used for description and profiling of clusters.

3 Results and discussion

Results of hierarchical clustering (see Figure 1) suggest that partitioning into two groups would be the most suitable, resulting in Cluster 1 with 15 articles and Cluster 2 with 37 articles. Cross classification of cluster number and source of the document (see Table 1) shows that Cluster 1 covers mostly news published by U.K. based broadcaster and Cluster 2 covers mostly news published by U.S. based broadcaster or news referenced on candidate's website. It seems that without any further detailed analysis, location (U.S. vs. U.K.) of the source is more important than whether the news are from the source which is supposed to be more independent (sample of news published by broadcasters or news linked from candidate's website). There is no substantial difference with respect to cluster membership between news published by U.S. based broadcaster and news referenced on candidate's website. Further analysis of another set of documents, published in different period of time, should follow to check if this holds over time.

For further description of clusters, the most frequent terms in the clusters were inspected. Both clusters have high frequency of terms like "republi-

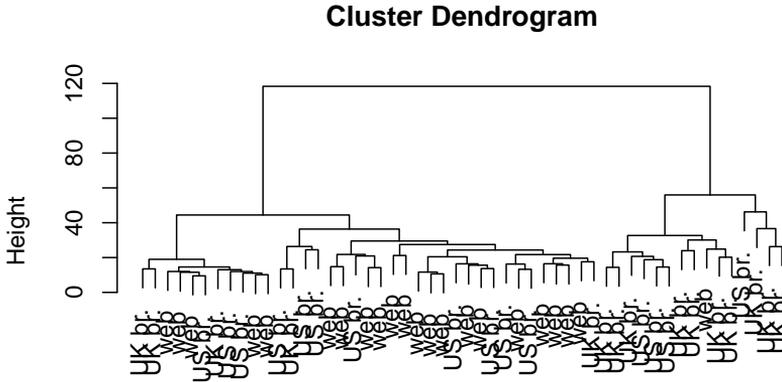


FIGURE 1. Dendrogram for hierarchical clustering, labels showing source of the document.

TABLE 1. Counts of documents by source and cluster.

	U.K. broadcaster	U.S. broadcaster	Website
Cluster 1	10	4	1
Cluster 2	5	11	21

can” and ”Romney” (as could have been expected). Cluster 1 has also high frequency of terms ”campaign” and ”Gingrich” (another Republican candidate), Cluster 2 has high frequency of terms like ”Mitt” and ”president”. When building prediction models for discrimination between clusters to get further insight into profiles of the clusters, terms like ”attack”, ”Ginrich” and ”primary” have been found as the most useful ones for identification of documents from Cluster 1. Classification tree model developed for cluster prediction was quite simple and accurate. Counts resulting form cross-classification of original cluster and cluster predicted using classification tree model are in Table 2. According to set of most frequent or discriminating words it seems that documents in Cluster 1 pay more attention to attacks of candidates on other candidates in the campaign.

TABLE 2. Counts of original by predicted cluster.

	Predicted cluster 1	Predicted cluster 2
Cluster 1	13	2
Cluster 2	1	36

4 Conclusion

Results of cluster analysis of news articles published within selected period of time which refer to U.S. presidential elections candidate have shown that there is a classification structure in the collection of documents. Two groups of documents, which were originally coming from three sources (U.S. based broadcaster, U.K. based broadcaster and documents referenced from the candidate's website), were identified. The classification structure which has been found in unsupervised classification quite closely corresponds with location of the source (U.S. or U.K.). Validation of this result using another set of documents, published in different period of time, should follow. The classification structure which has been obtained in cluster analysis is clear enough to be easily captured with simple classification tree model. According to set of terms used in model for discrimination between clusters it seems that documents in cluster which is represented mostly by news published by U.K. based broadcaster comparatively more often refer to attacks in the campaign.

References

- Feinerer, I., Hornik, K., and Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* 25/5.
- Junqué de Fortuny, E., De Smedt, T., Martens, D., Daelemans, W. (2012). Media coverage in times of political crisis: a text mining approach. *Expert Systems with Applications*, doi: 10.1016/j.eswa.2012.04.013.
- Lang, D.T. (2011). *Rstem: Interface to Snowball implementation of Porter's word stemming algorithm*. R package version 0.4-1. <http://CRAN.R-project.org/package=Rstem>
- Mittermayer, M.-A. and Knolmayer, G., F. (2006). *Text Mining Systems for Market Response to News: A Survey*. Working Paper 184, Institute of Information Systems University of Bern.
- 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, URL <http://www.R-project.org/>.

- Therneau, T. M. and Atkinson, B.. R port by Brian Ripley. (2012). *rpart: Recursive Partitioning*. R package version 3.1-52.
- Yu, W.-B., Lea B.-R., and Guruswamy, B. (2007). A theoretic framework integrating text mining and energy demand forecasting. *International Journal of Electronic Business Management*, **5/3**, 211–224.

Effects of different spatial modeling: Ruralisation of the NSDAP in the Weimar Republic?

André Klima¹, Helmut Küchenhoff¹, Paul W. Thurner²

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: Analyzing the results of seven elections from 1924 to 1933 in the Weimar Republic, we discovered an interesting modeling aspect when accounting for spatial effects. The focus of the analysis is to model the election behavior towards the NSDAP in districts. We present two different approaches to model potential spatial effects: an indicator and a smooth function. This spatial component was intended to be a confounder. Interestingly, one of the approaches reveals an effect for the degree of urbanization, thus allowing a proper interpretation. This example highlights the impact of modeling spatial effects on the the interpretation and the results in an important application.

Keywords: spatial effects; analysis of election; Weimar Republic; aggregated data

1 Introduction

1.1 Elections in the Weimar Republic

After the end of the first world war the first democracy in Germany was established. But the difficult situation in the Weimar Republic after the lost war and antidemocratic movements caused trouble right from the start for the young democracy. One and probably the most (in)famous of this enemies of the Weimar Republic was the NSDAP. The NSDAP developed from a minor party in the 1924 election to the biggest party in the 1933 election.

This last election and the takeover by Adolf Hitler symbolize also the end of the democratic Weimar Republic. A strong scientific interest in this period was caused by the world wide consequences of this takeover, but also by the fact that the clearly antidemocratic NSDAP could gain the power through democratic elections. One aspect of this research complex is the analysis of the elections and the election behavior.

1.2 Research Objective

The objective of the presented analysis is to investigate the elections and the election behavior. The question if certain social groups differed in their party preferences is one of the discussed topics in political science. (see King et al., 2008)

1.3 Data

One of the major difficulties in analyzing (past) elections is the data situation. It is often characterized through the missing of individual data, while the election results and also socio-demographic data aggregated over regional areas are available. With aggregated data alone there are mainly two options: Using a special method of ecological inference (e.g. King, 1997) or performing an analysis on the regional units. We used the second approach and therefore, have no direct individual behavior interpretation of our results.

The used data contains the results of the eight elections in the Weimar Republic, the primary results of the census 1925 and 1933 and other local or nationwide statistics (dataset: ZA8013, see Hänisch, 1988). While this dataset includes a lot of different information, the used information for the workforce of the farmer groups had to be added by us (Statistik des Deutschen Reichs - Volks-, Berufs- und Betriebszahlung, 1929). The level of urbanization, measuring the percentage of the inhabitants living in municipality with a size over 5000, was calculated using two sources in the data with differing collection periods. The data origins from 1925 and 1927, a fact, which caused minor errors while calculating the percentage.

The aggregation level of the regional units is mostly the district. In some cases it was necessary to merge district in order to have time constant regional units. In total, data for 849 districts is available. The membership to the state or province, and therefore in most cases also to a certain region, and the district centroids calculated from a map are available as spatial information.

For the analysis only seven of the eight elections were considered, in 1920 the NSDAP did not participate and therefore, this election was not analyzed. Preparatory testing showed that a grouping of the remaining seven elections into four groups is possible, the groups are the two elections 1924, the election 1928, the election 1930 and the three elections 1932 - 1933.

1.4 Model Specification

For the analysis we used a generalized additive model (GAM) (Fahrmeir et al., 2007) and two different approaches to model the spatial component. Model type one uses a regional indicator variable preserving mainly the political structure of the Weimar Republic, model type two uses the centroids and estimates a smooth surface. The spatial component itself was

added as confounder to the model, so we only tested a limited number of approaches. They were fitted with BayesX. (Berlitz et al., 2009) The models have the following formula:

$$\begin{aligned} \text{logit}(\pi) = & \beta_0 + [\beta_{elec}x_{elec}] + \\ & f_{B0-2ha}(x_{B0-2ha}) + \dots + f_{urb}(x_{urb}) + [f_{jl}(x_{jl})] + \\ & f_{spat}(z) \end{aligned}$$

π represents the share of the NSDAP in the analyzed election. For each election group smooth functions for the percentage of protestants, the workers total, the workers in industry and crafting, the workforce in four farm size classes (*B0-2ha*, ..., *B20-100ha*) and the urbanization (*urb*) are estimated. Terms in squared brackets symbolize parts of the formula, which are not included in every model. An election specific part is only added for the two election groups with more than one election. Because the percentage of jobless people (*jl*) origins from the census 1933, this variable is only included in the last election groups.

We present in detail the impact of different approaches incorporating spatial effects (z). The first is to include a regional indicator variable, representing the provinces of Prussia and the states of the Weimar Republic. The second uses a smooth surface estimation with districts centroids. The latter reveals an effect for the degree of urbanization, which allows a proper interpretation in the context of substantial theory.

2 Results

Because it is not possible to show all eight models in detail, only selected results are presented. To summarize the main results: The religion, like in other analyses before (King et al., 2008) is one of the major explanatory variables in the models.

2.1 Spatial Effects

Considering the spatial effects, between the two model types the general shape is similar. There are in particular strong regional effects in the first elections. Some regions, e.g. Bavaria, show a strong positive effect, while other regions have a negative effect. This spatial effects are not time constant for every region. Some regions with an estimated negative effect in the model for the first election group have a positive effect in the last election group. In general, even if in later elections the spatial effect becomes smaller and also if the differences between the regions are lower, the spatial effect itself is still one of the stronger effects in the model.

But there are recognizable differences between the estimated effects of the two approaches. When depicted in a map this differences indicate also the

limits of the chosen approaches. The smooth surface only badly performs in representing non-smooth differences at state borders. The regional indicator variable model is not able to capture differences in the regions.

2.2 Urbanization Effect

Comparing the estimated effects of the urbanization differences between the two model types are visible. In model type one no clear structure is visible (fig. 1). For the 1924 elections a positive effect for districts with a higher urbanization is visible, but there are also hardly to interpret bumps in the function, 1928 only three bumps are visible, while in the last two models almost no effect can be identified. In comparison, the models using the smooth surface for the regional effect show a different structure (fig. 2): A stronger positive effect of the level of urbanization is estimated. This effect is especially strong for the first two election groups, but decreases with time. For the last three elections in the years 1932 and 1933 no effect is estimated by the model.

With exception of 1924 the models with regional indicator variable have no straightforward interpretation of the estimated effects. The estimated smooth terms of model type two indicate better shares for the NSDAP in more urbanized areas at the beginning. This is an effect that decreases with time. This could be a sign for a ruralisation of the NSDAP over the

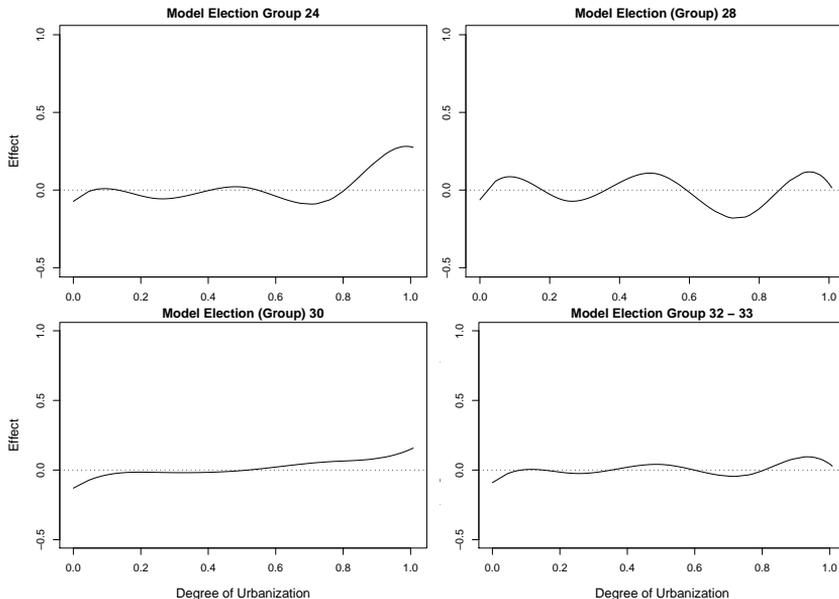


FIGURE 1. Estimated effect of urbanization in the model with regional indicator spatial component.

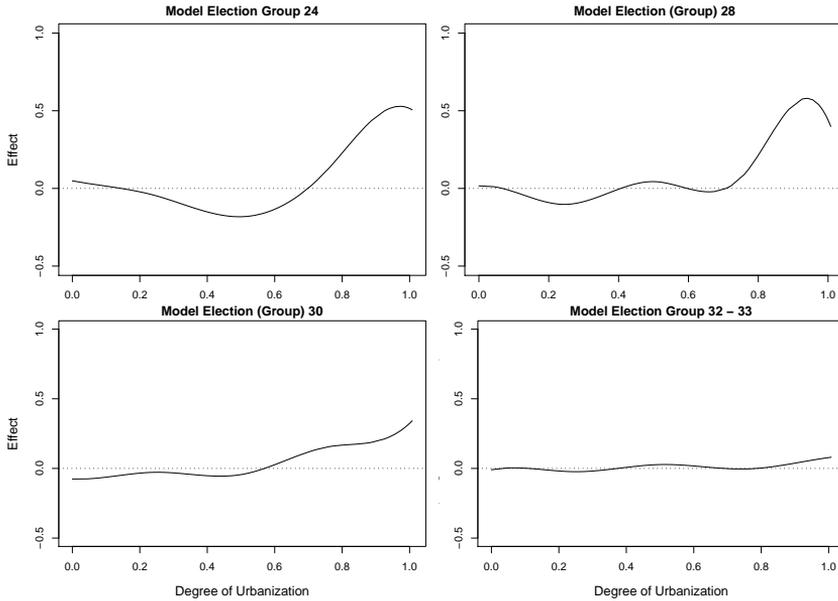


FIGURE 2. Estimated effect of urbanization in the model with smooth surface spatial component.

years. Because ruralisation can be seen as a local effect, - the spread from a town in the surrounding rural districts - it seems plausible that the more flexible smooth surface can better identify differences at the beginning of a ruralisation than a model with regional indicator variable.

3 Conclusions

The type of model influences the dependencies, which can be found in the model. While this fact is well known, it is not always trivial to take this into account. But it is important to remember that depending on the research question a different type of model can be necessary, while still analyzing the same data.

The results of our models indicate that in more urbanized districts the NSDAP performed better in the first elections, an effect that is reduced with the time. This would be a sign of a slow ruralisation. But this trend is only visible in one of the two considered models. This example shows that the model selection is an important part of the data analysis process and that this selection should also account for the analyzed issue.

References

- Belitz C., Brezger A., Kneib T., Lang S. (2009). *BayesX. Software for Bayesian Inference in Structured Additive Regressions Models, Version 2.0.1*. München: URL [http:// www.stat.uni-muenchen.de/ bayesx/ bayesx.html](http://www.stat.uni-muenchen.de/bayesx/bayesx.html) (20.04.2012).
- Fahrmeir L., Kneib T., Lang S. (2007). *Regression. Modelle, Methoden und Anwendungen*. Berlin: Springer.
- Hänisch D. (1988). *Benutzerhandbuch: Wahl- und Sozialdaten der Kreise und Gemeinden des Deutschen Reiches 1920-1933*.
- King G. (1997). *A Solution to the Ecological Inference Problem. Reconstructing Individual Behavior from Aggregate Data*. Princeton.
- King G., Rosen O., Tanner M. A., Wagner A. F. (2008). Ordinary Economic Voting Behavior in the Extraordinary Election of Adolf Hitler. In: *Journal of Economic History*, **68(4)**, 951–996.
- Statistik des Deutschen Reichs Volks-, Berufs- und Betriebszählung vom 16. Juni 1925 (1929). Landwirtschaftliche Betriebszählung Die Hauptergebnisse in den kleineren Verwaltungsbezirken der Länder des Deutschen Reichs, Band 412 I. Berlin.

Corporate financial performance and its predictors

Maria Králová¹, Alena Klapalová², Juraj Šiška¹

¹ Masaryk University, Faculty of Economics and Administration, the Czech Republic

² Masaryk University, Faculty of Economics and Administration, the Czech Republic

E-mail for correspondence: kralova@econ.muni.cz

Abstract: The purpose of the paper is to present results of an empirical survey focused on potential relations between the group of variables describing corporate financial performance and the group of other indicators such as: degree of centralization, innovation and knowledge management, market share, measure of uncertainties etc. Values of the indicators were perceived and recognized by managers of multinational companies' subsidiaries in the Czech Republic. Answers from 336 subsidiaries (quantitative on-line inquiry) were analyzed using various statistical methods including canonical correlation analysis.

Keywords: Canonical Correlation Analysis; Corporate Financial Performance.

1 Defining the groups of Variables

Corporate financial performance is one of essential indicators showing a degree of company's competitiveness and growth. The analysis of its association with other indicators describing strategies and further properties of companies may help to understand what makes companies to be good. Since in the survey there are two variables describing corporate financial performance (which could be mutually correlated) the authors decided not to use regression analysis for particular depending variable but to analyze the measure of association between two groups of variables, where in the first ("left") group there are indicators of financial performance.

1.1 Indicators of financial performance

Two basic financial strategies - strategy of an expansion and strategy of a rentability are represented by two variables *EA* and *ROA*. The first variable *EA* gives a percentage measure of growth of gross assets whereas the second variable *ROA* (return on asset) gives a percentage rate of economic results and gross assets. Due to various short-term fluctuation

and in order to comprehend a long-term financial standing of companies a five-year means of the variables *EA* and *ROA* were used in the study.

$$EA = \left(\frac{AF}{AI} - 1\right) \cdot 100 \quad ROA = \frac{ER}{AF} \cdot 100 \quad \text{where}$$

- ER* stands for the economic result
- AI* stands for the statement of initial assets
- AF* stands for the statement of final assets

1.2 Explanatory indicators

The list below presents variables which are considered to be "explanatory" in their effect on financial performance variables.

- * size of a company on a scale 1 to 3
- * degree of centralization on a scale 1 to 10
- * measure of satisfaction with a degree of centralization on a scale 1 to 10
- * measure of redistribution of finances on a scale 1 to 10
- * measure of satisfaction with redistribution of finances on a scale 1 to 10
- * knowledge impact on an innovation activity on a scale 1 to 10
- * presence of innovative centers binary
- * degree of cooperation with external organizations (universities, research institutes) on a scale 1 to 10
- * pace of change in the area of customers, suppliers, competitors and technologies on a scale 1 to 10
- * measure of uncertainty in the area of customers, suppliers, competitors and technologies on a scale 1 to 10
- * measure of competitive fight on a scale 1 to 10
- * measure of corruption in relevant market area on a scale 1 to 10
- * measure of accent on new products, services and new markets on a scale 1 to 10
- * position on the market on a scale 1 to 10
- * strategy of company (includes four categories transformed to 3 dummy variables) binary

2 Canonical Correlation Analysis

Let $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ stands for the left set of variables and $\mathbf{Y} = (Y_1, Y_2, \dots, Y_q)'$ stands for the right set of variables and let us assume that $p < q$. Let $\mathbf{a}_i = (a_{i1}, \dots, a_{ip})$, $i = j, \dots, p$ and $\mathbf{b}_i = (b_{i1}, \dots, b_{iq})$, $i = j, \dots, p$ stand for vectors of real numbers. Then the model is specified:

$$\begin{array}{ll} I. & a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p = U_1 & V_1 = b_{11}Y_1 + b_{12}Y_2 + \dots + b_{1q}Y_q \\ II. & a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p = U_2 & V_2 = b_{21}Y_1 + b_{22}Y_2 + \dots + b_{2q}Y_q \\ & \vdots & \vdots \\ p. & a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p = U_p & V_p = b_{p1}Y_1 + b_{p2}Y_2 + \dots + b_{pq}Y_q \end{array}$$

such that

1. \mathbf{a}_1 and \mathbf{b}_1 satisfy the condition that the random variables $\mathbf{a}_1\mathbf{X}$ and $\mathbf{b}_1\mathbf{Y}$ maximize the correlation $R_{(C)1} = \text{cor}(\mathbf{a}_1\mathbf{X}, \mathbf{b}_1\mathbf{Y})$. The random variables $U_1 = \mathbf{a}_1\mathbf{X}$ and $V_1 = \mathbf{b}_1\mathbf{Y}$ are said to be the *first pair of canonical variables*.
2. \mathbf{a}_r and \mathbf{b}_r , $r = 2, \dots, p$ satisfy the condition that the random variables $\mathbf{a}_r\mathbf{X}$ and $\mathbf{b}_r\mathbf{Y}$ maximize the correlation $R_{(C)r} = \text{cor}(\mathbf{a}_r\mathbf{X}, \mathbf{b}_r\mathbf{Y})$ and for $r \neq s$ it is true that $R(U_r, U_s) = 0 = R(V_r, V_s)$. (It can be proved that then $R(U_r, V_s) = 0$.) The random variables $U_r = \mathbf{a}_r\mathbf{X}$ and $V_r = \mathbf{b}_r\mathbf{Y}$ are said to be the *r-th pair of canonical variables*.

$R_{(C)r} = R(U_r, V_r)$ is said to be *r-th canonical correlation coefficient*, $r = 1, \dots, p$. Its significance can be tested via Bartlett χ^2 test under the assumption of $p + q$ multivariate normal distribution.

To assess the association between two sets of variables not only canonical correlation coefficients must be taken into account. A *total redundancy* is another information about association between two sets. It expresses the proportion of variability of the left (resp. right) set which can be explained by the canonical variables of the right (resp. left) set. $\sum_{r=1}^p \sum_{i=1}^p \frac{R^2(X_i, U_r)}{p} \cdot R_{(C)r}^2$ is a formula for the total redundancy for the left set of variables.

3 Results

The main purpose of applying canonical correlation analysis was to find out weather and to which extent financial performance can be explained by indicators listed in the paragraph 1.2.

EA and *ROA* represent the left set of variables, $p = 2$; other indicators from 1.2 represent the right set, $q = 17$. Quantitative character is met by all variables except those four which are binary; linear fit perform almost all pairs of variables, which may support assumption of multivariate normality. The large number of considered variables leads to undesirable sparse data matrix with only 180 wise cases (out of 336).

Referring to Table 1 it is clear that almost 15% of variance of the left set can be explained by the right set or, to be more exact, by right canonical variables. These represent only around 13% of variance of the right set which is comprehensible due to dissimilarity of the values p and q . Bearing this in mind the values of both canonical correlation coefficients can not be overrated even though the first coefficient $R_{(C)1} = 0,4159436$ is significant.

4 Conclusion

In the area of business economics the strong relationships are not very common. For that reason the result of 15% of explained financial performance

TABLE 1. Main results:

 $R_{(C)1} = 0,4159436$, $p = 0,0168$; $R_{(C)2} = 0,3471129$, $p = 0,15332$

	No. of vars.	Variance extracted	Total redundancy given the other set
Left set:	2	100,00000000%	14,7215%
Right set:	17	12,7688%	1,81552%

variability is powerful one. The soft pair correlations between any variable from the left set with any variable from the right set may suggest, that financial performance is not associated with above mentioned indicators. But canonical correlation analysis examining the complexity of relations shows that indicators from 1.2 can to some extent "predict" values of financial performance.

References

- Blažek, L. a kolektiv (2011). *Nadnárodní společnosti v České republice II*. Brno: Masarykova Univerzita.
- Hebák, P., Hustopecký, J., Pecáková, I., Průša, M., Řezanková, H., Svobodová, A., Vlach, P. (2007). *Vícerozměrné statistické metody (3)*. Praha: Informatorium.
- Meloun, M., Militký, J., Hill, M. (2005). *Počítačová analýza vícerozměrných dat v příkladech*. Praha: Academia.
- Monterio, L.F., Arvidsson, N. and Birkinshaw, J. (2008). Knowledge flows within multinational corporations: explaining subsidiary isolation and its performance implications. *Organization Science*, **19** No. 1, pp. 90–107.

Semi-parametric regression models with reliability data

Antonio Jesús López-Montoya ¹, M. L. Gámiz-Pérez ¹

¹ Departamento de Estadística e I.O., Universidad de Granada, Spain

E-mail for correspondence: ajlm@correo.ugr.es

Abstract: In this work we consider linear regression models to deal with filtered data in the context of Reliability analysis. Our motivation to evaluate the risk of failure in water supply networks. To do it we analyze a dataset consisting of breakdown data of the pipes in a network laid in a small city of the Mediterranean sea. Due to the nature of our data set, we firstly tried to fit a Cox model, however this model is not appropriate since in our case the proportional hazards assumption does not hold. In consequence, we propose a semi-parametric accelerated failure time model which allows estimation and inferences about the parameters of the model, without assuming a particular distribution for the failure time random variable. Furthermore, implementation and interpretation of the results is easy.

Keywords: Accelerated failure time model; Non-proportional hazards; Kaplan-Meier weights; Weighted least-squares; Jackknife estimator.

1 Introduction

In Reliability analysis, we usually have to deal with right censored and left truncated observations, see Lawless (1982), which do not allow us to use the general statistical methods with this type of data. As a consequence, specific models for failure times analysis have been developed in the recent literature. In this sense, Cox proportional hazard model (CPH) and accelerated failure time model (AFT) are frequently used in applications, for example, Mailhot et al. (2000), Christodoulou et al. (2010), Debón et al. (2010) and Carrión et al. (2010). The Cox model is based on the proportional hazards assumption which may not hold in many cases of reliability data. To account with such situations, AFT models are proposed. These models are of interest because they can be written specifying a direct relation between the failure time and the covariates. Their main disadvantage is that the estimation of these models is usually carried out by assuming a parametric distribution for the duration, which in most cases is very restrictive.

Stute (1996) considers a semi-parametric AFT model and presents a procedure to estimate the parameters of the model and to make inferences

without assuming any distribution family for the failure time variable. The main objective of this study is to promote the use of nonparametric tools for evaluating the risk of failure in a water supply system. To do so, we consider the procedure suggested by Stute (1996). That is, we consider a semi-parametric AFT that directly links the failure time of a pipe to its particular characteristics.

This paper is structured as follows. In Section 2 we present the semi-parametric model and the estimation procedure. Section 3 describes our dataset. In Section 4, we analyse the results obtained from fitting the AFT (semi-parametric and parametric) model. Finally, we discuss the conclusions of this work.

2 The model and estimation procedure

Let T denote the failure time whose distribution function is denoted by F . We assume that

$$\ln T = X\beta + \varepsilon. \quad (1)$$

where $X = [\mathbf{X}_1, \dots, \mathbf{X}_p]$, $\beta = (\beta_1, \dots, \beta_p)'$, and $\varepsilon = \sigma Z$ where Z is the random variable which describes the random behavior of $\ln T$.

In most cases, the data at hand are right censored. Which means that the actual failure time T is not always observable and instead we observe

$$Y_i = \min(T_i, C_i), \quad \delta_i = \begin{cases} 1; & \text{if } T_i \leq C_i \\ 0; & \text{if } T_i > C_i \end{cases}$$

where C_1, \dots, C_2 are the values of the censoring variable C , which is assumed independent of T , and δ_i is an indicator of whether T_i has been observed or not.

In Reliability analysis, the most commonly used parametric regression models are exponential, Weibull, log-normal, log-logistic and gamma, moreover, the exponential and Weibull regression models can be considered as particular cases of both the AFT and CPH models. In contrast, we use a similar method to the one proposed by Stute (1993), which requires very general hypotheses and where the estimators can be obtained using a weighted least squares method, that is, we use model (1) under the assumption that $E[\varepsilon|X] = 0$. We assume that the relation between the covariates and the duration, or some monotonic transformation of this, such as, for example, the logarithmic one, is considered linear. Under this model, the estimator of β can be obtained by minimizing

$$\sum_{i=1}^n W_{in} (\ln Y_{(i)} - \mathbf{X}_{[i]}\beta)^2 \quad (2)$$

where $\ln Y_{(i)}$ is the i -th ordered value of the observed response variable $\ln Y$, $\mathbf{X}_{[i]}$ is the covariable associated with $\ln Y_{(i)}$ and W_{in} are the Kaplan-Meier

weights. These weights can be computed as

$$W_{in} = \widehat{F}_n(\ln Y_{(i)}) - \widehat{F}_n(\ln Y_{(i-1)}) = \frac{\delta_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \left[\frac{n-j}{n-j+1} \right]^{\delta_{[j]}}$$

where \widehat{F}_n is the estimator of the distribution function F for the variable T by Kaplan and Meier (1958) and $\delta_{[i]}$ is the δ value associated with $\ln Y_{(i)}$. In this way, after calculating the W_{in} weights, the minimization of (2) leads to the estimator of β given by

$$\widehat{\beta} = (X'WX)^{-1} X'W \ln Y$$

where $\ln Y = (\ln Y_{(1)}, \dots, \ln Y_{(n)})'$, W is a diagonal matrix with the Kaplan-Meier weights and $X = [\mathbf{X}_1, \dots, \mathbf{X}_p]$. Stute (1996) studies the consistency of this estimator and its asymptotic normal distribution. The author also proposed the use of a simpler jackknife estimator to calculate the asymptotic variance.

3 Breakdown dataset of a water supply network

In order to assess the semi-parametric approach, we had access to data from a water supply company of a medium-sized Spanish city. The dataset includes 655 entries, that report the following information for each section of pipe: Identification number, section diameter, pressure, installation year, pipe material, date and type of failure, section length, traffic conditions. After a previous analysis of the dataset not reported here, we found it convenient to construct two artificial covariates: the first one is related to the installation year, is denoted as `x80` and takes the value of 1 if the pipe was installed after 1980, and 0 otherwise. The second one is defined starting from the length and diameter (physical dimensions). Thus we created a new covariate called `volume`, which is the volume of a section line that we must study as recommended in recent studies in the field of hydraulic engineering.

We have only considered the pipes that were installed after 1940. According to the dataset two types of different material have been employed: ductile cast iron and asbestos cement. The dataset also includes traffic conditions of the installation area of the pipes, considering three types of traffic: under sidewalk, normal traffic, and heavy traffic.

We have a medium censoring rate, up to 50%, and we assume an independent and non informative censoring scheme, see Fleming et al. (2002).

On the other hand, since we have only the year in which the fault occurs, the time is calculated as number of years, and therefore, failures with time equal to 0 represent a problem to operate with them have when implementing the algorithm, so we solve problem by replacing 0 by 1/365, assuming that the pipes have lasted for at least one day.

4 Results

We have used the R software environment, see R Development Core Team (2011) for the whole computations of our work. First of all we have explored the PH assumption. If we have a look at a FIGURE 1, we can say that the PH assumption does not hold in our case, which justifies that we use an AFT model to explain our data. In TABLE 1, we present the results

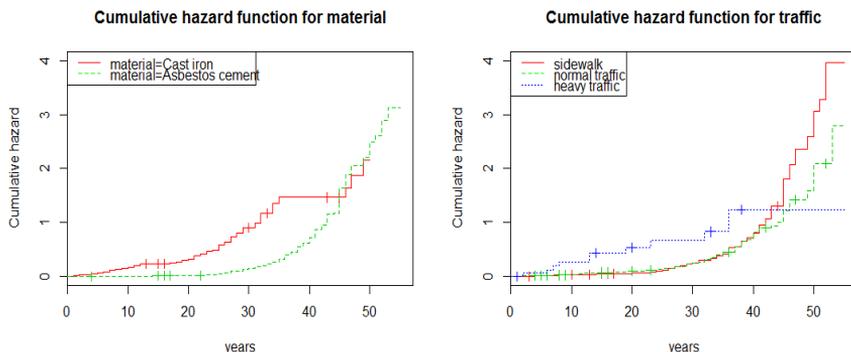


FIGURE 1. Cumulative hazard functions for risk groups by material and traffic.

of fitting different parametric AFT models as well as a semi-parametric model, including the value of the coefficients for each covariates and their standard error. These coefficients correspond to the covariates: **asbestos cement**, **normal traffic**, **x80** and **volume**.

The semi-parametric approach method seems to be a good option to address this problem. This method has advantages over the parametric models because the latter can be affected by a wrong specification of the probability distribution of the failure time. In TABLE 1, we can observe the slight difference presented by the standard errors of covariates between parametric model and the semi-parametric version.

We interpret the coefficients of the log-logistic model (for example) by its exponential form $\exp(\beta)$. In this way, for covariate **volume**, an increment of $1 m^3$ in a pipe means a failure time which is 1.61% shorter.

For covariate **asbestos cement**, the failure time of an asbestos cement pipe is a 27.39% longer than a pipe made of ductile cast iron.

With respect to covariate **normal traffic**, a normal traffic increases time by 7.03% compared to a pipe installed under heavy traffic.

For the covariate **x80** failure time for pipes installed before 1980 is 54.29% of the failure time of pipes installed after date.

To interpret the coefficients of the other parametric and semi-parametric models we can proceed similarly. The study has sought to provide insight into the impact of different covariates on the risk of failure in water supply networks. The analysis described above showed that pipes which were less

TABLE 1. Estimates of the parameters and their standard errors for parametric and semi-parametric AFT model.

covariate	log-normal AFT
Intercept	3.4616 (0.0601)
asbestos cement	0.2742 (0.0558)
normal traffic	0.0840 (0.0470)
volume	-0.0174 (0.0139)
x80	-0.9743 (0.0624)
log-logistic AFT	Semi-parametric AFT
3.4328 (0.0450)	3.2556 (0.0514)
0.2421 (0.0421)	0.2907 (0.0459)
0.0680 (0.0356)	0.0625 (0.0402)
-0.0163 (0.0105)	0.0024 (0.0131)
-0.7830 (0.0513)	-1.1465 (0.0747)

prone to failure had the following characteristics: smaller volumes, asbestos cement, installed under normal traffic and pipes that were installed after 1980.

5 Conclusions

When the PH assumption does not hold, the results of fitting a CPH model can lead us to wrong conclusions. In our case, the estimator obtained minimizing (2) has good properties and provides a better approach. The semi-parametric model has been compared to the parametric log-logistic and log-normal AFT models and as expected, the estimates of the semi-parametric case are less precise. It is nevertheless true, which this loss of precision is very small, in a sense, advocates the use of this methodology when the probability distribution is unknown, the most usual situation in practice.

Acknowledgments: We sincerely thanks Dr. A. Carrión, (Universidad Politécnica de Valencia, Spain) for providing the dataset.

References

Carrión A., Solano H., Gámiz M.L. and Debón A.(2010). Evaluation of the reliability of a water supply network from right-censored and left-truncated break data. *Water Resour Manage*, **24**, 2917–2935.

Christodoulou S. and Deligianni A. (2010). A neurofuzzy decision framework for the management of water distribution networks. *Water Resour Manage*, **24**, 139–156.

- Cox D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society Series B*, **34**, 187–220.
- Cox D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.
- Debón A., Carrión A., Cabrera E. and Solano H.(2010). Comparing risk of failure models in water supply networks using ROC curves. *Reliability Engineering and System Safety*, **95**, 43–48.
- Fleming, T. and Harriton, D.(2002). Counting Processes and Survival Analysis. Wiley, New York.
- Kaplan E.L. and Meier P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.
- Lawless J.F. (1982). Statistical Models and Methods for Lifetime Data Analysis. Wiley: New York.
- Mailhot A., Duchesne S., Musso E. and Villeneuve J.P.(2000). Modélisation de l'évolution de l'état structural des réseaux d'égout: application à une municipalité du Québec. *Can J Civ Eng*, **27**, 65–72.
- R Development Core Team. (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>.
- Stute W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, **45**, 89–103.
- Stute W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics*, **23**, 461–471.
- Stute W. (1996). The jack-knife estimate of variance of a Kaplan-Meier integral. *Annals of Statistics*, **24**, 2679–2704.

Inference of the symmetry point with different costs for the specificity and sensitivity

Mónica López-Ratón¹, Carmen Cadarso-Suárez¹, Elisa M. Molanes-López², Emilio Letón³

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: In modern medical practice, it is very important to know how to use a continuous biomarker to discriminate between non-diseased and diseased individuals, using an optimal cut-off value. There are several optimality criteria: the North-West corner, the Youden index, and the Symmetry point, among others. In this paper, we focus on the Symmetry point, giving a generalization that takes into account the prevalence and misclassification costs. We construct confidence intervals for this Generalized Symmetry point and the associated Specificity and Sensitivity using two approaches: one based on the Generalized Pivotal Quantity and the other on Empirical Likelihood. We perform a simulation study to check the practical behaviour of both methods and illustrate their use with a real biomedical dataset.

Keywords: Box-Cox transformation; Empirical Likelihood; Generalized Pivotal Quantity; Misclassification costs.

1 Introduction

In modern medical practice, it is very important to know how to use a continuous biomarker, Y , to discriminate between non-diseased and diseased individuals. This discrimination task will be usually based on a cut-off value, c , such that depending on whether $Y \geq c$ or $Y < c$, the individual will be classified as diseased or non-diseased, respectively.

In this context, the Receiver Operating Characteristic (ROC) curve is a commonly used tool for evaluating the discrimination accuracy of Y (see, among others, Metz, 1978). The ROC curve is obtained by plotting the pairs $(1 - p(c), q(c))$, with $-\infty < c < \infty$, and where $p(c)$ denotes the Specificity (true negative rate) and $q(c)$ the Sensitivity (true positive rate). Several

ways of estimating the ROC curve have been proposed in the literature (see, for instance, a review in López-Ratón, 2004). A key point in this context is to yield a binary classification rule to be used in practice. There are several methods to do so: the North-West corner, the Youden index, and the Symmetry point, among others (see, for example, Pepe, 2003). In this paper, we will focus on the Symmetry point, c_S , defined as the point where $p(c_S) = q(c_S)$.

A relevant feature on clinical practice is not only the biomarker accuracy but also its clinical efficacy. Therefore, it is important to take into account the prevalence of disease and the costs associated with the two misclassification errors (false positives and false negatives). Hence, we will define the Generalized Symmetry point, c_{GS} , as follows:

$$r(1 - p(c_{GS})) = 1 - q(c_{GS}),$$

where $r = \frac{1-\pi}{\ell\pi}$, π refers to the prevalence of disease, and $\ell = \frac{C_{F-}}{C_{F+}}$, with C_{F-} and C_{F+} denoting the misclassification costs associated to a false negative and a false positive, respectively. When $r = 1$, the Generalized Symmetry point yields the traditional Symmetry point. If the costs were not taken into account (and consequently assumed to be equal), then $\ell = 1$ and therefore $\pi = 0.5$ for the traditional Symmetry point, which may be far from reality. In Section 2, we construct confidence intervals for c_{GS} and the corresponding pair $(p(c_{GS}), q(c_{GS}))$ using two approaches: the first one will be based on the Generalized Pivotal Quantity (GPQ) and the second one on Empirical Likelihood (EL). These two methods have been recently applied for constructing confidence intervals for the Youden index and its corresponding optimal value, see, for instance, Lai et al (2011), where the GPQ method has been considered under the normality assumption, and Molanes-López and Letón (2011), where the nonparametric EL method has been used instead. In Section 3, we check the performance of the two new methods through a simulation study. Finally, in Section 4 we analyze a well-known real example to illustrate the new methodology discussed in the paper.

2 Two new methods

In this section, we describe how to obtain percentile confidence intervals for c_{GS} and $(p(c_{GS}), q(c_{GS}))$, using the GPQ and EL methods. To make inference on c_{GS} , we consider two independent samples of i.i.d. observations, $\{Y_{0k_0}\}_{k_0=1}^{n_0}$ and $\{Y_{1k_1}\}_{k_1=1}^{n_1}$, taken from the healthy and diseased populations, Y_0 and Y_1 , respectively, with sample sizes n_0 and n_1 .

2.1 Generalized pivotal confidence interval

The methodology of generalized confidence intervals was introduced by Weerahandi (1993) and it has recently applied in the context of diagnostic

studies to the Youden index in Lai et al. (2011). Under the normality assumption, using if necessary a monotone transformation of Box-Cox type, c_{GS} can be computed from the following equation

$$\Phi\left(\frac{\Phi^{-1}(1 - rt) - a}{b}\right) - t = 0, \tag{1}$$

where $a = \frac{m_1 - m_0}{s_1}$, $b = \frac{s_0}{s_1}$, Φ the standard normal cumulative distribution function (cdf) and $t = 1 - p(c_{GS})$, with m_i and s_i denoting the sample mean and standard deviation of the two populations, $i = 0, 1$. Once the root of (1) is obtained, \hat{t} , then $\hat{c}_{GS} = \Phi^{-1}(1 - \hat{t})$.

For computing the generalized confidence interval of the Generalized Symmetry point and of their corresponding Sensitivity and Specificity indexes, we follow the same reasoning as in Lai et al. (2011), based on substituting a and b with their generalized pivotal values, $R_a = \frac{R_{m_1} - R_{m_0}}{R_{s_1}}$ and $R_b = \frac{R_{s_0}}{R_{s_1}}$, into equation (1), where $R_{m_i} = m_i - t_i \frac{s_i}{\sqrt{n_i}}$, $R_{s_i} = \sqrt{\frac{(n_i - 1)s_i^2}{V_i}}$ with $V_i \sim \chi_{n_i - 1}^2$ and $t_i \sim t_{n_i - 1}$, for $i = 0, 1$. This procedure is repeated K times, that is, for $k = 1, \dots, K$, we first generate, for $i = 0, 1$, V_{ik} and t_{ik} , we then compute $R_{m_{ik}}$, $R_{s_{ik}}$, R_{ak} and R_{bk} , and we finally obtain the root of equation (1), \hat{t}_k , and set $\hat{c}_{GS_k} = \Phi^{-1}(1 - \hat{t}_k)$.

2.2 Empirical likelihood confidence interval

Empirical likelihood (EL) was first proposed by Thomas and Grunkemeier (1975) for constructing confidence intervals for the Kaplan-Meier estimator. Nowadays, it is an active area of research in several fields because the confidence intervals and regions given by this method have several good properties.

Taking into account that 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, we follow the same reasoning as in Molanes-López and Letón (2011), and derive the following log-likelihood function to make inference on c_{GS} :

$$\begin{aligned} \ell(c) &= -2 \log L(c) \\ &= 2n_0 \hat{F}_{0,g_0}(c) \log\left(\frac{\hat{F}_{0,g_0}(c)}{p(c)}\right) + 2n_0(1 - \hat{F}_{0,g_0}(c)) \log\left(\frac{1 - \hat{F}_{0,g_0}(c)}{1 - p(c)}\right) \\ &\quad + 2n_1 \hat{F}_{1,g_1}(c) \log\left(\frac{\hat{F}_{1,g_1}(c)}{r(1 - p(c))}\right) + 2n_1(1 - \hat{F}_{1,g_1}(c)) \log\left(\frac{1 - \hat{F}_{1,g_1}(c)}{1 - r(1 - p(c))}\right), \end{aligned}$$

where \hat{F}_{i,g_i} are kernel-type estimates of the cdf's F_i , of the two populations, $i = 0, 1$.

3 Simulation study

In this section we carry out a simulation study to compare the performance of the two approaches previously introduced in the former section. In this simulation study, we discuss the interval width and interval coverage of the GPQ and EL methods. The scenarios considered are similar to Fluss et al. (2005) and Molanes-López and Letón (2011), see Table 1, where we use the notation $Y = N$ to denote that Y is normally distributed, $Y = N^{-1/3}$ to denote that $Y^{-1/3}$ is normally distributed, $Y = LN$ to denote that $\ln Y$ is normally distributed, and $Y = G$ to denote that Y is gamma distributed. Note that the first three scenarios correspond to the binormal model, that is, either Y_0 and Y_1 follow normal distributions or a Box-Cox transformation of them (see, for example, Fluss et al., 2005 and Molodianovich et al., 2006). The estimation of the Box-Cox transformation is done as specified in Molanes-López and Letón (2011). The fourth scenario is outside the Box-Cox transformation family and has been included to study the robustness of the binormal model.

TABLE 1. Parameters under the binormal and bigamma models with $\pi = 1/3$

	μ_0	μ_1	σ_0^2	σ_1^2	(p, q) corresponding to $\ell(r)$		
					0.5 (4)	1 (2)	2 (1)
N	6.5	7.6	0.09	0.25	(0.97,0.87)	(0.95,0.89)	(0.92,0.92)
$N^{-\frac{1}{3}}$	3.5	2.5	0.09	0.25	(0.94,0.77)	(0.92,0.84)	(0.89,0.89)
LN	2.5	3.2	0.09	0.25	(0.93,0.70)	(0.88,0.76)	(0.81,0.81)
	β_0	β_1	α_0	α_1	(p, q) corresponding to $\ell(r)$		
					0.5 (4)	1 (2)	2 (1)
G	2	12.0	2	2	(0.95,0.81)	(0.92,0.84)	(0.88,0.88)

TABLE 2. c_{GS} : Coverage (%), width and MSE of $CI_{95\%}(c)$ under model N

Methods		GPQ			EL		
(n_0, n_1)	$\ell(r)$	<i>cov.</i>	<i>width</i>	<i>MSE</i>	<i>cov.</i>	<i>width</i>	<i>MSE</i>
(100,100)	0.5 (4)	94.9	0.161	0.002	93.2	0.207	0.004
	1 (2)	95.2	0.151	0.002	94.4	0.190	0.003
	2 (1)	94.4	0.147	0.001	95.5	0.185	0.002

The simulations were carried out in R. For every scenario specified in Table 1, 1000 trials were considered. For each trial, a sample of $n_0 = n_1 = 30, 100$ i.i.d. observations were independently drawn from Y_0 and Y_1 , respectively. For the EL method, the Gaussian kernel, K , was considered in the required kernel type estimates. For these kernel type estimates, we considered naive bandwidth selectors to choose the required smoothing parameters. From each pair of samples, we generated $B = 999$ bootstrapped resamples to

obtain 95%-confidence intervals for the optimal cut-off point c_{GS} and the corresponding pair of Specificity and Sensitivity, through the percentile method.

For the sake of brevity, we only present in Tables 2-4 the results for model N and $n_0 = n_1 = 100$ regarding the Generalized Symmetry point and its associated cost based Specificity and Sensitivity.

TABLE 3. $p(c_{GS})$: Coverage (%), width and MSE of $CI_{95\%}(p(c))$ under model N

Methods		GPQ			EL		
(n_0, n_1)	$\ell (r)$	<i>cov.</i>	<i>width</i>	<i>MSE</i>	<i>cov.</i>	<i>width</i>	<i>MSE</i>
(100,100)	0.5 (4)	94.1	0.025	0.000	96.2	0.031	0.000
	1 (2)	93.6	0.039	0.000	95.5	0.050	0.000
	2 (1)	93.6	0.061	0.000	95.1	0.080	0.000

TABLE 4. $q(c_{GS})$: Coverage (%), width and MSE of $CI_{95\%}(q(c))$ under model N

Methods		GPQ			EL		
(n_0, n_1)	$\ell (r)$	<i>cov.</i>	<i>width</i>	<i>MSE</i>	<i>cov.</i>	<i>width</i>	<i>MSE</i>
(100,100)	0.5 (4)	94.1	0.099	0.001	96.2	0.126	0.001
	1 (2)	93.6	0.079	0.000	95.5	0.101	0.001
	2 (1)	93.6	0.061	0.000	95.1	0.080	0.000

From Tables 2-4, we observe the following:

- In terms of coverage, both methods have a good behaviour when estimating the cut-off point, the Specificity and the Sensitivity.
- In terms of width, the EL gives slightly wider confidence intervals than the GPQ. This was expected since the EL is a nonparametric method and the GPQ is a parametric one.

As a concluding remark, the EL approach is competitive with the GPQ approach, with the advantage of not requiring any parametric assumption but with the disadvantage of being more time consuming.

4 Example

In this section, we analyze the dataset on prostate cancer studied in Le (2006), considering a prevalence of 0.15 and two different values for the costs: $\ell=0.5$ and $\ell=2$. Due to the fact that these data do not follow the Box-Cox family, we will only show the 95%-confidence intervals obtained with the EL approach. For $\ell=0.5$, the results are $CI(c_{GS}) = [76; 130]$, $CI(p(c_{GS})) = [0.91; 0.95]$, $CI(q(c_{GS})) = [0.00; 0.38]$. For $\ell=2$, the results are $CI(c_{GS}) = [67; 104]$, $CI(p(c_{GS})) = [0.73; 0.95]$ and $CI(q(c_{GS})) = [0.23; 0.85]$. We observe that as C_{F-} increases with respect to C_{F+} , the cutpoint decreases, and consequently, the Specificity decreases and the Sensitivity increases.

Acknowledgments: This research has been supported by several Grants from the Spanish Ministry of Science & Innovation. M. López-Ratón and C. Cadarso-Suárez acknowledges support to MTM2008-01603, 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 SEJ2007-64500, TIN2009-09158 and MTM2010-09213-E.

References

- Fluss, R., Faraggi, D. and Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal*, **47**, 458–472.
- 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.
- Le, C.T. (2006). A solution for the most basic optimization problem associated with an ROC curve. *Statistical Methods in Medical Research*, **15**, 571–584.
- López-Ratón, M. (2004). Revisión de las técnicas de estimación e inferencia de las Curvas ROC. Aplicación a sistemas de diagnóstico asistido por ordenador en Radiología. *Supervised Research Work*, Department of Statistics and Operations Research, USC, Santiago de Compostela.
- Metz, C.E. (1978). Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine*, **8**, 183–298.
- 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.
- Molodianovitch, K., Faraggi, D. and Reiser, B. (2006). Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. *Biometrical Journal*, **48**, 745–757.
- Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press.
- 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.

Model uncertainty and multimodel inference in reliability estimation

Annouschka Laenen¹

¹ Dept. of methodology and Statistics, Maastricht University, The Netherlands

E-mail for correspondence: `Annouschka.Laenen@med.kuleuven.be`

Abstract: Laenen et. al (2007, 2009) proposed a method to assess the reliability of rating scales in a longitudinal context. The methodology is based on hierarchical linear models and reliability coefficients are derived from the corresponding covariance matrices. Frequently, several models fit the data equally well, rising the problem of model selection uncertainty. When model uncertainty is high one may resort to model average, where inferences are not based on one but on an entire set of models. We explored the use of different model building strategies, including model averaging in reliability estimation. We found that the approach introduced by Laenen *et al* (2007, 2009) combined with such strategies may yield meaningful results in presence of high model selection uncertainty and when all models are misspecified, as far as some of them manage to capture the most salient features of the data. Nonetheless, when all models omit prominent regularities in the data, misleading results may be obtained.

Keywords: Reliability, Clinical Trials, Linear Mixed Models, Model Averaging

1 Introduction

The theory of parametric statistical inference assumes that the model used for the analysis is correct and known to the researcher beforehand. Nevertheless, in practice, many aspects of the model like its functional form, the number of random effects or the error structure are determined only after the data have been observed. Consequently, the actual statistical properties of estimators or inferential procedures following a model selection step, may differ substantially from the properties predicted by the traditional theory, Potscher (1991). Ignoring the additional uncertainty emanating from model selection has been called by Breiman (1992) the quiet scandal in statistics. The soundness of any reliability measure cannot surpass the soundness of the model it is predicated on. Ideally, one would like to fully establish a model based on substantive knowledge and theories. Nonetheless, this may be almost impossible when dealing with complex data structures in highly complex fields like psychiatry and psychology. In the present paper we study, via simulations, the impact of model building in the estimation of reliability within a longitudinal framework. We explore three different

types of alternative analyses: the naive approach that fully ignores the model building step, the procedure introduced by Buckland, Burnham and Augustin (1997) that takes this step into account when carrying out the inferences and model average methodology that uses all models to carry out inferences along the lines presented in Burnham and Anderson (2002). Notice that, given the huge difficulties one encounters when trying to settle these issues theoretically, the use of simulations that mimic as close as possible specific areas of application, are of utmost importance to get insight into the impact of model building on the estimation of reliability.

2 Reliability in a longitudinal context

Laenen *et al* (2007, 2009) proposed to assess reliability in a longitudinal context based on the model

$$\mathbf{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_i, \quad (1)$$

where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{ip})'$ and Y_{ij} denotes the response of subject i at time point j . Furthermore, $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{ip})'$ is a vector capturing the systematic evolution of the response over time and $\boldsymbol{\tau}_i = (\tau_{i2}, \tau_{i2}, \dots, \tau_{ip})'$ is a vector describing the time evolution of a latent subject-specific attribute. Finally, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{ip})'$ is the measurement error component. At a second level these components are further modeled as $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\tau}_i = \mathbf{Z}\mathbf{b}_i$. In this second level the systematic evolution is characterized using a set of q covariates, such as treatment, time, or hospital, that are contained in the $(p \times q)$ design matrix \mathbf{X}_i , with $\boldsymbol{\beta}$ a q -dimensional vector of fixed effects. Similarly, the evolution of the subject-specific attribute is modeled as a linear function of time by considering the $(p \times r)$ design matrix \mathbf{Z} and a r -dimensional vector \mathbf{b}_i of subject-specific effects.

Furthermore, these authors split the error term in two parts $\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}_{(1)i} + \boldsymbol{\varepsilon}_{(2)i}$, where the first one contains a serial correlation component and the second one a residual error. The model is completed by assuming that $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, $\boldsymbol{\varepsilon}_{(1)i} \sim N(\mathbf{0}, \tau^2 \mathbf{H}_\rho)$, $\boldsymbol{\varepsilon}_{(2)i} \sim N(\mathbf{0}, \mathbf{R})$ and \mathbf{b}_i , $\boldsymbol{\varepsilon}_{(1)i}$ and $\boldsymbol{\varepsilon}_{(2)i}$ independent.

Marginally the previous model implies $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V})$ where $\mathbf{V} = \boldsymbol{\Sigma}_D + \boldsymbol{\Sigma}$ with $\text{Var}(\boldsymbol{\tau}_i) = \boldsymbol{\Sigma}_D = \mathbf{Z}\mathbf{D}\mathbf{Z}'$ and $\text{Var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma} = \tau^2 \mathbf{H}_\rho + \mathbf{R}$. Notice that the total variability is decomposed into two parts with the first one ($\boldsymbol{\Sigma}_D$) accounting for the variability induced by the subject-specific attribute and the second one ($\boldsymbol{\Sigma}$) including all the remaining sources of variability. This decomposition is crucial for the study of reliability, which is essentially based on the relative size of these two sources of variability.

Laenen *et al* (2007, 2009) defined reliability in a longitudinal context using an axiomatic approach. Two instances that emanate from this definition are the so-called R_T and R_Λ coefficients. Assuming model 1, the first one

can be written as

$$R_T = 1 - \frac{\text{tr}(\boldsymbol{\Sigma})}{\text{tr}(\mathbf{V})}$$

and should be interpreted as the *average* reliability over the different measurement occasions. This coefficient quantifies the average amount of information that a single measurement conveys about the subject-specific attribute (Laenen *et al* 2007). The second coefficient is given by the expression

$$R_\Lambda = 1 - |\boldsymbol{\Sigma}\mathbf{V}^{-1}|,$$

and bears a quite different interpretation. Indeed, unlike R_T that only uses the diagonal element of the the covariance matrices, R_Λ is based on both the diagonal and off diagonal elements. Therefore, R_Λ quantifies the amount of information about the subject-specific attribute conveyed not only by the variances but also by the association structure. In that sense it may be said that R_Λ expresses the total reliability of the entire longitudinal sequence (Laenen *et al* 2009).

2.1 Simulation Study

When modeling psychological and psychiatric outcomes, it might frequently occur that the data generating mechanism (DGM) is rather complex, with influencing factors ranging from very important to nearly irrelevant. Usually, the statistical analysis will detect the most prominent effects, while smaller effects will only be discerned when sufficiently large data sets are available. Actually, some minor sources of variability might be nearly imperceptible and, therefore, be systematically omitted from the analysis. To mimic this scenario we considered the following DGM

$$\begin{aligned} Y_{ij} &= \mu_{ij} + \tau_{ij} + \varepsilon_{ij1} + \varepsilon_{ij2}, \\ \mu_{ij} &= \beta_0 + \beta_1 t_j + \beta_2 t_j^2 + \beta_3 t_j^3 + \beta_4 \log(t_j), \\ \tau_{ij} &= b_{0i} + b_{1i} t_j + b_{2i} t_j^2 + b_{3i} \log(t_j), \end{aligned}$$

where Y_{ij} denotes the response of subject i at time t_j , the vector of systematic effects is fixed at $\boldsymbol{\beta} = (40, 20, -3, -0.05, -5)'$ and

$$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}) \quad \text{with} \quad \mathbf{D} = \begin{pmatrix} 15 & 8 & -5 & 0.5 \\ 8 & 15 & -2.1 & 0.2 \\ -5 & -2.1 & 6 & 0.2 \\ 0.5 & 0.2 & 0.2 & 0.8 \end{pmatrix}.$$

Further, $\varepsilon_{i1} \sim N(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = \text{diag}(\sigma_j^2)$ where $\sigma_1^2 = 15$ and $\sigma_{j+1}^2 = \sigma_j^2 + 5$ and $\varepsilon_{i2} \sim N(\mathbf{0}, \tau^2 \mathbf{H}_\rho)$ where $\tau^2 = 20$ and \mathbf{H}_ρ is a first order autoregressive correlation matrix with $\rho = 0.2$.

There are six simulation settings, considering three different sample sizes (50, 100, 300) together with two different numbers of repeated measurements (4, 6) and in each setting 500 data sets are generated.

TABLE 1. Five models fitted to the data.

Model	Fixed eff. structure	Rand. eff. structure	Error structure
1	saturated	intercept	simple
2	saturated	intercept	autoregressive
3	saturated	intercept	heterogeneous autoregressive
4	saturated	intercept, time	simple
5	saturated	intercept, time	autoregressive

2.2 Analysis of the Simulation Study

Table 1 displays the five models fitted to each of the simulated data sets. Notice that all these models assume a saturated mean structure in order to reduce bias in the estimation of the covariance parameters. Furthermore, they comprise covariance structures of different complexity, trying to find a balance between the model for the time evolution of the subject-specific attribute and the error component.

Table 2 presents the true values of R_T and R_Λ , as well as their average point estimate over the 500 data sets. These results are based on the best model (BM) and model averaging (MA) strategies. Besides the average point estimates, we also present the coverage probabilities (CP) for the approximate 95% confidence interval. CP is computed as the percentage of times in which the true value of the coefficient lies within the estimated 95% confidence interval. Based on the best model the confidence intervals were computed using the naive approach (CP_{na}), that fully ignores the model building step, and the method of Buckland, Burnham and Augustin (1997) that does take this step into account (CP_{ad}).

The point estimates reveal a slight underestimation in most of the settings, what may be ascribed to the fact that all models were misspecified. Remarkably, this underestimation was in general relatively mild. We should take into account that in some settings oversimple models, which represent severe misspecifications of the true DGM, were chosen and still the estimation of the reliability coefficients seems to be relatively robust in these situations.

In addition, the coverage probabilities of the approximate confidence intervals are below the nominal 95% level, with results worsening as the sample size increases. Once more, these results may be better understood when they are placed into a general theoretical framework. Two important factors may contribute to the undercoverage observed for the confidence intervals. First, notice that all the models used for the analysis are misspecified and under misspecification the asymptotic distribution of the maximum likelihood estimators is not center at the true value of the parameters (White 1982). Second, the model building step may also decrease the coverage

TABLE 2. Average point estimates and coverage probabilities (CP) for R_T and R_Λ in various settings and based on the best model (BM) and model averaging (MA).

p		4			6		
n		50	100	300	50	100	300
	R_T	.606	.606	.606	.780	.780	.780
BM	\widehat{R}_T	.601	.573	.531	.744	.733	.744
	CP_{na}	73.6	69.0	69.6	92.0	89.6	66.8
	CP_{ad}	84.8	82.0	83.6	94.0	92.6	73.6
MA	\widehat{R}_T	.596	.573	.536	.745	.735	.745
	CP	83.0	81.4	83.4	94.0	82.8	74.2
	R_Λ	.888	.888	.888	.976	.976	.976
BM	\widehat{R}_Λ	.861	.828	.787	.952	.943	.957
	CP_{na}	69.6	61.2	56.4	91.6	86.8	45.6
	CP_{ad}	86.2	82.2	80.2	94.0	96.2	83.6
MA	\widehat{R}_Λ	.854	.827	.793	.953	.944	.957
	CP	81.2	77.6	78.4	93.8	96.2	83.8

probabilities (Leeb and Pötscher 2005).

Let us then finally look at the differences between the two inference strategies under investigation. Regarding the point estimates the best model and the model average approach perform similarly, even in the settings where a lot of model uncertainty was observed. Nevertheless, when we compare the coverage probabilities from the model average approach and the *naive* best model approach we observe that the former leads to better results. In addition, the results from the two approaches, MA and adjusted model selection, are largely equivalent. Burnham and Anderson (2002) concluded that, when model averaging is used, it often gives a more honest measure of precision and reduces bias compared to estimators from just the selected best model. The results of our simulation study confirms the first but not the second conclusion for the specific case of reliability estimation through the R_T and R_Λ coefficients.

Finally, Table 3 presents the average R_T estimates for the models ranked according to their model fit (AIC). Similar estimates are obtained from the best and the second best model and the results emanating from them give a good general idea about the true reliability of the measurements. However, from the third best model onwards the estimates deviate more from the true values and the results are no longer trustworthy.

TABLE 3. Average \hat{R}_T for ranked models.

p	4			6		
	50	100	300	50	100	300
\hat{R}_T	.606	.606	.606	.780	.780	.780
Best	.601	.573	.531	.744	.733	.744
Second best	.569	.561	.583	.755	.773	.776
Third best	.332	.260	.206	.441	.453	.459
Fourth best	.377	.359	.289	.455	.444	.471
Fifth best	.467	.476	.449	.494	.499	

References

- Akaike, H. (1992). Information theory and extension of the maximum likelihood principle. In Kotz, S. and Johnson N.L. (eds) *Breakthroughs in Statistics*, vol. 1. London: Springer-Verlag.
- Breiman, L. (1992), The little bootstrap and other methods for dimensionality selection in regression: X-Fixed predictor error. *Journal of the American Statistical Association* **87**, 738-754.
- Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1997). Model selection: an integral part of inference. *Journal of Applied Ecology*, **30**, 478-495.
- Burnham, K.P., and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach*. 2d ed. New York: Springer-Verlag.
- Laenen, A., Alonso, A., and Molenberghs, G. (2007). A measure for the reliability of a rating scale based on longitudinal clinical trial data. *Psychometrika*. **72**, 443-448.
- Laenen, A., Alonso, A., Molenberghs, G., and Vangeneugden, T. (2009). Reliability of a longitudinal sequence of scale ratings. *Psychometrika*. **74**, 49-64.
- Leeb, H. and B. M. Pötscher (2005): Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 21-59.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.
- Pötscher, B. M. (1991): Effects of model selection on inference. *Econometric Theory* **7**, 163-185.

Generative linear mixture modelling

Antony Lawson¹, Jochen Einbeck¹

¹ Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, England

E-mail for correspondence: `jochen.einbeck@durham.ac.uk`

Abstract: For multivariate data with a low-dimensional latent structure, a novel approach to linear dimension reduction based on Gaussian mixture models is proposed. A generative model is assumed for the data, where the mixture centres (or ‘mass points’) are positioned along lines or planes spanned through the data cloud. All involved parameters are estimated simultaneously through the EM algorithm, requiring an additional iteration within each M-step. Data points can be projected onto the low-dimensional space by taking the posterior mean over the estimated mass points. The compressed data can then be used for further processing, for instance as a low-dimensional predictor in a multivariate regression problem.

Keywords: EM; Dimension Reduction; Mixture Modelling.

1 Introduction

Mixtures of exponential family distributions are often used to model complex data structures, with finite Gaussian mixtures being the most common representant of such models. In this article we are interested in situations where a multivariate data set, $x_i \in \mathbb{R}^m$, $i = 1, \dots, n$, possesses a latent structure of lower dimension $d < m$ (these ‘data’ may play the role of a ‘predictor space’ in a multivariate regression problem, but this is not relevant for the moment). The objective, for now, is to recover the latent structure, and to compress the original data by projecting them (in some form) onto the estimated latent space. As a first step towards a more general handling of this problem, we consider a simplified scenario in which the latent structure is thought to be a straight line, say $\alpha + \beta z$, with $\alpha, \beta \in \mathbb{R}^m$, $z \in \mathbb{R}$, through an m -dimensional space. The variable z is considered as a random effect, and represented by a discrete distribution with mass points $z_k \in \mathbb{R}$ and masses π_k , $k = 1, \dots, K$. The data are assumed to be generated by adding Gaussian noise $\varepsilon_i \sim N(0, \Sigma)$ to mixture centres $\alpha + \beta z_k \in \mathbb{R}^m$ positioned along this line, yielding the generative linear mixture model

$$x_i = \alpha + \beta z_k + \varepsilon_i. \tag{1}$$

The variance matrix $\Sigma \in \mathbb{R}^{m \times m}$ is assumed to be of diagonal form $\text{diag}(\sigma_j^2)_{\{1 \leq j \leq m\}}$, and to be the same for all K components of the mixture.

2 The EM Algorithm

As for univariate mixtures, the data likelihood, L , can be written in the form

$$L = \prod_{i=1}^n \sum_{k=1}^K f_{ik} \pi_k$$

where, for model (1),

$$f_{ik} = f(x_i | z_k) = \frac{1}{|\Sigma|^{\frac{1}{2}}} \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}(x_i - \alpha - \beta z_k)^T \Sigma^{-1} (x_i - \alpha - \beta z_k)\right).$$

In order to setup an EM algorithm, we need to consider the complete data likelihood, which is the likelihood of the data given that we know the component each x_i belongs to. However, the components each datum belongs to are unobservable, so we must use the posterior probabilities that x_i belongs to component k , which are obtained as

$$\omega_{ik} = \frac{f_{ik} \pi_k}{\sum_{l=1}^K f_{il} \pi_l}.$$

The complete data likelihood therefore takes the form

$$L^* = \prod_{i=1}^n \prod_{k=1}^K (f_{ik} \pi_k)^{\omega_{ik}},$$

giving the complete log-likelihood

$$\begin{aligned} \ell^* = \log(L^*) &= \sum_{i=1}^n \sum_{k=1}^K \omega_{ik} \log(\pi_k) + \sum_{i=1}^n \sum_{k=1}^K -\frac{1}{2} \omega_{ik} \log(|\Sigma|) \\ &+ \sum_{i=1}^n \sum_{k=1}^K -\omega_{ik} \frac{m}{2} \log(2\pi) + \sum_{i=1}^n \sum_{k=1}^K -\frac{1}{2} \omega_{ik} (x_i - \alpha - \beta z_k)^T \Sigma^{-1} (x_i - \alpha - \beta z_k) \end{aligned}$$

Score equations were obtained by partially differentiating ℓ^* with respect to each of the variables. Although an analytical solution was not obtained for α, β and z_k , we were able to find an iteration process involving these parameters. Solving the score equations for α and z_k give

$$\hat{z}_k = \frac{1}{m} \sum_{j=1}^m \left(\frac{\sum_{i=1}^n \omega_{ik} x_{ij}}{\sum_{i=1}^n \omega_{ik}} - \hat{\alpha}_j \right) / \hat{\beta}_j$$

(with the subscript j denoting the j -th component of the respective vector), and

$$\hat{\alpha} = \frac{1}{n} \left(\sum_{i=1}^n x_i - \hat{\beta} \sum_{i=1}^n \sum_{k=1}^K \omega_{ik} \hat{z}_k \right).$$

Substituting $\hat{\alpha}$ into the equation for $\hat{\beta}$ and solving gives

$$\hat{\beta} = \frac{\sum_{i=1}^n \sum_{k=1}^K \omega_{ik} x_i \hat{z}_k - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n \sum_{k=1}^K \omega_{ik} \hat{z}_k \right)}{\sum_{i=1}^n \sum_{k=1}^K \omega_{ik} \hat{z}_k^2 - \frac{1}{n} \left(\sum_{i=1}^n \sum_{k=1}^K \omega_{ik} \hat{z}_k \right)^2}.$$

To implement this in the EM algorithm, at each M-step there will be an internal iteration loop involving these parameters. First, the \hat{z}_k will be calculated using the values of the previous internal iteration. Then $\hat{\beta}$ will be calculated using the newly calculated values of \hat{z}_k . Then finally $\hat{\alpha}$ will be calculated using the new values of $\hat{\beta}$ and \hat{z}_k . The initial $\hat{\beta}$ and $\hat{\alpha}$ values used will be those from the previous E-step.

Given the new values of $\hat{\alpha}$, $\hat{\beta}$ and \hat{z}_k , the score equation for $\hat{\sigma}_j$ solves very easily to

$$\hat{\sigma}_j = \sqrt{\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \omega_{ij} (x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k)^2}$$

Using a Lagrange multiplier, $\ell^* - \lambda(\sum_{k=1}^K \pi_k - 1)$, one obtains

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^n \omega_{ik}.$$

3 Results

Analysis was carried out on the mussels data set (Bura and Cook, 2001; available from R package **dr**), considering initially the data frame constituted by the variables shell length, shell width, shell height, and shell mass. Applying the above methodology, Table 1 shows how the disparity, $-2 \log L$, of the model varies with number of components, K . The disparity decreases considerably with each component added, until the 8th component, where the disparity stabilizes at a value of 2608.088.

A bootstrapping method was required to test for a sensible number of components. Testing a model with 5 components against one with 6 returned a p-value of 0.31, and testing 4 components against 5 returned a p-value of 0.01, implying a 5 component model is a good representation of the data. The iteration loop in the M-step converges very fast, with not more than 5 iterations initially, quickly falling to 3 iterations after a few EM cycles. The number of EM iterations taken for the variables to convergence was also observed and the $\hat{\sigma}_j$ were generally the fastest to converge, with $\hat{\beta}$ converging slower, $\hat{\alpha}$ and \hat{z}_k a little slower than $\hat{\beta}$, and $\hat{\pi}_k$ considerably

TABLE 1. Table of measurements for a variety of components

K	Disparity	RSS	R^2	# Iterations for disparity to converge
2	2881.936	7.057	0.6421	7
3	2804.671	5.574	0.7767	26
4	2676.017	5.222	0.8041	13
5	2645.342	5.073	0.8151	34
6	2630.438	5.010	0.8196	102
7	2623.526	4.783	0.8356	126
8	2608.088	4.759	0.8373	145
9	2608.088	4.759	0.8373	194

slower. The disparity of the models converge somewhat faster than any of the components.

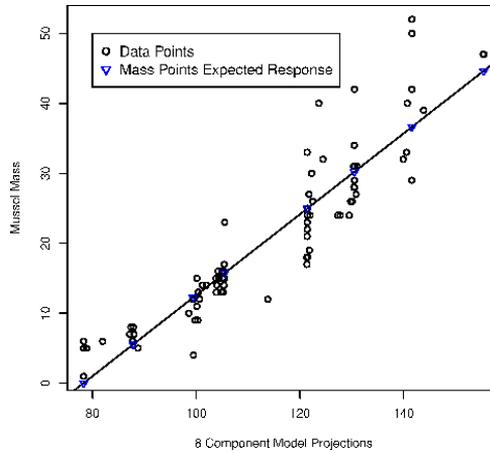
The next step taken in the analysis was projecting the data points onto the fitted line. For each data point x_i , projected (or compressed) data are obtained as ‘posterior means’ (Aitkin, 1996)

$$x_i^P = \sum_{k=1}^K \omega_{ik} \hat{z}_k.$$

These ‘projections’ are not orthogonal, and hence are of fundamentally different character as those in PCA, for instance. To verify the usefulness of this type of compression, we considered now the additional variable mussel mass as response variable, y , and fitted a simple linear regression model for y_i versus x_i^P . The resultant line is shown along with $(x_i^P, y_i), i = 1, \dots, n$ in Figure 1 and appears to represent the data reasonably well. The RSS and R^2 values for the fitted linear model were recorded for each model and are included in Table 1. It is clear that the model improves as number of components is increased. Comparing these results to the ‘parametric inverse regression’ method by Bura and Cook (2001), with RSS = 6.051 and $R^2 = 0.741$, we find the proposed mixture-based approach to perform considerably better.

4 Discussion

This article has reported on the results of a pilot study using the most simple of all latent model scenarios, namely a straight line spanned through the data which carries the mixture centres. This research has been tentatively extended in two directions: Firstly, the case of a bivariate latent structure (i.e., a plane), and secondly, the case of a ‘staggered line’ which is allowed to change its direction at each mass point. In both cases, the likelihood

FIGURE 1. Graph of mussel mass (response) modelled by projection index (x_i^P).

equations were still tractable and the algorithms converged in reasonable time, though the issue of starting point selection for the EM algorithm requires more attention with increasing complexity of the model.

The presented work could be considered as a generalization of the (linear version of the) ‘Generative topographic mapping’ (Bishop et al, 1998), where the z_k form a fixed grid, and $\pi_k = 1/K$. Using a grid to capture the latent variable distribution may require a quite large value of K , especially when considering multivariate latent structures. Since our method recovers adaptively the latent variable distribution, the value K can be kept on a far smaller level (say, 6 or 7) even for a bivariate latent space (i.e., a plane). A further interesting aspect of the proposed technique is that, due to the generative model structure, it would allow additionally for inclusion of covariates in model (1). This would be attractive, for instance, for the computation of league tables from multivariate index data. This is a matter of further research.

Acknowledgments: The first author was supported by Nuffield Bursary URB/39353.00.

References

- Aitkin, M. (1996) Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In: *Proceedings of the 11th IWSM 1996*, 87–94, Orvieto, Italy.
- Bishop, C.M J., Svensen, M. and Williams, C.K.I(1998) The Generative

Topographic Mapping. *Neural Computation*, **10**, 215–234.

Bura, E. and Cook, R.D.(2001) Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society, Series B*, **63**, 393–410.

Refined information retrieval and frequency distribution

Kenan M. Matawie¹, Sargon Hasso²

¹ School of Computing, Engineering and Mathematics, University of Western Sydney, Kingswood NSW Australia. email: K.Matawie@uws.edu.au

² Wolters Kluwer, Law & Business, Chicago, IL. USA. email: dshasso@gmail.com

Abstract: This paper presents the process of refining the document and their terms in Information Retrieval. It also shows the significance of this process prior to applying any of the information retrieval applications including probability models on the actual terms distribution. This is an important issue in language models approach, it also helps and show the effectiveness and efficiency in term of minimizing the amount of time and space required to process the data. This is also very important for probabilistic approaches such as Single Poisson, double Poisson, Binomial and Multinomial distributions which are used to define the weights in the document matching process. This approach is applied on specific data sources rather than Web pages.

Keywords: Information Retrieval; Terms Refinement; Frequency distribution.

1 Introduction

The amount of information produced, saved and available to us every day justifies the necessity for effective and efficient IR systems which can accurately meet and identify users' relevant and requested information. No doubt these IR systems also need intelligent refined techniques that will significantly contribute to reduce the amount of time and space required to process the data and the relevant information.

2 Terms Refinement

We used American National Corpus Data [openANC] to generate term matrix. In a term document matrix, there is a collection of N documents. Each document is viewed as a collection of term vectors. Each component in a document vector corresponds to each term along with its frequency count. A term document matrix, therefore, is represented as $M \times N$ matrix, call it C , whose i^{th} component represents the term, and j^{th} component represents the document [Manning et al., 2008]. The value of C_{ij} entry is the frequency of $term_i$ in $document_j$. We used Apache's Lucene search engine

[ApacheLucene] to process the ANC data and generate the term document matrix in a two-step process: indexing, where term vectors are generated for each document, and data extraction and refinement. We adapted software algorithms from McCandless et al. [McCandless] for our processes. The ANC data we have chosen consists of 6424 documents, ranges over different domains such as government, technical, travel guides, fiction, journal articles, etc., which includes over 14 million words (terms). The initial size of the M component corresponds to the total number of unique terms out of the 14 million words in the corpus. So each document's term vector, the columns in the term document matrix, will consist not only of the terms that exist in that document but also inflated with terms that don't. Obviously, this results in sparse matrix with many zero elements and containing about 1.2 billion entries. Corpus data was analyzed and several refinement processes were chosen as candidates to reduce their dimensionality. Here is a summary of the initial processes applied:

1. Raw Data: This is the baseline and constitutes the total number of all words. Reduction rate: 0%.
2. Initial Refinement: Extract unique terms, eliminate common words (stop words), and apply stemming. Stop words, like "the" and "a" are eliminated due to their prevalence. We used Porter's stemming algorithm [Porter] so that words like "happen", "happened", "happening" all reduce to one term, i.e. "happen". Reduction rate: 43%.
3. Secondary Refinement: Eliminate terms with 0 or 1 frequencies. Since each document term vector was inflated, see discussion above, it is possible that certain terms will be absent, i.e. have 0 frequency, across all documents; therefore, these terms are eliminated. Reduction rate: 58%.

3 Probability Models

Successes of the retrieval process depend on many factors including the algorithms, techniques that deal with the data and their relevance, and focuses on the users and information needed. Probability of relevance for a document given a query is one of the main task and basis for the IR models. There are many models of IR based on probability theory, such as the widely used Binomial, poisson and two poisson models [Harter 1975a,b]. This model use "relevance" as an element of the algebra of events and possibly satisfy the probability ranking principle, that ranks the documents in decreasing order with respect to the probability of relevance in relation to a given query. The significance of a word in a document collection can be identified by using the probability (Poisson)distribution. These early models for automatic indexing were based on the observation that the distribution of

Class	Base Line	No 0's or 1's	Class	Base Line	No 0's or 1's
< 10	45173608	17896608	60-70	183	183
10-20	9156	9156	70-80	95	95
20-30	2599	2599	80-90	105	105
30-40	1075	1075	100	69	69
40-50	576	576	>100	225	225
50-60	309	309			

TABLE 1. Frequency distribution profile across refinement process.

informative content words, called by Harter “specialty” words (the technical vocabulary and more informative words that is assumed to appear more densely in a few documents) over a text collection deviates from the distributional behavior of “non-specialty” words (usually are included in a stop list) are randomly distributed over the document collection. The non-specialty words are modelled by a Poisson distribution with some mean. The informative word then can be mechanically detected by measuring its deviation from a Poisson distribution. Harter’s model advanced to include second Poisson model for specialty words but on a smaller set he called them “elite” documents, with mean greater than mean of the first Poisson distribution [Harter 1975a]. In future paper we will use the Poisson models on the refined data, and redefine the “elite” document, “speciality” and “non-speciality” words then compare the outcome of the relevance before and after the data reduction.

4 Preliminary Results

This section will focus on the results obtained from the term refinement process particularly the reduction in the amount of irrelevant data, the change in data frequency distribution will contribute to more efficient parameter estimations of the probability models as discussed in the previous section. Here we will only show the refining process without going to probability models and estimations.

For the openANC data, the original documents were 6424 documents and the total terms were about 187000 terms and the matrix contained more than 1.2 billion entries. This is a very large collections, and not all the terms in the document are used and useful for indexing. Some terms have to be removed and this is usually done by the elimination of stopwords and the use of stemming as it was described above. This data is still considered too large for modeling and other purposes, so we reduced the documents to 1000 randomly selected documents from the original collection. After stopwords and stemming elimination we will have a 1000×45188 matrix. To further refine the matrix we suggest to delete all the terms that have only 0 and 1 frequencies as these terms are not contributing to any indexing, ranking

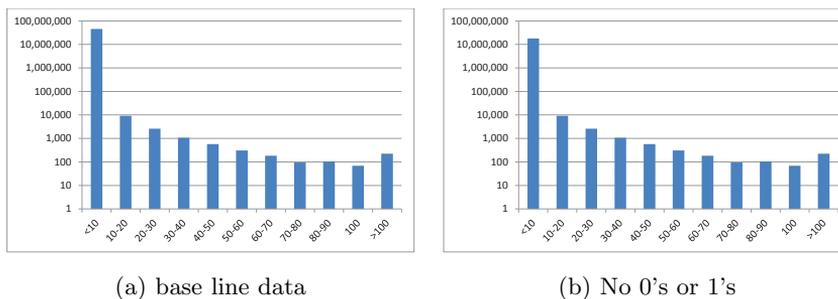


FIGURE 1. Frequency distribution diagrams corresponding to Table 1 (log scale used for y-axis).

and comparisons [Matawie, 2011]. This elimination will reduce the number of terms by almost 40% to 17910 terms. Frequency table and associated graphs is given in Table 1 and Figure 1. This reduction is significant and can help for better indexing, ranking and parameter estimations associated with probability models. Table 2 and Figure 2 below show the distribution of the frequencies in the cells, this is very useful to show and identify the single and multiple frequency nature of each term. It was clear that the single frequency term and the >14 categories formed more than 55% of the data (26.3% and 28.7% respectively), this indicate the need for further reduction investigation can be considered particularly for the single and large multiple frequencies.

5 Related Work

In our data experiment, we have already reduced or refined our terms by using stop list and stemming techniques. Further reduction requires care taken so we don't throw out terms that are part of collocational text units. A collocation is a phrase expression consisting of two or more words that has non-compositional meaning by its constituent parts, e.g. New York vs. Black Sea [Manning et al., 1999]. A related terminology to collocation is *ngrams*: sequence of n -adjacent words, e.g. *bigrams* (two-word phrases) as in "New York", *trigrams* (three-word phrases) as in "degrees of freedom". We are not interested in collocational words because these tend to pair-up (in case of bigrams) or cluster up (in case of trigrams) frequently. In other words, we are interested in finding terms that are not part of collocations. So our task is to identify these in the text corpus so that we can avoid refining, eliminating, these n -word phrases. There are several collocation discovery techniques and we list few of them [Manning et al., 1999]: 1) Frequency. This technique is good for identifying fixed phrases. It uses a simple quantitative method (frequency filter) and a small amount of linguistic knowledge (part of speech taggers) [Justeson et al. 1995]; 2) Mean

Cells frequency	(%)	Cells frequency	(%)
1	26.3%	8	2.6%
2	11.0%	9	2.3%
3	7.1%	10	1.9%
4	5.4%	11	1.7%
5	3.9%	12	1.4%
6	3.3%	13	1.3%
7	2.8%	>14	28.68%

TABLE 2. Single and Multiple Terms Frequency Distributions.

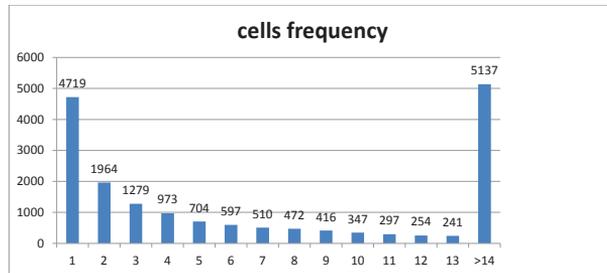


FIGURE 2. Single and Multiple Terms Frequency Distributions corresponding to Table 2.

and Variance. This technique looks at two or more words, not necessarily adjacent, and computes the mean and variance between their distance from each other in either direction, i.e. to the left or right of one word. High deviation from the mean is a sign the two words are not related. If the mean is close to zero, this also indicates that we have a uniform distribution; 3) Hypothesis Testing. We formulate a null hypothesis to test for independence between two words (bigrams). Apply *t-test* for collocation discovery and reject the null hypothesis if they don't fail the test for significance, otherwise if the null hypothesis is true, then the bigrams are not collocational; 4) Mutual Information (MI). MI for a bigram of words x and y , $I(x;y)$, computes the log of the ratio of joint probability (probability of observing words x and y together) to the probabilities of observing words x and y independently. If x and y are collocational then $I(x;y) \gg 0$, otherwise they are not, i.e. $I(x;y) \cong 0$ [Church et al. 1991].

6 Conclusion and Future Work

This paper focused on the refining process, no doubt these process will have a significant impact on the various approaches to the information retrieval particularly the probability models, such as modeling word frequencies,

two-poisson model and BM25. Further refinement using probability models is subject for future research.

References

- ApacheLucene (2011). *The Apache Software Foundation. Apache Lucene*. <http://lucene.apache.org/>. Accessed Jan 2011.
- Church, K., William, G., Hanks, P., Hindle, D.(1991). Using Statistics in Lexical Analysis. In *Lexical Acquisition: Exploiting On-Line Resources to build a Lexicon*, Hilldale, NJ (Uri Zernik, ed.), pp. 115–164.
- Dalgaard, P.(2008). *Introductory Statistics with R*. Springer. 2nd Edition. Springer.
- Harter, S.(1975a). *A probabilistic Approach to Automatic Keyword Indexing. Part I: On the Distribution of Specialty Words in a Technical Literature*. *J. ASIS*,**26**, 197–216.
- Harter, S.(1975b). *A probabilistic Approach to Automatic Keyword Indexing. Part II: An Algorithm for Probabilistic Indexing*. *J. ASIS*,**26**, 280–289.
- Hasti, T., Tibshirani, R., and Friedman, J.(2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition. Springer.
- Justeson, J. S., and Katz, S. M.(1995). Technical Terminology: Some Linguistic Properties and an algorithm for identification in Text. *Natural Language Engineering*, **1**, 9–27.
- Maning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Maning, and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.
- McCandless, M., Hatcher, E., Gospodnetić, O.(2010). *Lucene in Action*. 2nd edition. Manning Publications Co: Stamford, CT.
- Matawie, K. M. (2011). *Document Terms with Same Statistical Properties*. WASET 2011, Thailand.
- OpenANC (2010). *American National Corpus Project. Open ANC*. <http://www.americannationalcorpus.org/index.html>. Accessed Jan 2012.
- Porter, M. F.(1980). *An algorithm for Suffix Stripping*. *Program*,**14**(**3**), 130–137.
- Widdows, D.(2004). *Geometry and Meaning*. Stanford, California: CSLI Publications.

Likelihood based inference for linear and nonlinear mixed-effects models with censored response using the multivariate- t distribution

Larissa A. Matos¹, Marcos O. Prates¹, Ming H. Chen², Victor H. Lachos¹

¹ Departamento de Estatística, Universidade Estadual de Campinas, Brazil

² Department of Statistics, University of Connecticut, U.S.A

E-mail for correspondence: larissam@ime.unicamp.com

Abstract: Mixed models are commonly used to represent longitudinal or repeated measures data. An additional complication arises when the response is censored, for example, due to limits of quantification of the assay used. Normal distributions for random effects and residual errors are usually assumed, but such assumptions make inferences vulnerable to the presence of outliers. Motivated by a concern of sensitivity to potential outliers or data with tails longer-than-normal, we aim to develop a likelihood based inference for linear and nonlinear mixed effects models with censored response (NLMEC/LMEC) based on the multivariate Student- t distribution, being a flexible alternative to the use of the corresponding normal distribution. We propose an ECM algorithm for computing the maximum likelihood estimates for NLMEC/LMEC with standard errors of the fixed effects and likelihood function as a by-product. The proposed algorithm is implemented in the R package [tlmec](#).

Keywords: Censored data; ECM Algorithm; Linear mixed models.

1 Introduction

Linear and nonlinear mixed effects models (LME/NLME) are frequently used to analyze grouped data because they model flexibly the within-subject correlation often present in this type of data (Pinheiro and Bates, 2000). Examples of grouped data include longitudinal data, repeated measures, and multilevel data. However, in many longitudinal studies, such as studies on environmental pollution and infection diseases, measurement of some variables may be subjects to a detection limit, i.e., a certain threshold value below or above which the measurement are not quantifiable. For instance, viral load measures the amount of actively replicating virus and depending upon the diagnostic assays used, its measurement may be subjected to some upper and lower detection limits (hence, left or right censored), below or above which they are not quantifiable.

The proportion of censored data in these studies may not be trivial and considering crude/adhoc methods, namely, substituting threshold value or some arbitrary point such as midpoint between zero and cutoff for detection (Vaida and Liu, 2009) might lead to biased estimates of fixed effects and variance components (Wu, 2010). As alternatives to crude imputation methods, Hughes (1999) proposed a likelihood-based Monte Carlo expectation-maximization (MCEM) algorithm for LME with censored responses (LMEC). Vaida et al. (2007) proposed a hybrid EM (HEM) algorithm for linear and nonlinear mixed effects models with censored response (LMEC/NLMEC) using a more efficient implementation of Hughes algorithm based on an efficient block-sampling scheme. Vaida et al. (2007) proposed an exact EM-type algorithm for LMEC/NLMEC which uses closed-form expressions at the E-step, as opposed to Monte Carlo Simulation, leading to an improvement in the speed of computation of up to an order of magnitude. More recently, Matos et al. (2011) provided some additional tools, including influence diagnostics analyses for LMEC/NLMEC.

In the framework of LMEC/NLMEC, the random effects and the within-subject errors are routinely assumed to have a normal distribution for mathematical convenience. However, such normality assumptions may not always be realistic because they are vulnerable to the presence of atypical observations. To deal with the problem of atypical observations in LME with complete responses, some proposals have been made in the literature by replacing the assumption of normality by a more flexible class of distributions. For instance, Pinheiro et al. (2011) proposed a multivariate-t linear mixed model (t-LME) and demonstrated its robustness against outliers through an application to orthodontic data and extensive simulations. Lin and Lee (2007) developed some additional tools for t-LME from a Bayesian perspective. Rosa et al. (2003) advocate the use of a subclass of elliptical distributions, called normal/independent (NI) distributions (Liu, 1996) and adopted a Bayesian framework to carry out posterior analysis for heavy-tailed LME (NI-LME). Further elaborations in t-LME have been studied by Song et al. (2007) and Wang and Fan (2011). More recently, in the context of heavy-tailed LMEC/NLMEC, Lachos et al. (2011) advocate the use of the NI class of distributions and adopted a Bayesian framework to carry out posterior analysis. Even though, some works with elliptical distributions has recently appeared in the literature, there are no studies on censored LMEC/NLMEC under the Student-t family from a frequentist perspective. In this work we propose a robust parametric modeling of LMEC/NLMEC based on the multivariate-t distribution so that the t-LMEC/t-NLMEC is defined and a fully likelihood based approach is considered, including the implementation of an exact ECM algorithm for maximum likelihood (ML) estimation.

2 Mixed effects models with censored response

The classical LME is given by:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

with the assumption that

$$\begin{pmatrix} \mathbf{b}_i \\ \boldsymbol{\epsilon}_i \end{pmatrix} \stackrel{ind.}{\sim} t_{n_i+q} \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}_{n_i} \end{pmatrix}, \nu \right), i = 1, \dots, n, \tag{2}$$

where $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ denotes the p -variate t distribution with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom ν ; the subscript i is the subject index; \mathbf{I}_p denotes the $p \times p$ identity matrix; $\mathbf{y}_i = (Y_{i1}, \dots, Y_{in_i})^\top$ is a $n_i \times 1$ vector of observed continuous responses for sample unit i , \mathbf{X}_i is the $n_i \times p$ design matrix corresponding to the fixed effects, $\boldsymbol{\beta}$ is a $p \times 1$ vector of population-averaged regression coefficients called fixed effects, \mathbf{Z}_i is the $n_i \times q$ design matrix corresponding to the $q \times 1$ vector of random effects \mathbf{b}_i , $\boldsymbol{\epsilon}_i$ is the $n_i \times 1$ vector of random errors, and the dispersion matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$ depends on unknown and reduced parameters $\boldsymbol{\alpha}$.

From (2), it is clear that marginally

$$\mathbf{b}_i \stackrel{iid}{\sim} t_q(\mathbf{0}, \mathbf{D}, \nu) \quad \text{and} \quad \boldsymbol{\epsilon}_i \stackrel{iid}{\sim} t_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}, \nu), \quad i = 1, \dots, n. \tag{3}$$

Note that \mathbf{b}_i and $\boldsymbol{\epsilon}_i$ are uncorrelated, once $Cov(\mathbf{b}_i, \boldsymbol{\epsilon}_i) = E\{\mathbf{b}_i\boldsymbol{\epsilon}_i^\top\} = E\{E\{\mathbf{b}_i\boldsymbol{\epsilon}_i^\top | U_i\}\} = \mathbf{0}$. Classical inference on the parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\alpha}^\top, \nu)^\top$ is based on the marginal distribution for \mathbf{y}_i , which are marginally distributed as

$$\mathbf{y}_i \stackrel{ind.}{\sim} t_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu), \tag{4}$$

for $i = 1, \dots, n$, where $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top$.

Following Vaida and Liu (2009), we consider the case in which the response Y_{ij} is not fully observed for all i, j . Thus, let the observed data for the i -th subject be $(\mathbf{Q}_i, \mathbf{C}_i)$, where \mathbf{Q}_i represents the vector of uncensored reading or censoring level, and \mathbf{C}_i the vector of censoring indicators:

$$y_{ij} \leq Q_{ij} \quad \text{if } C_{ij} = 1, \quad \text{and} \quad y_{ij} = Q_{ij} \quad \text{if } C_{ij} = 0, \tag{5}$$

so that, the t-LMEC is defined. For simplicity we will assume that the data are left-censored. The extensions to arbitrary censoring are immediate. It follows that for responses with censoring pattern as in (5), we have that marginally $\mathbf{y}_i \sim Tt_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu; \mathbb{A}_i)$, where $\mathbb{A}_i = A_{i1} \times \dots \times A_{in_i}$, with A_{ij} as the interval $(-\infty, \infty)$ if $C_{ij} = 0$ and the interval $(-\infty, Q_{ij}]$ if $C_{ij} = 1$. Partition \mathbf{y}_i into the observed and censored parts: $\mathbf{y}_i = vec(\mathbf{y}_i^o, \mathbf{y}_i^c)$, that is, $C_{ij} = 0$ for all elements in \mathbf{y}_i^o , and 1 for all elements in \mathbf{y}_i^c ; write accordingly $\mathbf{Q}_i = vec(\mathbf{Q}_i^o, \mathbf{Q}_i^c)$, where $vec(\cdot)$ denote the function which

stacks vectors or matrices of the same number of columns. We have that $\mathbf{y}_i^o \sim t_{n_i^o}(\mathbf{X}_i^o \boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo}, \nu)$, $\mathbf{y}_i^c | \mathbf{y}_i^o \sim t_{n_i^c}(\boldsymbol{\mu}_i^{co}, \mathbf{S}_i^{co}, \nu + n_i^o)$, where

$$\boldsymbol{\mu}_i^{co} = \mathbf{X}_i^c \boldsymbol{\beta} + \boldsymbol{\Sigma}_i^{co} \boldsymbol{\Sigma}_i^{oo-1} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta}), \tag{6}$$

$$\mathbf{S}_i^{co} = \left(\frac{\nu + Q(\mathbf{y}_i^o)}{\nu + n_i^o} \right) \boldsymbol{\Sigma}_i^{cc.o}, \tag{7}$$

with $\boldsymbol{\Sigma}_i^{cc.o} = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co} \boldsymbol{\Sigma}_i^{oo-1} \boldsymbol{\Sigma}_i^{oc}$ and $Q(\mathbf{y}_i^o) = (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})^\top \boldsymbol{\Sigma}_i^{oo-1} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})$. Thus, the likelihood for cluster i is given by

$$L_i(\boldsymbol{\theta} | \mathbf{y}) = P(\mathbf{Q}_i | \mathbf{C}_i, \boldsymbol{\theta}) = P(\mathbf{y}_i^c \leq \mathbf{C}_i^c | \mathbf{y}_i^o = \mathbf{Q}_i^o, \boldsymbol{\theta}) P(\mathbf{y}_i^o = \mathbf{Q}_i^o | \boldsymbol{\theta}), \tag{8}$$

$$= t_{n_i^o}(\mathbf{Q}_i^o | \mathbf{X}_i^o \boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo}, \nu) T_{n_i^c}(\mathbf{Q}_i^c | \boldsymbol{\mu}_i^{co}, \mathbf{S}_i^{co}, \nu + n_i^o) = L_i. \tag{9}$$

Therefore, the log-likelihood function for the observed data is given by $\ell(\boldsymbol{\theta} | \mathbf{y}) = \sum_{i=1}^n \{\log L_i\}$. This can be computed at each step of the EM-type algorithm without additional computational burden, because L_i 's are computed at the E-step. In this work we will maintain fixed the degrees of freedom and the shape parameters for Student-t, and we will use a model selection procedure based on the AIC or BIC to choose the most appropriate values of ν .

On the other hand, the EM algorithm originally proposed by Dempster, Laird and Rubin (1977) has several appealing features such as stability of monotone convergence with each iteration increasing the likelihood and simplicity of implementation. However, ML estimation in model (1)-(2) and (5) is complicated such that the EM algorithm is less advisable due to a computational difficulty in the M-step. To cope with this problem, we apply an extension of EM algorithm, called the ECM algorithm, which shares the appealing features of the EM and has a typically faster convergence rate than the EM in the sense of a small amount of iterations or actual computer time. This algorithm uses closed-form expressions at the E-step, which relies on formulas for the mean and variance of a truncated multivariate-t distribution, and can be computed using available software. Finally, we also discuss how to implement the EM algorithm for a variety of structures for the scale matrix \mathbf{D} .

We also studied a general nonlinear mixed-effects model in which the random terms are assumed to follow a multivariate-t distribution (t-NLME). Let $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^\top$ denote the (continuous) response vector for subject i and $\boldsymbol{\eta} = (\eta(\mathbf{X}_{i1}, \boldsymbol{\phi}_i), \dots, \eta(\mathbf{X}_{in_i}, \boldsymbol{\phi}_i))^\top$ be a nonlinear vectorvalued differentiable function of the individuals random parameter $\boldsymbol{\phi}_i$ and a vector of covariates \mathbf{X}_i . The t-NLME can then be expressed as:

$$\mathbf{y}_i = \boldsymbol{\eta}(\boldsymbol{\phi}_i, \mathbf{X}_i) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\phi}_i = \mathbf{A}_i \boldsymbol{\beta} + \mathbf{B}_i \mathbf{b}_i, \tag{10}$$

where the joint distribution of $(\mathbf{b}_i, \boldsymbol{\epsilon}_i)$ is as in (2), \mathbf{A}_i and \mathbf{B}_i are known design matrices of dimensions $r \times p$ and $r \times q$ respectively, possibly depending on some covariable values, $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of fixed effects, \mathbf{b}_i

is the $(q \times 1)$ vector of random effects. Thus, from the properties of the multivariate-t distribution, we have that marginally,

$$\phi_i \stackrel{ind}{\sim} t_r(\mathbf{A}_i\boldsymbol{\beta}, \mathbf{B}_i\mathbf{D}\mathbf{B}_i^\top, \nu) \text{ and } \boldsymbol{\epsilon}_i \stackrel{ind}{\sim} t_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}, \nu), \quad (11)$$

and as in the linear case, they are uncorrelated because $\text{Cov}(\phi_i, \boldsymbol{\epsilon}_i) = \mathbf{0}$. In the normal case, various approximations (viz. first-order Taylor series expansion of the model function around the conditional mode of \mathbf{b}_i , says $\tilde{\mathbf{b}}_i$) have been proposed to achieve tractable numerical optimizations (Wu, 2010). Most algorithms for computing the approximate MLE $\hat{\boldsymbol{\beta}}$ and empirical Bayes estimators (predictors) for the random effects $\hat{\mathbf{b}}_i$ considers iterative maximization of the approximate log-likelihood functions $\ell(\boldsymbol{\theta}, \tilde{\mathbf{b}}) = \sum_{i=1}^n \log f(\mathbf{y}_i|\boldsymbol{\theta}, \tilde{\mathbf{b}}_i)$. Following Taylor series expansions, we have two theorems. The first uses a point in a neighborhood of the conditional mode $\tilde{\mathbf{b}}_i$ as the expansion point and it has been proven useful for implementation of model selection, in a Bayesian context (Lachos et al., 2011). The second, useful for the implementation of the EM algorithm, uses simultaneously neighborhood of \mathbf{b}_i and $\boldsymbol{\theta}$ as expansions points, with the advantage that the likelihood is completely linearized (in \mathbf{b}_i and $\boldsymbol{\theta}$). We call these LME approximations and can be considered as extensions of the result given in Lindstrom and Bates(1990) and Pinheiro and Bates(2000) for the Student-t case.

We propose a empirical Bayes estimates of the random effects $\tilde{\mathbf{b}}$ which can be used iteratively in the linearization procedure. As commented by Vaida and Liu (2009), the linearization (L) procedure to obtain the approximate MLE of $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\alpha}^\top)^\top$ consists to iteratively solving the LME model (L-step). For censored response the linearized model is an LME with censored data, with same structure as (1)-(2), which is then solved as indicated in the previous section. The algorithm iterates to convergence between L-, E-, and CM-steps.

Acknowledgments: The authors acknowledges the partial financial support from FAPESP and CAPES-Brazil.

References

Hughes, J. (1999). Mixed effects models with censored data with application to HIV RNA levels. *Biometrics*, **55** (2), 625–629.

Lachos, V., D. Bandyopadhyay, and D. Dey. (2011). Linear and nonlinear mixed-effects models for censored hiv viral loads using normal/independent distributions. *Biometrics*, **67**, 1594–1604.

Lin, T., H. Ho, H. Chen, and W. Wang. (2011). Some results on the truncated multivariate t distribution. *Journal of Statistical Planning and*

Inference.

- Lin, T. and J. Lee. (2007). Bayesian analysis of hierarchical linear mixed modeling using the multivariate t distribution. *Journal of Statistical Planning and Inference*, **137** (2), 484–495.
- Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association*, **91**, 1219–1227.
- Matos, L. A., V. Lachos, N. Balakrishnan, and F. Labra. (2011). Influence diagnostics in linear and nonlinear mixed-effects models with censored data. *Universidade Estadual de Campinas Technical Report*, **10-11**, 484–496.
- Pinheiro, J. C., C. H. Liu, and Y. N. Wu. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using a multivariate t-distribution. *Journal of Computational and Graphical Statistics*, **10**, 249–276.
- Rosa, G. J. M., C. R. Padovani, and D. Gianola. (2003). Robust linear mixed models with normal/independent distributions and bayesian mcmc implementation. *Biometrical Journal*, **45**, 573–590.
- Song, P., P. Zhang, and A. Qu. (2007). Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. *Statistica Sinica*, **17**, 929–943.
- Vaida, F., A. Fitzgerald, and V. DeGruttola. (2007). Efficient hybrid EM for linear and nonlinear mixed effects models with censored response. *Computational statistics & data analysis*, **51** (12), 5718–5730.
- Vaida, F. and L. Liu. (2009). Fast Implementation for Normal Mixed Effects Models With Censored Response. *Journal of Computational and Graphical Statistics*, **18** (4), 797–817.
- Wang, W. and T. Fan. (2011). Estimation in multivariate t linear mixed models for multivariate longitudinal data. *Statistica Sinica*, **21**, 1857–1880.
- Wu, L. (2010). Mixed Effects Models for Complex Data. *Boca Raton, FL: Chapman & Hall/CRC*.

Modelling tracks of cabbage root fly larvae in a novel study of crop protection

Chris R. McLellan¹, Bruce J. Worton¹, William Deasy^{2,3},
A. Nicholas E. Birch³

¹ School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK

² Scottish Agricultural College, King's Buildings, Edinburgh EH9 3JG, UK

³ The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, UK

E-mail for correspondence: `Bruce.Worton@ed.ac.uk`

Abstract: We consider modelling the movements of cabbage root fly larvae by flexible diffusion processes. The aim is to characterise the behaviour of the larvae when exposed to attractant and repellent compounds. Practical aspects of the model estimation and inference are considered on extensive data collected in a study of novel approaches for the management of cabbage root fly.

Keywords: Bioassay; *Brassica*; *Delia radicum*; Diffusion; Larvae movement.

1 Introduction

The larva of the cabbage root fly (*Delia radicum* L.) is a serious pest that causes damage to *Brassica* host-plants by feeding on their roots. The use of organophosphate insecticide for controlling larvae is restricted to a single pre-planting application and novel alternative treatments are currently being investigated. So far, the mechanism by which larvae locate host-plant roots is only poorly understood. Studies suggest that cabbage root fly larvae respond to the odour of *Brassica* plants, and use the presence of plant-specific chemicals excreted by the roots of host plants to locate suitable hosts (Košťál, 1992; Baur et al., 1998; Deasy, 2011). The larvae have also been shown to be repelled by sufficiently high concentrations of plant-specific chemicals (Ross and Anderson, 1992; Ewan, 2011). If repellent chemicals in the host plants can be identified, it will be possible to develop a control system using plant extracts containing repellent compounds as soil amendments to deter the cabbage root fly larvae. Often data are collected in the form of the locations of the larvae after a given time and involve analysis using angular data, but in this paper we consider the more challenging problem of modelling the tracks of the larvae. The former method gives only one locational observation per bioassay,

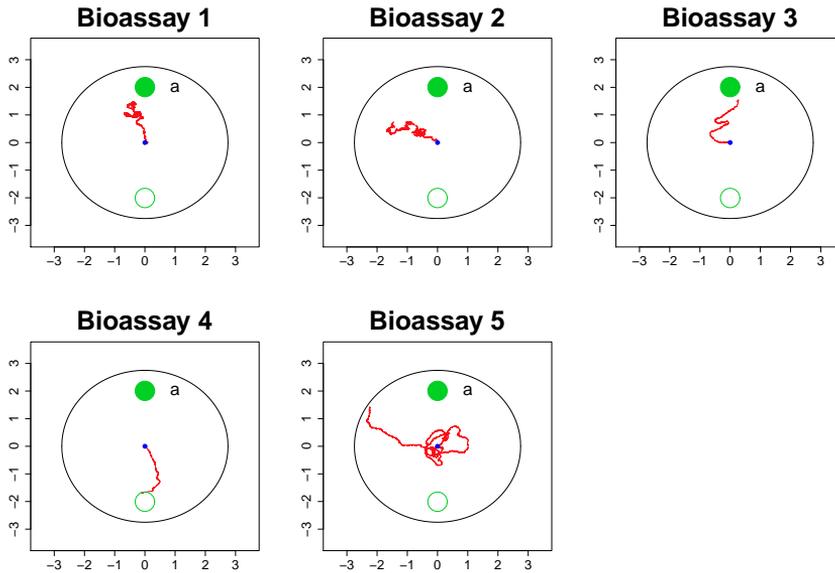


FIGURE 1. Tracks of cabbage root fly larvae for 5 bioassays. Each track starts at the origin (small dot) and the location of the larva is recorded every 0.2 seconds for 30 minutes using EthoVision 3.1. Each bioassay has a nominal 9,000 observations. The outer circle is the arena, and the upper and lower dots are the attractant (repellent for Bioassay 4) and control regions respectively.

whereas the number of observations obtained by the latter approach is in the thousands and this leads to new problems of how to model these highly correlated data. In this paper we study such models, together with appropriate methods of statistical analysis. The aim is then to use the parameters of the models as a way of summarising the complex patterns of the tracks of the larvae.

2 Experiments and data collection

Bioassays were conducted at The James Hutton Institute in a research project with the aim of developing *novel* approaches to pest management of cabbage root fly. In each bioassay, a newly hatched neonate cabbage root fly larva was placed in an arena within a 9 cm diameter petri dish half filled with agar. A zone of attraction (or repulsion) was placed within the arena, and a control zone was used. Positions of the larvae were detected by infrared camera (Sanyo) and recorded using the EthoVision 3.1 software system (Noldus et al., 2001) at intervals of 0.2 seconds for 30 minutes. Figure 1 shows plots of the tracks of the larvae for 5 of the bioassays. In Bioassays 1–3 and 5 the upper solid dot corresponds to an attractant of

damaged broccoli roots, while in Bioassay 4 the solid dot corresponds to allyl isothiocyanate from which the larva is repelled.

3 Proposed diffusion models for larvae tracks

The general form of the model we propose is defined by the conditional distribution of the position of the larva \mathbf{X}_{s+t} at time $s+t$ given the larva's position \mathbf{X}_s at time s . For example, one possibility is to use a bivariate normal distribution

$$\mathbf{X}_{s+t}|\mathbf{X}_s \sim N\{\mathbf{a} + \Gamma(\mathbf{X}_s - \mathbf{a}), \Phi\}, \tag{1}$$

where \mathbf{a} is the point of attraction. This is similar to a bivariate Ornstein-Uhlenbeck diffusion process which has been used to model movements of radio tracked animals (Dunn and Gipson, 1977; Worton, 1995). However, in the current context, an equilibrium distribution may not exist and we may relax this assumption within our modelling framework. In general, we can use mixtures of diffusion processes, with the possibility of allowing the parameters of the models to vary with time, and this gives an extremely flexible class of models.

4 Estimation and inference

4.1 Likelihood approach

As each larva is released at the origin, \mathbf{x}_0 , and generates a subsequent observed sample path, $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, with a constant time interval between observations of 0.2 seconds, the log likelihood in the special case of the bivariate Ornstein-Uhlenbeck type diffusion, up to an additive constant, is

$$-\frac{n}{2} \ln |\Phi| - \frac{1}{2} \sum_{i=1}^n \{\mathbf{x}_i - \mathbf{a} - \Gamma(\mathbf{x}_{i-1} - \mathbf{a})\}^T \Phi^{-1} \{\mathbf{x}_i - \mathbf{a} - \Gamma(\mathbf{x}_{i-1} - \mathbf{a})\}.$$

In simple cases such as this it is possible to obtain explicit estimates of the parameters, but in more complex mixed processes the likelihood may be difficult to write down. The maximum likelihood estimates of the parameters of the bivariate Ornstein-Uhlenbeck diffusion type process can be shown to be

$$\hat{\Gamma} = \left\{ \sum_{i=1}^n (\mathbf{x}_i - \mathbf{a})(\mathbf{x}_{i-1} - \mathbf{a})^T \right\} \left\{ \sum_{i=1}^n (\mathbf{x}_{i-1} - \mathbf{a})(\mathbf{x}_{i-1} - \mathbf{a})^T \right\}^{-1},$$

$$\hat{\Phi} = \frac{1}{n} \sum_{i=1}^n \{\mathbf{x}_i - \mathbf{a} - \hat{\Gamma}(\mathbf{x}_{i-1} - \mathbf{a})\} \{\mathbf{x}_i - \mathbf{a} - \hat{\Gamma}(\mathbf{x}_{i-1} - \mathbf{a})\}^T.$$

TABLE 1. The maximum likelihood estimates obtained for the parameters of diffusion process (1) fitted to each of the five bioassays.

Parameter	Bioassays				
	1	2	3	4	5
$\Gamma_{1,1}$	0.9996	1.0000	0.9985	1.0029	1.0003
$\Gamma_{1,2}$	-0.0001	-0.0002	-0.0006	-0.0011	-0.0002
$\Gamma_{2,1}$	-0.0004	0.0000	0.0000	0.0037	0.0007
$\Gamma_{2,2}$	0.9997	0.9998	0.9993	0.9992	0.9999
$\Phi_{1,1}^\dagger$	4.2554	4.1341	7.6770	3.4717	5.5888
$\Phi_{1,2}^\dagger$	0.0369	0.2327	0.2573	0.4795	0.3289
$\Phi_{2,2}^\dagger$	4.5125	3.9588	4.7225	5.2004	5.0342

[†] Values multiplied by 10^5 .

Table 1 gives the maximum likelihood estimates for Bioassays 1–5. For Bioassays 1–3 the matrix $\hat{\Gamma}$ suggests an attraction towards \mathbf{a} . However, for Bioassays 4 and 5, $\hat{\Gamma}$ indicates no attraction to \mathbf{a} . Similar results are obtained for the model with \mathbf{a} estimated from the data, but BIC values provide consistent support for the model with fixed \mathbf{a} for Bioassays 1–5. As the likelihood analysis is fairly limited in scope we consider a Bayesian approach.

4.2 Bayesian approach

Bayes estimates were calculated using WinBUGS (Lunn et al., 2000). The following prior distributions were used,

$$\Gamma_{ij} \sim \text{Normal}_{TRUNC[0,2]}(1, 10^2), \quad i, j = 1, 2,$$

$$\Phi \sim \text{Inverse-Wishart}(10^{-6}\mathbf{I}, 2),$$

and represent fairly vague prior information. Bayes estimates are shown in Table 2 for Bioassay 1.

TABLE 2. Bayes estimates of the parameters of diffusion process (1) for Bioassay 1. Only the most important parameter estimates are presented.

Parameter	Mean	Median	SD	2.5%	97.5%
$\Gamma_{1,1}$	0.9999	0.9999	0.0002	0.9995	1.0000
$\Gamma_{2,2}$	0.9998	0.9998	0.0001	0.9997	1.0000
$\Phi_{1,1}^\dagger$	4.2585	4.2580	0.0645	4.1340	4.3840
$\Phi_{2,2}^\dagger$	4.5163	4.5160	0.0664	4.3870	4.6480

[†] Values multiplied by 10^5 .

5 A more biologically realistic model

The diffusion process (1) model does not provide an adequate characterisation of the movements of the larva as it is only based on a drift. To also account for the small localised movements we considered a Hidden Markov Model (HMM). This model comprises

- (i) a component related to diffusion process (1); and
- (ii) a component accounting for small localised movements of larvae (resulting from body movements, etc), $\mathbf{X}_{s+t}|\mathbf{X}_s \sim N(\mathbf{X}_s, \Sigma)$, with a parameter $0 < \pi < 1$, representing the probability of an observation being generated from diffusion process (1).

The results are presented in Table 3. Again we use vague prior information for all the parameters. Comparing the BIC values for the models corresponding to Tables 2 and 3, we see that the more realistic HMM is a great improvement over the much simpler model. It is possible to fit more complex models but the two component model has captured the main features of the track data and is easy to interpret.

TABLE 3. Bayes estimates of the HMM for Bioassay 1. Only the most important parameter estimates are presented.

Parameter	Mean	Median	SD	2.5%	97.5%
$\Gamma_{1,1}$	0.9996	0.9996	0.0011	0.9975	1.0020
$\Gamma_{2,2}$	0.9993	0.9993	0.0003	0.9987	0.9999
$\Phi_{1,1}^\dagger$	1.8323	1.8310	0.0562	1.7260	1.9450
$\Phi_{2,2}^\dagger$	1.9368	1.9360	0.0596	1.8260	2.0580
$\Sigma_{1,1}^\ddagger$	1.4489	1.4490	0.0248	1.4010	1.4980
$\Sigma_{2,2}^\ddagger$	1.4487	1.4480	0.0249	1.4010	1.4990
π	0.2329	0.2328	0.0045	0.2241	0.2418

[†] Values multiplied by 10^4 .

[‡] Values multiplied by 10^{10} .

6 Conclusions

In this paper we have seen that diffusion processes may be used to model the tracks of cabbage root fly larvae. By doing this we can gain a greater understanding of the processes that govern the movements of the larvae. It is clear, even from the plots in Figure 1, that behaviour of the larvae in the presence of allyl isothiocyanate (Bioassay 4) is very different from the behaviour in the case of damaged broccoli roots (Bioassays 1–3 and 5).

However, we can use the estimated model to characterize the movements more quantitatively. This is much more efficient than the earlier research that simply used the final locations of larvae after a given time. Although it is possible to collect extensive sets of data, care is needed to incorporate realistic features of the movement into our models of path movement of larvae.

Acknowledgments: The experimental study was conducted at The James Hutton Institute, Invergowrie, Dundee. Chris McLellan is supported by an EPSRC studentship, and William Deasy is supported by an HDC studentship (FV 364 Novel approaches for the management of cabbage root fly).

References

- Baur, R., Städler, E., Monde, K., and Takasugi, M. (1998). Phytoalexins from *Brassica* (Cruciferae) as oviposition stimulants for the cabbage root fly, *Delia radicum*. *Chemoecology*, **8**, 163–168.
- Deasy, W. (2011). Novel approaches for the management of cabbage root fly. *Horticultural Development Company (HDC) Technical Seminar: Protecting Your Field Veg Crop*, 30th June 2011, Stockbridge Technology Centre, North Yorkshire, UK.
- Dunn, J.E., and Gipson, P.S. (1977). Analysis of radio telemetry data in studies of home range. *Biometrics*, **33**, 85–101.
- Ewan, A. (2011). How to put root fly larvae off their food. *HDC News Field Vegetables Review*, Annual Supplement, Second Edition, 11.
- Košťál, V. (1992). Orientation behavior of newly hatched larvae of the cabbage maggot, *Delia radicum* (L.) (Diptera: Anthomyiidae), to volatile plant metabolites. *Journal of Insect Behavior*, **5**, 61–70.
- 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.
- Noldus, L.P.J.J., Spink, A.J., and Tegelenbosch, R.A.J. (2001). EthoVision: a versatile video tracking system for automation of behavioral experiments. *Behavior Research Methods, Instruments and Computers*, **33**, 398–414.
- Ross, K.T.A., and Anderson, M. (1992). Larval responses of three vegetable root fly pests of the genus *Delia* (Diptera: Anthomyiidae) to plant volatiles. *Bulletin of Entomological Research*, **82**, 393–398.

Worton, B.J. (1995). Modelling radio-tracking data. *Environmental and Ecological Statistics*, **2**, 15–23.

Modeling and forecasting customer behavior for revolving credit facilities

Radoslava Mirkov¹, Holger Thomae¹, Michael Feist², Thomas Maul¹, Gordon Gillespie¹, Bastian Lie¹

¹ TriSolutions GmbH, Hamburg, Germany,

² DZ BANK AG, Frankfurt am Main, Germany

E-mail for correspondence: radoslava.mirkov@trisolutions.de

Abstract: Revolving credit facilities are studied with an aim to forecast customer behavior and thus minimize liquidity and income risk. Historical data provide information about characteristics and withdrawal patterns of borrowers, which enables a deeper insight into their behavior. The withdrawal patterns are expressed in terms of the relative withdrawal and its variance over the contractual period, and the relative duration. Possible explanatory variables are for e.g. a customer's rating, a credit line limit, a lifespan of the credit facility, the economic sector of the customer, etc. We study the dependence of the target and explanatory variables utilizing statistical modeling techniques integrated in data mining methodology. Multivariate recursive partitioning enables recognition of the customer behavior, and the prediction of future withdrawals for new customers.

Keywords: revolving credit facilities; customer behavior; data mining; multivariate recursive partitioning; prediction.

1 Introduction and Motivation

Lines of credit are the focus of our investigation. In this paper we concentrate on revolving credit facilities (RCF) and want to forecast customer behavior taking the RCF. The RCF, often referred to as a revolver, is a flexible loan which allows the borrower to use the funds when they are needed. It can be taken out by both corporations and individuals. The bank guarantees the customer a loan up to the credit limit during a lifespan of the credit facility, without having to reapply each time the cash is needed. As the borrower repays the money, it is available to be borrowed again. This is a flexible loan available to clients which has been designed so clients can repay their loan balance to zero but still have the facility available to borrow again. This facility is useful for clients who may not want to use a loan facility at all times but appreciate the possibility that the agreed limit will be available up to a contractual maturity of the facility. In other

words, the borrower is under no obligation to actually take out a loan at any particular time, but may take part or all of the funds at any time over a period of several years. This agreement is common in situations in which a business must pay obligations but its operating cashflow varies, for e.g. seasonally. At any given time, the balance due may fluctuate from zero to the maximum credit limit. The interest rate paid on this kind of credit is usually floating, i.e. it is linked to EURIBOR or LIBOR so that customers receive market rate. In general, it is available in all major currencies.

In the case a bank grants the RCF, it is faced to several very specific risks beside the common credit risks. The most obvious risk is the uncertainty up to which extent of the facility will the borrower draw a loan. This depends usually on the operational needs of the customer, so the RCF might be used just for back up reasons to bridge liquidity gaps in stressed market situations, or it might be permanently drawn to some extent, due to his general operational needs. If the bank would hedge the liquidity requirements fully, the RCF would just in time when the money is drawn, it might happen that there is no sufficient liquidity available in the markets, especially in times of financial crisis, and it is not warranted that this liquidity is accessible at reasonable prices, and the loan can cause high losses when interest rate levels are higher than they were at the time the RCF was granted and its conditions were fixed. These risks are also not hedged completely, but the conditions would no longer be competitive for the customer. If the bank decides, on the other hand, to refinance the loans under the RCF two characteristics of liquidity risk: availability and term transformation.

Beside these liquidity risks the customer is granted several options that have the potential to have a high impact on the revenues of the bank. For e.g. the withdrawal date can be chosen, or in which currency and for what term the loan is drawn. The hedging of these options is also very difficult for the banks and causes additional opportunistic costs. We will also focus on the rights of cancellation of the RCF. Usually (and especially in the German jurisdiction) every borrower has the right to cancel the facility at the end of each thereunder withdrawn loan. In the case the RCF was hedged for a longer term and cancelled prematurely the bank might be faced with the problem to reinvest the former acquired liquidity at lower interest rate which would again imply losses. All these issues influence the loan conditions, and imply the risk adjusted conditions for customers. One might suggest to develop an option pricing model to deal with all of the above mentioned customer options. But option pricing models are not adequate when one takes into account that the behavior of the customer does not strictly depend on the current market conditions (interest rates) and not only on the specific credit prices (credit spreads) for liquidity. One could assume that most of the small and medium sized customers behave just according to their specific operational needs. Figure 1 illustrates the described situation.

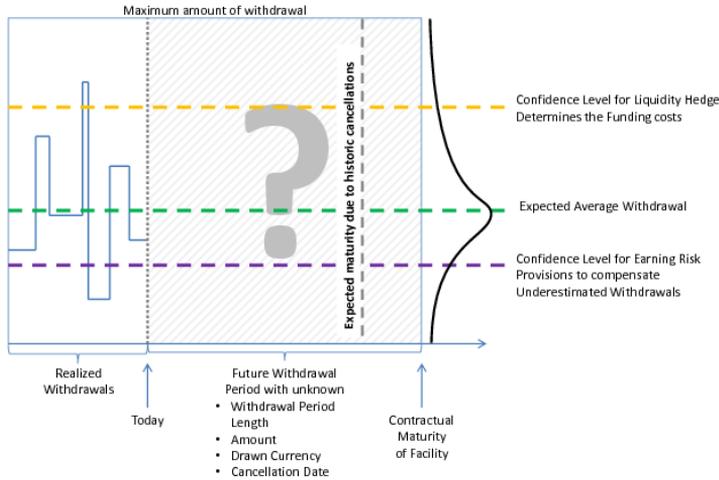


FIGURE 1. The Principle of the RCF.

This drives the authors to investigate the historic behavior of a large portfolio of RCFs with the intention to identify behavioral clusters for the drawings under RCFs that allows the unique assignment of a cluster to each RCF. The aim is to predict up to a certain degree the likely future behavior of a specific client in a way that allows the bank to find an appropriate strategy for refinancing and therefore minimize liquidity risks and earnings risks.

2 Data Description

Data for this study is provided by one of the largest German banks. It contains historical data related to the RCF customers and their behavior. Mean withdrawals over the lifetime of the RCF and their variance as well as the relative duration of the actual withdrawal compared to the contractual duration of $n = 128$ agreements from one specific business branch with different customers are included. Further facts about these agreements, such as the duration of the contract, the agreed limit and the currency are also provided. As we are interested in the customer behavior, the information about customers like the rating, the probability of default, the type of financing, their economic and risk group, is also relevant. Additionally, the data describes the purpose and some properties of the loan.

3 Data Mining Modeling Approach for the RCF Forecasting Customer Behaviour

Data mining, the process of automatically discovering patterns in large data sets, includes predictive modeling approach. Models for the continuous target variable as a function of the explanatory variables based on the regression techniques are built. We chose recursive partitioning (RPART) to build regression models of a very general structure

$$y_i = \sum_{i=1}^n x_i^k, \quad n = 128, \quad k = 15,$$

based on the ANOVA method, as y is a numeric vector. The splitting criteria is $SST - (SSL + SSR)$, where SST is the sum of squared errors for the node, and SSR and SSL are the sums of squared errors of the right and the left child-node, respectively.

In De'ath, G. (2002) the generalization of this approach to multivariate target variables, which also includes cross-validation of the results is proposed. Thus, the solution with the best predictive power is retained. Multivariate regression trees (MRT) turn out to be a powerful and robust method that handles a variety of situations, such as missing values, non-linear or higher-order interactions among explanatory variables, as well as categorical or quantitative predictors.

As we are interested in the mean relative withdrawal and its standard deviation simultaneously, we choose the MRT method to analyse the data. The measure of predictive error is also of great value.

In the RCF we investigate which of the above mentioned explanatory variables influence the mean relative withdrawal over the lifespan of the loan and its variance, i.e. standard deviation. The type of financing of the RCF customers and the risk group seem to influence the target variable to the great extent.

In our model

$$y_i = (w_i, sd_i), \quad n = 128,$$

where w_i stands for the mean relative withdrawal, and sd_i for its standard deviation. The variables x_i^k with the best predictive value turn out to be the type of financing, denoted by $X1$, and the risk group of the borrower, given by the variable $X2$. Both explanatory variables are categorical. $X2$ describes how risky is a specific group of RCF consumers, whereas each level of the variable implies different levels of risk. The levels of the financing type $X1$ are coded and define project (1) and acquisition financing (2), credits for trading (4) and structured products (5), and other (9) types. The default probability is denoted by $X3$.

The results of the MRT procedure are shown in Figure 2. The minimal cross-validation error is attained for seven nodes, as shown in 2 (left). The relative and cross-validation error in dependence of the tree size can be

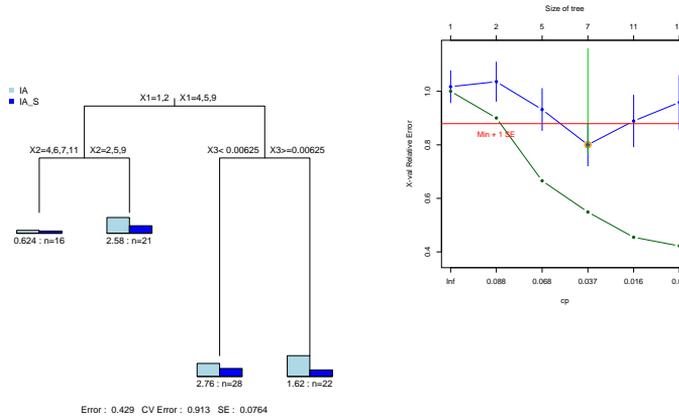


FIGURE 2. Multivariate Decision Tree and its Cross-Validation Error for the Mean Relative Withdrawal and its Standard Deviation.

TABLE 1. Mean Relative Withdrawal and its Standard Deviation for Obtained Cluster.

cluster mean	C1	C2	C3
w	0.10	0.50	0.60
sd	0.07	0.25	0.20

seen in Figure 2 (left). The red dot denotes the optimal solution based on the cross-validation. We note that the cross-validation error drops to 0.8, whereas the relative error drops below 0.5. We refer to Table 1 for mean values of the possible resulting cluster. The models are implemented in R using the packages RPART and MVPART.

4 Conclusions

We investigate prediction of the customer behavior based on the multivariate data mining models. This seems to be the best available method for our purpose, as the model with the best predictive properties is selected and the multivariate target variables can be analysed. It turns out that the type of financing and the risk category of the consumer are of particular importance for the behavior of the loan takers. The probability of default seems to influence their behavior, too.

The cross-validation error, which indicates the stability of the prediction

drops to 0.8. This means that the predictive power of the model for the given subset of data could be improved. Further steps include more detailed analysis based on a complete data set of better quality, and an investigation of other explanatory variables.

References

- Tan, P-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education.
- Atkinson, E.J., and Therneau, T.M. (2000). An Introduction to Recursive Partitioning Using the RPART Routines. *Technical Report*, Mayo Foundation.
- Kordichev, A., Powell, J.G. and Trippe, D.W. (2005). Structural Models of Revolving Consumer Credit Risk. *Credit Scoring and Credit Control Conference Proceedings*, 1–13.
- De'ath, G. (2002). Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships. *Ecology*, **83**(4), 1105–1117.
- Borcard, D., Gillet, F., and Legendre, P. (2011). *Numerical Ecology with R*. Springer.

Multivariate copula models in ROC analysis

Elisa M. Molanes-López¹, Emilio Letón²

¹ Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain.

² Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia, Madrid, Spain.

E-mail for correspondence: emilio.leton@dia.uned.es

Abstract: In classification studies, when there are several continuous biomarkers available, it is very important to know how to combine them into an improved composite one. In the literature, there are several methods to do so: the linear combination of the available biomarkers that maximizes the empirical ‘Area Under the ROC Curve’ or the smoothed version of this index. These methods can be far superior to logistic regression in certain situations. However, the strict assumption of linearity is often unable to capture informative nonlinear structures in the real world. For that reason, using the fact that the likelihood ratio function is the optimal combination, we have recently proposed to combine multiple biomarkers using a semiparametric estimate of their likelihood ratio function. Due to some computational aspects, we only considered there the case of two normally distributed biomarkers. Now, we extend here the algorithm to deal with multivariate copulas and non normally distributed biomarkers. Through a simulation study, we check its performance for moderate dimension. Finally, we analyze an example with eight biomarkers.

Keywords: Archimedean copula; *AUC*; Copula density; ROC curve; Youden index.

1 Introduction

The accuracy of a continuous biomarker, for the classification into two groups (healthy versus diseased), is usually described graphically through the ‘Receiver Operating Characteristic’ (ROC) curve. In this setting, the ‘Area Under the ROC Curve’ (*AUC*) and the Youden index (*J*) are commonly used measures of classification performance (see, for example, Pepe, 2003, and Molanes-López and Letón, 2011).

In practice, it is usual to have several continuous biomarkers available, let’s say $\mathbf{Y} = (Y_1, \dots, Y_p)^T$, with \mathbf{Y}_0 and \mathbf{Y}_1 referring to the p -dimensional biomarker in the healthy and diseased population with \mathbf{f}_k and \mathbf{F}_k , the multivariate density function and the multivariate cumulative distribution function (cdf), respectively, of \mathbf{Y}_k , for $k = 0, 1$.

In the literature, the problem of combining multiple biomarkers has been tackled considering the linear combination of the available biomarkers that

maximizes the empirical AUC (see Pepe and Thompson, 2000) or the smoothed AUC (see Ma and Huang, 2007). The AUC -based method can be far superior to logistic regression in certain situations (Pepe et al., 2006). However, the strict assumption of linearity is often unable to capture informative nonlinear structures in the real world (Komori, 2009). For that reason, Letón and Molanes-López (2010) proposed to combine multiple biomarkers using a semiparametric estimate of their likelihood ratio function via copula functions and relative curves, based on the Neyman-Pearson lemma as a theoretical basis.

In Section 2, we describe some computational aspects that are necessary to have in mind when generating multivariate copulas. In Section 3, we show a simulation study for moderate dimension and non-normal marginals. Finally, in Section 4, we present a well known real example with $p = 8$, where the computational aspects of Section 3 have been applied.

2 Computational aspects in the generation of multivariate copulas

A p -dimensional copula, $\mathbf{C} : [0, 1]^p \rightarrow [0, 1]$, is a multivariate distribution from the unit ‘hypercube’ to the unit interval with uniform marginals in $[0, 1]$. We describe several examples of copulas that we have used in the simulation: Clayton, Gumbel, Frank, and Normal. These copulas have been previously considered in, for example, Lawless and Yilmaz (2011), although we have also included the rotated (survival) versions of them to cope with different kinds of asymmetry.

Our approach is flexible in the sense that you fit your data using as many copula functions as you want to model the dependence structure, and then you use the Akaike Information Criterion (AIC) to select the best fit. In order to do so, it is necessary to have closed form expressions for their densities. If this is not the case, one possibility is to use symbolic differentiation.

In our case, the Clayton, Gumbel and Frank copulas belong to the Archimedean family, which are defined in terms of a generator function φ with

$$\mathbf{C}(u_1, \dots, u_p) = \varphi^{-1}(\varphi(u_1) + \dots + \varphi(u_p)),$$

see, for example, Wu et al. (2007), for the conditions on the generator φ . For this family, it is easy to see that the copula density is given by the general expression:

$$\mathbf{c}(u_1, \dots, u_p) = \frac{\partial^p}{\partial u_1 \dots \partial u_p} \mathbf{C}(u_1, \dots, u_p) = \varphi^{-1(p)}(x) \prod_{j=1}^p \varphi'(u_j),$$

where $x = \sum_{j=1}^p \varphi(u_j)$. Using the expressions for $\varphi^{-1(p)}$ given in Wu et al. (2007) for the Clayton, Gumbel and Frank copulas, we collect below the closed form expression for their corresponding copula densities.

- Density of the Clayton copula with $\varphi(u) = \frac{1}{\theta}(u^{-\theta} - 1)$ and $\theta > 0$:

$$\mathbf{c}(u_1, \dots, u_p) = (-1)^p (1 + \theta x)^{-\left(\frac{1}{\theta} + p\right)} \prod_{j=1}^p (1 + (j - 1)\theta) \prod_{j=1}^p \varphi'(u_j),$$

with $\varphi'(u_j) = -u_j^{-(\theta+1)}$.

- Density of the Gumbel copula with $\varphi(u) = (-\log(u))^\theta$ and $\theta > 1$:

$$\mathbf{c}(u_1, \dots, u_p) = (-1)^p \alpha e^{-x^\alpha} x^{-p+\alpha} \Psi_{p-1}(x^\alpha) \prod_{j=1}^p \varphi'(u_j),$$

with $\alpha = \frac{1}{\theta}$, $\varphi'(u_j) = -\frac{\theta}{u_j} (-\log u_j)^{\theta-1}$.

- Density of the Frank copula with $\varphi(u) = -\log\left(\frac{e^{-\theta u} - 1}{e^{-\theta} - 1}\right)$ and $\theta > 0$:

$$\mathbf{c}(u_1, \dots, u_p) = -\frac{1}{\theta} \Psi_{p-1}\left(\frac{1}{1 + e^{-x}(e^{-\theta} - 1)}\right) \prod_{j=1}^p \varphi'(u_j),$$

with $\Psi_0(x) = x - 1$, $\Psi_j(x) = x(1 - x)\Psi'_{j-1}(x)$, for $j = 1, \dots, p - 1$, and $\varphi'(u_j) = \frac{\theta e^{-\theta u_j}}{e^{-\theta u_j} - 1}$, for $j = 1, \dots, p$.

For the Normal copula, its density is given by

$$\mathbf{c}(u_1, \dots, u_p) = \frac{1}{|R|^{\frac{1}{2}}} e^{-\frac{1}{2}\zeta^T (R^{-1} - I)\zeta},$$

where R is a correlation matrix and $\zeta = (\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p))$, with Φ denoting the cdf of a univariate standard Normal random variable, as can be seen in Cherubini et al. (2004). Since all the previous multivariate copulas are defined in terms of a one-dimensional parameter θ , we have considered R with pairwise correlation θ , $-1 \leq \theta \leq 1$.

3 Simulation study

In this section, we consider different scenarios to generating values of the multidimensional biomarker \mathbf{Y} , using copula functions to model the dependence structure existing between the marginals of \mathbf{Y} . To construct these scenarios, it is necessary to specify the marginals and the copula function. We have considered two values for the dimension parameter p , $p = 4$ with $(J_1, \dots, J_4) = (0.4, 0.5, 0.6, 0.7)$, and $p = 8$ with $(J_1, \dots, J_8) = (0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7)$.

We have used three possible scenarios for $p = 4$:

- S_1^4 : $Y_{0\ell} \sim N(\mu_{0\ell}, \sigma_{0\ell}^2)$, $\ell = 1, \dots, 4$, for the healthy group, and $Y_{1\ell} \sim N(\mu_{1\ell}, \sigma_{1\ell}^2)$, $\ell = 1, \dots, 4$, for the diseased group.
- S_2^4 : $Y_{0\ell} \sim N(\mu_{0\ell}, \sigma_{0\ell}^2)$, $\ell = 1, 3$, $Y_{0\ell} \sim \text{Log}N(\mu_{0\ell}, \sigma_{0\ell}^2)$, $\ell = 2, 4$, for the healthy group, and $Y_{1\ell} \sim N(\mu_{1\ell}, \sigma_{1\ell}^2)$, $\ell = 1, 3$, $Y_{1\ell} \sim \text{Log}N(\mu_{1\ell}, \sigma_{1\ell}^2)$, $\ell = 2, 4$, for the diseased group.
- S_3^4 : $Y_{0\ell} \sim \text{Log}N(\mu_{0\ell}, \sigma_{0\ell}^2)$, $\ell = 1, \dots, 4$, for the healthy group, and $Y_{1\ell} \sim \text{Log}N(\mu_{1\ell}, \sigma_{1\ell}^2)$, $\ell = 1, \dots, 4$, for the diseased group.

Analogously, for $p = 8$ we have considered three possible scenarios with a similar structure: normals for all the marginals; normals for the odd marginals and lognormals for the even marginals; and lognormals for all the marginals.

For the copula functions, the parameter θ has been chosen in such a way that their pairwise Kendall's tau is $\tau = 0.494$ or $\tau = 0.713$. For example, for the Frank copula, $\theta = 5.622$ yields a Kendall's tau of $\tau = 0.494$ and $\theta = 12.025$ yields a Kendall's tau of $\tau = 0.713$.

The simulations have been carried out in MATLAB using 1000 trials per scenario with sample sizes $n_0 = n_1 = 100, 200$. For the sake of brevity, we only present results for the Frank copula with $p = 4$ and sample sizes $n_0 = n_1 = 100$. See Table 1, where we collect the averaged Youden index, J , for both the composite biomarker and the **LR** biomarker (\hat{J}_c and \hat{J}_{LR} , respectively), and the averaged AUC values for both the composite biomarker and the theoretical **LR** biomarker (\widehat{AUC}_c and $\widehat{AUC}_{\text{LR}}$, respectively). In brackets we give the sample standard deviation to assess the accuracy of the estimates.

TABLE 1. Average J and AUC over 1000 trials with $(J_1, J_2, J_3, J_4) = (0.4, 0.5, 0.6, 0.7)$, Frank copula and $n_0 = n_1 = 100$.

	(θ_0, θ_1)	\hat{J}_c	\hat{J}_{LR}	\widehat{AUC}_c	$\widehat{AUC}_{\text{LR}}$
S_1^4	(5.6, 5.6)	0.866(0.033)	0.835(0.036)	0.976(0.009)	0.965(0.011)
	(5.6, 12.0)	0.865(0.034)	0.835(0.037)	0.976(0.009)	0.965(0.012)
	(12.0, 5.6)	0.930(0.027)	0.912(0.027)	0.991(0.005)	0.987(0.006)
	(12.0, 12.0)	0.924(0.026)	0.910(0.028)	0.990(0.006)	0.986(0.007)
S_2^4	(5.6, 5.6)	0.869(0.036)	0.836(0.037)	0.976(0.010)	0.966(0.012)
	(5.6, 12.0)	0.862(0.034)	0.835(0.038)	0.974(0.010)	0.965(0.012)
	(12.0, 5.6)	0.933(0.026)	0.914(0.027)	0.991(0.006)	0.988(0.006)
	(12.0, 12.0)	0.926(0.027)	0.911(0.028)	0.989(0.006)	0.986(0.007)
S_3^4	(5.6, 5.6)	0.868(0.034)	0.834(0.036)	0.975(0.010)	0.965(0.012)
	(5.6, 12.0)	0.857(0.033)	0.834(0.036)	0.973(0.010)	0.965(0.012)
	(12.0, 5.6)	0.934(0.025)	0.910(0.027)	0.991(0.006)	0.987(0.006)
	(12.0, 12.0)	0.925(0.027)	0.910(0.028)	0.989(0.007)	0.986(0.007)

From the results obtained in this simulation study we conclude that:

- The averaged AUC and J of the combined biomarker approximate well the AUC and J of the likelihood ratio based biomarker, although with a slight overestimation.
- The combined biomarker always outperforms each biomarker alone.
- The stronger the dependence between the marginals, the larger the probability of selecting the right copula.
- The larger the sample sizes, the more accurate the results.
- The above remarks are valid for both dimensions used in the simulation.

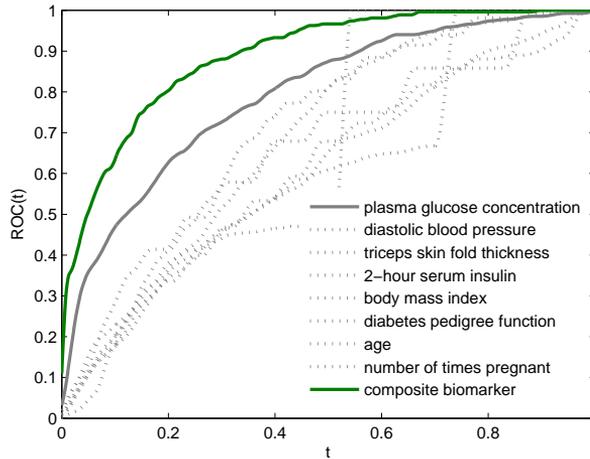


FIGURE 1. Pima Indians dataset: ROC curve for each variable (grey lines; in solid the best marginal) and for the composite biomarker (green line).

4 Example

In this section, we illustrate this methodology to combine a moderate number of multiple biomarkers into a composite one, using the Pima Indians dataset. This example comes from the Pima Indians Diabetes Study, which includes 268 patients with signs of diabetes and 500 patients without signs of diabetes. Eight variables were considered in this study: number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, and age. The raw data is available at the UCI Machine Learning repository, <http://archive.ics.uci.edu/ml/datasets.html>. The smoothed ROC curves are shown in Figure 1. It can be seen that the plasma glucose concentration yields a better discrimination between the two groups than any of the other variables alone does ($J = 0.437$ and $AUC = 0.792$ for the plasma glucose concentration). With the help of our approach, the

rotated Gumbel copula with $\theta = 1.214$ and $\theta = 1.148$ have been selected for the healthy and diseased groups, respectively, yielding $J = 0.610$ and $AUC = 0.888$ for the composite biomarker.

Acknowledgments: This research has been supported by several Grants from the Spanish Ministry of Science & Innovation. E.M. Molanes-López acknowledges support to MTM2010-09213-E, ECO2011-25706 and MTM2011-28285-C02-02. E. Letón acknowledges support to SEJ2007-64500, TIN2009-09158 and MTM2010-09213-E.

References

- Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula methods in finance*. New York: John Wiley & Sons.
- Komori, O. (2009). A boosting method for maximization of the area under the ROC curve. *Annals of the Institute of Statistical Mathematics* **63**, 961–979.
- Lawless, J. F. and Yilmaz, Y. E. (2011). Comparison of semiparametric maximum likelihood estimation and two-stage semiparametric estimation in copula models. *Computational Statistics and Data Analysis* **55**, 2446–2455.
- Letón, E. and Molanes-López, E. M. (2010). Copula based estimate of the likelihood ratio function for combining continuous biomarkers. *IWSM 2010, Glasgow*.
- Ma, S. and Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics* **63**, 751–757.
- 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.
- Pepe, M.S., Cai, T. and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62**, 221–229.
- Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123–140.
- Wu, F., Valdez, E. and Sherris, M. (2007). Simulating from exchangeable Archimedean copulas. *Communications in Statistics – Simulation & Computation* **36**, 1019–1034.

Multiple testing based on depth

Elisa M. Molanes-López¹, Juan Romo¹

¹ Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain.

E-mail for correspondence: elisamaria.molanes@uc3m.es

Abstract: A multiple testing technique is aimed to test a fixed number of hypotheses by both controlling a suitable Type I error rate and, simultaneously, maximizing the power of each univariate test. In this paper, we propose several procedures based on the concept of depth and bootstrap resampling. Through a simulation study, we show that these new approaches yield asymptotically balanced rejection regions in a natural way, without requiring the use of studentized univariate test statistics, and that they prove to be competitive with other existing methods recently proposed in the literature.

Keywords: balanced control; false-discovery-exceedance; false-discovery rate; generalized family-wise-error rates; modified band depth.

1 Introduction

The construction of multiple testing procedures (MTPs) is nowadays an important area of research, mainly motivated by real data applications emerging in different fields such as genomics, brain imaging, spatial analysis, applied economics and clinical trials, among others. See Dudoit and van der Laan (2008) and Farcomeni (2008), for more details.

Two types of errors may occur when solving a multiple testing problem. On one hand, a Type I error or false positive appears when a true null hypothesis is rejected. On the other hand, a Type II error or false negative happens when a false null hypothesis is not rejected. When a large number of hypotheses need to be tested simultaneously, it is required to use a procedure that makes an adjustment for multiple comparisons and avoids many false discoveries. Moreover, the criterion of not making any false positive may be too stringent. Taking into account these ideas, different Type I error rates have been proposed in the literature and different procedures that aim to control a specific Type I error rate have been designed to solve (low- and high-dimensional) multiple testing problems. For instance, the generalized family-wise-error rates (or k -FWER, where k is any prefixed positive integer, $k \geq 1$), the false-discovery rate (FDR), and the false-discovery-exceedance (γ -FDX, where γ is any prefixed real number, $\gamma \in (0, 1)$), are now common generalized Type I error rates used in

applications involving many hypotheses. We refer the reader to Dudoit and van der Laan (2008), for more details.

2 New multiple testing procedures

Let H_i be a null hypothesis concerned with a test of a real-valued parameter θ_i and consider that we have s null hypotheses. For simplicity, let H_i be formulated by $H_i : \theta_i = 0$, where $i \in S = \{1, \dots, s\}$. Let $\hat{\theta}_{n,i}$ be an estimate of θ_i based on an independent and identically distributed (i.i.d.) sample, $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$, with $\mathbf{X}_k = (X_{k,1}, \dots, X_{k,s})$, for $k = 1, \dots, n$, taken from a random vector of dimension s , $\mathbf{X} = (X_1, \dots, X_s)$. Let $d_{n,i}$ denote the univariate test statistic used to test the i null hypothesis, either basic (that is, the difference between the estimate and the true parameter θ_i under the null, $d_{n,i} = \hat{\theta}_{n,i} - \theta_i$) or studentized (that is, $d_{n,i} = (\hat{\theta}_{n,i} - \theta_i)/\sigma_{n,i}$, where $\hat{\sigma}_{n,i}$ denotes the sample standard deviation of $\hat{\theta}_{n,i}$), for $i = 1, \dots, s$.

The objective here is to test simultaneously the s null hypotheses against two-sided alternatives at a global significance level α , fixed in advance. To do so, we introduce now a new single-step MTP for control of the k -FWER, based on bootstrap resampling and on the finite dimensional version of the modified band depth (MBD), recently introduced for functional data by López-Pintado and Romo (2007, 2009). The steps required to implement this new single-step MTP for control of the k -FWER are as follows. First, we approximate the sampling distribution of $\mathbf{d}_n = (d_{n,1}, \dots, d_{n,s})$ by the empirical distribution of its bootstrap analogues. Specifically, we take B resamples with replacement from the n s -dimensional i.i.d. observations, and we compute the bootstrap analogues of \mathbf{d}_n by $\mathbf{d}_n^{*,b} = (d_{n,1}^{*,b}, \dots, d_{n,s}^{*,b})$, where $d_{n,i}^{*,b} = \hat{\theta}_{n,i}^{*,b} - \hat{\theta}_{n,i}$, $b = 1, \dots, B$, if basic statistics are used or $d_{n,i}^{*,b} = (\hat{\theta}_{n,i}^{*,b} - \hat{\theta}_{n,i})/\hat{\sigma}_{n,i}^{*,b}$, $b = 1, \dots, B$, if studentized statistics are used instead, with $\hat{\theta}_{n,i}^{*,b}$ and $\hat{\sigma}_{n,i}^{*,b}$ denoting the bootstrap analogues of $\hat{\theta}_{n,i}$ and $\hat{\sigma}_{n,i}$, respectively, obtained with the resample b . Considering the s -dimensional bootstrap estimates of \mathbf{d}_n , we then compute their s -dimensional MBDs and obtain the convex envelope of the deepest ones, say $(\mathbf{a}_n^*, \mathbf{b}_n^*)$, satisfying the following condition:

‘At least $(1 - \alpha) \times 100\%$ of the bootstrap estimates are such that $a_{n,i}^* \leq d_{n,i}^{*,b} \leq b_{n,i}^*$, for all but at most $(k - 1)$ of the s null hypotheses, where $\mathbf{a}_n^* = (a_{n,1}^*, \dots, a_{n,s}^*)$ and $\mathbf{b}_n^* = (b_{n,1}^*, \dots, b_{n,s}^*)$.’

Finally, the null hypothesis H_i is rejected in favour of the alternative if $d_{n,i} < a_{n,i}^*$ or $b_{n,i}^* < d_{n,i}$, for $i \in S$. Otherwise, H_i is accepted, for $i \in S$.

Based on this single-step MTP for control of the k -FWER, it is easy to design stepdown procedures for control of either the k -FWER or the γ -FDX, similarly to Romano and Wolf (2010).

3 Simulation study

A simulation study has been carried out in MATLAB to compare the new MTPs with those recently introduced by Romano and Wolf (2010), using their scenarios. This simulation study proves that the new MTPs are competitive with the balanced MTPs proposed by Romano and Wolf (2010), and that they yield asymptotically balanced confidence bands without requiring the use of studentized univariate test statistics.

4 Example

We illustrate now the use of our new methodology using the Wisconsin dataset, publicly available at <http://archive.ics.uci.edu/ml/machine-learning-databases/>. This dataset consists of 569 individuals, 212 out of them suffering from malignant breast cancer and 357 without the disease. For every individual, there is available information on 30 features. After applying our stepdown MTP aimed to control the k -FWER with $k = 1$ and using basic univariate test statistics, it turns out that only 5 of the 30 features are identified as not differentially expressed at a nominal alpha level of $\alpha = 10\%$.

Acknowledgments: E.M. Molanes-López acknowledges support to MTM-2010-09213-E, ECO2011-25706 and MTM2011-28285-C02-02. J. Romo acknowledges support to ECO2011-25706, SEJ2007-64500 and UC3M-ECO-05-072.

References

- Dudoit, S., and van der Laan, M.J. (2008). *Multiple testing procedures with applications to genomics*. New York: Springer.
- Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, **17**, 347–388.
- López-Pintado, S., and Romo, J. (2007). Depth-based inference for functional data. *Computational Statistics and Data Analysis*, **51**, 4957–4968.
- López-Pintado, S., and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, **104**, 718–734.
- Romano, J.P., and Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics*, **38**, 598–633.

Smoothing parameter selection for spatiotemporal models with application to the analysis of contaminants in groundwater

Daniel Alberto 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: When using P-Splines to model spatiotemporal data, the choice of the smoothing parameter is a critical issue. Standard procedures for this task involve selecting the value of the parameter that minimises some optimisation criterion like AIC. These procedures may lead to anomalous predicted values not supported by the data. We propose a Bayesian approach to tackle this problem, which has proved to be effective in some case studies.

Keywords: Smoothing; Spatiotemporal; Bayesian.

1 Background

P-Splines (Eilers & Marx, 1996) form the basis of a non-parametric technique widely used to fit spatiotemporal data by means of a smooth function using the model $\mathbf{Y} = f(\mathbf{x}) + \boldsymbol{\epsilon} = \mathbf{B}(\mathbf{x})\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$ where $\mathbf{B}(\mathbf{x})$ is an $n \times m$ matrix of basis functions. The estimators of the coefficients are obtained by minimising the objective function:

$$S(\boldsymbol{\alpha}) = \|\mathbf{y} - \mathbf{B}(\mathbf{x})\boldsymbol{\alpha}\|^2 + \lambda \|\mathbf{D}\boldsymbol{\alpha}\|^2 = (\mathbf{y} - \mathbf{B}(\mathbf{x})\boldsymbol{\alpha})' (\mathbf{y} - \mathbf{B}(\mathbf{x})\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}' \mathbf{D}' \mathbf{D} \boldsymbol{\alpha}$$

where λ is a non-negative smoothing parameter and \mathbf{D} is a difference matrix of order 2. For a given value of λ it is $\hat{\mathbf{y}} = \hat{\mathbf{f}}(\mathbf{x}) = \mathbf{B}\hat{\boldsymbol{\alpha}} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'\mathbf{y} = \mathbf{H}\mathbf{y}$. As there is a one-to-one (decreasing) correspondence between the penalisation parameter λ and the trace of \mathbf{H} , usually known as *degrees of freedom (df)*, sometimes the model is described in terms of the latter because it has a more intuitive interpretation.

2 On the choice of the penalisation parameter λ

When trying to fit a smooth function to a given data set using the P-Splines approach, the choice of the penalisation parameter λ is a crucial matter as it will determine the trade-off between smoothness and capturing the signal.

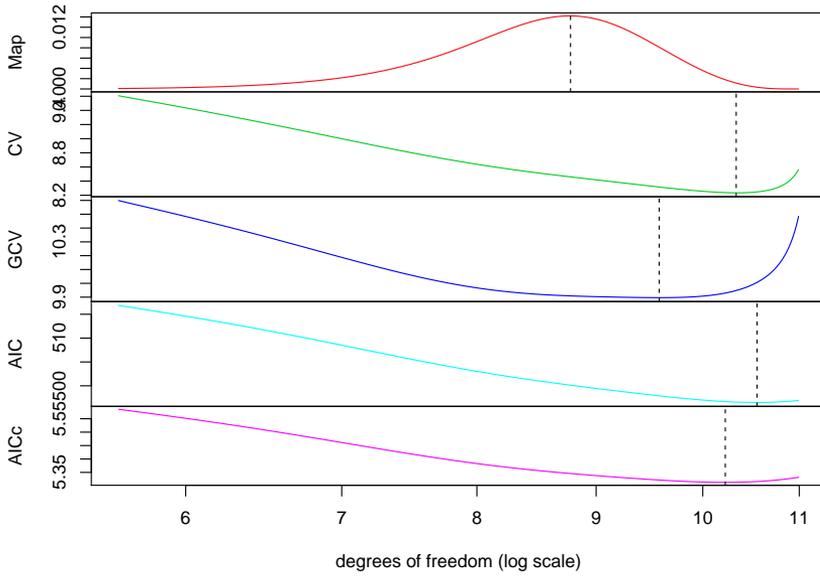


FIGURE 1. Optimal degrees of freedom determination for one-dimensional simulation.

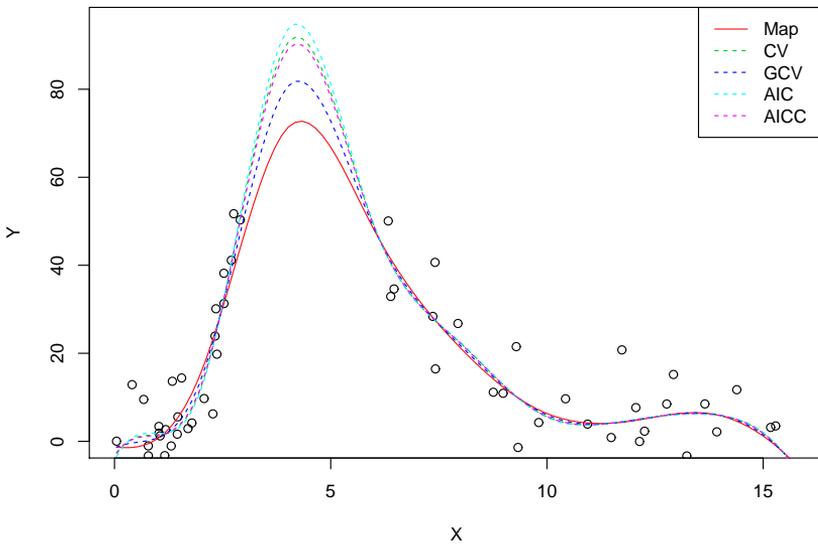


FIGURE 2. Predictions for one-dimensional simulation.

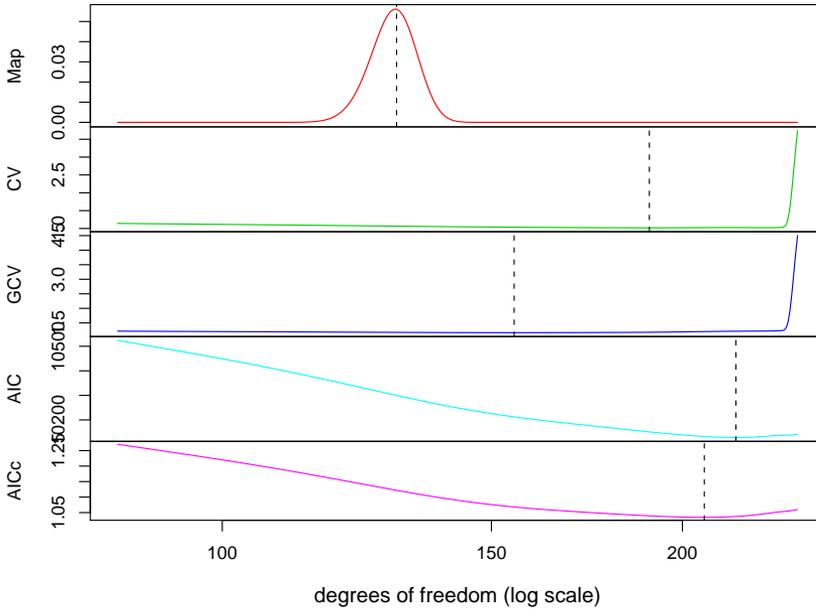


FIGURE 3. Optimal degrees of freedom determination in case study.

A typical approach to deal with this issue is to select a value of λ that minimises some sensible optimisation criterion such as AIC , $AICc$, CV or GCV . But although these methods asymptotically give good prediction error performance, their convergence to optimal values may be slow and they are prone to yielding small values of λ which lead to severe undersmoothing (Wood, 2011). In practice, this can result in "balloonings", as observed in the four upper plots in Figure 4, which are typically triggered by unevenly spaced observations with big gaps in between.

As an alternative, we propose a Bayesian approach. If we denote by M_λ the model resulting for a particular value of the penalisation parameter λ , the initial set-up of the model described in the previous section can be rewritten as $\mathbf{Y}|\boldsymbol{\alpha}, \sigma^2, M_\lambda \sim \mathcal{N}_n(\mathbf{B}\boldsymbol{\alpha}, \sigma^2 \mathbf{I}_n)$ with $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{B} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{\alpha} \in \mathbb{R}^m$. A normal-inverse-gamma prior distribution was placed on the parameters $\boldsymbol{\alpha}, \sigma^2 | M_\lambda$. Because in practice we will be dealing with a finite set of values for λ , its prior distribution has been assumed to be discrete and for simplicity we consider it to be uniform. The derived posterior distribution of the penalisation parameter $w_\lambda = f_{M_\lambda | \mathbf{Y}}$ is then

$$w_\lambda = f_{M_\lambda | \mathbf{Y}} = \frac{\lambda^{\frac{m-2}{2}} G(\lambda)}{\sum_\lambda \lambda^{\frac{m-2}{2}} G(\lambda)} \quad \text{with}$$

$$G(\lambda) = \frac{|B'B + \lambda D'D|^{-\frac{1}{2}}}{\left\{2b + \mathbf{y}' \left[\mathbf{I}_n - B(B'B + \lambda D'D)^{-1} B' \right] \mathbf{y} \right\}^{a + \frac{n}{2}}} f_{M_\lambda}$$

where f_{M_λ} represents the prior distribution placed on λ , which as stated, in our case was assumed to be a discrete uniform.

If we denote by $\hat{\mathbf{y}}_\lambda$ the posterior expectation of a particular fitted value for a given value of the penalisation parameter λ , the Bayesian approach proceeds by means of “*model averaging*”. That is, it estimates the posterior mean across models using $\tilde{\mathbf{y}} = \sum_\lambda w_\lambda \hat{\mathbf{y}}_\lambda$. Alternatively, sometimes $\hat{\mathbf{y}}_\lambda$ is estimated by simply considering the value of $\hat{\mathbf{y}}_\lambda$ which corresponds to the maximum value of w_λ known as *MAP (maximum a posteriori)*.

As an illustration, a one-dimensional simulation was carried out with data having a large gradient for the observations in the neighborhood of a “gap” or “hole”. Figure 2 shows that traditional methods for the choice of the penalisation parameter tend to produce high predictions in the “uncertain” area whereas the Bayesian maximum a posteriori criterion yields a smoother fitting function. Figure 1 shows how in this example, the latter penalises overfitting more seriously while the other methods are prone to severe undersmoothing.

The same analysis was repeated on a real spatiotemporal case study. The data correspond to contamination levels on groundwater due to an industrial operation. Their logarithm was modeled as a function of the position of the wells (represented by black dots in Figure 4 and whose location is largely due to external constraints) and the time in which the measurements were carried out. Figure 3 again shows that the classical methods for the selection of the penalisation parameter favour small values in comparison with MAP. The effect of undersmoothing can be observed in the four upper plots in Figure 4 where high unexpected values of contamination are predicted whereas those corresponding to the MAP criterion present a smoother pattern. It should be mentioned that 24064.9 is the maximum observed value of contamination whereas the maximum predicted values (after transformation onto original scale) is 1010034.0 using the MAP criterion and 5.0e54 in the AIC case.

In conclusion we believe that the Bayesian approach for selecting the smoothing parameter when using P-Splines, can yield better results than the classical methods, in particular when dealing with irregularly spaced data and with a large gradient in the neighbourhood of the separating gaps.

References

- Denison D., Holmes C., Mallick B., Smith A. (2002). *Bayesian Methods for Nonlinear Classification & Regression*. Wiley.

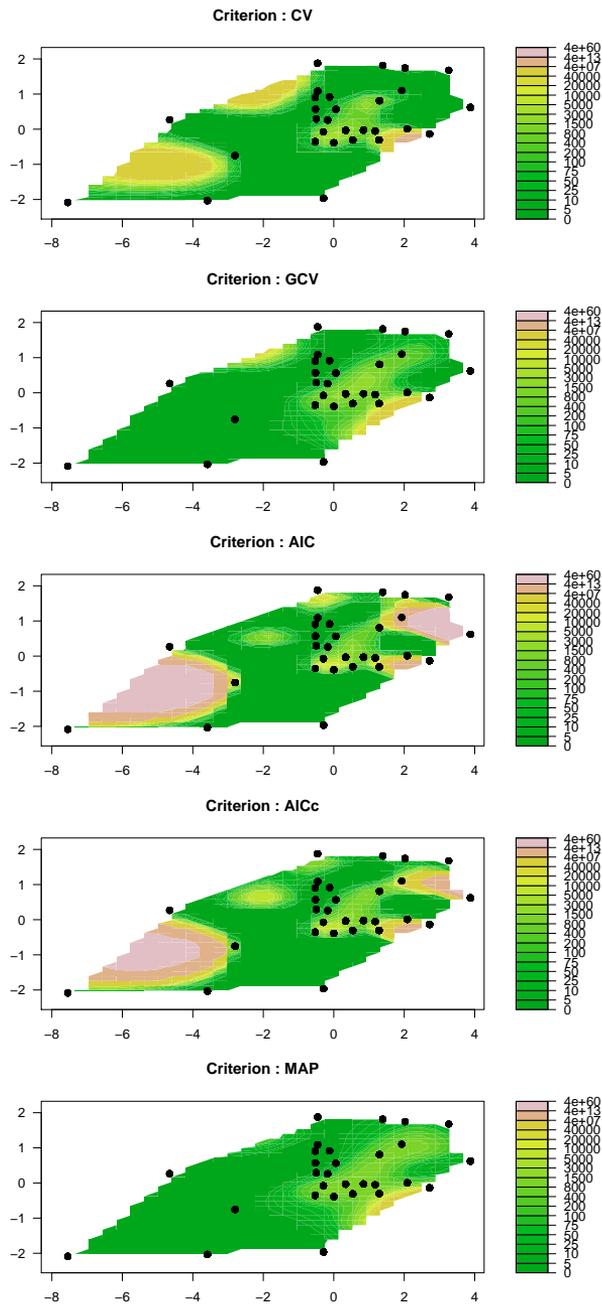


FIGURE 4. Predicted values in case study with penalisation parameter chosen with traditional selection methods and MAP criteria.

Eilers P., Marx B. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science*, **11**, 89–102.

Wood S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation on semiparametric generalized models. *Journal of the Royal Statistical Society: Series B*, **73**, 3–36.

Estimation of the bivariate distribution function: A comparative study

Ana Moreira, Artur Agostinho Araújo, Luís Machado¹

¹ Department of Mathematics and Applications, University of Minho

E-mail for correspondence: a.moreira.cris@gmail.com

Abstract: Let (X, T) be a random vector where the response variable T denotes a lifetime, which is subject to random right censoring, and X denotes a covariate. In this paper we compare several estimators for the bivariate distribution of (X, T) through a simulation study. The methods are applied to bone marrow transplant data where we use one qualitative predictor. The proposed estimators can be applied to ROC curves for censored data as explained by Heagerty, Lumley and Pepe (2000).

Keywords: Beran estimator; Bivariate distribution; Conditional Survival; Kaplan-Meier; gap times.

1 Introduction

Let T be the survival time of individuals and let X be a quantitative random covariate (like blood pressure, age, diagnostic test or marker, etc.). Because of censoring, T is subject to random right censoring which we denote by C and assume to be independent of T . Because of this, we only observe (X_i, Y_i, Δ_i) , which are n independent replications of (X, Y, Δ) , where $Y = \min(T, C)$ and $\Delta = I(T \leq C)$. Since the censoring time is assumed to be independent of the process, the distribution of T , say F may be consistently estimated by the Kaplan-Meier estimator based on the (Y, Δ) .

In this paper we are interested in estimating the bivariate distribution function: $F(x, t)$, that can be computed for any quantitative predictor x and any time t . The manuscript is organized as follows. In the next section, the methods will be briefly introduced. Some results are given for a simulation experiment. Illustrative real data applications are provided in Section 4. The main body of the paper ends with a discussion section (Section 5).

2 Methods

The times between consecutive events (gap times) are often of interest and lead to problems that have received much attention recently. A recent package called *survivalBIV* was recently developed that consider the estimation

of the bivariate distribution function for censored gap times. The problem that we consider in this paper is more simpler since we are assuming that the first gap time is uncensored. The aim of this paper is to compare several methods for the bivariate distribution of the pair (X, T) where T is a lifetime subject to right random censoring and X is assumed to be a quantitative covariate, for example a diagnostic marker. The choice of an appropriate method for this setting can be very important in several applications, in particular to time-dependent ROC curves. Heagerty, Lumley and Pepe (2000) propose the use of two estimators for the bivariate distribution function of (X, T) . The first estimator (CKM) is based on using the (conditional) Kaplan-Meier estimator for each possible subset of X . However, this estimator does not guarantee the necessary condition that sensitivity and specificity are monotone in X . An alternative estimator that does guarantee monotonicity is based on a nearest neighbor estimator (NNE). In this paper we compare these two estimators with four additional estimators. A brief description of these estimators is given in the paper by Moreira and Meira-Machado (2012).

The bivariate distribution function of (X, T) can be obtained using the estimators developed for censored gap times. Two estimators were recently proposed using the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data (de Uña-Álvarez and Meira-Machado, 2008; de Uña-Álvarez and Amorim, 2011). Difference between these two methods (hereafter denoted by KMW and KMPW) is that the more recent paper uses a presmoothed version of the Kaplan-Meier estimator. The estimator proposed by Lin et al. (1999), which is based on inverse probability of censoring weighted can also be used to this context (Lin).

A valid estimator of the bivariate distribution function, was recently provided by Van Keilegom, de Uña-Álvarez and Meira-Machado (2011). This estimator (LS) is based on the work of Akritas (1994). This methodology assumes that the vector (X, T) satisfies the nonparametric location-scale regression model, allowing for the transfer of tail information from lightly censored areas to heavily ones. This estimator was proposed as an attempt to solve this inconsistency problem of the Beran estimator (Beran, 1981). To compute the kernel bandwidth required for the Beran estimator we use the `dpik` function from the `KernSmooth` R package.

3 Simulation Study

In this section, we compare by simulations the six estimators (CKM, NNE, KMW, KMPW, Lin and LS), for the bivariate distribution function of (X, T) . We consider the same scenario as 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, $F(x, y) = F_1(x)F_2(y)[1 + \delta(1 - F_1(x))(1 -$

	t	0.2231	0.5108	0.9163	1.6094	2.3026	2.9957
	x						
CKM	0.2231	0.00063	0.00112	0.00147	0.00163	0.00170	0.00172
	0.5108	0.00109	0.00182	0.00232	0.00261	0.00272	0.00279
	0.9163	0.00140	0.00226	0.00280	0.00299	0.00311	0.00324
	1.6094	0.00158	0.00250	0.00296	0.00295	0.00298	0.00314
	2.3026	0.00164	0.00257	0.00296	0.00278	0.00275	0.00301
	2.9957	0.00167	0.00261	0.00295	0.00266	0.00255	0.00286
NNE	0.2231	0.00061	0.00109	0.00142	0.00156	0.00163	0.00167
	0.5108	0.00108	0.00180	0.00228	0.00250	0.00269	0.00295
	0.9163	0.00138	0.00223	0.00277	0.00290	0.00328	0.00407
	1.6094	0.00158	0.00250	0.00293	0.00289	0.00342	0.00521
	2.3026	0.00164	0.00258	0.00295	0.00278	0.00338	0.00592
	2.9957	0.00167	0.00261	0.00293	0.00267	0.00325	0.00624
Lin	0.2231	0.00073	0.00125	0.00153	0.00160	0.00161	0.00160
	0.5108	0.00134	0.00226	0.00267	0.00267	0.00259	0.00250
	0.9163	0.00166	0.00295	0.00361	0.00340	0.00306	0.00278
	1.6094	0.00199	0.00298	0.00386	0.00378	0.00324	0.00262
	2.3026	0.00370	0.00318	0.00352	0.00361	0.00322	0.00252
	2.9957	0.00751	0.00485	0.00392	0.00362	0.00329	0.00249
KMW	0.2231	0.00061	0.00107	0.00151	0.00196	0.00223	0.00237
	0.5108	0.00106	0.00182	0.00270	0.00409	0.00520	0.00606
	0.9163	0.00138	0.00227	0.00337	0.00562	0.00811	0.01048
	1.6094	0.00163	0.00251	0.00330	0.00538	0.00865	0.01273
	2.3026	0.00177	0.00270	0.00311	0.00404	0.00637	0.01038
	2.9957	0.00187	0.00293	0.00325	0.00316	0.00404	0.00685
KMPW	0.2231	0.00051	0.00101	0.00155	0.00216	0.00248	0.00264
	0.5108	0.00093	0.00185	0.00304	0.00484	0.00612	0.00694
	0.9163	0.00121	0.00221	0.00359	0.00624	0.00881	0.01082
	1.6094	0.00144	0.00230	0.00303	0.00512	0.00829	0.01135
	2.3026	0.00159	0.00249	0.00270	0.00346	0.00580	0.00866
	2.9957	0.00173	0.00281	0.00287	0.00251	0.00369	0.00582
LS	0.2231	0.00051	0.00137	0.00216	0.00252	0.00269	0.00278
	0.5108	0.00123	0.00329	0.00529	0.00638	0.00705	0.00756
	0.9163	0.00195	0.00501	0.00835	0.01097	0.01256	0.01389
	1.6094	0.00268	0.00658	0.01104	0.01607	0.01901	0.02151
	2.3026	0.00298	0.00735	0.01231	0.01879	0.02255	0.02574
	2.9957	0.00308	0.00769	0.01294	0.02022	0.02439	0.02793

TABLE 1. Estimated mean square error values for the bivariate exponential distribution. Sample size of $n = 100$, uniform censoring $C \sim U[0, 3]$.

$F_2(y))$ 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 X and T) for the bivariate exponential distribution is obtained for $\delta = 1$ with bound equal to 0.25. The uniform censoring time C was generated according to model $U[0, 3]$. For this scenario we have considered several sample sizes, but we only show the results for $n = 100$, based on 10000 generated samples. For each setting we computed the mean square error and standard deviations for the bivariate estimators at several pairs of time points (x, t) . Results are shown in Table 1. Results reveal that the NNE method and the CKM are the ones with less MSE. The method based on presmoothing leads to good results for some quantiles.

4 Example of Application

To illustrate our methods we will use the Bone-Marrow transplant data (Copelan et al., 1991) dealing with bone marrow transplants for leukemia patients. For more details see Klein & Moeschberger (page 3). The data set is available in the KMsurv package of R. In this data set, T is set as the disease free survival time (time to relapse, death or end of study) and assume X to be the time to acute graft-versus-host disease (GVHD). For

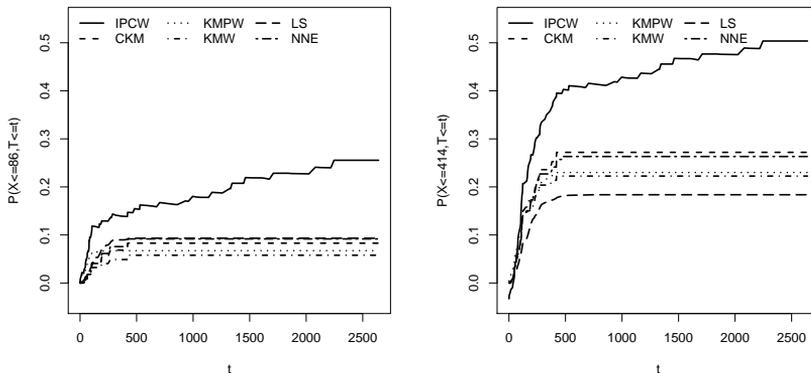


FIGURE 1. Bivariate distribution, $F(x, t)$ for the Bone Marrow Transplant data for $x = 86$ and $x = 414$ (X is time to GVHD).

illustration purposes we show in Figure 1 the plot for $F(x, t)$ for all methods by fixing $X = 86$ and $X = 414$ (first and second quantiles). From this plot we can see the behavior of all methods.

5 Discussion

In this paper we present several estimators for the bivariate distributions of (X, T) where T is observed subject to right random censoring and X denotes a quantitative covariate. These estimators have several applications and extensions. One of these applications is the use of these methods to time-dependent ROC curves. ROC curves can be used to display sensitivity and specificity of a continuous diagnostic test or marker, X , for a binary disease variable of disease status (D). In these cases, the outcome variable is time dependent and we can define a time dependent sensitivity and specificity and plot time dependent ROC curve $ROC(t)$ as sensitivity(t) againsts 1- specificity(t). The proposed estimators permit a useful extension of their methodology to the case where the quantitative covariate is also subject to censoring.

Acknowledgments: The authors acknowledge financial support by Grant *PTDC/MAT/104879/2008* (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education. Ana Moreira acknowledges financial support by grant *SFRH/BD/62284/2009* of the Portuguese Ministry of Science, Technology and Higher Education. This research was financed by FEDER Funds through Programa Operacional Fac-

tores de Competitividade COMPETE and by Portuguese Funds through FCT - Fundação para a Ciência e a Tecnologia, within the Project Est -C/MAT/UI0013/2011.

References

- Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*, **22**, 1299–1327.
- Beran, R. (1981). *Nonparametric regression with randomly censored survival data*. Technical report, University of California, Berkeley.
- Copelan et al. (1991) Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with Bu/Cy. *Blood*, **78**, 838–843.
- de Uña-Álvarez, J. and Meira-Machado, L. (2008). A Simple Estimator of the Bivariate Distribution Function for Censored Gap Times. *Statistics and Probability Letters*, **78**, 2440–2445.
- de Uña-Álvarez, J. and Amorim, A. P. (2011). A semiparametric estimator of the bivariate distribution function for censored gap Times. *Biometrical Journal*, **53**, 113–127.
- Heagerty, P. J., Lumley, T. and Pepe, M. S. (2000) Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*, 337–344.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.
- Klein, J. P. and Moeschberger, M. L. (1997) *Survival analysis - techniques for censored and truncated data*. Springer-Verlag, New-York.
- 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.
- Moreira, A., Meira-Machado, L. (2012) survivalBIV: Estimation of the Bivariate Distribution Function for Sequentially Ordered Events Under Univariate Censoring. *Journal of Statistical Software*, March Volume 46, Issue 13.
- Van Keilegom, I., de Uña-Álvarez, J. and Meira-Machado, L. (2011). Nonparametric location-scale models for successive survival times under dependent censoring *Journal of Statistical Planning and Inference*, **141**, 1118–1131.

Spatial regression of quantiles based on parametric distributions

Chenjerai Kathy Mutambanengwe¹, Christel Faes¹, Marc Aerts¹

¹ Hasselt University, Belgium

E-mail for correspondence: chenjerai.mutambanengwe@uhasselt.be

Abstract: Inadequate hormone production adversely affects the tissues, resulting in many diseases, the most devastating of which are the consequences on the developing human brain. As such, it is necessary to be able to determine which factors may be causing such extreme values so that control measures may be put into place where applicable. It is defined that $< 3\%$ of TSH concentrations $> 5\mu IU/mL$ indicates iodine sufficiency in a population. In this regard, several models are fitted in order to determine the quantiles for thyroid stimulating hormone (TSH) levels in a population in Spain. First the mean (natural parameters) of the response is estimated and the resulting parameters are used to estimate the quantile of interest; then the quantiles are modelled directly using Bayesian methods; and the models are compared.

Keywords: CAR; Quantiles; Spatial; Thyroid stimulating hormone.

1 Introduction

In the beginning of the millennium, iodine deficiency was still spread throughout Spain. Mostly, the iodine deficiency is mild to moderate degree, but some areas still have severe iodine deficiency (Vitti *et al*, 2001). Studies in different parts of Catalonia in Spain have suggested that most school children and general population have acceptable iodine nutrition, however, about 50% of the pregnant women in this region had insufficient iodine supply (Serra-Prat *et al* 2004; Vila *et al* 2006). Since neonatal TSH concentration is increased when the supply of thyroid hormone and iodine from the maternal circulation to the foetus has been compromised, measuring the TSH levels in newborns can be an indicator of iodine sufficiency in a population. Although WHO has suggested that the frequency of moderately elevated thyrotropin concentrations in newborn screening programs can be used to assess the severity of iodine deficiency in a population, the cutoff values for defining severity are uncertain. Current cut-off values suggest that $< 3\%$ of TSH concentrations $> 5\mu IU/mL$ indicates iodine sufficiency in a population (WHO, 2007). However, multiple factors affect

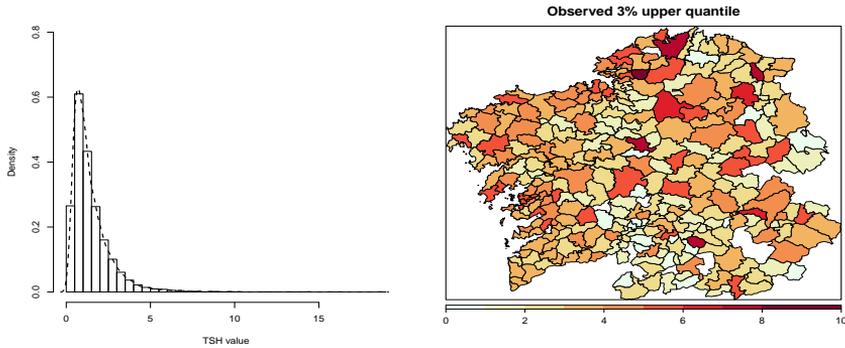


FIGURE 1. Left: Histogram of observed data with empirical density function. Right: Observed 3% upper quantile.

the measurement of newborn TSH including prematurity, the timing of heel prick for sample to be taken, among others (Li and Eastman 2010). Due to this, it is of interest to explore the underlying distribution of this 97% quantile in order to more clearly understand the dynamics behind the TSH concentrations in the newborns. Naturally, differences may exist in certain characteristics of a population when space is taken into account; with locations closer to each other being more similar than locations further apart. Recently, research in spatial quantile regression has been expanding as this topic gathers interest in the statistical world. In this paper, quantile regression is used to investigate the population based characteristics of TSH. We demonstrate how Bayesian methods can be used with a choice of parametric distributions in order to estimate the quantiles of the distribution whilst taking into account spatial dependencies using conditionally autoregressive random effects on the response.

2 Data

TSH measurements were taken in newborns across 297 municipalities in Galicia, Spain. A total of almost 15000 observations were obtained in 2009. Covariates available for the data include gender, birth weight (kg), time to feeding (hours), type of feeding, time to heel prick (hours). Figure 1 shows a histogram of the observed TSH levels as well as the corresponding 97% quantile by municipality. The TSH responses shows a right-skewed distribution, and hence logical choices of a parametric distribution for the response include lognormal, loglogistic, log double exponential, gamma, weibull, skew-normal, skew-t, inverse gamma, and so forth.

3 Methodology

3.1 Quantile estimation

A distribution is usually characterised by some natural parameters θ such that its probability density function is a function $f(X, \theta)$, for example a normal distribution has parameters mean μ and variance σ^2 . Usually quantiles are derived as a function of these parameters as $Q_X(p) = F^{-1}(p)$. In this paper, two methods are used; first the natural parameters θ are modelled and the relevant quantiles are estimated from the model parameters as is usually done. The parameters θ are allowed to depend on covariates as well as random effects as explained in the next section. In the second case, two quantiles, $Q_1(p_1)$ and $Q_2(p_2)$ are modelled directly, and it is shown how any other desired quantile can be estimated from the model parameters (Cook, 2010). As in the first case, both quantiles are allowed to depend on covariates as well as random effects.

3.2 Dealing with heterogeneity

Normally distributed random effects on the model parameters (either mean and variance, or quantiles) may be assumed where, for example, each model parameter is fitted with a random effect, b_0 and b_1 , each with its own variance. An example of such a random effect structure is

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim N \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}$$

Spatially correlated random effects may also be assumed, and in this case a conditionally autoregressive (CAR) structure is considered where

$$\phi_i | \phi_j, \tau_\phi \sim N(\bar{\phi}_i, 1/m_i \tau_\phi)$$

where m_i is the number of neighbours for location i , $\bar{\phi}_i$ is an average of effects from the neighbours j in the vicinity of location i , and the variance of the random effects is $1/\tau_\phi$.

4 Results

4.1 Model selection

The results showed that the location-scale models (Lognormal and Loglogistic) always had a better fit than the Weibull models. Models that incorporate random effects fit the data better than the models with fixed effects only, meaning that one common structure is not applicable for all the municipalities in the study region. Furthermore, models with spatially

TABLE 1. Summary of DIC fit statistics for lognormal distribution

Type of effects	Structure	Lognormal	
		Natural	Quantile
Fixed (no random)	No covariates	36961.6	36961.4
	Main effects	36362.5	36361.7
	Interactions	36272.8	36292.8
Normal random	No covariates	36603.4	36628.6
	Main effects	36167.6	36176.5
	Interactions	36090.2	36107.8
CAR	No covariates	36565.6	36573.3
	Main effects	36140.0	36138.4
	Interactions	36064.3	36071.3

correlated random effects fit better than models with independent normal random effects, meaning that incorporating spatial effects is necessary. Whilst the Loglogistic model seemed to fit better for simpler models, we faced some convergence issues as the models got more complex. Results are therefore based on the Lognormal model and fit statistics are shown in Table 1.

4.2 Spatial effects

Using the parameter estimates from the final models, the 97% quantile was estimated and the plots are shown in Figure 2 and 3. As can be seen, there are slight differences between the maps using the two different methods, but the overall spatial trends remain similar. According to the fit statistics in Table 1, the model which estimates the distribution's natural parameters fits slightly better as it has a lower DIC than the model which estimates the quantiles directly. However, the difference is not very large so the two models may be considered to be behaving in a similar manner.

5 Discussion and Conclusion

Estimation of quantiles has been carried out using two different models. In the first case the natural parameters of a distribution are modelled, and the desired 97% quantile is then estimated. In the second case two quantiles are modelled directly and then the desired quantile is estimated from the model. Inclusion of spatially correlated random effects in the models significantly improved the model fit for both cases, indicating that neighbouring regions may be quite similar in terms of TSH levels, and as a consequence in terms of iodine deficiency. Velilla *et al* (2010) assessed TSH concentrations in three other regions of Spain (Huelva, Seville and Cordoba) and they also

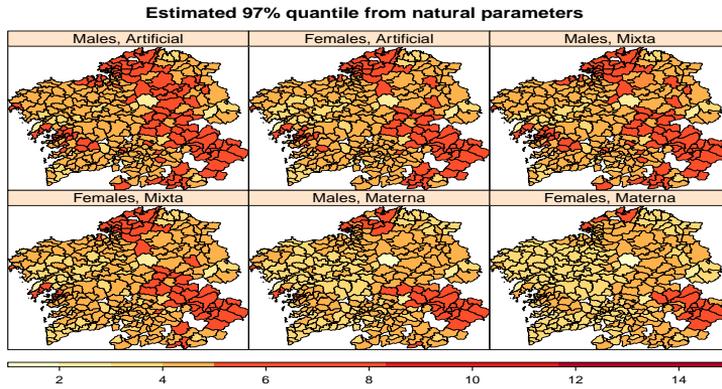


FIGURE 2. 97% quantile estimates from modelling natural parameters.

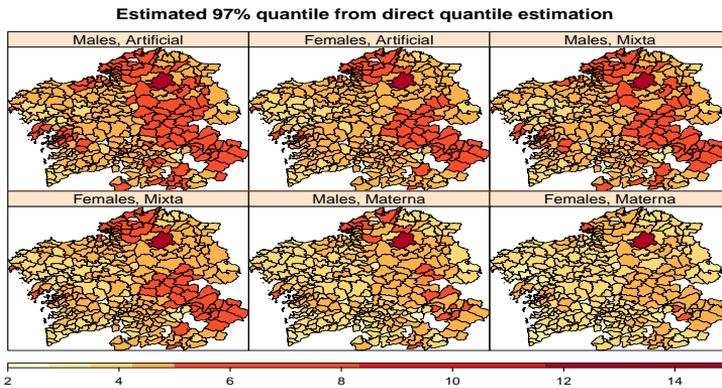


FIGURE 3. 97% quantile estimates from modelling direct quantiles.

found a heterogeneous distribution which appeared to indicate an irregular and deficient iodine intake, thus supporting the use of random effects in our study. It was noted that in the case where homoscedasticity is assumed, there is minimal difference between modelling the natural parameters of the distribution and modelling the quantiles directly, as you obtain similar

models. On the other hand, when the parameters are allowed to vary up to a constant, the difference between quantiles is a simple shift on the scale of the response. In addition, if we condition on all other covariates, and find an effect of one covariate insignificant in the variance estimator, then we expect the same effect on all quantiles of the distribution. This can be viewed as homoscedasticity in the quantile space.

In conclusion, there is minimal difference between modelling quantiles directly as opposed to modelling natural parameters and then estimating the quantiles. However, modelling quantiles has added advantage of direct interpretation for covariate effects readily available, and spatial effects are estimated directly on the desired quantile as well. In addition, any third quantile may be derived from a fitted model estimating two arbitrary quantiles.

References

- Cook, J.D. (2010). *Determining distribution parameters from quantiles*. UT MD Anderson Cancer Center Department of Biostatistics Working Paper Series. Working Paper 55.
- Li, M and Eastman, C.J. (2010). Neonatal TSH screening: is it a sensitive and reliable tool for monitoring iodine status in populations?. *Best Practice & Research Clinical Endocrinology & Metabolism*, **24**, 63–75.
- Serra-Prat, M., De Castro, A., Palomera, E., Casamitjana, R., Vila, L., and Puig-Domingo, M. (2004). Prevalence of iodine deficiency in pregnancy and effects of its replacement: results of the Mataró study. *Endocrinol Nutr.*, **51**, 21–22.
- Velilla, T.A., Rodríguez, C.G., Sánchez, A.B., Portillo, C.M., de la Vega, J.A., Cerrato, S.B., Baldrich, A.G., Montávez, J.M., Pérez, A.S., Arriero, J.M. and Ortiz, R.G. (2010). Using newborn congenital hypothyroidism screening specimens to detect iodine deficiency in three regions of Spain. *An Pediatr (Barc)*, **72**, 121–127.
- Vila, L., Castell, C., Wengrovicz, S., de Lara, N. and Casamitjana, R. (2006). Urinary Iodide Assessment of the Adult Population in Catalonia. *Med Clin (Barc)*, **127**, 730–733.
- Vitti, P., Rago, T., Aghini-Lombardi, F. and Pinchera, A. (2001). Iodine deficiency disorders in Europe. *Public Health Nutrition*, **4**, 529–535.
- World Health Organization (WHO) (2007). *Assessment of iodine deficiency disorders and monitoring their elimination: a guide for program managers*. 3rd ed. Geneva, Switzerland: World Health Organization.

Analysis of pseudo-panel data with dependent samples

Ainhoa Oguiza ¹, Inmaculada Gallastegui ¹, Vicente Núñez-Antón ¹

¹ Departamento de Econometría y Estadística (E.A. III), Universidad del País Vasco UPV/EHU, Lehendakari Aguirre, 83, E-48015 Bilbao, Spain.

E-mail for correspondence: ainhoa.oguiza@ehu.es

Abstract: In this paper we discuss a model for pseudo-panel data when some but not all of the individuals stay in the sample for more than one period. We use data on the labor market of the Basque Country from 1993 to 1999 treated through Fortran-77 programming. We construct economically reasonable age cohorts for active population and use gender, qualification and social status as explanatory variables in our model. Given the class of data we use, we analyze the properties of the random error and estimate the model through maximum likelihood, finding significant results from an applied point of view.

Keywords: panel data; pseudo-panels; time-related samples.

1 Introduction

When we have observations on a set of individuals along different periods of time, we say that we have a ‘panel of data.’ However, we may have observations on sets of individuals that change from one period to another, which do not constitute a panel of data. An example of this are the data obtained at the Family Expenditure Surveys which are held by many countries.

Deaton (1985) established a procedure to convert these independent samples into ‘pseudo-panels’ by constructing a ‘pseudo-individual’ through the mean of individuals who share a given characteristic (for instance, having an age between 20 and 25 years). If the number of individuals belonging to each cohort is or can be assumed to be very large, the usual covariance estimator can still be used; otherwise, Deaton proposed a new instrumental variables estimator. Following this line of reasoning, some authors have worked with panel data assuming that the number of individuals is sufficiently large.

The main difference in our approach with respect to previous research is that we do not consider the case of independent samples, but rather introduce time dependence between them. Thus, while using pseudo-panels, the work presented in this paper follows a different approach. We consider

the case of having observations on individuals at different periods of time, but not all individuals change from one period to another, staying in the sample for a given number of periods. Therefore, successive samples are not independent for different periods as the individuals rotate and, thus, do not stay in the sample for all periods.

2 Methodological issues

In this section we propose a model for a pseudo-panel of serially correlated samples. Starting with Deaton's model:

$$y_{i(t)t} = \alpha_{i(t)} + \mathbf{x}'_{i(t)t}\boldsymbol{\beta} + u_{i(t)t}, \quad t = 1, \dots, T; \quad i(t) = 1(t), \dots, N(t), \quad (1)$$

where $\alpha_{i(t)}$ is the individual fixed effect; $\boldsymbol{\beta}$ are the slope parameters; $y_{i(t)t}$ is the value of the dependent variable for individual i in period t ; $\mathbf{x}_{i(t)t}$ are the explanatory variables for individual i in period t ; $u_{i(t)t}$ is the error term, assumed to be independent and identically distributed; $i(t)$ denotes the i th individual and the subindex (t) is used to indicate that the i th individual is different in each period.

The data represented above do not constitute a panel. Deaton proposed to group individuals in cohorts using some common characteristic, for instance, age, and calculate the mean of individuals i belonging to a given cohort (or age group), c . Denoting the number of individuals in cohort c as n_{ct} . We can write the model as:

$$\bar{y}_{ct} = \bar{\mathbf{x}}'_{ct}\boldsymbol{\beta} + \alpha_c + \bar{u}_{ct}, \quad c = 1, \dots, C; \quad t = 1, \dots, T, \quad (2)$$

where \bar{y}_{ct} is the mean of the observed values of $y_{i(t)t}$, in cohort c in period t , C is the number of cohorts and T the number of time periods. We can see that $\bar{\alpha}_{ct}$ is the mean of the individual fixed effects for individuals belonging to cohort c in period t . As in Deaton, we consider that the number of individuals in each cohort, n_{ct} , is sufficiently large, so that its number remains constant in time (i.e., $n_{ct} \simeq n_c$), and so $\bar{\alpha}_{ct} \simeq \alpha_c$.

2.1 A model for interrelated samples

On the basis of the model above, we can now propose a new model that can accommodate different assumptions. We start by analyzing the model for the cohort means:

$$\bar{y}_{ct} = \bar{\alpha}_{ct} + \bar{\mathbf{x}}'_{ct}\boldsymbol{\beta} + \bar{u}_{ct}, \quad c = 1, \dots, C; \quad t = 1, \dots, T \quad (3)$$

On the one hand, the number of individuals in each cohort, n_{ct} , will be, in our case, sufficiently large, so that we may assume, as in Deaton, that

$n_{ct} \simeq n_c$. On the other hand, since we do not really have a random sample of individuals in each period, we propose to use the model:

$$\bar{\alpha}_{ct} \approx \frac{1}{n_c} \sum_{i=1}^{n_c} \alpha_i \approx \frac{1}{n_c} \sum_{i=1}^{n_c} (\alpha_c + \epsilon_{it}) = \alpha_c + \bar{\epsilon}_{ct}, \text{ with } \epsilon_{it} \sim N(0, \sigma_\epsilon^2), \quad (4)$$

where we separate the individual effect, α_i , into two terms, α_c and ϵ_{it} . In model (4), α_c is a fixed effect representing the specificity of each cohort and is assumed to be constant over time, whereas ϵ_{it} represents the random effect in the sample due to the presence of individuals which remain several periods in the same cohort. Therefore, we propose a mixed model: on the one hand, α_c represents the specific fixed effect of each cohort, on the other, $\bar{\epsilon}_{ct}$ (the cohort mean over the individuals), is a random effect that allows for the existence of correlation between different cohorts in time. Replacing the term in equation (4) into equation (3), and writing it in a vectorial form we have:

$$\bar{y}_c = e \quad \alpha_c + \bar{X}_c \beta + \bar{v}_c, \quad c = 1, \dots, C \quad (5)$$

The total random error is distributed with mean and variance given by:

$$\begin{aligned} E(\bar{v}_c) &= \mathbf{0}, \text{ and} \\ E(\bar{v}_c \bar{v}_c') &= \begin{pmatrix} \sigma_{c,11}^2 & \sigma_{12}^c & \dots & \sigma_{1T}^c \\ \sigma_{12}^c & \sigma_{c,22}^2 & \dots & \sigma_{2T}^c \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1T}^c & \sigma_{2T}^c & \dots & \sigma_{c,TT}^2 \end{pmatrix} = V_c, \text{ for each } c = 1, \dots, C, \end{aligned}$$

where the elements of V_c , the variance-covariance matrix, are variances and covariances between the different cohorts.

2.2 Assumptions on the covariance matrix

As can be seen in the variance-covariance, the number of parameters in the covariance matrix is very large. Thus, it would be interesting to reduce it while keeping enough diversity to accommodate the structure of the data. More specifically, and given that we work with a set of data in which some individuals remain in the sample for a number of periods, we know that the random error will clearly be correlated over time. For each cohort c , $c = 1, \dots, C$, and two periods t and s , $t \neq s$, we will have:

$$\text{Cov}(\bar{\epsilon}_{ct}, \bar{\epsilon}_{cs}) = \frac{n_{c,ts}}{n_{ct}n_{cs}} E(\epsilon_{it}\epsilon_{is}), \quad (6)$$

where n_{ct} and n_{cs} are the number of individuals in cohort c in periods t and s , and $n_{c,ts}$ is the number of individuals in cohort c which are common to periods t and s . Moreover, this last number (i.e., $n_{c,ts}$) will become smaller as periods are farther away, and, therefore, the aforementioned covariance will decrease, so that, for instance, $n_{c,12} > n_{c,15}$. On the one hand, since individuals remain in the sample for a number of periods and then drop out, one could think of a moving average structure for the random term. On the other hand, autoregressive schemes cannot be discarded as they introduce a correlation which decreases in time. The actual scheme to be selected will depend on the characteristics of the data and will be discussed.

3 Labor market data

The data we use come from a large data base obtained from EUSTAT (1986) (The Basque Statistics Office). The information we have on this survey corresponds to the period going from the second quarter of 1993 to the fourth quarter of 1999. The sample unit is the family and there are observations on different variables, with the following structure: between 1993-2 (i.e., second quarter of 1993) and 1997-4 (i.e., fourth quarter of 1999) there are 5,000 families in the sample (between 16,000 and 16,500 individuals) and they remain in the sample for 8 time periods; between 1998-1 and 1999-4 there are 3,750 families (11,500-12,000 individuals) and they remain in the sample for 6 time periods. A more detailed descriptive analysis and description of the data can be found in Oguiza et al. (2003) and Oguiza et al. (2004).

The variables selected for the analysis are: active population, which will be the response variable, and gender, qualifications and familiar status, which will be used as explanatory variables. It should be noted that all variables are included in the sample as Bernoulli variables. This means that, when the cohort mean is computed to construct the pseudo-panels, each of these variables will represent the proportion of individuals in the cohort which belongs to each of the categories, and will take values in the interval $[0, 1]$, a fact that should be taken into account when interpreting the results.

4 Modelling proposal application

First of all, we must construct the cohorts. As stated above, the variable age has been selected to construct the cohorts, which will consist of individuals in different age groups. the cohorts selected correspond to the following intervals of age: $[16 - 20]$, $[21 - 24]$, $[25 - 27]$, $[28 - 30]$, $[31 - 35]$, $[36 - 40]$, $[41 - 45]$, $[46 - 50]$, $[51 - 55]$, $[56 - 60]$, $[61 - 65]$.

As already stated, the covariance matrix of the disturbances contains a large number of unknown parameters, which would be very convenient to

reduce. There are two questions that we could explore, involving assumptions on the variances and on the covariances of the error term.

First, we will model the covariance structure. As a first step, we will compute the correlogram and partial autocorrelation function of the data in order to be able to identify a suitable error term covariance structure. Some criteria will be needed to help in the choice of a suitable structure. Akaike's Information Criterion (AIC) is a good choice to compare non-nested models. The Bayesian Information Criterion (BIC) adds a penalty for excessive parametrization. Both of them, along with the Likelihood Ratio Test (LRT) for nested models will be used. This exploratory analysis has been performed for all 11 cohorts.

In our view and on the basis of the results we have obtained, there are two possible model selection criteria to follow in order to be able to select the best or more adequate covariance model for each of the cohorts in the study. The first criterion could be to use the best selected models for each cohort and fit them to the data set under study. The second criterion could be the one based on practical guidelines and practitioner's convenience of the model selection process. In this sense, one may decide to fit one very general single process, that somehow encompasses most of the best selected processes for each individual cohort, to all of them. Therefore, under this second approach and without loss of generality, we fit a moving average process of order 3 (i.e., an MA(3) process) to all of the cohorts in the study.

Now, we will analyze the behavior of the variances. We would like to test for the possibility that variances within a given cohort remains constant, or if variances for a given time period are the same for the different cohorts. The specific tests for these assumptions make use of Harris' test. We conclude that, for each cohort, its variance does not vary over time, whereas, for a given time period, variances are different for the different cohorts.

5 Model Estimation

We will now propose a complete model for our data, the response variable is the Active Population in the Basque Country, whereas the explanatory variables to be used are Gender, Qualification and Status. Therefore, and based on our proposed methodology, the model will be:

$$\bar{y}_{ct} = \boldsymbol{\mu}_{[c]} + \beta_1 \bar{x}_{ct} + \beta_2 \bar{z}_{ct} + \beta_3 \bar{s}_{ct} + \bar{v}_{ct}, \quad c = 1, \dots, 11, \quad t = 1, \dots, 27,$$

As suggested by the covariance model selection procedure, we will estimate the model under the two aforementioned approaches. Estimation of the general model is performed by using maximum likelihood estimation methods. Under both approaches (i.e., the one that assumes the best fitting covariance model for each cohort, and the one where a common covariance

model is assumed for all cohorts), we have fitted the models where cohort variances are assumed to be different and also equal. To test for the best fitting model under each specific approach, likelihood ratio tests for nested models were performed, where the null hypothesis of equal cohort variances was tested. Once the best model under each approach is selected, the best models are compared by using goodness-of-fit model selection criteria for non-nested models, such as AIC and BIC.

If we compare the results for the mean model parameter estimates under both approaches, we observe that the coefficients associated to the different intercepts for each cohort are statistically significant. Those coefficients capture the mean value of the proportion of active population in the group of men, with no qualification and head of the household, for each age group or cohort. Thus, the sign of the intercepts as well as their values are, from a practical point of view, reasonable. We could thus say that, for men in the age group between 25 and 55 years, the proportion of mean active population estimated for men with no qualification and head of the household remains between 84 – 86% and 68%, while for younger (i.e., 16-20 years), or older men (i.e., 61-65 years), this percentage is considerably reduced.

Acknowledgments: This research was supported by Ministerio Español de Ciencia e Innovación, FEDER, and Departamento de Educación del Gobierno Vasco (UPV/EHU Econometrics Research Group) under research grants MTM2010-14913 and IT-334-07.

References

- Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics* **30**, 109-126.
- EUSTAT (1986). Encuesta Continua de la Población en Relación con la Actividad. *Eusko Jaurlaritz/Gobierno Vasco: Vitoria/Gasteiz*.
- Oguiza A., Gallastegui I. and Núñez-Antón V. (2003). Búsqueda de empleo en la CAPV (1993-1999). Cualificación y Género. *Cuadernos de Economía* **26**, 233-257.
- Oguiza A., Gallastegui I. and Núñez-Antón V. (2004). La población ocupada en la CAPV (1993-1999). Género y formación como características relevantes. *Estadística Española* **46**, 229-292.

Shrinkage estimation when calibrating in the presence of random effects

Samuel D. Oman¹

¹ Department of Statistics, Hebrew University, Mount Scopus, Jerusalem 91905, Israel

E-mail for correspondence: oman@mscc.huji.ac.il

Abstract: We wish to use a calibration set of (x, Y) values to estimate the linear relation between Y , an inexact measurement, and x , a precise measurement which is, however, more expensive or difficult to obtain. At the prediction step, only Y will be observed and we wish to estimate the corresponding unknown x , which we denote by ξ . Assume that at the calibration step we have repeated observations for different sampling units, and that Y at the prediction step will be for a new sampling unit. We show that if ξ is centered about a known value c , for example the mean x -value of the calibration set, then the estimator of Oman (1998) now shrinks to c ; and that this can result in substantially more accurate predictions. We illustrate this method on a set of measurements of true bladder-volumes (x) and ultrasound measurements (Y).

Keywords: Calibration; Mixed model; Prediction; Shrinkage.

1 Introduction

Suppose the scalar variable x is a precise value of a quantity of interest, and Y is an imprecise value which is, however, cheaper or more easily obtained. We wish to use a calibration set, comprising measurements of both x and Y , to estimate the relation between the variables. At the future, prediction step, only Y will be obtained, and we wish to estimate the corresponding unobserved x . If the calibration data may be modeled by a simple linear regression $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ independently, and if $Y_0 = \beta_0 + \beta_1 \xi + \epsilon_0$ denotes the new observation with ξ the unknown x , then clearly $Z \equiv (Y_0 - \beta_0)/\beta_1 \sim N(\xi, \sigma^2)$. The classical estimator of ξ (Eisenhart, 1939; Osborne, 1991) is then

$$\hat{Z} = (Y_0 - \hat{\beta}_0)/\hat{\beta}_1, \tag{1}$$

where $\hat{\beta}_i$ are the estimates of β_i from the calibration set.

In many applications, however, the calibration data must be interpreted as clustered observations obtained from different sampling units such as subjects, experiments or machines, while at the prediction step we wish to

estimate ξ for a new sampling unit. For example, we might want to estimate particle-mass concentration ξ using a particular machine which gives an infrared thermal imaging measurement Y_0 , and have available a calibration set comprising measurements Y , together with more precise measurements x obtained using filter weighing, for a number of machines of this type. In the application discussed below (Haylen et al, 1989), we wish to use a transvaginal ultrasound measurement Y_0 to estimate a woman's bladder volume ξ , for example before and after surgery for urinary incontinence. At the calibration step, each of 23 different patients in a urodynamic clinic had up to 8 known volumes x induced by catheterization, and the corresponding ultrasound measurements Y were obtained.

In such cases, subject-specific random effects need to be included at both the calibration and prediction steps. This was done by Oman (1998), who discussed point estimation and obtained confidence intervals for ξ using an analogue of Fieller's (1954) method. The main emphasis there was on the properties of the confidence intervals; here, we concentrate on point estimation. We show that if ξ is centered about a particular x -value c (for example, the mean of the calibration set, or a value of ξ whose detection is particularly important), then the estimate proposed in Oman (1998) now shrinks towards c , while the analogue to (1) does not. In the next section we define this estimator. In Section 3 we numerically examine its accuracy in the context of the bladder ultrasound data, and show that centering can lead to substantially improved predictions. Section 4 contains some concluding remarks and suggestions for further research.

2 The estimator

For n subjects, assume the observations at the calibration step are given by

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i} x_{ij} + \epsilon_{ij}, \quad \begin{array}{l} i = 1, \dots, n \\ j = 1, \dots, m_i \end{array}$$

where $(b_{0i}, b_{1i})^t = \mathbf{b}_i \sim_{\text{ind}} N(\mathbf{0}, \mathbf{\Phi})$ and $\epsilon_{ij} \sim_{\text{ind}} N(0, \sigma^2)$ independently. Letting

$$Y_0 = \beta_0 + \beta_1 \xi + b_0 + b_1 \xi + \epsilon_0 \quad (2)$$

denote the observation at the prediction step, we have that

$$Y_0 \sim N(\beta_0 + \beta_1 \xi, \sigma^2 + (1, \xi) \mathbf{\Phi} (1, \xi)^t)$$

so that now

$$Z \equiv \frac{Y_0 - \beta_0}{\beta_1} \sim N(\xi, Q(\xi)) \quad (3)$$

where

$$Q(\xi) = (\sigma^2 + (1, \xi) \mathbf{\Phi} (1, \xi)^t) / \beta_1^2. \quad (4)$$

Thus, in contrast to the case with simple linear regression, both the mean and variance of Y_0 contain information on ξ ; in particular, the simple estimator Z will tend to be farther from ξ for large values of ξ . This led Oman (1998) to propose the following contraction estimator, obtained by minimizing $E(\gamma Z - \xi)^2$ over γ :

$$\hat{\xi} = \frac{\xi^2}{\xi^2 + Q(\xi)} Z. \tag{5}$$

In practice, of course, the unknown parameters β, Φ and σ^2 in Z and Q are replaced by their estimates from the calibration step, and ξ in (5) is replaced by the estimated Z .

Now suppose that in (2) we center ξ about a known value c , giving

$$\begin{aligned} Y_0 &= (\beta_0 + \beta_1 c) + \beta_1(\xi - c) + (b_0 + b_1 c) + b_1(\xi - c) + \epsilon_0 \\ &= \alpha_0 + \alpha_1 \xi^* + a_0 + a_1 \xi^* + \epsilon_0 \end{aligned}$$

where $\xi^* = \xi - c$ and now $(a_0, a_1)^t \sim N(\mathbf{0}, \Phi^*)$ for $\Phi^* = \mathbf{G}\Phi\mathbf{G}^t$ with $\mathbf{G} = \begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix}$. If we estimate ξ^* by Z^* defined as in (3), but using α instead of β , it is immediate that the corresponding estimate of $\xi, c + Z^*$, reduces to Z . Thus, centering has no effect when the simple estimator Z is used. However, if we estimate ξ^* using the analogue to (5) (where now Q^* is defined using Φ^* and α_1), we obtain the following estimate of ξ :

$$\begin{aligned} \hat{\xi}_c &= c + \frac{(\xi^*)^2}{(\xi^*)^2 + Q^*(\xi^*)} Z^* \\ &= c + \frac{(\xi - c)^2}{(\xi - c)^2 + Q(\xi)} (Z - c); \end{aligned} \tag{6}$$

here, we have used the fact that $Q^*(\xi^*) = Q(\xi)$. Thus, centering gives an estimator which contracts Z towards c . Again, the unknown parameters in (6) must be replaced by their estimates from the calibration step, and ξ in the shrinkage factor must be replaced by the estimated Z . Although (6) resembles a Bayesian estimator, one does not require a Bayesian framework to use it. The shrinkage target c is not a prior mean for ξ , but rather a value for which it is important to get more precise estimates. Moreover, the amount of shrinkage does not depend on a prior variance, but rather on the relationship of the distance between ξ to c , to the contribution of ξ to the variance of Z in (4).

3 Numerical performance

Figure 1a shows the bladder-volume data described in the Introduction, following the transformation and rescaling suggested by Brown (1993) and

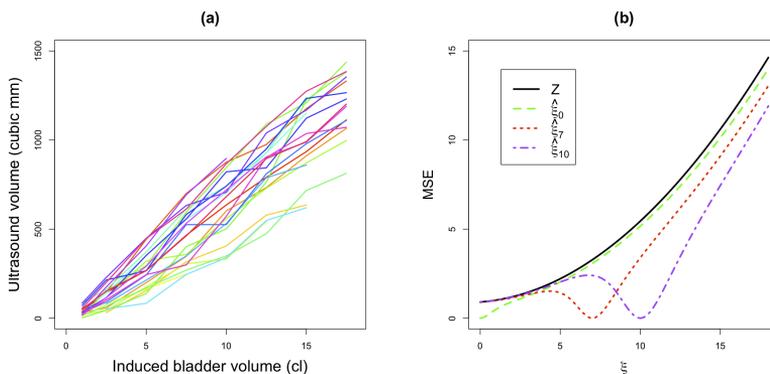


FIGURE 1. (a) Ultrasound bladder volume vs induced bladder volume. (b) Idealized mean squared error (7) for several estimators.

TABLE 1. REML parameter estimates and approximate 95% confidence intervals ($\rho = \phi_{12}/\sqrt{\phi_{11}\phi_{22}}$).

Parameter	Estimate	Lower limit	Upper limit
β_0	-54.479	-76.946	-32.011
β_1	69.192	63.319	75.065
$\sqrt{\phi_{11}}$	37.945	20.924	68.812
$\sqrt{\phi_{22}}$	13.582	9.632	19.153
ρ	0.318	-0.388	0.789
σ	54.017	47.614	61.282

Oman (1998); and Table 1 gives the REML estimates and asymptotic confidence intervals (from the R function lme) for the parameters β , Φ and σ^2 . In the calculations below, these are taken to be the true parameter values. If it were possible to use Z and $\hat{\xi}_c$ defined by (3) and (6), then a simple calculation gives

$$\text{MSE}(\hat{\xi}_c, \xi) = \frac{(\xi - c)^2}{(\xi - c)^2 + Q(\xi)} Q(\xi). \tag{7}$$

Since $\text{MSE}(Z, \xi) = Q(\xi)$, $\hat{\xi}_c$ would dominate Z , with greatest improvement for ξ near c . Figure 1b, which graphs these idealized mean squared errors for several values of c , shows that the MSE improvement would be substantial. In practice, of course, the unknown quantities in Z and $\hat{\xi}_c$ must be replaced by estimates; and then, as with the classical estimator (1) in the context of simple linear regression, division by $\hat{\beta}_1$ results in infinite MSE for both estimators. However, in typical applications there is a minimal probability of $\hat{\beta}_1$ being close to zero (note the confidence interval in Table 1), so as

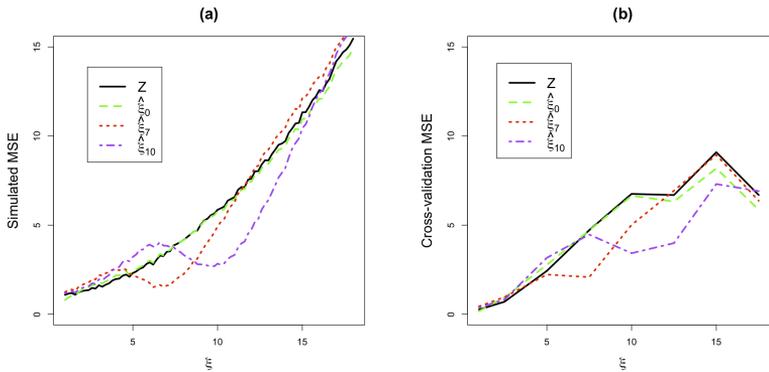


FIGURE 2. (a) Mean squared errors based on 1000 simulations. (b) Cross-validation mean squared errors.

with the classical estimator (Berkson, 1969; Shukla, 1972) one can consider asymptotic MSE, conditional on $|\hat{\beta}_1| > 0$. It is difficult to get an analytic approximation to the asymptotic $MSE(\hat{\xi}_c, \xi)$, so we have used simulations from the asymptotic distributions of $\hat{\beta}$, $\hat{\Phi}$ and $\hat{\sigma}^2$ (for the variance components, we used their natural parameterization (Pinheiro and Bates, 2000) and sampled from the appropriate multivariate normal distribution).

The results, in Figure 2a, are qualitatively similar to those in Figure 1b. Although $\hat{\xi}_c$ no longer dominates Z , $\hat{\xi}_5$ and $\hat{\xi}_7$ substantially decrease the MSE for ξ in a large region near the contraction origins, at the expense of a minimal increase for points farther away. We then did a cross-validation, in which we removed each of the 23 women from the calibration set, estimated the model parameters, and then used her Y values to predict the corresponding ξ . Figure 2b, which shows the average squared errors of these estimates, leads to the same qualitative conclusion as Figure 2a (the apparent decrease in MSE for $\xi = 17.5$ may be due to a relatively large number of missing observations: 8, as opposed to a maximum of 4 for the other ξ values).

4 Concluding remarks

In the context of the bladder-volume data, we have shown that centering about a value c can substantially improve prediction accuracy in a large neighborhood of c , with minimal decrease in accuracy outside the neighborhood. It would clearly be worthwhile to consider other examples as well. We have focused on point estimation. Regarding confidence intervals, for simple linear regression the Fieller (1954) interval, being exact, is generally preferred to one based on \hat{Z} and its asymptotic moments (Osborne, 1991). In the present context, however, Oman's (1998) analogue to Fieller's

interval is not exact, since when forming the pivot we must now use the asymptotic distribution of the mixed-model estimates $\hat{\beta}_i$ in (1). There is thus room to compare the two approaches. Simulation results for the bladder data, not presented here, indicate that using \hat{Z} and its asymptotic moments somewhat decreases the interval length, but at the price of a coverage probability below the nominal level; while the Fieller-like intervals, although a bit wider, give the required coverage.

Acknowledgments: I would like to thank Micha Mandel and David Zucker for numerous helpful conversations during the course of this research.

References

- Berkson, J. (1969). Estimation of a linear function for a calibration line; consideration of a recent proposal. *Technometrics*, **11**, 649–660.
- Brown, P.J. (1993). *Measurement, Regression, and Calibration*. Oxford: Clarendon Press.
- Eisenhart, C. (1939). The interpretation of certain regression methods and their use in biological and industrial research. *Annals of Mathematical Statistics*, **10**, 162–186.
- Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B*, **16**, 175–185.
- Haylen, B.T., Frazer, M.I., Sutherst, J.R. and West, C. R. (1989). Transvaginal ultrasound in the assessment of bladder volumes in women. *British Journal of Urology*, **63**, 149–151.
- Oman, S. D. (1998). Calibration with random slopes. *Biometrika*, **85**, 439–449.
- Osborne, C. (1991). Statistical calibration: a review. *International Statistical Review*, **59**, 309–336.
- Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-effects Models in S and S-PLUS*. New York: Springer.
- Shukla, G. K. (1972). On the problem of calibration. *Technometrics*, **14**, 547–553.

ANOSIM test revisited

Marek Omelka¹

¹ Department of Probability and Statistics, MFF UK, Charles University in Prague, Czech Republic

E-mail for correspondence: `omelka@karlin.mff.cuni.cz`

Abstract: The problem of comparing K independent groups is a common task in statistics. While the tests for univariate data are already well established, the choice of an appropriate test for multivariate data is much more delicate. Moreover the assumptions of the traditional multivariate analogues of the univariate tests are usually considered too strict for most ecological multivariate data sets. Analysis of Similarities (ANOSIM) is a distance-based test that is often considered to be a nonparametric analogue of a standard F -test that is suitable for multivariate data. This paper compares the two most common versions of ANOSIM test that have appeared in literature.

Keywords: ANOSIM, ANOVA, Distance matrix, K -sample problem

1 Introduction

In ecology, analysts often have to face very complex multivariate data. For instance, the distributions of abundances of individual species are usually highly aggregated or skewed. It is also common that datasets contain large proportion of zeroes and are discrete rather than continuous. As parametric modelling is often very difficult to justify, analysts often use a suitable distance measure (for a list of mostly used distance measures see Legendre and Legendre (1998)) between each pair of observations to calculate a distance matrix. That distance matrix is then used as the basis for statistical inference. The computation of the distance matrix is also the first step of cluster analysis, which is popular in the ecology community. That is why there has been a strong interest in distance-based testing methods. Among others let us mention the work Mantel and Valand (1970), Clarke (1993), Legendre and Anderson (1999), McArdle and Anderson (2001) and Mielke and Berry (2007) and the references therein.

Although methods for complex designs are already available in the literature, in this note we will concentrate on a simple one-way ANOVA design. ANOSIM (ANalysis Of SIMilarities) is a K -sample test that is based on pairwise distances and that, similarly as e.g. F -test or Kruskal-Wallis test, aims at detecting differences between K -independent groups.

The paper is organized as follows. Two different versions of ANOSIM test are described in Section 2 and compared in a small simulation study in Section 3. In Section 4 we review the existing literature with respect to the versions of ANOSIM test explicitly or implicitly used.

With no loss of generality, we talk only about distance matrices in this paper with the understanding that similarity matrices can be handled in the same fashion.

2 ANOSIM

Let \mathbf{Y}_i be the vector corresponding to the i -th observation and suppose we have a random sample $\mathbf{Y}_1, \dots, \mathbf{Y}_N$. We will assume the simple one-way ANOVA design, that is each observation is associated with exactly one of the K treatments and the sample sizes corresponding to the treatments are n_1, \dots, n_K . The different treatments may correspond to different types of soils, locations of samples, seasons of a year, etc. Further, let us fix a distance (dissimilarity) measure d (e.g. Euclidean, Manhattan, Bray-Curties, ...) and compute the distance matrix \mathbb{D} of type $N \times N$ with the elements $d_{ij} = d(\mathbf{Y}_i, \mathbf{Y}_j)$.

The original ANOSIM test was developed by Clarke (1993) in order to test the null hypothesis

$$H_0 : \text{There are no differences between treatments.} \quad (1)$$

The test statistic is given by

$$R = \frac{\bar{r}_B - \bar{r}_W}{N(N-1)/4}, \quad (2)$$

where \bar{r}_B is the mean of the ranks of the *between sample distances* (\mathbf{Y}_i and \mathbf{Y}_j receive different treatments) and \bar{r}_W is the mean of the ranks of the *within sample distances* (\mathbf{Y}_i and \mathbf{Y}_j receive the same treatment). Using ranks instead of the original distances is not a fundamental requirement for the procedure, but it follows Clarke's recommendation that the test statistic should reflect the patterns formed by multidimensional scaling methods, which preserve ranks of distances.

The rationale behind the test statistics (2) is that if the hypothesis (1) does not hold then one would expect that the between sample distances are larger than within sample distances.

The statistical significance of the test statistic is assessed through permutation methods. Note, that a permutation of the original observations $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ can be easily implemented by permuting both columns and rows of the distance matrix \mathbb{D} .

In order to keep things simpler, we will not switch to ranks for this moment and use the original distances d_{ij} . The ANOSIM statistics is now given by

$$t = \bar{d}_B - \bar{d}_W, \quad (3)$$

where \bar{d}_B and \bar{d}_W are corresponding means of *between distances* and *within distances*, that is

$$\bar{d}_W = \frac{1}{N_W} \sum_{k=1}^K \sum_{i,j \in G_k, i < j} d_{ij}, \quad N_W = \sum_{k=1}^K \binom{n_k}{2}$$

$$\bar{d}_B = \frac{1}{N_B} \sum_{k=1}^{K-1} \sum_{l=k+1}^K \sum_{i \in G_k} \sum_{j \in G_l} d_{ij}, \quad N_B = N(N-1)/2 - N_W,$$

where the set G_k contains the indices of the observations that belongs to the k -th group.

As shown in Mielke and Berry (2007), the test statistic t of (3) can be rewritten as

$$t = \frac{1}{N_B} \sum_{i < j}^N d_{ij} - \left(\frac{1}{N_B} + \frac{1}{N_W} \right) \sum_{k=1}^K w_k \bar{d}_k, \tag{4}$$

where

$$\bar{d}_k = \frac{1}{\binom{n_k}{2}} \sum_{i,j \in G_k, i < j} d_{ij}, \quad \text{and} \quad w_k = \binom{n_k}{2}, \quad k = 1, \dots, K. \tag{5}$$

As the first term at the right-hand side of the equation (4) is invariant to permutations of observations, one can concentrate on the following test statistic

$$t_1 = \sum_{k=1}^K w_k \bar{d}_k \tag{6}$$

and reject the null hypothesis, if t_1 is ‘too small’.

Note that there is something suspicious about the test statistic t_1 given by (6). While one would expect that a within sample mean (\bar{d}_k) should be given a weight that is proportional to the size of the corresponding group (n_k), in fact it gets a weight that is proportional to the squared size of the group. To correct for that we will consider the test statistics t_1 with the weights $w_k = \frac{n_k - 1}{N - K}$ for $k = 1, \dots, K$ and we will call this test ANOSIM2. Note that ANOSIM and ANOSIM2 coincide if the sample sizes of all groups are equal, that is for balanced designs.

One may wonder why to use $w_k = \frac{n_k - 1}{N - K}$ instead of $w_k = \frac{n_k}{N}$. The reason is that, if the observations are univariate and the squared Euclidean distance is used, then the ANOSIM2 test coincides with the traditional (permutation) F -test (see Mielke and Berry (2007)).

3 Comparing ANOSIM and ANOSIM2

In this section we compare ANOSIM and ANOSIM2 in two very simple situations so that we can understand what is happening. Only two groups are considered and the observations are univariate.

TABLE 1. Power of different versions of ANOSIM for location alternatives.

(n_1, n_2) test	(10,15)		(10,25)		(10,50)	
	Orig.	Rank	Orig.	Rank	Orig.	Rank
ANOSIM	0.55	0.50	0.56	0.49	0.51	0.50
ANOSIM2	0.61	0.56	0.62	0.54	0.78	0.68

3.1 Difference in locations

The observations follow the model

$$Y_i = e_i, \quad \text{for } i = 1, \dots, n_1, \quad Y_i = 1 + e_i, \quad \text{for } i = n_1 + 1, \dots, n_1 + n_2,$$

where e_1, \dots, e_N are independent random variables with a standard normal distribution. The Euclidean distance is used to compute the distances d_{ij} . We generate 2000 samples. The number of permutations is 1999. The prescribed level of the test is 0.05.

The power of ANOSIM and ANOSIM2 for different sample sizes can be found in Table 1. The names of the columns distinguish if the original distances or their ranks are used. From Table 1 one can clearly see that the power of the test ANOSIM2 dominates the power of ANOSIM. It is also interesting to note that when increasing the sample size of the second treatment, the power of ANOSIM remains approximately constant. The same pattern was observed for different sample sizes and different (both univariate as well as multivariate) distributions, provided the treatments give rise to differences in locations of the original distributions and not in scales.

3.2 Difference in scales

The following model is considered

$$Y_i = e_i, \quad \text{for } i = 1, \dots, n_1, \quad Y_i = 2e_i, \quad \text{for } i = n_1 + 1, \dots, n_1 + n_2,$$

with all the quantities being the same as in the previous section. The results are to be found in Table 2. Note that the power of the test of the null hypothesis (1) strongly depends on the fact, whether the more variable observations are in a smaller or larger sample. This is in particular true for ANOSIM which has a larger power if the larger sample goes with the larger variance, but its power is even below the level of the test if it is the other way around. The power of the ANOSIM2 test is more stable in this aspect. Note also that rank-based version of the ANOSIM2 test now produces a substantial increase in power over the version with original distances.

TABLE 2. Power of different versions of ANOSIM for scale alternatives.

(n_1, n_2) test	(10,15)		(15,10)	
	Orig.	Rank	Orig.	Rank
ANOSIM	0.51	0.51	0.01	0.02
ANOSIM2	0.20	0.31	0.08	0.21

4 Conclusions

Although the range of the simulations in this paper is rather limited we hope that we have made sufficiently clear that ANOSIM test in its original version as suggested in Clarke (1993) cannot be recommended (unless the sample sizes are balanced) and ANOSIM2 should be used instead. Unfortunately, this is still not well known in ecology literature and ANOSIM test in its original form is still used and sometimes even recommended in recent papers, see e.g. Ramette (2007) and Kropf et al. (2007).

On the other hand, the distance-based procedures that use the weights of ANOSIM2 can be found in Excoffier et al. (1992), Pillar and Orlóci (1996), Gower and Krzanowski (1999), McArdle and Anderson (2001), Baringhaus and Franz (2004) and Rizzo and Székely (2010).

Finally, we would like to stress that although we generally recommend ANOSIM2, the aim of the paper is not to argue for distance-based methods in favour of variable-based methods like MANOVA. Criticism of the distance-based methods can be found in Warton and Hudson (2004) and Warton et al. (2012).

Acknowledgments: The research was supported by the grant GAČR P201/11/P290.

References

- Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, **88**, 190–206.
- Clarke, K.R. (1993). Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology*, **18**, 117–143.
- Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.
- Gower, J.C. and Krzanowski, W.J. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of

- variance. *Journal of the Royal Statistical Society. Series C.*, **48**, 505–519.
- Kropf, S., Lux, A., Eszlinger, M., Heuer, H., and Smalla, K. (2007). Comparison of independent samples of high-dimensional data by pairwise distance measures. *Biometrical Journal*, **49**, 230–241.
- Legendre, P. and Anderson, M.J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, **69**, 1–24.
- Legendre, P. and Legendre, L. (1998). *Numerical Ecology*. Amsterdam: Elsevier Science.
- Mantel, N. and Valand, R. (1970). A technique of nonparametric multivariate analysis. *Biometrics*, **26**, 547–558.
- McArdle, B.H. and Anderson, M.J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, **82**, 290–297.
- Mielke, J.P.W. and Berry, K.J. (2007). *Permutation Methods, A distance function Approach*. New York: Springer.
- Pillar, V.D.P. and Orlóci, L. (1996). On randomization testing in vegetation science: multifactor comparisons of relevé groups. *Journal of Vegetation Science*, **7**, 585–592.
- Ramette, A. (2007). Multivariate analyses in microbial ecology, *FEMS microbiology ecology*, **62**, 142–160.
- Rizzo, M.L. and Székely G.J. (2010) DISCO analysis: A nonparametric extension of analysis of variance, *The Annals of Applied Statistics*, **4**, 1034–1055.
- Smouse, P.E., Long, J.C., and Sokal, R.R. (1986). Multiple regression and correlation extension of mantel test of matrix correspondence. *Syst. Zool.*, **35**, 627–632.
- Warton, D.I. and Hudson, H.M. (2004). A Manova statistic is just as powerful as distance-based statistics, for multivariate abundances. *Ecology*, **85**, 858–874.
- Warton, D.I., Wright, S.T. and Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects, *Methods in Ecology and Evolution*, **3**, 89–101.

Bayesian variable selection method for modeling dose-response microarray data under simple order restrictions

Martin Otava¹, Adetayo Kasim², Ziv Shkedy¹, Dan Lin¹,
Bernet S. Kato³

¹ Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Center for Statistics, Universiteit Hasselt, Belgium

² Wolfson Research Institute, Durham University, United Kingdom

³ National Heart and Lung Institute, Imperial College London, United Kingdom

E-mail for correspondence: martin.otava@uhasselt.be

Abstract: Bayesian modeling of dose-response microarray data offers the possibility to jointly establish the dose-response relationships between gene expression and increasing doses of therapeutic compound, and to determine the nature of the relationships wherever it exist. Moreover, correction for multiplicity adjustment for Bayesian modeling of dose-response microarray data can be based on the direct posterior probability of the null model. The posterior probabilities are obtained by translating the inequality constraints for monotone relationship into Bayesian variable selection problem.

Keywords: Microarray Data; Bayesian Analysis; Dose-Response Relationship; False Discovery Rate; Direct Posterior Probability.

1 Introduction

Dose-response microarray experiments are a growing area in biomedical and pharmaceutical research to study the relationship between increasing doses of a therapeutic compound and the activity of entire genome at once. The primary goal of such an experiment is to identify genes with significant dose-response relationship under the monotone constraints (Lin et al., 2012). Secondly, it is necessary to determine the nature of the relationship wherever it exists. Denote the mean gene expression of a gene under the placebo dose as μ_0 . Similarly, we consider an increasing doses of a therapeutic compound and μ_i , $i = 1, \dots, K$ be an the mean gene expression under dose i . Therefore, the primary interest is to test the null hypothesis

$$H_0 : \mu_0 = \mu_1 = \mu_2 = \dots = \mu_K, \quad (1)$$

TABLE 1. The set of seven possible monotonic dose-response models for an experiment with three dose levels. The mean response of dose level i is denoted as μ_i . The model g_0 represents the null model of no dose effect.

Model	Non-decreasing profile	Non-increasing profile
g_1	$\mu_0 = \mu_1 = \mu_2 < \mu_3$	$\mu_0 = \mu_1 = \mu_2 > \mu_3$
g_2	$\mu_0 = \mu_1 < \mu_2 = \mu_3$	$\mu_0 = \mu_1 > \mu_2 = \mu_3$
g_3	$\mu_0 < \mu_1 = \mu_2 = \mu_3$	$\mu_0 > \mu_1 = \mu_2 = \mu_3$
g_4	$\mu_0 < \mu_1 = \mu_2 < \mu_3$	$\mu_0 > \mu_1 = \mu_2 > \mu_3$
g_5	$\mu_0 = \mu_1 < \mu_2 < \mu_3$	$\mu_0 = \mu_1 > \mu_2 > \mu_3$
g_6	$\mu_0 < \mu_1 < \mu_2 = \mu_3$	$\mu_0 > \mu_1 > \mu_2 = \mu_3$
g_7	$\mu_0 < \mu_1 < \mu_2 < \mu_3$	$\mu_0 > \mu_1 > \mu_2 > \mu_3$

against the alternative hypotheses

$$\begin{aligned}
 H_a^{up} : \mu_0 \leq \mu_1 \leq \mu_2 \leq \dots \leq \mu_K, \\
 \text{or} \\
 H_a^{dn} : \mu_0 \geq \mu_1 \geq \mu_2 \geq \dots \geq \mu_K
 \end{aligned}
 \tag{2}$$

with at least one strict inequality. The choice between H_a^{up} and H_a^{dn} depends on the direction of the ordered constraints. Note that the determination of the nature of the dose-response relationship is related to the further decomposition of the alternative hypotheses into their basic hypotheses. This process results in $2^K - 1$ hypotheses under each of the monotone directions. For a dose-response microarray experiments with one control dose and $K = 3$ (i.e. three increasing doses of a therapeutic compound), the alternative hypotheses can be decomposed into further basic hypotheses as shown in Table 1. Note that each alternative hypothesis corresponds to a monotone model. In particular the null hypothesis corresponds to the null model for which $\mu_0 = \mu_1 = \mu_2 = \mu_3$.

Bayesian modeling of dose-response microarray data offers a framework to simultaneously establish a dose-response relationship and to determine the nature of the relationship by providing posterior probability for each of the models g_i , $i = 1, \dots, K$, given the data. The posterior probability of the null model is particularly interesting, because it is also a probability of false positives findings, i.e. of genes that are wrongly assigned to the alternative hypotheses. Hence, posterior probability allows for adjustment for false discovery rate (Newton et al., 2007), to identify few important genes in a pool of potential false positives. However, the estimation of the required parameters to obtain posterior probability for the models requires estimation under equality constraints between two or more parameters which could not be estimated with the standard approach of Gelfand et al. (1991). Therefore, the Bayesian variable selection approach offers elegant solution how to identify the relationship and correct for multiplicity simultaneously using conditional false discovery rate.

2 Methodology

The Bayesian inequality models (Klugkist and Hoijtink, 2005) cannot be used in our framework because of the equality constraints specified in the models. The equality constraints would cause that standard estimation approach assigns zero probabilities to each of our models except g_7 . Therefore, we propose the following parametrization. We consider the following linear model,

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad i = 0, \dots, K, \quad j = 0, 1, 2, \dots, n_i, \quad (3)$$

where $\mathbf{Y} = (Y_{01}, Y_{02}, \dots, Y_{Kn_K})$ are gene expression levels and n_i represents the number of observations at the i th dose level. Reparameterize the mean response such that

$$E(Y_{ij}) = \mu_i = \begin{cases} \mu_0, & i = 0, \\ \mu_0 + \sum_{\ell=1}^i \delta_\ell, & i = 1, \dots, K, \end{cases} \quad (4)$$

with the constraints that $\delta_\ell \geq 0$ for an upward trend or $\delta_\ell \leq 0$ for a downward trend. The difference in the mean structures of the different models therefore depends on which of the components in $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_K)$ are set to be equal to zero. The problem of model estimation is equivalent to decision which columns in the full design matrix of model 4 are selected or deleted. This is related to the Bayesian variable selection (BVS) approach (George and McCulloch, 1993), which is used to determine an optimal model from a priori set of R known plausible models. In our setting the BVS model allows us to calculate the posterior probability of each model, $p(g_r|\text{data})$ and in particular the posterior probability of the null model, $p(g_0|\text{data})$. Let $z_i, i = 1, \dots, K$ be an indicator variable such that

$$z_i = \begin{cases} 1, & \delta_i \text{ is included in the model,} \\ 0, & \delta_i \text{ is not included in the model,} \end{cases} \quad (5)$$

and let $\theta_i = \delta_i \cdot z_i$. Hence, we can reformulate the mean structure in (4) (O’Hara and Sillanpää, 2009) in terms of θ_i and z_i as

$$E(Y_{ij}) = \mu_0 + \sum_{\ell=1}^i \theta_\ell = \mu_0 + \sum_{\ell=1}^i z_\ell \delta_\ell, \quad i = 1, \dots, K. \quad (6)$$

For K dose levels experiment, the vector $\mathbf{z} = (z_1, \dots, z_K)$ defines uniquely each one of the 2^K plausible models. For example for $K = 3$ and $\mathbf{z} = (z_1 = 1, z_2 = 0, z_3 = 0)$ we obtain $E(Y_{ij}|\mathbf{z}) = (\mu_0, \mu_0 + \delta_1, \mu_0 + \delta_1, \mu_0 + \delta_1)$, which corresponds to the mean of model g_3). We assume that z_i and δ_i are independent, and use truncated normal prior distribution for δ_i and

$$\begin{aligned} z_i &\sim \text{Bernoulli}(\pi_i), \\ \pi_i &\sim \text{U}(0, 1). \end{aligned} \quad (7)$$

As pointed out by O'Hara and Sillanpää (2009) the posterior inclusion probability of δ_i into the model equals the posterior mean of z_i . The posterior probability of each model can be straightforwardly obtained by using the transformation of \mathbf{z} instead of the entire vector \mathbf{z} itself. Denote $M_R = 1 + \mathbf{z}\mathbf{c}$, where $\mathbf{c} = (1, 2, \dots, 2^{K-1})^T$, then M_R has unique value for each of the plausible models (for example: $M_R = 2$ only for the model g_3). Thus, the posterior probability of $M_R = r$, $r = 1, \dots, R$, defines uniquely the posterior probability of the r th model,

$$p(M_R = r|\text{data}) = p(g_r|\text{data}), \quad (8)$$

and in particular, the posterior probability of the null model is given by,

$$p(M_R = 1|\text{data}) = p(g_0|\text{data}). \quad (9)$$

Assume that there are $m = 1, \dots, M$ genes in the experiment and the aim is to find the differentially expressed ones with respect to dose. In our framework, the problem is translated to the determination if the gene follows any other model than g_0 . Assume that the genes satisfying $p_m(g_0|\text{data}) \leq \alpha$ for given threshold α are considered differentially expressed. Hence, according to Newton et al. (2007), $p_m(g_0|\text{data})$ represents probability of such statement being false. Let I_m be an indicator variable of $p_m(g_0|\text{data}) \leq \alpha$. Since $p_g(g_0|\text{data})$ is also the probability that the considering the m th gene differentially expressed is incorrect, the expected number of false discoveries (cFD) is

$$\text{cFD}(\alpha) := \text{E}(\text{cFD}) = \sum_{m=1}^M p_m(g_0|\text{data})I_m. \quad (10)$$

Newton et al. (2007) defined the conditional (on the data) false discovery rate as

$$\text{cFDR}(\alpha) = \frac{\text{cFD}(\alpha)}{N(\alpha)}, \quad (11)$$

where $N(\alpha)$ is the number of genes declared differentially expressed for a given threshold α . Note that $\text{cFDR}(\alpha)$ is interpreted as the average error that is made by considering any gene as differentially expressed. Hence, the value of α is selected is such a way that $\text{cFDR}(\alpha)$ does not exceed a pre-specified threshold τ .

3 Results

We apply the direct posterior probability approach discussed above for multiplicity adjustment. The framework enables adjustment for false discovery rates among the significant genes. We use the R2WINBUGS package to fit a gene specific model and to obtain the posterior probability of the null model. For each gene an MCMC simulation of 20000 iterations (from which

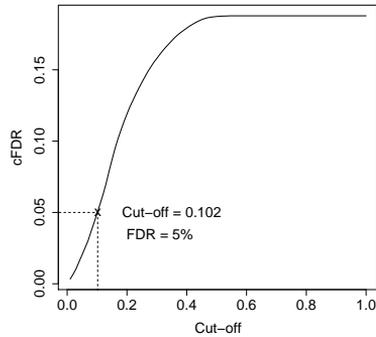


FIGURE 1. Adjustment for multiplicity. The relationship between the conditional false discovery rate (cFDR) and the cut-off values.

5000 are used as burn-in period) was used to fit the BVS model. Figures 1 and 2 show the relationship between false discovery rate (cFDR), number of significant genes and cut-off value α . Figure 1 shows that an increase in cut-off values results in an increase in false discovery rate. However, the false discovery rate reaches its maximum of 0.2 at the cut-off of about 0.5. Figure 2 also shows an increase in the number of significant genes with an increase in cut-off values. The implication of the finding is that, as expected, the higher the cut-off value, the larger the number of significant genes and consequently, the higher the proportion of false positives among the significant genes. Similar to the frequentists practice, one may wish to control for false discovery rate at 1% or 5%, which corresponds to cut-off values of 0.029 and 0.102, respectively. Based on these cut-off values, the corresponding numbers of significant genes are 609 and 3295 genes, respectively.

4 Discussion

There are two main challenges in Bayesian analysis of dose-response microarray data. The first is the presence of strictly equality relationship between differences in gene expressions at different doses of a therapeutic compound and the second is the question how to adjust for multiplicity. The BVS method is useful as an approach to circumvent the first problem by replacing strict equality between doses by a common parameter. The BVS model estimates equal means for two successive dose levels, i and $i - 1$ whenever the corresponding binary variable for the i th dose level $z_i = 0$. Further, the posterior probability of the null model can be estimated and can be used for multiplicity adjustment. In summary, the BVS methodology offers the tools how to handle the differentially expressed genes finding

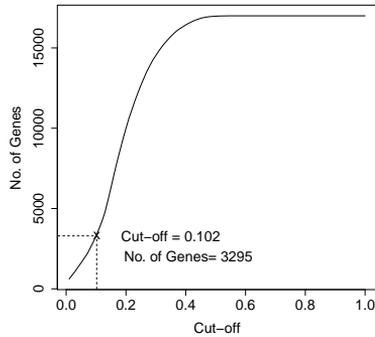


FIGURE 2. Adjustment for multiplicity. The relationship between number of significant genes and the cut-off values.

in elegant and efficient way.

References

- Gelfand, A.E., Smith, A.F.M., and Lee, T.M. (1992). Bayesian Analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, **87**, 523–532.
- George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.
- Klugkist, I. and Hoijtink, H. (2007). The Bayes Factor for Inequality and About Equality Constrained Models. *Computational Statistics and Data Analysis*, **51**, 6367–6379.
- Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijmens, L., (editors)(2012). *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R - Order Restricted Analysis of Microarray Data*. Springer.
- Newton, M.A., Wang, P., and Kendziorski, C. (2007). Hierarchical mixture models for expression profiles. In: *Do, K.M., Müller, P. and Vannucci, M. (Editors): Bayesian Inference for gene expression and proteomics*, Cambridge university press, pp. 40–52,
- O’Hara, R. B. and Sillanpää, M. J. (2009). Review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, **4**, 85–118.

Saddlepoint-based bootstrap inference in parametric and semiparametric models

Robert L. Paige¹, A. Alexandre Trindade²

¹ Missouri University of Science and Technology, Rolla, U.S.A.

² Texas Tech University, Lubbock, U.S.A.

E-mail for correspondence: `alex.trindade@ttu.edu`

Abstract: An overview and examples are presented showcasing the usefulness of the Saddlepoint-Based Bootstrap (SPBB) methodology (Paige *et al.*, 2009) as a tool for improving small-sample inference under Gaussian-based modeling. Extensions are proposed generalizing SPBB to the case of inference under estimating equations which: (i) can either be well-approximated by Gaussian linear-quadratic forms; or (ii) are non-Gaussian quadratic forms, but have a tractable expression for the moment generating function.

Keywords: Estimating Equation; Quadratic Form; Moment Generating Function; Elliptically-Contoured Family; Skew-Normal Distribution.

1 Introduction

Many statistical applications involve inference primarily on a (scalar) parameter, θ , in the presence of a finite dimensional nuisance parameter, $\boldsymbol{\lambda}$. Paige, Trindade, & Fernando (2009) considered a class of problems in which the estimator $\hat{\theta}$ of θ is an appropriate root of a *quadratic estimating equation (QEE)*, i.e. $\hat{\theta}$ is a solution of

$$\Psi(\theta) = \mathbf{y}^\top Q_\theta \mathbf{y} = 0, \quad (1)$$

where Q_θ is a conformable symmetric matrix whose entries are functions of θ , and $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$. The multivariate normality immediately furnishes a closed-form expression for the moment generating function (mgf) of the QEE. If the QEE is monotone in θ , then it is possible to relate the cumulative distribution function (cdf) of $\hat{\theta}$ to that of $\Psi(\theta)$. The vector of nuisance parameter(s) $\boldsymbol{\lambda}$ is dealt with by substituting a conditional maximum likelihood estimator (MLE), $\hat{\boldsymbol{\lambda}}_\theta$. Using saddlepoint approximations, it is then possible to accurately approximate the distribution of the estimator of interest. Confidence intervals for θ can be produced by pivoting the saddlepoint cdf approximation.

The landmark paper by Paige *et al.* (2009) developed this *SaddlePoint-Based Bootstrap* (SPBB) methodology in a coherent manner. As the name

suggests, the technique is identical to a parametric bootstrap, but with (slow) Monte Carlo simulation replaced by (fast) saddlepoint approximation. The essential elements of SPBB are: (i) a monotone estimating equation whose unique root is the estimator of interest; (ii) substitution of conditional MLEs for any nuisance parameters present resulting in a profile estimating equation; (iii) accurate saddlepoint approximations to the cdf of the estimator through that of the QEE; and (iv) pivoting of this cdf to produce confidence intervals (c.i.s) for the parameter of interest.

2 SPBB: Key Accomplishments and Examples

A key point emphasized in Paige *et al.* (2009) is that SPBB c.i.'s are 2nd-order accurate; the coverage probability of a nominal $(1 - \alpha)100\%$ c.i. is $\alpha + O(n^{-1})$. Also, monotonicity (in θ) of the QEE is not a serious restriction, and can be relaxed to just assuming that there exists an interval where this happens, and within which the MLE $\hat{\theta}$ lies with high probability. (A simple diagnostic plot can be constructed to assess the validity of this assumption for small n . For large enough n the existence of such an interval is assured by the usual conditions which guarantee consistency and asymptotic normality of the MLE.)

The SPBB seminal work, over the 2008-2012 period, resulted in the following key accomplishments and applications in a variety of inferential settings: Maximum Likelihood (ML); Method of Moments (MoM); REstricted Maximum Likelihood (REML); Akaike's Information Criterion (AIC); Generalized Cross-Validation (GCV).

- MoM and ML inference for AR(1) and MA(1) time series models (Paige & Trindade, 2008);
- ML inference for the nonlinear parameter in (conditionally linear) nonlinear regression and ratios of regression parameter problems (Paige *et al.*, 2009, Paige & Fernando, 2008);
- ML, REML, AIC, and GCV inference for the smoothing parameter in penalized spline models with independent errors (Paige & Trindade, 2012);
- ML, REML, AIC, and GCV inference for the smoothing parameter in penalized spline models with correlated errors (Paige & Trindade, 2010).

Indications concerning SPBB from the above applications are as follows:

- (i) c.i. lengths and coverage probabilities compare favorably with those obtained from various competing methods, many of which have 2nd or 3rd order accuracy, and some of which are exact;

- (ii) is typically easier to implement;
- (iii) enjoys faster computational speeds, by at least an order of magnitude;
- (iv) may be the only alternative to a full-blown parametric bootstrap, in cases which lack exact, asymptotic, or nonparametric bootstrap procedures.

2.1 Example: SPBB inference in penalized spline models.

We illustrate the methodology by drawing from Paige & Trindade (2010, 2012), which compare the performance of exact vs. SPBB methods for inference on the smoothing parameter (α) in penalized spline models. Common methods for inference on α include ML, REML, AIC, and GCV. A key insight in Paige & Trindade (2010, 2012) was a unification of all these methods where the estimator, $\hat{\alpha}$, is viewed as the root of QEE $\Psi(\alpha) \equiv \mathbf{y}'Q_\alpha\mathbf{y}$, with \mathbf{y} multivariate normal, and Q_α a conformable matrix that depends on α . For example, in the case of GCV, if we let $S_\alpha = B(B^\top B + \alpha D)^{-1}B^\top$ denote the *smoothing matrix* in the linear mixed model representation of a penalized spline (where B and D are respectively, the matrix of basis functions, and the penalty matrix), and $\dot{S}_\alpha = \partial S_\alpha / \partial \alpha$, then,

$$Q_\alpha = (I_n - S_\alpha)\dot{S}_\alpha \{1 - \text{Tr}(S_\alpha)/n\} - (I_n - S_\alpha)^2\text{Tr}(S_\alpha)/n.$$

At present (simulation-based) exact inference has only been devised for ML and REML. (Asymptotic methods perform poorly here.) Although Crainiceanu & Ruppert (2004) and Crainiceanu *et al.* (2005) devised a fast algorithm to sample from the null distribution of the RLRT/LRT statistic, it is still an order of magnitude slower than SPBB. Table 1, abstracted from Paige & Trindade (2012), provides a qualitative assessment of Exact-REML, SPBB-REML, and the parametric/nonparametric bootstrap c.i. construction procedures, as well as benchmark comparisons in comparable computing times.

TABLE 1. Empirical comparison of Exact-REML, SPBB-REML, and semiparametric bootstrap 95% c.i.s for the smoothing parameter α in 200 simulated datasets from a penalized spline model (Paige & Trindade, 2012).

Method	Compute Time (minutes/c.i.)	Coverage Probability	Interval Lengths		
			Min.	Median	Max.
SPBB-REML	15	0.925	5.25	6.31	6.80
Exact-REML	105	0.915	5.13	6.09	8.86
Bootstrap	2,100	1.000	8.84	15.48	28.57

3 SPBB Confidence Intervals: Scalar Parameter Case

Let $\boldsymbol{\mu}_{\theta_0, \boldsymbol{\lambda}_0} \equiv \boldsymbol{\mu}_0$ and $\Sigma_{\theta_0, \boldsymbol{\lambda}_0} \equiv \Sigma_0$ denote hypothesized values for the mean and covariance of vector \mathbf{y} in QEE (1), and let $M_{\Psi(\theta)}(s; \theta_0, \boldsymbol{\lambda}_0)$ denote its mgf (valid for all s in a sufficiently small neighborhood of zero). To derive saddlepoint approximations for the distribution of $\hat{\theta}$, first assume $\Psi(\theta)$ is monotone in $\theta \in \Theta$. If \mathcal{Y} denotes the sample space corresponding to \mathbf{y} , we then have that either

$$(i) \hat{\theta}_{\text{obs}} \leq \theta \Leftrightarrow \Psi(\theta) \leq 0, \quad \text{or} \quad (ii) \hat{\theta}_{\text{obs}} \leq \theta \Leftrightarrow \Psi(\theta) \geq 0, \quad (2)$$

where $\hat{\theta}_{\text{obs}}$ denotes the (uniquely) observed root of $\Psi(\theta)$ for a given sample. Without loss of generality we may assume condition (i) is satisfied, since $\hat{\theta}_{\text{obs}}$ is invariant under sign changes of $\Psi(\theta)$. Since (i) holds for every $\theta \in \Theta$ and $\mathbf{y} \in \mathcal{Y}$, we have the device

$$P(\hat{\theta} \leq \theta) = P(\Psi(\theta) \leq 0), \quad (3)$$

which, noting the dependence on the parameter hypothesized values, we can express in terms of the respective cdf's at $\theta = t$ as

$$F_{\hat{\theta}}(t; \theta_0, \boldsymbol{\lambda}_0) = F_{\Psi(t)}(0; \theta_0, \boldsymbol{\lambda}_0). \quad (4)$$

A saddlepoint approximation for the cdf in (4) can now be obtained from the formula of Lugannani and Rice (1980), and a corresponding formula for the pdf follows from Daniels (1983). These formulas are functions of the cumulant generating function of $\Psi(t)$, $K_{\Psi(t)}(s; \theta_0, \boldsymbol{\lambda}_0)$, and its first and second derivatives with respect to both s and t . The most computationally expensive step involves finding \hat{s} at each point t in the support of $\hat{\theta}$, by solving the (nonlinear) *saddlepoint equation*

$$\frac{\partial}{\partial t} K_{\Psi(t)}(\hat{s}; \theta_0, \boldsymbol{\lambda}_0) = 0.$$

The result, upon substitution of the conditional MLE $\hat{\boldsymbol{\lambda}}_{\theta_0}$ for the nuisance parameter $\boldsymbol{\lambda}_0$, is the cdf approximation

$$F_{\hat{\theta}}(t; \theta_0, \boldsymbol{\lambda}_0) \approx \hat{F}_{\hat{\theta}}(t; \theta_0, \hat{\boldsymbol{\lambda}}_{\theta_0}) = \hat{F}_{\Psi(t)}(0; \theta_0, \hat{\boldsymbol{\lambda}}_{\theta_0}),$$

which is third-order accurate over sets of bounded central tendency. This leads immediately to an automatic approximate percentile confidence set construction method, where the lower and upper bounds of the desired $(1 - \alpha)100\%$ (equi-tailed) c.i. for θ_0 , (θ_L, θ_U) , are determined by solving

$$\hat{F}_{\hat{\theta}}(\hat{\theta}_{\text{obs}}; \theta_L, \hat{\boldsymbol{\lambda}}_{\theta_L}) = 1 - \alpha/2, \quad \text{and} \quad \hat{F}_{\hat{\theta}}(\hat{\theta}_{\text{obs}}; \theta_U, \hat{\boldsymbol{\lambda}}_{\theta_U}) = \alpha/2,$$

or equivalently,

$$\hat{F}_{\Psi(\hat{\theta}_{\text{obs}})}(0; \theta_L, \hat{\boldsymbol{\lambda}}_{\theta_L}) = 1 - \alpha/2, \quad \text{and} \quad \hat{F}_{\Psi(\hat{\theta}_{\text{obs}})}(0; \theta_U, \hat{\boldsymbol{\lambda}}_{\theta_U}) = \alpha/2.$$

4 Generalizing SPBB: Estimating Equations With Tractable MGFs

Consider the QEE $\Psi(\theta) \equiv \Psi_0(\theta)$ defined in (1), with $\mathbf{y} \sim N(\boldsymbol{\mu}_0, \Sigma_0)$. Generalizations of SPBB as outlined in the previous section are possible by considering other forms for $\Psi(\theta)$, and other distributions for \mathbf{y} .

Linear quadratic form for the QEE. A straightforward generalization occurs if $\Psi(\cdot)$ contains linear and constant terms,

$$\Psi(\theta) = a_\theta + \mathbf{b}_\theta^\top \mathbf{y} + \mathbf{y}^\top Q_\theta \mathbf{y} = 0, \quad (5)$$

where a_θ and \mathbf{b}_θ are scalar and vector-valued constants (possibly depending on θ). This case leads also to an explicit expression for the mgf of $\Psi(\theta)$, $M_{\Psi(\theta)}(s; \boldsymbol{\mu}_0, \Sigma_0)$. An application of this is if the underlying estimating equation $g(\mathbf{y}; \theta)$ is a function that is itself not a QEE, but can be approximated as such via a Taylor series expansion to second order terms, so that $g(\mathbf{y}; \theta) \approx \Psi(\theta)$. An application is provided by Feuerverger & Wong (2000) who show how the tractable mgf of (5) leads directly to a saddlepoint approximation for a scalar-valued function of a Gaussian vector of returns. This kind of scenario is subsumed as a special case of the SPBB approach, which can handle instances when only the estimating equation giving rise to the estimator of interest is known.

Inference under Skew-Normal families. Generalizations of the Gaussian distribution with tractable QEE mgf promise include the *elliptically-contoured* (EC) and *skew-normal* (SN) families. Gupta & Huang (2002) show that the distribution of a quadratic form in SN's is the same as its distribution in the multivariate normal mean zero case, and provide a closed-form expression for the joint mgf of a linear and a quadratic form in SN's. Huang & Chen (2006) show that some of these results holds more generally for skew-symmetric distributions when defined with respect to a continuous random variable symmetric about zero. Wang *et al.* (2009) define SN random vectors, and present a corresponding closed-form expression for the mgf of a quadratic form.

Inference under Elliptically-Contoured families. Provost and Cheong (2002) present a one-dimensional integral expression (in t) for the pdf of an EC random vector with location $\boldsymbol{\mu}_0$ and scale Σ_0 that involves a “weighting” function $w(t)$ (defined through the inverse Laplace transform of the EC generator function), and the pdf of an n -dimensional normal with mean $\boldsymbol{\mu}_0$ and covariance matrix Σ_0/t . With this construct, and if the weighting function can be analytically obtained, it is possible to derive an integral representation for the mgf of a QEE in EC random variables. A *Laplace approximation* may then be utilized to furnish a closed-form expression for the mgf. Any of a large number of fast numerical integration routines could also be used to approximate this mgf to a very high degree of precision. Either method of mgf approximation then yields a possible starting point for SPBB computations.

References

- Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. Roy. Statist. Soc. Ser. B*, **66**, 165–185.
- Crainiceanu, C., Ruppert, D., Claeskens, G. and Wand, M. (2005). Exact likelihood ratio tests for penalized splines. *Biometrika*, **92**, 91–103.
- Daniels, H.E. (1983). Saddlepoint approximations for estimating equations. *Biometrika*, **70**, 89–96.
- Feuerverger, A. and Wong, A. (2000). Computation of Value-at-Risk for nonlinear portfolios. *Journal of Risk*, **3**, 37–55.
- Gupta, A.K. and Huang, W.J. (2002). Quadratic forms in skew normal variates. *J. Math. Anal. Appl.*, **273**, 558–564.
- Huang, W-J. and Chen, Y-H. (2006). Quadratic forms of multivariate skew normal-symmetric distributions. *Stat. Prob. Lett.*, **9**, 871–879.
- Lugannani, R. and Rice, S.O. (1980). Saddlepoint approximations for the distribution of sums of independent of random variables. *Adv. Appl. Prob.*, **12**, 475–490.
- Paige, R.L. and Trindade, A.A. (2008). Practical small sample inference for single lag subset autoregressive models. *J. Statist. Planning and Inference*, **138**, 1934–1949.
- Paige, R.L. and Fernando, P. (2008). Robust inference in conditionally linear nonlinear regression models. *Scand. J. Statist.*, **35**, 158–168.
- Paige, R.L., Trindade, A.A. and Fernando, P.H. (2009). Saddlepoint-Based bootstrap inference for quadratic estimating equations. *Scand. J. Statist.*, **36**, 98–111.
- Paige, R.L. and Trindade, A.A. (2010). The Hodrick-Prescott Filter: a special case of penalized spline smoothing. *Electron. J. Statist.*, **4**, 856–874.
- Paige, R.L. and Trindade, A.A. (2012). Fast and accurate inference for the smoothing parameter in semiparametric models. *Austral. N. Zeal. J. Statist.*, to appear.
- Provost, S.B. and Cheong, Y-H. (2002). The distribution of hermitian quadratic forms in elliptically contoured random vectors. *J. Statist. Planning and Inference*, **102**, 303–316.
- Wang, T., Li, B. and Gupta, A.K. (2009). Distribution of quadratic forms under skew normal settings. *J. Multivariate Anal.*, **100**, 533–545.

Projection-based nonparametric checks of regressions with functional covariates

Valentin Patilea¹, César Sánchez-Sellero², Matthieu Saumard³

¹ CREST (Ensaï) & IRMAR, France

² Facultad de Matemáticas, Universidad de Santiago de Compostela, Spain

³ INSA-IRMAR, France

E-mail for correspondence: patilea@ensai.fr

Abstract: This paper studies the problem of nonparametric testing for the effect of a random functional covariate on a real-valued error term. The covariate takes values in $L^2[0, 1]$, the Hilbert space of the square-integrable real-valued functions on the unit interval. The error term could be directly observed as a response or *estimated* from a functional parametric model, like for instance the functional linear regression. Our test is based on the remark that checking the no-effect of the functional covariate is equivalent to checking the nullity of the conditional expectation of the error term given a sufficiently rich set of projections of the covariate. Such projections could be on elements of norm 1 from finite-dimension subspaces of $L^2[0, 1]$. Next, the idea is to search a finite-dimension element of norm 1 that is, in some sense, the least favorable for the null hypothesis. Finally, it remains to perform a nonparametric check of the nullity of the conditional expectation of the error term given the scalar product between the covariate and the selected least favorable direction. For such finite-dimension search and nonparametric check we use a kernel-based approach. As a result, our test statistic is a quadratic form based on univariate kernel smoothing and the asymptotic critical values are given by the standard normal law. The test is able to detect nonparametric alternatives, including the polynomial ones. The error term could present heteroscedasticity of unknown form. We do not require the law of the covariate to be known. The test could be implemented quite easily and performs well in simulations and real data applications. We illustrate the performance of our test for checking the functional linear regression model.

Keywords: functional data regression; Kernel smoothing; Nonparametric testing

1 Introduction

Consider a sample of independent copies $(U_1, X_1), \dots, (U_n, X_n)$ of (U, X) where U is a real-valued random variable and X is a square-integrable random function defined on the unit interval. The problem we investigate herein is the test of the hypothesis

$$H_0 : \mathbb{E}(U|X) = 0 \quad \text{almost surely (a.s.)} \quad (1)$$

against the nonparametric alternative $\mathbb{P}[\mathbb{E}(U|X) = 0] < 1$. We consider two cases: (a) U is observed; and (b) U is estimated as a residual of a parametric model for functional covariates and scalar responses.

The monographs of Ramsay and Silverman (2002, 2005) and Ferraty (2011) provide a comprehensive landscape of the importance of the statistical methods for functional data. Estimation and prediction with functional covariates received substantial attention in the literature, while the goodness-of-fit problem we address here seems to be much less explored. There is a large literature on model checks like (1) against nonparametric alternatives when X takes values in a finite-dimension space, see for instance Härdle and Mammen (1993), Guerre and Lavergne (2005) and the references therein. In the case of functional covariate X , much little work was accomplished for testing against general types of alternatives. To our best knowledge, the only contribution considering the problem of testing H_0 against nonparametric alternatives in the cases (a) and (b) is the recent paper of Delsol et al. (2011) who extend the idea of Härdle and Mammen (1993) to the functional covariate case. However, it is not clear how the test of Delsol et al. (2011) could be applied in practice, for instance for testing the goodness-of-fit of the functional linear model. More substantial work was done for testing for no effect in a functional linear model, see Cardot et al. (2003), or for testing the functional linear model against quadratic alternatives, see Horváth and Reeder (2011). By construction, such procedures are not able to detect general alternative hypotheses.

The test we introduce herein is based on a dimension reduction idea used by Lavergne and Patilea (2008) in a finite dimension setup. Our test is able to detect *nonparametric* alternatives, including the polynomial ones. The variable U could present heteroscedasticity of unknown form. We do not require the law of the covariate X to be given or to be of a certain type, like for instance Gaussian. The test could be implemented quite easily and performs well in simulations and real data applications.

We also apply our projection-based approach for nonparametric checks of the functional regression models. We will focus on the linear functional model, although the methodology we propose also adapts to other models, like for instance the generalized functional linear models introduced by Müller and Stadtmüller (2005). In the functional regression case the variable U is the unobserved error term of the regression model and hence the test statistic is based on the estimated residuals. We still obtain standard normal critical values and consistency against nonparametric alternatives, fixed or approaching the null hypothesis. However, more restrictive conditions on the bandwidths are required due to the estimation of the slope of the functional linear model. This induces restrictions on the rate the directional alternatives may approach the null hypothesis. More difficult the estimation of the slope parameter is, slower the rate the directional alternative approach the null hypothesis should be. To estimate the slope parameter we use the functional principal component analysis approach.

2 Dimension reduction in nonparametric testing

For any $p \geq 1$, let $\mathcal{S}^p = \{\gamma \in \mathbb{R}^p : \|\gamma\| = 1\}$ denote the unit hypersphere in \mathbb{R}^p . Let $L^2[0, 1]$ be the space of the square-integrable real-valued functions defined on the unit interval and let

$$\langle X_1, X_2 \rangle = \int_0^1 X_1(t)X_2(t)dt, \quad X_1, X_2 \in L^2[0, 1].$$

Let $\|\cdot\|_{L^2}$ be the associated norm. Hereafter $\mathcal{R} = \{\rho_1, \rho_2, \dots\}$ be an arbitrarily fixed orthonormal basis of the function space $L^2[0, 1]$, that is $\langle \rho_i, \rho_j \rangle = \delta_{ij}$. Then the predictor process X can be expanded into $X(t) = \sum_{j=1}^\infty x_j \rho_j(t)$, where the random coefficients x_j are given by $x_j = \langle X, \rho_j \rangle$. For a fixed positive integer p , $X^{(p)} \in L^2[0, 1]$ will be the projection of X on the subspace generated by the first p elements of the basis \mathcal{R} , that is $X^{(p)}(t) = \sum_{j=1}^p x_j \rho_j(t)$. By abuse we identify $X^{(p)}$ with the p -dimension random vector (x_1, \dots, x_p) . In the following $\beta = \sum_{j=1}^\infty b_j \rho_j(t)$ will denote a non random element of $L^2[0, 1]$. Our approach relies on the following lemma that extends Lemma 2.1 of Lavergne and Patilea (2008) to Hilbert space-valued random variables.

Lemma. *Let $X \in L^2[0, 1]$ and $Z \in \mathbb{R}$ be random variables. Assume that $\mathbb{E}|Z| < \infty$ and $\mathbb{E}(Z) = 0$.* (A) *The following statements are equivalent:*

1. $\mathbb{E}(Z | X) = 0$ a.s.
2. $\mathbb{E}(Z | \langle X, \beta \rangle) = 0$ a.s. $\forall \beta \in L^2[0, 1]$ with $\|\beta\|_{L^2} = 1$.
3. for any integer $p \geq 1$, $\mathbb{E}(Z | \langle X, \gamma \rangle) = 0$ a.s. $\forall \gamma \in \mathcal{S}^p$.
4. for any integer $p \geq 1$, $\mathbb{E}(Z | X^{(p)}) = 0$ a.s.

(B) *Suppose in addition that for any positive real number s ,*

$$\mathbb{E}(|Z| \exp\{s\|X\|\}) < \infty. \tag{2}$$

If $\mathbb{P}[\mathbb{E}(Z | X) = 0] < 1$, then there exists a positive integer $p_0 \geq 1$ such that for any integer $p > p_0$, the set $\{\gamma \in \mathcal{S}^p : \mathbb{E}(Z | \langle X, \gamma \rangle) = 0$ a.s. $\}$ has Lebesgue measure zero on the unit hypersphere \mathcal{S}^p and is not dense.

Point (A) is a cornerstone for proving the behavior of our test under the null and the alternative hypothesis. Point (B) shows that in applications it will not be difficult to find directions γ able to reveal the failure of the null hypothesis (1). Under the additional assumption (2) such directions represent almost all the points on the unit hyperspheres \mathcal{S}^p , provided p is sufficiently large. The assumption (2) is not restrictive for testing purposes.

3 Testing the effect of a functional covariate

We introduce a general approach for nonparametric testing of the effect of a functional covariate X on a real-valued random variable U . Without loss of generality, let $\mathbb{E}(U) = 0$. Our approach is based on the Lemma above and *univariate* kernel smoothing. In this way we avoid the problem of smoothing in infinite-dimension, in particular we avoid using the small ball function required in the kernel regression with functional covariates, see Ferraty and Vieu (2006), Delsol et al. (2011). For any $\gamma \in \mathbb{R}^p$, let

$$Q(\gamma) = \mathbb{E}\{U \mathbb{E}[U \mid \langle X, \gamma \rangle] f_\gamma(\langle X, \gamma \rangle)\} = \mathbb{E}\{\mathbb{E}^2[U \mid \langle X, \gamma \rangle] f_\gamma(\langle X, \gamma \rangle)\}.$$

For any $p \geq 1$, let $B_p \subset \mathcal{S}^p$ be a set with strictly positive Lebesgue measure in \mathcal{S}^p . From the Lemma above, the null hypothesis (1) holds true iff

$$\forall p \geq 1, \quad \max_{\gamma \in B_p} Q(\gamma) = 0. \tag{3}$$

3.1 The test statistic

In view of equation (3), with at hand a sample of (U, X) , define

$$Q_n(\gamma) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} U_i U_j \frac{1}{h} K_h(\langle X_i - X_j, \gamma \rangle), \quad \gamma \in \mathcal{S}^p,$$

where $K_h(\cdot) = K(\cdot/h)$, where $K(\cdot)$ is a kernel and h a bandwidth. For fixed p and $\gamma \in \mathcal{S}^p$, it is well-known that $nh^{1/2}Q_n(\gamma)$ has asymptotic centered normal distribution under H_0 . We will show that the asymptotic normal distribution is preserved even when p grows at a suitable rate with the sample size. On the other hand, Lemma point (B) indicates that if p is large enough, the maximum of $Q(\gamma)$ over γ stay away from zero.

For a fixed p the statistic $Q_n(\gamma)$ is expected to be close to $Q(\gamma)$ uniformly in γ . Then a natural idea would be to build a test statistic using the maximum of $Q_n(\gamma)$ with respect to γ . However, there is an additional difficulty one faces in the functional data framework since then one has to let p to grow to infinity with the sample size, and hence the closeness between $Q_n(\gamma)$ and $Q(\gamma)$ requires a more careful investigation. On the other hand, like in the finite dimension covariate case, under H_0 one expects $Q_n(\gamma)$ to converges to zero for any p and γ and thus the objective function to be flat.

We will choose a direction γ as the least favorable direction for the null hypothesis H_0 obtained from a penalized criterion based on a standardized version of $Q_n(\gamma)$. (Lavergne and Patilea (2008) considered a similar idea.) More precisely, fix some $\beta_0 \in L^2[0, 1]$ that could be interpreted as an initial *guess* of an unfavorable direction for H_0 . Let $b_{0j}, j \geq 1$, be the coefficients in the expansion of β_0 in the basis \mathcal{R} . For any given $p \geq 1$ such that $\|(b_{01}, \dots, b_{0p})\| > 0$, let $\gamma_0^{(p)} = (b_{01}, \dots, b_{0p}) / \|(b_{01}, \dots, b_{0p})\|$. Let $\hat{v}_n^{(2)}(\cdot)$ be

as estimate of the variance of $nh^{1/2}Q_n(\cdot)$. Given $B_p \subset \mathcal{S}^p$ with strictly positive Lebesgue measure in \mathcal{S}^p that contains $\gamma_0^{(p)}$, the least favorable direction γ for H_0 is defined as

$$\hat{\gamma}_n = \arg \max_{\gamma \in B_p} \left[nh^{1/2}Q_n(\gamma)/\hat{v}_n(\gamma) - \alpha_n \mathbb{I}_{\{\gamma \neq \gamma_0^{(p)}\}} \right],$$

where \mathbb{I}_A is the indicator function of a set A , and $\alpha_n, n \geq 1$ is a sequence of positive real numbers increasing to infinity at an appropriate rate that depends on the sample size and the rates of h and p and that will be made explicit below. Let us notice that the maximization used to define $\hat{\gamma}_n \in \mathcal{S}^p$ is a finite dimension optimization problem. The choice of β_0 , and thus of $\gamma_0^{(p)}$, is theoretically irrelevant, it does not affect the asymptotic critical values and the consistency results. However, in practice the choice of β_0 could be related to some prior information on a class of alternatives.

We prove that if $h \rightarrow 0$ and $(nh^2)^\alpha / \ln n \rightarrow \infty$ for some $\alpha \in (0, 1)$, there exists a constant $\lambda > 0$ such that $p \ln^{-\lambda} n$ is bounded, and $\alpha_n / \{p^{3/2} \ln n\} \rightarrow \infty$, the probability of the event $\{\hat{\gamma}_n = \gamma_0^{(p)}\}$ tends to 1 under H_0 . Hence $Q_n(\hat{\gamma}_n)/\hat{v}_n(\hat{\gamma})$ behaves asymptotically like $Q_n(\gamma_0^{(p)})/\hat{v}_n(\gamma_0^{(p)})$, even when p grows with the sample size. Therefore the test statistic we consider is

$$T_n = nh^{1/2} \frac{Q_n(\hat{\gamma}_n)}{\hat{v}_n(\hat{\gamma}_n)}, \text{ with } \hat{v}_n^2(\gamma) = \frac{2}{n(n-1)h} \sum_{j \neq i} U_i^2 U_j^2 K_h^2(\langle X_i - X_j, \gamma \rangle).$$

We will show that an asymptotic α -level test is given by $\mathbb{I}(T_n \geq z_{1-a})$, where z_a is the $(1 - a)$ -th quantile of the standard normal distribution.

We show that our test is omnibus and detects directional alternatives like

$$H_{1n} : U = U^0 + r_n \delta(X), \quad n \geq 1, \quad \text{with } \mathbb{E}(U^0 | X) = 0,$$

as soon as $r_n^2 nh^{1/2} / \{p^{3/2} \ln n\} \rightarrow \infty$.

4 Testing the functional linear model

Let us briefly explain how the new testing approach applies to functional regression models. Let U be a real-valued random variable and X be a random variable with values in $L^2[0, 1]$. The model we want to test is the functional linear model defined by

$$Y = a + \langle b, X \rangle + U,$$

with $b \in L^2[0, 1]$ and $a \in \mathbb{R}$ unknown parameters. The null hypothesis is $H_0 : \mathbb{E}(U|X) = 0$ a.s. The test statistic is similar to the one we proposed for testing the effect of a functional covariate. Let $\beta_0, \gamma_0^{(p)}, \mathcal{S}^p$ and B_p be defined as above. Let $\hat{b} \in L^2[0, 1]$ denote a generic estimator of b and let

$$\hat{a} = \bar{Y}_n - \int_0^1 \hat{b}(t) \bar{X}_n(t) dt = a - \int_0^1 \{\hat{b}(t) - b(t)\} \bar{X}_n(t) dt + \bar{U}_n,$$

where $\bar{U}_n = n^{-1} \sum_{i=1}^n U_i$. Let $\hat{U}_i = Y_i - \hat{a} - \langle \hat{b}, X_i \rangle$ be the residuals and let

$$Q_n(\gamma; \hat{a}, \hat{b}) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \hat{U}_i \hat{U}_j \frac{1}{h} K_h(\langle X_i - X_j, \gamma \rangle), \quad \gamma \in \mathcal{S}^p,$$

where recall $K(\cdot)$ is a kernel, h a bandwidth and $K_h(\cdot) = K(\cdot/h)$. Let $\hat{v}_n^2(\cdot; \hat{a}, \hat{b})$ be an estimate of the variance of $nh^{1/2}Q_n(\cdot; \hat{a}, \hat{b})$ like in section 3. The least favorable direction γ for H_0 is defined as

$$\hat{\gamma}_n = \arg \max_{\gamma \in B_p} \left[nh^{1/2}Q_n(\gamma; \hat{a}, \hat{b}) / \hat{v}_n(\gamma; \hat{a}, \hat{b}) - \alpha_n \mathbb{I}_{\{\gamma \neq \gamma_0^{(p)}\}} \right].$$

The test statistic is then

$$T_n = nh^{1/2} \frac{Q_n(\hat{\gamma}_n; \hat{a}, \hat{b})}{\hat{v}_n(\hat{\gamma}_n; \hat{a}, \hat{b})}.$$

We will show that, under suitable conditions, an asymptotic α -level test is given by $\mathbb{I}(T_n \geq z_{1-\alpha})$. Moreover, we show that the test is omnibus and we investigate its power against directional alternatives.

References

- Cardot, H., Ferraty, F., Mas, A., and Sarda, P. (2003). Testing Hypotheses in the Functional Linear Model. *Scand. J. Statist.*, **30**, 241–255.
- Delsol, L., Ferraty, F., and Vieu, P. (2011). Structural test in regression on functional variables. *J. of Multivariate Analysis*, **102**, 422–447.
- Ferraty, F. (Ed.) (2011). *Recent Advances in Functional Data Analysis and Related Topics*. Springer-Verlag Berlin Heidelberg.
- Guerre, E., and Lavergne, P. (2005). Data-driven rate-optimal specification testing in regression models. *Annals of Statistics*, **33**, 840–870.
- Härdle, W., and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, **21**, 1296–1947.
- Horváth, L., and Reeder, R. (2011). A test of significance in functional quadratic regression. arXiv:1105.0014v1 [math.ST].
- Lavergne, P. and Patilea, V. (2008). Breaking the curse of dimensionality in nonparametric testing. *Journal of Econometrics*, **143**, 103–122.
- Müller, H.G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, **33**, 774–805.
- Ramsay, J., and Silverman, B.W. (2005). *Functional Data Analysis* (2nd ed.). Springer-Verlag, New York.

Trend analysis of snow avalanche accidents in Tyrol within the years 1989–2010

Christian Pfeifer¹, Achim Zeileis¹

¹ Institut für Statistik, Universität Innsbruck, Universitätsstraße 15, 6020 Innsbruck, Austria

E-mail for correspondence: christian.pfeifer@uibk.ac.at

Abstract: Considering fatal snow avalanche events due to backcountry skiing in Tyrol we take notice of a positive trend within the last 22 years.

Keywords: Time series, fatal snow avalanche events.

1 Introduction

In Tyrol, which is known to be the part of Austria with the highest number of snow avalanche accidents, approximately 10–15 fatal incidents of avalanches occur every year. However, this number includes avalanche accidents of different kinds: Casualties in buildings, on traffic roads, outdoors without skiing, casualties due to skiing on slopes and backcountry skiing. It is reported that in Alpine countries (including Austria) the number of avalanche fatalities seems to be more or less constant over time [2]. Further on, it is reported that there is some sort of seasonality in the data in terms of higher frequencies of accidents within a distance of 5 or 6 years ([4][11]). 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. We are considering annual count data of fatal backcountry skiing avalanches in Tyrol within the years 1989 and 2010 which is resulting in a time series of length 22 ([1][6]).

2 Model

We propose the following model for capturing the:

$$\log(\mathbf{y}_t) = f(t) + x_t$$

	OLS	GAM	ARMA
Intercept	1.5478 (0.1829)	<i>smooth</i> –	1.5824 (0.1545)
Trend	0.0580 (0.0121)	<i>smooth</i> –	0.0555 (0.0127)
MA(1)			–0.3814 (0.2257)
log-likelihood	–19.7050	–19.7050	–18.3284
AIC	45.4101	45.4101	44.6568
BIC	48.6832	48.6832	49.0209

TABLE 1. Model estimation results for log-deaths in Tyrol: Parameter estimates, standard errors (in parentheses), log-likelihood, and information criteria. For the OLS model heteroskedasticity and autocorrelation consistent (HAC) standard errors are reported. For the GAM a smooth trend is employed (although it is selected to be essentially linear) and hence no coefficient estimates and standard errors are reported.

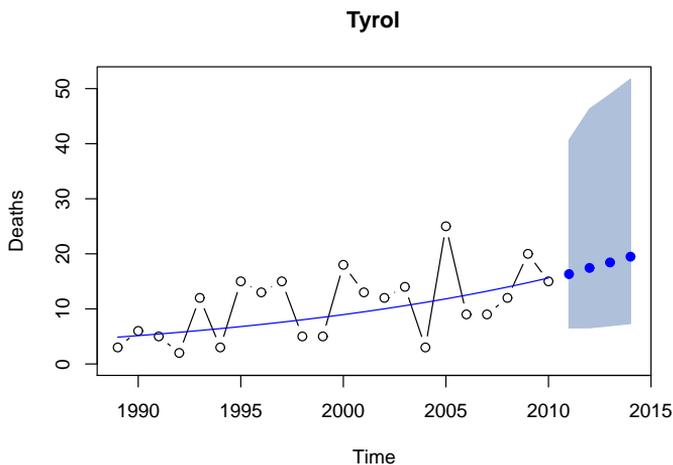


FIGURE 1. Observed annual backcountry avalanche fatalities in Tyrol 1989–2010 and forecasted fatalities 2011–2014 along with 90% prediction intervals (computed on log-scale).

where \mathbf{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 in the remainder, we consider autoregressive moving-average (ARMA) effects.

More precisely, we consider estimation of a simple linear trend $f(t) = \alpha + \beta \cdot t$

	OLS	GAM	ARMA
Intercept	3.1585 (0.1100)	<i>smooth</i> –	3.1771 (0.0736)
Trend	–0.0104 (0.0059)	<i>smooth</i> –	–0.0118 (0.0048)
MA(1)			–0.4867 (0.1963)
log-likelihood	–13.7157	–13.7157	–11.4660
AIC	33.4314	33.4314	30.9319
BIC	37.5332	37.5332	36.4011

TABLE 2. Model estimation results for log-deaths in Switzerland: Parameter estimates, standard errors (in parentheses), log-likelihood, and information criteria. For the OLS model HAC standard errors are reported. For the GAM a smooth trend is employed (although it is selected to be essentially linear) and hence no coefficient estimates and standard errors are reported.

with intercept α and trend/slope β , estimated by ordinary least squares (OLS), as well as a generalized additive model (GAM) with a smooth spline-based trend. Additionally, an ARMA model is selected using the best AIC fit from ARMA(p, q) models with a linear trend and $p = 0, \dots, 5$ and $q = 0, \dots, 5$.

The estimation results are reported in Table 1 which shows that all models closely agree on a linear trend (because the GAM selects an essentially linear trend) with about 5.5% growth in fatalities per year. Additionally, the ARMA model selects an MA(1) process for x_t with a moving average coefficient of -0.3814 , corresponding to an autocorrelation of -0.333 at lag 1 (and 0 at all higher lags).

Figure 1 shows the data and the trend function on the original scale including forecasts and confidence bands at the 90% level for the years 2011–2014. For comparison, we also assess the number of annual avalanche fatalities within the years 1980–2009 in Switzerland ([3][11]). The estimation results are reported in Table 2. Interestingly, all models agree again on a linear trend – however, the number of deaths *decreases* slightly by about 1% per year. Again, the ARMA model performs slightly better, exhibiting again moderate negative autocorrelation of -0.393 at lag 1 (corresponding to the MA coefficient of -0.4867), matching the results for Tyrol rather closely. In Figure 2, the data and trend on the original scale are depicted along with 90% prediction intervals.

3 Discussion

Modeling the number of annual backcountry avalanche fatalities in Tyrol within 1989 and 2010 on the log scale, we employed the sum of a linear trend

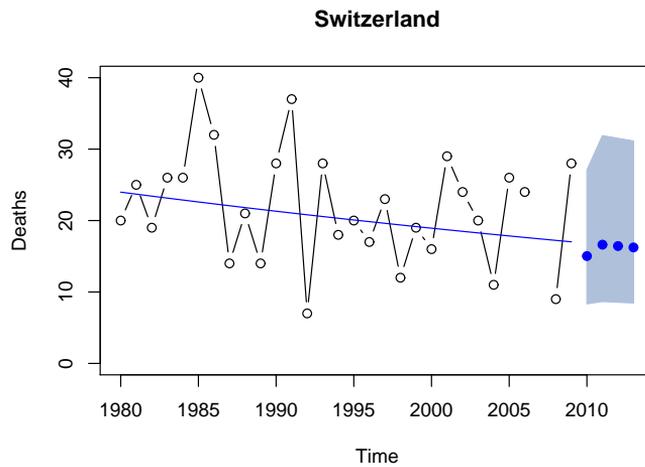


FIGURE 2. Observed annual backcountry avalanche fatalities in Switzerland 1980–2009 and forecasted fatalities 2010–2013 along with 90% prediction intervals (computed on log-scale).

model and a moving average MA(1) process (see Table 1 and Figure 1). There is a significant positive linear trend of backcountry skiing avalanche fatalities depending on time t by taking notice a growth rate of about 5.5% of annual fatalities. If our focus is only on the trend of avalanche fatalities we notice that the simple OLS modeling leads almost to the same result. And as far we can see, there is no evidence for seasonality in the data.

It is discussed that the positive trend is an effect of a possibly larger number of backcountry skiers in the last years. But unfortunately we do not have any reliable information about the number of backcountry skiers in Tyrol. At least, this is in contradiction to data of Switzerland and other countries in Europe ([2][11]). In the case of backcountry avalanche fatalities in Switzerland model selection leads to (as we can see in Table 2 and Figure 2) a very similar model with the only difference: There is a slightly negative trend of about 1% decrease per year.

In conclusion from our point of view, it is necessary to discuss steps to be taken against backcountry avalanches in Tyrol (e.g., employment of security services operating on off-piste slopes).

Computational details

All results have been obtained using the R system for statistical computing, version 2.15.0 ([10]). The GAMs have been fitted using `mgcv` 1.7-13 ([12], using default settings for `s()`). The ARMA models and their predictions

have been obtained using `forecast` 3.19 ([5]) and the HAC standard errors have been computed by `vcovHAC()` from `sandwich` 2.2-9 ([13]).

References

- [1] Amt der Tiroler Landesregierung (1994–2010): Schnee und Lawinen; Jahresberichte.
- [2] Brugger H. Durrer B. Adler-Kastner L. Falk M. Tschirky F. (2001): Field management of avalanche victims; *Resuscitation* 51(2001) 7–15.
- [3] Etter H.J. Zweifel B. Dürri L (2011) Schnee und Lawinen in den Schweizer Alpen. *Hydrologisches Jahrbuch 2008\2009*; WLS-Institut für Schnee- und Lawinenforschung SLF Davos.
- [4] Höller P. (2009): Avalanche cycles in Austria: an analysis of the major events in the last 50 years; *Natural Hazards* 48(2009) 399–424.
- [5] Hyndman R.J. Khandakar Y. (2008). Automatic time series forecasting: The `forecast` package for R; *Journal of Statistical Software* 27(3) 1–22.
- [6] Kuratorium für alpine Sicherheit (1989–201): Sicherheit im Bergland; Jahrbücher des Kuratoriums für alpine Sicherheit.
- [7] Pfeifer C. Rothart V. (2004): On probabilities of avalanches triggered by alpine skiers. An application of models for counts with extra zeros; *Proceedings International Workshop of Statistical Modelling 2004 Florence*.
- [8] Pfeifer C. (2010): On probabilities of avalanches triggered by alpine skiers. An empirically driven decision strategy for backcountry skiers based on these probabilities; *Presentation at the International Workshop of Statistical Modelling 2010 Glasgow*.
- [9] Pfeifer C. (2011): On probabilities of avalanches triggered by alpine skiers. Models with random effects; *Proceedings International Workshop of Statistical Modelling 2011 Valencia*.
- [10] 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.
- [11] Tschirky F. Bräber B. 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>

- [12] Wood S.N. (2006) Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.
- [13] Zeileis A. (2004). Econometric computing with HC and HAC covariance matrix estimators; Journal of Statistical Software 11(10) 1–17.

Address for correspondence

Christian Pfeifer
Institut für Statistik, Universität Innsbruck
Universitätsstraße 15, A-6020 Innsbruck
E-Mail: christian.pfeifer@uibk.ac.at

Standard statistics as likelihood statements

Charles Rohde¹

¹ Johns Hopkins University, USA

E-mail for correspondence: crohde@jhsph.edu

Abstract: P-values do not measure evidence but evidence is a useful way to describe what the data have to say. I explain how the results of a standard statistical analysis can be given an evidential explanation using the likelihood paradigm. The likelihood paradigm also is shown to satisfy Birnbaum's confidence concept. Comments about a Bayesian presentation are also made.

Keywords: likelihood, P-values, evidence

1 Introduction

As is well known P-values do not measure evidence; Schervish (1996), Royall (1997). Nevertheless it is widely said, particularly in elementary texts, that a P-value measures evidence against a null hypothesis; Freedman et al. (2007), Moore and McCabe (2005). We can partially rectify this disagreement using the likelihood paradigm. It is possible to give a crude evidential interpretation to P-values and confidence intervals that is correct and a better way of describing what the data has to say about parameter values. Recall that the likelihood paradigm is based on the Law of Likelihood which states that the (statistical) evidence for parameter value θ_1 vs θ_0 is captured (represented) by

$$\frac{\mathcal{L}(\theta_1; \hat{\theta})}{\mathcal{L}(\theta_0; \hat{\theta})}$$

where for any θ

$$\mathcal{L}(\theta; \hat{\theta}) = \frac{f(x; \theta)}{f(x; \hat{\theta})}$$

and $\hat{\theta}$ is the best supported value of θ (the maximum likelihood estimate). We are assuming the standard statistical model representing x as an observed value of random X which has sample space \mathcal{X} , parameter space Θ and density $f(x; \theta)$ at x for a given value of θ . Royall (1997)

In the presence of nuisance parameters the preceding Law of Likelihood is based on the profile likelihood

$$\mathcal{L}_P(\theta; \hat{\theta}) = \frac{f(x; \theta, \hat{\lambda}(\theta))}{f(x; \hat{\theta}, \hat{\lambda})}$$

where $\widehat{\lambda}(\theta)$ is the MLE of λ for fixed θ and $\widehat{\theta}$, $\widehat{\lambda}$ are the MLE's of θ and λ . Royall (2000) has shown that under the usual regularity conditions for the asymptotic normality of maximum likelihood estimators that the profile likelihood may be used as if it were a likelihood.

2 Standard Results

Suppose the results of a statistical analysis are reported as

Parameter	Estimate	Std Error	Lower Bound	Upper Bound	P-value
θ_1	$\widehat{\theta}_1$	se ₁	θ_{1L}	θ_{1U}	p_1
θ_2	$\widehat{\theta}_2$	se ₂	θ_{2L}	θ_{2U}	p_2
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
θ_k	$\widehat{\theta}_k$	se _k	θ_{kL}	θ_{kU}	p_k

where

- (1) θ_i is the parameter.
- (2) $\widehat{\theta}_i$ is the estimate of the parameter.
- (3) se_{*i*} is the estimated standard error of the estimate.
- (4) θ_{iL} is the lower confidence bound (typically a 95% interval)
- (5) θ_{iU} is the upper confidence bound (typically a 95% interval)
- (6) p_i is the P-value (assumed to be testing that the parameter is 0).

This would be the kind of analysis reported as the result of regression, analysis of covariance, logistic regression, Poisson regression etc. by SAS, R, STATA and other statistical packages. The estimates are typically maximum likelihood and are approximately normally distributed.

As Royall (2000) and then Royall and Tsou (2002) have argued we may view the estimates and standard errors as providing input to a normal (profile) likelihood. From this we may crudely interpret the results of the analysis from a likelihood perspective. That is we may consider the likelihood of θ_i as approximately normal with the estimate as center and with curvature specified by the estimated standard error i.e.

$$\mathcal{L}(\theta_i; \widehat{\theta}_i) = \exp \left\{ -\frac{(\theta_i - \widehat{\theta}_i)^2}{2(\text{se}_i)^2} \right\}$$

3 Confidence Intervals to Likelihood Intervals

Consider first the confidence interval. We have that the lower and upper confidence bounds occur at

$$\theta_L = \hat{\theta} - \text{se}(\hat{\theta})z_{1-\alpha/2} \quad \text{and} \quad \theta_U = \hat{\theta} + \text{se}(\hat{\theta})z_{1-\alpha/2}$$

Thus the likelihood at the lower bound is given by

$$\mathcal{L}(\theta_L; \hat{\theta}) = \exp \left\{ -\frac{(\hat{\theta} - \hat{\theta} - \text{se}z_{1-\alpha/2})^2}{2(\text{se})^2} \right\} = \exp \left\{ -\frac{z_{1-\alpha/2}^2}{2} \right\}$$

Similarly the likelihood at the upper bound is given by

$$\mathcal{L}(\theta_U; \hat{\theta}) = \exp \left\{ -\frac{z_{1-\alpha/2}^2}{2} \right\}$$

and hence the likelihood ratio for comparing either θ_L or θ_U to the maximum likelihood value is

$$\exp \left\{ \frac{z_{1-\alpha/2}^2}{2} \right\}$$

This allows us to construct a short table relating confidence intervals to likelihood intervals.

Confidence	z	k
90	1.645	3.9
95	1.960	6.8
99	2.576	27.6
99.9	3.291	224.4
99.99	3.891	1935.9

4 P-values to Likelihood Evidence

Consider now the P-value (assumed two-sided)

$$\begin{aligned} \frac{p_i}{2} &= \mathbb{P}_{\theta_i=0}(\hat{\theta}_i \geq c_i) \\ &= \mathbb{P}_{\theta_i=0} \left(\frac{\hat{\theta}_i}{\text{s.e.}(\hat{\theta}_i)} \geq \frac{c}{\text{se}} \right) \\ &= \mathbb{P} \left(Z \geq \frac{c}{\text{se}_i} \right) \end{aligned}$$

It follows that

$$c_i = \text{se}_i z_{1-p_i/2}$$

where $z_{1-p_i/2}$ is the $1 - p_i/2$ quantile of the normal distribution Hence the observed value of $\hat{\theta}_i$ is given by

$$\hat{\theta}_i = z_{1-p_i/2} \text{se}_i$$

The likelihood at this value of θ_i is 1 and the likelihood at $\theta_i = 0$ is

$$\mathcal{L}(0, \hat{\theta}_i) = \exp \left\{ -\frac{(0 - \text{se}_i z_{1-p_i/2})^2}{2[\text{se}_i]^2} \right\} = \exp \left\{ -\frac{z_{1-p_i/2}^2}{2} \right\}$$

It follows that the support for $\theta_i = \hat{\theta}_i$ vs $\theta_i = 0$ is

$$\frac{\mathcal{L}(\hat{\theta}_i; \hat{\theta}_i)}{\mathcal{L}(0; \hat{\theta}_i)} = \exp \left\{ \frac{z_{1-p_i/2}^2}{2} \right\}$$

This provides a connection between of k and P-values. Here is a short table

P-value	z -one sided	z -two sided	k one sided	k two sided
.1	1.28	1.64	2.27	3.87
.05	1.64	1.96	3.87	6.83
.01	2.33	2.58	14.97	27.59
.001	3.09	3.29	118.48	224.48
.0001	3.72	3.89	1007.82	1935.95

- (1) Thus, for example, a P-value of .01 corresponds to the observed value $\hat{\theta}_i$ being 15 times better supported than $\theta_i = 0$.
- (2) It follows that there is evidence at level 15 for some value of θ_i (the MLE) vis a vis $\theta_i = 0$.
- (3) Note that this does not say that there is evidence against $\theta = 0$ as P-values are commonly interpreted.
- (4) Nor does it say that an important fact has been discovered, this must be judged by the magnitude of $\hat{\theta}_i$ i.e. is its magnitude indicative of an effect of potential subject-matter importance?

5 Birnbaum's Confidence Concept

Alan Birnbaum introduced the "confidence concept" to the foundations of statistics in 1977.

Confidence Concept

The **confidence concept**: A concept of statistical evidence is not plausible unless it finds "strong evidence" for H_2 as against H_1 with small probability (α) when H_1 is true, and with much larger probability ($1 - \beta$) when H_2 is true.

The Law of Likelihood (LL) has the following property: The probability of misleading evidence, mis , satisfies the **universal bound** i.e.

$$\text{mis} = \mathbb{P}_{\theta_0} \left\{ \frac{f(x; \theta_1)}{f(x; \theta_0)} \geq k \right\} \leq \frac{1}{k}$$

for $k > 1$ and any $\theta_1 \neq \theta_0$.

In most situations the probability of misleading evidence is much smaller than the universal bound. For a random sample from the normal distribution we have

$$\text{mis} = \Phi \left\{ -\frac{\sqrt{n}(\theta_1 - \theta_0)}{2\sigma} - \frac{\sigma \ln(k)}{\sqrt{n}(\theta_1 - \theta_0)} \right\} \leq \Phi \left(-\sqrt{2 \ln(k)} \right)$$

This bound is achieved when

$$\frac{\theta_1 - \theta_0}{\sigma} = \sqrt{\frac{2 \ln(k)}{n}}$$

Royall has shown that, again, under the usual regularity conditions for maximum likelihood, if X_1, X_2, \dots, X_n are iid as $f(x; \theta)$ then

$$\text{mis} \approx \Phi \left(-\frac{c}{2} - \frac{\ln(k)}{c} \right)$$

where

$$c = \frac{\sqrt{n}(\theta_1 - \theta_0)}{i(\theta_0)}$$

and $i(\theta)$ is Fisher's information at $\theta = \theta_0$ i.e.

$$i(\theta_0) = -\mathbb{E}_{\theta_0} \left\{ \frac{\partial^2 \ln[f(X; \theta)]}{\partial \theta^2} \right\} \Bigg|_{\theta=\theta_0}$$

This result implies that for large samples the probability of misleading evidence has approximately the bound given by the normal distribution. It follows that defining statistical evidence by the LL produces a measure of statistical evidence which satisfies the first of Birnbaum's plausibility criteria in CC.

The second plausibility criterion is more subtle since it depends on sample size and not just the magnitude of the statistical evidence. Consider the probability of obtaining strong statistical evidence when the alternative is true using the Law of Likelihood as the definition of statistical evidence.

First notice that if H_1 is true then with iid X 's as $f(x; \theta)$ and $\theta_1 \neq \theta_0$

$$\mathbb{P}_{\theta_1} \left(\frac{f(x_n; \theta_1)}{f(x_n; \theta_0)} \geq k \right) \rightarrow 1$$

by the Law of Large Numbers. Thus Birnbaum's second plausibility requirement is satisfied at least for large samples. Rather routine calculations show that for a variety of distributions the sample size can be selected so as to achieve large probability of finding strong evidence in favor of the θ_1 when θ_1 is true.

6 Bayesian Calibration

Similar results obtain for a Bayesian interpretation since the estimators in section 1 may be viewed as having an approximate joint normal posterior distribution; Berger (1985). Thus confidence intervals may be viewed as approximate posterior intervals provided prior information is “vague”.

7 Conclusions

Most of what we do in beginning statistics courses can be rephrased in terms of a likelihood or Bayesian analysis with current software. All we need is a way to draw likelihood functions and or posterior densities, easy in R or SAS, not so easy in STATA. If a graphical capability is not available or lack of space prohibits the presentation of a graph the notation of Louis and Zeger (2007) can be used in which $_{1.21}2.32_{3.43}^{-(1/8)}$ indicates a $1/8$ likelihood interval centered at 2.32 with upper endpoint at 3.43 and lower endpoint at 1.21.

While not the neatest notation this type of reporting is better than the convoluted interpretation of confidence intervals students are required to learn and statisticians are forced to explain to their scientific clients. Indeed, as Bayarri, et al. (1988) pointed out

Fisher (1973) uses this normalized likelihood to evaluate the reasonableness of different values of the parameter and states (p.76) that values for which the likelihood is less than $1/15$ “are obviously open to grave suspicion”. It is to be regretted that the use of this cutoff point of $1/15$ has not become nearly as popular as Fisher’s other famous proposal of using 0.05 as a cutoff for the significance level of a test. Had this proposal become widely accepted, statistical practice would have been significantly changed for the better.

References

- Bayarri, M., DeGroot, M. and Kadane, J. (1988) What is the Likelihood Function?, *Statistical decision Theory and related Topics IV*, Springer.
- Berger, J.O (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer.
- Freedman, D., Pisani, R. and Purves, R. (2007) *Statistics Fourth Edition* Norton, New York and London, page 480
- Birnbaum, A. (1977) The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory, *Synthese* 1, 19-49

- Goodman, S.N. and Royall, R.M.(1988),Evidence and Scientific Research, *American Journal of Public Health*,**78**
- Louis, T.A. and Zeger, S.L.(2007), Effective Communication of Standard Errors and Confidence Intervals, *Johns Hopkins University, Dept. of Biostatistics Working Papers*
- Moore, D.S. and McCabe, G.P.(2005), *Introduction to the Practice of Statistics; Fifth Edition*, W.H. Freeman & Co., page 405
- Royall, R. M.(1997), *Statistical Evidence: A Likelihood Paradigm*,Chapman & Hall/CRC
- Royall, R.M.(2000). On the Probability of Observing Misleading Statistical Evidence. *Journal of the American Statistical Association*, **95** , 760–768
- Royall, R.M. and Tsou, (2002). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions,*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 391–404
- Schervish, M.J.(1996) P values; What They are and What They are Not, *The American Statistician*, **50** 203–206

Bayesian approximation methods for pseudo-posterior distributions in the presence of nuisance parameters

Erlis Ruli¹, Laura Ventura¹

¹ Department of Statistics, Padova, Italy

E-mail for correspondence: `ruli@stat.unipd.it`

Abstract: This paper discusses recent developments in higher-order asymptotics for obtaining accurate approximations to pseudo-posterior distributions, e.g. posterior distributions based on suitable pseudo-likelihood functions, and to the corresponding tail area probabilities, for practical use in Bayesian analysis. The methodology proposed rely on pseudo-likelihood quantities and the prior distribution for the parameter of interest. An example is illustrated in the context of survival models with a real-life dataset (concerning malignant mesothelioma).

Keywords: Asymptotics expansion; Cox regression; Laplace approximation; Pseudo-likelihood; Tail area probability.

1 Introduction

Several pseudo-likelihood functions – such as the composite, the empirical, the quasi and the partial likelihoods – may be used successfully as a basis for Bayesian inference. In recent years, there have been considerable developments and applications of the corresponding pseudo-posterior distributions. For instance, in semiparametric models or in the presence of minimal assumptions on the model, a posterior distribution may be derived from the combination of prior information with a suitable likelihood function derived from estimating equations (see, e.g., Lazar, 2003, Lin, 2006, Greco *et al.*, 2008), or in models with complicated dependence structures, when the full likelihood is impractical to compute or even analytically unknown, a posterior distribution may be derived using a composite likelihood (see Smith and Stephenson, 2009, Pauli *et al.*, 2011, Ribatet *et al.*, 2012). A general pseudo-likelihood function $\tilde{L}(\theta) = \tilde{L}(\theta|y)$ is a function of the parameter of interest θ and of the data with properties similar to those of a genuine likelihood function (see e.g. Pace and Salvani, 1997, Ch. 4). By considering a pseudo-likelihood function $\tilde{L}(\theta)$ in the Bayesian framework, a pseudo-posterior distribution can be defined as

$$\tilde{\pi}(\theta|y) \propto \pi(\theta) \tilde{L}(\theta), \tag{1}$$

where $\pi(\theta)$ is a given prior on $\theta \in \Theta \subseteq \mathbb{R}^p$ and $y = (y_1, \dots, y_n)$ is the observed random sample of size n . Pseudo-posterior distributions of the form (1) have been discussed in the Bayesian literature for the elimination of nuisance parameters (see, e.g., Ventura *et al.*, 2009 and Ventura and Racugno, 2011), to achieve robustness with respect to the presence of outliers or model misspecifications (Greco *et al.*, 2008, Agostinelli and Greco, 2012) or to relieve some assumptions on the model (see, e.g., Lazar 2003, Lin 2006, Pauli *et al.*, 2011).

Under broad regularity conditions on $\tilde{L}(\theta)$, similar to those required for asymptotic normality of the maximum likelihood estimator (MLE) and under regularity conditions on the prior, by the technique of Laplace approximation for (1) we get (see e.g. Brazzale *et al.*, 2007, Pauli *et al.*, 2011)

$$\tilde{\pi}(\theta|y) = \frac{|\tilde{j}(\tilde{\theta})|^{1/2}}{(2\pi)^{p/2}} \exp \left\{ \tilde{\ell}(\theta) - \tilde{\ell}(\tilde{\theta}) \right\} \frac{\pi(\theta)}{\pi(\tilde{\theta})} (1 + O_p(n^{-1})), \quad (2)$$

where $\tilde{\theta}$ is the pseudo-MLE, $\tilde{j}(\theta)$ is the pseudo-observed information, and $\tilde{\ell}(\theta) = \log \tilde{L}(\theta)$. Expansion (2) is the basis to derive a normal approximation for $\tilde{\pi}(\theta|y)$ centered at $\tilde{\theta}$, with covariance matrix approximated by $\tilde{j}(\theta)$ at $\tilde{\theta}$ (see Lazar, 2003, Greco *et al.*, 2008, Pauli *et al.*, 2011).

In many applications of practical interest, the parameter θ is partitioned as $\theta = (\psi, \lambda)$, where ψ is a scalar interest parameter and λ is a $(p - 1)$ -dimensional nuisance parameter. When nuisance parameters are present, Bayesian inference about ψ may be based on the marginal pseudo-posterior distribution for ψ , given by

$$\tilde{\pi}_m(\psi|y) = \int \tilde{\pi}(\theta|y) d\lambda = \frac{\int \pi(\psi, \lambda) \tilde{L}(\psi, \lambda) d\lambda}{\int \int \pi(\psi, \lambda) \tilde{L}(\psi, \lambda) d\lambda d\psi}. \quad (3)$$

The aim of this paper is to discuss a Laplace approximation for the posterior distribution (3), paralleling the results for posteriors based on a proper likelihood function (see Reid, 2003, Ventura and Racugno, 2011). Moreover, we derive a tail area approximation from (3), which can be used to compute accurate credible sets for ψ , even for small sample sizes.

The outline of the paper is as follows. Section 2 discusses the higher-order approximations for (3). An application to survival models is reported in Section 3. Some final remarks conclude the paper.

2 Higher-order approximations

Let us focus on the marginal pseudo-posterior distribution (3). It is a ratio of two integrals, and the denominator can be approximated by the Laplace method, as in (2). To approximate the numerator, we expand $\tilde{\ell}(\psi, \lambda)$ as a function of λ about $\tilde{\lambda}_\psi$, the value satisfying $\partial \tilde{\ell}(\psi, \tilde{\lambda}_\psi) / \partial \lambda = 0$, to get

$$\exp \left\{ \tilde{\ell}(\psi, \tilde{\lambda}_\psi) \right\} |\tilde{j}_{\lambda\lambda}(\psi, \tilde{\lambda}_\psi)|^{-1/2} (2\pi)^{(p-1)/2} \pi(\psi, \tilde{\lambda}_\psi),$$

where $\tilde{j}_{\lambda\lambda}(\psi, \lambda) = -\partial^2 \tilde{\ell}(\psi, \lambda) / \partial \lambda \partial \lambda^T$. Combining this expansion with the Laplace approximation to the denominator, we have

$$\tilde{\pi}_m(\psi|y) \doteq c |\tilde{j}_P(\tilde{\psi})|^{1/2} \exp \left\{ \tilde{\ell}_P(\psi) - \tilde{\ell}_P(\tilde{\psi}) \right\} \rho(\psi, \tilde{\psi}) \frac{\pi(\psi, \tilde{\lambda}_{\psi})}{\pi(\tilde{\psi}, \tilde{\lambda})}, \tag{4}$$

where the symbol “ \doteq ” indicates that the approximation is accurate to $O(n^{-3/2})$ (see Reid, 2003), c is a suitable constant,

$$\rho(\psi, \tilde{\psi}) = \frac{|\tilde{j}_{\lambda\lambda}(\tilde{\psi}, \tilde{\lambda})|^{1/2}}{|\tilde{j}_{\lambda\lambda}(\psi, \tilde{\lambda}_{\psi})|^{1/2}},$$

$\tilde{\ell}_P(\psi) = \tilde{\ell}(\psi, \tilde{\lambda}_{\psi})$ is the pseudo-profile likelihood and $\tilde{j}_P(\psi)$ denote the pseudo-observed profile information. Remark that approximation (4) holds also for $\dim(\psi) > 1$.

Starting from (4), it is possible to derive a tail area approximation paralleling results in Reid (2003) and in Ventura and Racugno (2011). More precisely, accurate tail probabilities are directly computable by direct integration of (4). Let us consider the posterior tail area probability

$$\int_{-\infty}^{\psi_0} \tilde{\pi}_m(\psi|y) \, d\psi \doteq \int_{-\infty}^{\psi_0} c \tilde{j}_P(\tilde{\psi})^{1/2} \exp \left\{ \tilde{\ell}_P(\psi) - \tilde{\ell}_P(\tilde{\psi}) \right\} \frac{\pi(\psi, \tilde{\lambda}_{\psi})}{\pi(\tilde{\psi}, \tilde{\lambda}_{\psi})} \, d\psi. \tag{5}$$

Considering the change of variable in (5) from ψ to

$\tilde{r}_P = \tilde{r}_P(\psi) = \text{sign}(\tilde{\psi} - \psi) \sqrt{2(\tilde{\ell}_P(\tilde{\psi}) - \tilde{\ell}_P(\psi))}$ i.e. to the signed pseudo-likelihood root (see, e.g., also Ventura and Racugno, 2011), whose Jacobian is $d\tilde{r}_P(\psi)/d\psi = \tilde{\ell}'_P(\psi)/\tilde{r}_P(\psi)$, we obtain

$$\begin{aligned} \int_{-\infty}^{\psi_0} \tilde{\pi}_m(\psi|y) \, d\psi &\doteq \int_{-\infty}^{\tilde{r}_0} c \exp \left\{ -\frac{1}{2} \tilde{r}_P^2 \right\} \left(\frac{\tilde{r}_P}{\tilde{q}} \right) \, d\tilde{r}_P \\ &= \Phi(\tilde{r}_0) + \phi(\tilde{r}_0) \left(\frac{1}{\tilde{r}_0} - \frac{1}{\tilde{q}_0} \right) \\ &= \Phi \left(\tilde{r}_0 + \frac{1}{\tilde{r}_0} \log \frac{\tilde{q}_0}{\tilde{r}_0} \right) \\ &= \Phi(\tilde{r}_0^*), \end{aligned} \tag{6}$$

with $\tilde{r}_0 = \tilde{r}_P(\psi_0)$,

$$\tilde{r}_0^* = \tilde{r}^*(\psi_0) = \tilde{r}_P(\psi_0) + \frac{1}{\tilde{r}_P(\psi_0)} \log \frac{\tilde{q}(\psi_0)}{\tilde{r}_P(\psi_0)}, \tag{7}$$

and

$$\tilde{q}_0 = \tilde{q}(\psi_0) = -\tilde{\ell}'_P(\psi_0) |\tilde{j}_P(\tilde{\psi})|^{-1/2} \rho(\psi_0, \tilde{\psi})^{-1} \frac{\pi(\tilde{\psi}, \tilde{\lambda})}{\pi(\psi_0, \tilde{\lambda}_{\psi_0})}, \tag{8}$$

Remark that, from a practical point of view, this result may be used to compute the highest posterior density (HPD) credible set $H(b_\alpha) = \{\psi : \log \tilde{\pi}_m(\psi|y) \geq b_\alpha\}$ for ψ as $|\tilde{r}^*(\psi)| \leq z_{1-\alpha/2}$, where $z_{1-1/\alpha}$ is the $1 - 1/\alpha$ quantile of the standard normal distribution. Moreover, the value of ψ such that $0.5 = \Phi(\tilde{r}^*(\psi))$, i.e. the value of ψ such that $\tilde{r}^*(\psi) = 0$, gives the posterior median of $\tilde{\pi}_m(\psi|y)$.

3 Application to the Cox model

The Cox proportional hazards model (Cox, 1972) is widely used for survival data modelling. In its simplest form the failure times T_1, \dots, T_n for n independent individuals have survival functions

$$P(T_i > t) = \exp\{\xi_i H(t)\}, \quad i = 1, \dots, n, \quad (9)$$

where $\xi_i = x_i^T \beta$ is a non-negative function with parameters β , x_i is a vector of covariates for the i th individual, and $H(t)$ is the so-called baseline cumulative hazard function. Suppose that the data are n pairs (t_i, δ_i) , $i = 1, \dots, n$, where t_i denotes the observed lifetimes for the i th individual and δ_i is an indicator assuming value 1 if t_i is uncensored and 0 if it is censored. The partial likelihood function for β is given by (see e.g. Cox, 1972)

$$\tilde{L}(\beta) = \prod_{i=1}^m \frac{e^{x_i \beta}}{\sum_{j \in \mathcal{R}(t_{(i)})} e^{x_j \beta}}, \quad (10)$$

where $t_{(i)}$ denotes the ordered failure times and $\mathcal{R}(t_{(i)})$ is the set of the indexes of the individuals at risk in the instant $t_{(i)}$, that is, $\mathcal{R}(t_{(i)}) = \{(i), (i+1), \dots, (n)\}$ and $m = \sum_i \delta_i$.

Suppose it is of interest to focus on the scalar parameter β_j , the j th component of β . Let then $\beta = (\psi, \lambda)$, with $\psi = \beta_j$, the interest parameter and with $\lambda = \beta_{-j}$, i.e. β except β_j , the nuisance parameter.

For the application we consider a real dataset concerning a clinical study on malignant mesothelioma (MM) (Fassina *et. al*, 2012). This dataset reports survival times (**Surv**) for 109 individuals, with the censoring variable **Death** which takes value 1 if the individual died and zero otherwise. Some other covariates, like the gender (**Sex**), the type of malignant mesothelioma, i.e. type epithelioid, biphasic or sarcomatoid, and a set of genetical markers, are provided as well. The main goal here is to study the relationship between the survival time and type of MM using Cox regression. The Cox model for the hazard function in our case is

$$\lambda_T(t; x_i) = \lambda_0(t) \exp\{\beta_1 d_{B_i} + \beta_2 d_{S_i}\}, \quad (11)$$

where $\lambda_0(t)$ is the baseline hazard function, d_{B_i} is a dummy variable which takes value 1 if the MM is biphasic and zero otherwise and d_{S_i} is a dummy

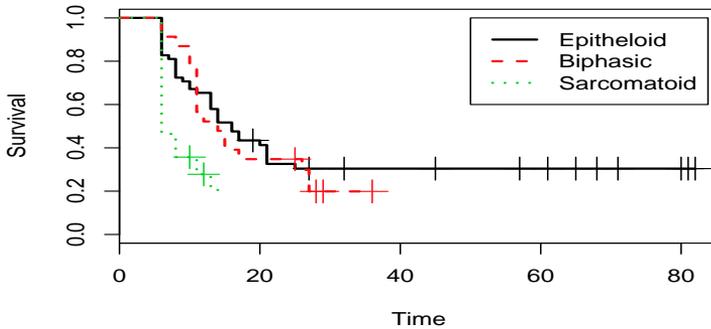


FIGURE 1. Kaplan-Meier curves for the three types of MM (+ indicates censored observations).

TABLE 1. Estimates for β_1 and β_2 from the Cox model (11) by the first-order Normal approximation, the \tilde{r}^* approximation and the profile partial likelihood.

Approximation		Median	95% HPD	Length
Normal (1 ^o order)	β_1	0.0990	(-0.4646, 0.6627)	1.1273
Normal (1 ^o order)	β_2	0.9795	(0.4138, 1.5451)	1.1313
\tilde{r}^* (3 ^o order)	β_1	0.0987	(-0.4919, 0.6331)	1.1251
\tilde{r}^* (3 ^o order)	β_2	0.9798	(0.4158, 1.4996)	1.0839
		MLE	95% CI	
$\tilde{L}_P(\beta_1)$	β_1	0.0988	(-0.4871, 0.6480)	1.1351
$\tilde{L}_P(\beta_2)$	β_2	0.9797	(0.4034, 1.5408)	1.1374

variable for sarcomatoid MM. From the Kaplan-Meier curves for the three types of MM we can state that individuals with sarcomatoid type of MM have a lower survival curve (see Fig. 1).

For β we assume an uninformative normal prior distribution with large variance, i.e. $\pi(\beta) \sim N_2(0, kI_2)$, with $k = 100$ and I_2 a (2×2) identity matrix. The parameter estimates along with the 95% HPD respective credible sets are given in Table 1. Only for comparison purposes we also report the median and the quantiles obtained from a first-order normal approximation of the posterior distribution. The table shows also the MLEs along with its 95% confidence interval (CI) based on the profile partial likelihood function. Note that, the credible sets computed with \tilde{r}^* have shorter length compared with those based on the first-order Normal approximation and on the profile partial likelihood.

4 Final remarks

Other examples of pseudo-likelihood functions, for instance, robust likelihood functions derived from M-estimators as well as composite likelihood functions are currently under study.

References

- Agostinelli, C. and Greco, L. (2012). A weighted strategy to handle likelihood uncertainty in Bayesian inference. *Comput. Statist.*, to appear.
- Brazzale, A.R., Davison, A.C. and Reid, N. (2007). *Applied asymptotics*. Cambridge: Cambridge University Press.
- Cox, R.D. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B*, **34**, pp. 187–220.
- Fassina, A., Cappellesso, O., Guzzardo, V., Dalla Via L. Piccolo, S., Ventura, L. and Fassan, M. (2012). Epithelial–mesenchymal transition in malignant mesothelioma. *Modern Pathology*, **25**, pp. 86–99.
- Greco, L., Racugno, W. and Ventura, L. (2008). Robust likelihood functions in Bayesian inference. *J. Statist. Plan. Inf.*, **138**, pp. 1258–1270.
- Lazar, N.A. (2003). Bayesian empirical likelihood. *Biometrika*, **90**, pp. 319–326.
- Lin, L. (2006). Quasi Bayesian likelihood. *Statist. Methodol.*, **90**, pp. 444–455.
- Pace, L. and Salvan, A. (1997). *Principles of statistical inference*. Singapore: World Scientific.
- Pauli, F. Racugno, W. and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statist. Sinica*, **21**, pp. 149–164.
- Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.*, **31**, pp. 1695–1731.
- Ribatet, M., Cooley, D. and Davison, A.C. (2012) Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statist. Sinica*, **22**, pp. 813–845.
- Smith, E.L. and Stephenson, A.G. (2009). An extended Gaussian max-stable process model for spatial extremes. *J. Statist. Plan. Inf.*, **139**, pp. 1266–1275.
- Ventura, L., Cabras, S., Racugno, W. (2009). Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *J. Amer. Statist. Assoc.*, **104**, pp. 768–774.
- Ventura, L. and Racugno, W. (2011). Recent advances on Bayesian inference for $P(X < Y)$. *Bayesian Analysis.*, **6**, pp. 411–428.

A penalized elliptical mixture partially nonlinear mixed effects model

Cibele M. Russo¹, Emmanuel Lesaffre^{2,3}, Gilberto A. Paula⁴

¹ Departamento de Matemática Aplicada e Estatística, ICMC, Universidade de São Paulo, Caixa Postal 668, CEP 13560-970, São Carlos, SP, Brazil, e-mail: cibele@icmc.usp.br

² Department of Biostatistics Erasmus Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands, e-mail: e.lesaffre@erasmusmc.nl

³ L-BioStat, Catholic University of Leuven, Kapucijnenvoer 35 blok d - box 7001 3000 Leuven, Belgium

⁴ Departamento de Estatística, IME, Universidade de São Paulo, Caixa Postal 66281 (Ag. Cidade de São Paulo), CEP 05311-970, São Paulo, SP, Brazil, e-mail: giapaula@ime.usp.br

E-mail for correspondence: `cibele@icmc.usp.br`

Abstract: Our proposal is to generalize a partially nonlinear mixed effects model assuming that the random effects follow a mixture of elliptical distributions. The most usual assumption in the literature is that the random effects and errors jointly follow a multivariate normal distribution. This may not be the most adequate choice when the random effects distribution is not unimodal and/or have heavier- or lighter- tails than the normal distribution. The mixture of elliptical distributions provides an interesting alternative to the Gaussian model, delivering a better fit to the data and more robust estimates against outliers.

Keywords: elliptical models; correlated data; nonlinear models; random effects

1 Introduction

Mixed effects models are a useful statistic tool to deal with correlated data, as longitudinal data or repeated measurements. It has been discussed in the literature that the normal distribution is not always the best choice for the random effects distribution when, e.g., their tails are heavier or lighter than the normal or when there is bimodality. The consequence of the non-normality of the random effects may be, for instance, distorted empirical Bayes estimates. A penalized Gaussian mixture linear mixed effects model was proposed in Ghidry et al. (2004) and a simulation study comparing different approaches for flexible random effects distributions was presented by Ghidry et al. (2010). Here we propose the use of a penalized elliptical mixture partially nonlinear mixed effects model, which may provide interesting alternatives to the normal random effects distribution.

2 The model

A partially nonlinear mixed effects models can be written as

$$\mathbf{Y}_i = \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) + Z_i \mathbf{b}_i + \boldsymbol{\epsilon}, \quad i = 1, \dots, n \quad (1)$$

where the $(n_i \times 1)$ vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$ is the response vector, $\boldsymbol{\beta}_i = (\beta_0, \beta_1, \dots, \beta_p)$ is the $([p + 1] \times 1)$ fixed-effects vector, the random effects $(d \times 1)$ vector is represented by $\mathbf{b}_i = (b_{i1}, \dots, b_{id})^T$, $\boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta})$ is an m_i -dimensional nonlinear function of $\boldsymbol{\beta}$, \mathbf{X}_i is a matrix of explanatory variable values. Inspired by Ghidey et al. (2004), the random effects may be represented by $\mathbf{b}_i = R\mathbf{s}_i$, such that $RR^T = D$ and \mathbf{s}_i is the standardized form of \mathbf{b}_i . It is commonly assumed that the joint distribution of $\boldsymbol{\beta}_i$ and $\boldsymbol{\epsilon}$ is multivariate normal. A more general assumption, considered for example by Russo et al. (2009), would be that the joint distribution of \mathbf{Y}_i and \mathbf{b}_i is given by

$$\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{b}_i \end{bmatrix} \sim \text{El}_{n_i+d} \left\{ \begin{bmatrix} \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} Z_i D Z_i^T + \sigma^2 \mathbf{I}_{n_i} & Z_i D \\ D Z_i^T & D \end{bmatrix} \right\}, \quad (2)$$

where $\boldsymbol{\Sigma}_i = Z_i D Z_i^T + \sigma^2 \mathbf{I}_{n_i}$, D and $Z_i D$ are proportional to the variance-covariance $\text{Var}(\mathbf{Y}_i)$, $\text{Var}(\mathbf{b}_i)$ and $\text{Cov}(\mathbf{Y}_i, \mathbf{b}_i)$, respectively, by nonnegative factors.

The notation $\mathbf{W} \sim \text{El}_m(\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W)$ indicates that an m -dimensional vector \mathbf{W} follows a multivariate elliptical distribution, with mean $\boldsymbol{\mu}_W \in \mathbb{R}^m$ and scale matrix $\boldsymbol{\Sigma}_W$ (positive definite), with probability density function given by $f(\mathbf{w}) = |\boldsymbol{\Sigma}_W|^{-\frac{1}{2}} g[(\mathbf{w} - \boldsymbol{\mu}_W)^T \boldsymbol{\Sigma}_W^{-1} (\mathbf{w} - \boldsymbol{\mu}_W)]$, where $g: \mathbb{R} \rightarrow [0, \infty)$ is known as density generating function, with $\int_0^\infty u^{\frac{m}{2}-1} g(u) du < \infty$. The elliptical class of distributions encompasses multivariate symmetric distributions, such as Student- t , power exponential, logistic, contaminated normal, the normal itself, among others. The use of this class has been recently discussed in the literature and includes the possibility of obtaining robust estimates against outlying observations and less sensitive estimates when the model is perturbed. Denoting by u the quantity $(\mathbf{w} - \boldsymbol{\mu}_W)^T \boldsymbol{\Sigma}_W^{-1} (\mathbf{w} - \boldsymbol{\mu}_W)$, the normal and Student- t distributions are obtained when $g(u) = (2\pi)^{-m/2} \exp(-\frac{u}{2})$ and $g(u) = q \left(1 + \frac{u}{\nu}\right)^{-(\nu+m)/2}$, respectively, with q a constant.

In a slightly more general but very useful way, we want to deal with the case that the distribution of the random effects \mathbf{b}_i is given by a mixture of elliptical distributions

$$\mathbf{b}_i \sim \sum_{j=1}^J \sum_{l=1}^L c_{jl} \text{El}_d(\boldsymbol{\mu}_{jl}, D_b),$$

where $c_{jl} = \exp(a_{jl}) / \sum_{k=1}^J \sum_{m=1}^L \exp(a_{km})$, and $\sum_{j=1}^J \sum_{l=1}^L c_{jl} = 1$, with $a = (a_{11}, \dots, a_{JL})^T$ are the smoothing parameters, $\boldsymbol{\mu}_{jl}$ is the mean vector

of \mathbf{b}_i and D_b is the scale matrix, usually assumed to be diagonal. Here we consider the bivariate case, that is, $\boldsymbol{\mu}_{jl} = (\mu_{1j}, \mu_{2l})^T$ and $D_b = \text{diag}(\tau_1, \tau_2)$. Since not all Gaussian distribution properties are extendable to elliptical distributions and to enable the extension of Ghidey et al. (2004) for the elliptical case, one possibility is to make an assumption on the joint distribution of \mathbf{Y}_i and \mathbf{b}_i ,

$$\begin{bmatrix} \mathbf{Y}_i \\ \mathbf{b}_i \end{bmatrix} \sim \sum_{j=1}^J \sum_{l=1}^L c_{jl} \text{El}_{n_i+d} \left\{ \begin{bmatrix} \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) \\ Z_i \boldsymbol{\mu}_{jl} \end{bmatrix}, \begin{bmatrix} Z_i D Z_i^T + \sigma^2 \mathbf{I}_{n_i} & Z_i D \\ D Z_i^T & D \end{bmatrix} \right\}, \tag{3}$$

which leads to the marginal model given by

$$\mathbf{Y}_i \sim \sum_{j=1}^J \sum_{l=1}^L c_{jl} \text{El}_{n_i}(\boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) + Z_i \boldsymbol{\mu}_{jl}, Z_i D Z_i^T + \sigma^2 \mathbf{I}_{n_i}), \quad i = 1, \dots, m.$$

The idea is to extend the modifications of Ghidey et al. (2004) on the Eilers and Marx (1996) approach, by replacing the B-spline base functions with elliptical densities. Thus the penalized log-likelihood function is given by

$$\ell(\boldsymbol{\theta}; \mathbf{Y} | \boldsymbol{\lambda}) = \sum_{i=1}^K \ell(\boldsymbol{\theta}; \mathbf{Y}_i) - \left[\frac{\lambda_1}{2} \sum_j \sum_l (\Delta_1^k a_{jl})^2 + \frac{\lambda_2}{2} \sum_j \sum_l (\Delta_2^k a_{jl})^2 \right]$$

where Δ_i^k , $i = 1, 2$ is a difference operator of order k for the i th dimension, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}, \mathbf{a})^T$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ is a vector of penalty parameters, one for each of the two dimensions. Following the arguments of Ghidey et al. (2004) we consider $k = 3$.

2.1 Estimating $\boldsymbol{\theta}$

The estimating process consists in maximising the penalized likelihood with respect to the parameters for a given $\boldsymbol{\lambda}$ using for example the Newton-Raphson algorithm. One possibility to choose the penalty coefficient $\boldsymbol{\lambda}$ is by considering an information criterion, for instance the AIC. As in Ghidey et al. (2004), one of a_{jl} is kept fixed (to 0) to ensure identifiability.

3 Numerical illustration: Dialyzer revisited

Russo et al. (2009) proposed an elliptical mixed effects model for a hemodialysis problem with correlated measurements. In this example, the response variable is the ultrafiltration rate (Y , in ml/hr), the explanatory variable is the transmembrane pressure (X , in mmHg). Russo et al. (2009) fitted the model

$$\mathbf{Y}_i = \boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) + Z_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad \text{with } Z_i = \left[\frac{\partial \boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_0}, \frac{\partial \boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_1} \right] \Big|_{\tilde{\boldsymbol{\beta}}}$$

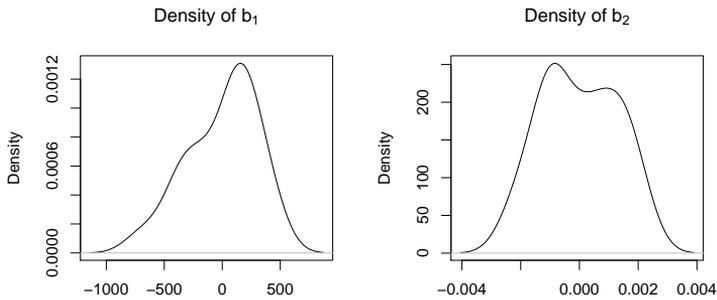


FIGURE 1. Estimated random effects density in Russo et al. (2009) approach.

$i = 1, \dots, n$, and $\tilde{\boldsymbol{\beta}}$ is the ordinary least squares estimate. The joint distribution of \mathbf{Y}_i and \mathbf{b}_i was assumed to be multivariate elliptical. For that model, the estimated random effects density show that the usual assumption may not be appropriate (see Figure 1).

Here $\mathbf{b}_i = (b_{1i}, b_{2i})^T$ is assumed to have a smooth bivariate distribution and the nonlinear function is written as $\boldsymbol{\eta}(\mathbf{X}_i, \boldsymbol{\beta}) = (\eta(X_{i1}, \boldsymbol{\beta}), \dots, \eta(X_{in_i}, \boldsymbol{\beta}))^T$, with $\eta(X_{ij}, \boldsymbol{\beta}) = \beta_0 \{1 - \exp[-\beta_1(X_{ij} - \beta_2)]\}$, for the j th measurement of the i th dialyzer. The parameter β_0 represents the maximum UFR one can attain due to protein polarization, β_1 is a hydraulic permeability transport rate, and β_2 is the transmembrane pressure required to offset patient oncotic pressure. In that experiment, the ultrafiltration rate was measured in seven different transmembrane pressure levels in 20 high flux membrane dialyzers.

The discussion about estimating or keeping fixed the degrees of freedom in the Student-t distribution recurs in the literature and it is known that the estimating approach may increase sensitivity to the model. For the Student-t model, ν was fixed at 5, but other choices could be considered without problems. Also, the values of J and L were fixed to 5. The estimated random effects density under the mixture of normal distributions are presented in Figure 2.

4 Concluding remarks and discussion

The most common assumption for the random effects in mixed models is the normality. Although often this assumption is appropriate, in some situations, especially in non-linear mixed models, the parameter estimates maybe sensitive to this assumption. One alternative is to allow the random effects to have more flexible distributions, for example a mixture of normal or elliptical distributions. Considering this approach, we propose a

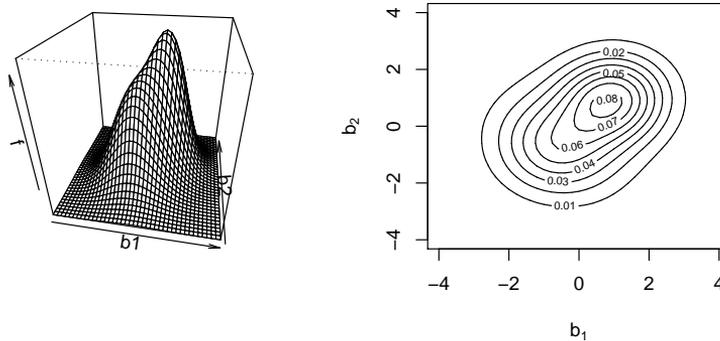


FIGURE 2. Estimated random effects density.

penalized elliptical mixture partially nonlinear mixed effects model, which is suitable for the hemodialysis problem. One limitation of our approach is that it assumes an elliptical distribution for the joint distribution of \mathbf{Y}_i and \mathbf{b}_i , the same elliptical distribution is assumed for the response variable and the random effects, but our limited results showed that this assumption may improve the classical model, providing a more adequate fit to the data.

Acknowledgments: The authors thank to Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil.

References

- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties *Statistical Science*. **11**, 89–121.
- Ghidey, W., Lesaffre, E. and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics* **60**, 945–953.
- Ghidey, W., Lesaffre, E. and Verbeke, G. (2010). A comparison of methods for estimating the random effects distribution of a linear mixed model. *Statistical Methods in Medical Research* **19**, 575–600.
- Russo, C. M., Paula, G. A. and Aoki, R. (2009). Influence diagnostics in nonlinear mixed-effects elliptical models. *Computational Statistics & Data Analysis* **11**, 4143–4156.

Methods to control for the concurvity in spatial ecological regressions (IneqCities project)

Marc Saez^{1,2}, Annibale Biggeri^{3,4}, Dolores Catelan^{3,4}, Maria Antònia Barceló^{1,2}, Laura Grissoto^{3,4}, Alberto Allepuz⁵

¹ Research Group on Statistics, Applied Economics and Health (GRECS), University of Girona, Spain

² CIBER of Epidemiology and Public Health (CIBERESP), Spain

³ Department of Statistics 'G. Parenti', University of Florence, Italy

⁴ Biostatistics Unit, ISPO Cancer Prevention and Research Institute, Florence, Italy

⁵ CRESA, Autonomous University of Barcelona, Spain

⁶ Department of Animal Health and Anatomy, Autonomous University of Barcelona, Spain

E-mail for correspondence: marc.saez@udg.edu

Abstract: Estimates of ecological regression coefficients may be biased when the ecological exposure is correlated with other factors which explain the spatial aggregation and a spatially autocorrelated model is specified. The problem is probably due to a near concurvity between the clustering term and the covariate. In order to assess the consequences and also possible solutions of the concurvity, we simulated a Poisson response using a neighbourhood matrix based on the Girona Metropolitan Area, Spain, by census tracts (9 municipalities and 85 census tracts). In particular, we considered three scenarios (high, moderate and low concurvity) and three sub-scenarios for each scenario (spatial dependence equal to heterogeneity, much greater and much lower). Then, we fitted four different models for the spatial dependence: intrinsic CAR; proper CAR; space varying regression; and a Gaussian random field representation of Markov (GMRF) constructed explicitly from a stochastic partial differential equation (SPDE) whose solution is a Gaussian field (GF) with covariance function Matérn.

In presence of concurvity, not only the variance of the estimators is inflated, but also the estimates of the parameters are biased. These effects increase as the higher the correlation between the explanatory variable and the spatial random effect (i.e. the concurvity), as well as the more this (spatial) effect in relation to the heterogeneity. However, when we standardized the explanatory variable, however, the estimates of the parameters were not biased and their variances reduced considerably. We show that, with the explanatory variable standardized or not, all four models considered (iCAR, pCAR, Space Varying Regression and SPDE) provided similar results. SPDE models were those with a better goodness-of-fit.

Keywords: concurvity, intrinsic and proper CAR, SPDE models

1 Introduction

The inclusion of the clustering term in ecological analysis has been considered as a way to adjust for unmeasured confounders that vary locally over the study region (Clayton *et al.*, 1993).

The problem is that estimates of ecological regression coefficients may be biased when the ecological exposure is correlated with other factors which explain the spatial aggregation and a spatially autocorrelated model is specified (Breslow and Clayton, 1993)

Catelan *et al.* (2009) show that the problem is probably due to a near concurvity between the clustering term and the covariate. In fact, they conclude that the parameter ρ , that controls the overall strength of the spatial dependence, could be thought of as a neg-penalty: when the penalty is low (i.e. ρ is near to 1), the spatially structured term tends to overfitting, capturing part of the covariate effect and leading to biased estimates of the covariate effect.

2 Methods

2.1 Simulation

We simulated the following response,

$$Y_i \sim \text{Poisson}(\mu_i \text{Pop}_i)$$

where Pop was the population of males in the census tract i using a neighbourhood matrix based on the Girona Metropolitan Area by census tracts (9 municipalities and 85 census tracts).

μ_I was

$$\log(\mu_i) = -5.1096 + 8.7128 \text{Pop65}m_i + \log(\text{Pop}_i) + \eta_i + S_i + \beta \text{index}_i$$

where $\text{Pop65}m$ was the percentage of males aged 65 or older in census tract i over all males in the census tract i ; $\beta = 1$; $S_i \sim N(0.1522, 0.0648)$; and $\text{index}_i \sim \text{Normal}$.

These figures are based on an estimation of a Besag, York and Mollié (BYM) model (Besag *et al.*, 1991; Mollié, 1996) with real data (incidence of prostate cancer in the Girona Metropolitan Area, 1993-2006).

We considered three scenarios (high, moderate and low concurvity) and three sub-scenarios for each scenario (spatial dependence equal to heterogeneity, spatial dependence much greater than heterogeneity and spatial dependence much lower than heterogeneity):

TABLE 1. Results of the simulation.

Linear effect	Scenarios	Sub-scenarios
$\sigma_\eta = \sigma_S$		Cor($S_i, index_i$)=0.9
		Cor($S_i, index_i$)=0.5
		Cor($S_i, index_i$)=0.3
$\sigma_\eta = \frac{1}{5}\sigma_S$		Cor($S_i, index_i$)=0.9
		Cor($S_i, index_i$)=0.5
		Cor($S_i, index_i$)=0.3
$\sigma_\eta = 5\sigma_S$		Cor($S_i, index_i$)=0.9
		Cor($S_i, index_i$)=0.5
		Cor($S_i, index_i$)=0.3

2.2 Models

Then, we fitted four different models for the spatial dependence:
 Intrinsic CAR (iCAR)

$$\begin{aligned}
 Y_i &\sim \text{Poisson}(\mu_i \text{Pop}_i) \\
 \log(\mu_i) &= \alpha + \gamma_1 \text{Pop4564}_i + \gamma_2 \text{Pop65m}_i + \log(\text{Pop}_i) + \eta_i + \\
 &\quad + S_i + \beta \text{index}_i \\
 S_i | S_j &\sim N\left(\frac{1}{n_i} \sum_{i \sim j} S_j, \frac{1}{n_i}\right) \quad \forall i \neq j
 \end{aligned}$$

Proper CAR (pCAR)

$$\begin{aligned}
 Y_i &\sim \text{Poisson}(\mu_i \text{Pop}_i) \\
 \log(\mu_i) &= \alpha + \gamma_1 \text{Pop4564}_i + \gamma_2 \text{Pop65m}_i + \log(\text{Pop}_i) + \eta_i + \\
 &\quad + S_i + \beta \text{index}_i \\
 S_i | S_j &\sim N\left(\frac{1}{d + n_i} \sum_{i \sim j} S_j, \frac{1}{\tau(d + n_i)}\right) \quad \tau = \frac{1}{\rho} \quad \forall i \neq j
 \end{aligned}$$

Space Varying Regression Model

$$\begin{aligned}
 Y_i &\sim \text{Poisson}(\mu_i \text{Pop}_i) \\
 \log(\mu_i) &= \alpha + \gamma_1 \text{Pop4564}_i + \gamma_2 \text{Pop65m}_i + \log(\text{Pop}_i) + \eta_i + S_i \text{index}_i \\
 S_i | S_j &\sim N\left(\frac{1}{n_i} \sum_{i \sim j} S_j, \frac{1}{n_i}\right) \quad \forall i \neq j
 \end{aligned}$$

SPDE

$$\begin{aligned}
 Y_i &\sim \text{Poisson}(\mu_i \text{Pop}_i) \\
 \log(\mu_i) &= \alpha + \gamma_1 \text{Pop4564}_i + \gamma_2 \text{Pop65m}_i + \log(\text{Pop}_i) + \eta_i + \\
 &\quad + s(x_coord, y_coord) + \beta \text{index}_i
 \end{aligned}$$

Following the recent work of Lindgren *et al.* (2011), we specify a Matérn structure (Stein, 1999) for the spatial dependence.

$$\text{corr}(d_{ii'}) = M(|i - i'|, r_i^2 \sigma_i^2, \rho_i, v_i) + (1 - r_i^2) \sigma_i^2 I(i = i')$$

In short, we use a Gaussian random field representation of Markov (GMRF) constructed explicitly from a stochastic partial differential equation (SPDE) whose solution is a Gaussian field (GF) with covariance function Matérn (Lindgren *et al.*, 2011). Instead of using a regular lattice, as was the standard practice, which would imply an estimate with a high computational cost and also very little efficiency (Lindgren *et al.*, 2011), we specify a Matérn spatial covariance structure in a triangulation (triangulation of Delaunay - Hjelle and Daehlen, 2006 -) of the Girona Metropolitan Area, with very little computational cost and, most importantly in our context, much more efficient.

3 Results

In presence of concurvity, not only the variance of the estimators is inflated, but also the estimates of the parameters are biased. These effects increase as the higher the correlation between the explanatory variable and the spatial random effect (i.e. the concurvity), as well as the more this (spatial) effect in relation to the heterogeneity (Tables 2 and 4).

However, when we standardized the explanatory variable, however, the estimates of the parameters were not biased and their variances reduced considerably (Tables 3 and 5).

We show that, with the explanatory variable standardized or not, all four models considered (iCAR, pCAR, Space Varying Regression and SPDE) provided similar results. SPDE models were those with a better goodness-of-fit.

Acknowledgments: This work was partially supported by the European Union, DG-SANCO, Second Programme of Community action in the field of Health (2008-2013), project A/101156 and for the FIS (Health Research Fund), Spanish Ministry of Science and Innovation, project FIS-08/0142.

TABLE 2. Linear effect of the fixed effect. Standard deviation of η = Standard deviation of S . Correlation between fixed effect and $S = 0.9$.

	ICAR	Proper CAR	Space Varying regresión	spde
Fixed effects mean (sd) median (2.5%,97.5%)	1.8136 (1.3380) 1.8143 (-0.8176, 4.4412)	1.8220 (1.2751) 1.8222 (-0.6831, 4.3265)	1.8257 (1.2063) 1.8255 (-0.5411, 4.1946)	1.7500 (1.2361) 1.7534 (-0.6901, 4.1710)
Random effects ¹ (sd)				
Spatial	0.1310 (0.0481)	0.1429 (0.0584)	0.2008 (0.1209)	
Unstructured	0.1218 (0.0400)	0.1221 (0.0394)	0.1251 (0.0494)	
Matérn				
Range (sd)				0.2764 (0.4366)
τ				1.4249
$\hat{\sigma}_w$				0.0210
τ proper CAR (2.5%,97.5%)		1.0573 (0.0905, 3.7615)		
DIC	482.5275	482.1563	481.5180	481.1769
-log(mean(cpo))	2.8409	2.8386	2.8357	2.8352
Effective number of parameters	18.9622	18.2990	15.4653	10.1082

¹Standard deviation of the random effects

TABLE 3. Linear effect of the fixed effect. Standard deviation of η = Standard deviation of S . Correlation between fixed effect and $S = 0.9$. Standardized.

	ICAR	Proper CAR	Space Varying regresión	spde
Fixed effects mean (sd) median (2.5%,97.5%)	1.0228 (0.0340) 1.0227 (0.9563, 1.0900)	1.0231 (0.0328) 1.0229 (0.9590,1.0878)	1.0220 (0.0697) 1.0220 (0.8822, 1.1614)	1.0234 (0.0309) 1.0232 (0.9630, 1.0844)
Random effects ¹ (sd)				
Spatial	0.1265 (0.0450)	0.1356 (0.0526)	0.1173 (0.0390)	
Unstructured	0.1235 (0.0344)	0.1238 (0.0344)	0.1041 (0.0350)	
Matérn				
Range (sd)				0.2791 (0.4820)
τ				1.5060
$\hat{\sigma}_w$				0.0199
τ proper CAR (2.5%,97.5%)		1.0526 (0.0881,3.7238)		
DIC	483.0446	482.7316	483.3604	481.5959
-log(mean(cpo))	2.8487	2.8465	2.8514	2.8407
Effective number of parameters	19.1698	18.5076	18.7011	9.9610

¹Standard deviation of the random effects

TABLE 4. Linear effect of the fixed effect. Standard deviation of $\eta = 5$ Standard deviation of S . Correlation between fixed effect and $S = 0.3$.

	ICAR	Proper CAR	Space Varying regresión	spde
Fixed effects mean (sd) Median (2.5%,97.5%)	1.0177 (0.2302) 1.0175 (0.5664, 1.4703)	1.0251 (0.2276) 1.0249 (0.5790, 1.4727)	1.0309 (0.2444) 1.0306 (0.5498, 1.5132)	1.0143 (0.2117) 1.0143 (0.5991, 1.4300)
Random effects ¹ (sd)				
Spatial	0.1241 (0.0426)	0.1360 (0.0523)	0.1988 (0.1187)	
Unstructured	0.4735 (0.1266)	0.4877 (0.1262)	0.4911 (0.1262)	
Matérn				
Range (sd)				0.2599 (0.3863)
τ				1.3203
$\hat{\sigma}_w$				0.0247
τ proper CAR (2.5%,97.5%)		1.0650 (0.0932, 3.7637)		
DIC	510.1949	509.8430	509.1537	508.6438
-log(mean(cpo))	3.0050	3.0030	2.9999	2.9979
Effective number of parameters	21.3413	20.6545	17.6926	11.2774

¹Standard deviation of the random effects

TABLE 5. Linear effect of the fixed effect. Standard deviation of $\eta = 5$ Standard deviation of S . Correlation between fixed effect and $S = 0.3$. Standardized.

	ICAR	Proper CAR	Space Varying regresión	spde
Fixed effects mean (sd) Median (2.5%,97.5%)	1.0026 (0.030) 1.0025 (0.9437, 1.0622)	1.0036 (0.0298) 1.0035 (0.9454, 1.0624)	1.0057 (0.0640) 1.0055 (0.8781, 1.1340)	1.0004 (0.0267) 1.0004 (0.9480, 1.0529)
Random effects ¹ (sd)				
Spatial	0.1163 (0.0371)	0.1266 (0.0449)	0.1081 (0.0332)	
Unstructured	0.4790 (0.1220)	0.4796 (0.1221)	0.4806 (0.1228)	
Matérn				
Range (sd)				0.2640 (0.4161)
τ				1.1137
$\hat{\sigma}_w$				0.0235
τ proper CAR (2.5%,97.5%)		1.0574 (0.0900,3.7052)		
DIC	511.4242	511.1443	511.8711	510.1313
-log(mean(cpo))	3.0198	3.0182	3.0244	3.0108
Effective number of parameters	21.3350	20.6732	20.3386	11.4185

¹Standard deviation of the random effects

References

- Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**:1-59.
- Breslow, N.E. and Clayton, D.J. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**:9-25.
- Catelan, D., Biggeri, A. and Lagazio, C. (2009). On the clustering term in ecological analysis: how to different prior specifications affect results? *Statistical Methods and Applications* **18**:49-61.
- Clayton, D.J., Bernardinelli, L. and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology* **22**:1193-1202.
- Hjelle, O. and Daehlen, M. (2006). *Triangulations and Applications*. Berlin: Springer.
- Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B* **73**(4):423-498 [Available in: http://www.rss.org.uk/uploadedfiles/userfiles/files/Lindgren_16_March_2011.pdf accessed on January, 28, 2012].
- Mollié, A. (1996). Bayesian mapping of disease. In Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds). *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall, pp. 359-379.
- Stein, M.L. (1999) *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

Analysis of residuals in quantile regression: An application to income data in Brazil

Bruno R. Santos ¹, Silvia N. Elian ¹

¹ Institute of Mathematics and Statistics - University of Sao Paulo, Brazil

E-mail for correspondence: bramos@ime.usp.br

Abstract: Analysis of residuals is a very important analysis usually performed in the classical regression diagnostics framework. In this paper, we propose a similar kind of analysis, but in quantile regression models. We make use of quantile residuals defined by Dunn and Smyth (1996) to verify the assumption of asymmetric Laplace distribution (Yu and Zhang, 2005) to the errors in a quantile regression model. To illustrate the method we used data from the National Household Sample Survey, performed in Brazil. We were able to visualize a better approximation of the asymmetric Laplace assumption only in the log-linear model fitted to describe income as a function of other variables.

Keywords: Analysis of Residuals; Quantile Residuals; Quantile Regression; Income; Equivariance Property.

1 Introduction

Analysis of residuals has been a common way of verifying some model assumptions in the classic regression analysis, usually the normal distribution assumption for the errors. In this article, we propose a similar method using the asymmetric Laplace distribution and a definition of quantile residuals to analyze residuals of a quantile regression model fit.

In Section 2, we give a brief summary of the asymmetric Laplace distribution used in this article, then in Section 3 we provide the main concepts of quantile regression models and in Section 4 we define the quantile residuals used in the analysis of residuals. We finish this paper with an application of this method in Section 5 and in Section 6, we give our last remarks on the subject.

2 Asymmetric Laplace Distribution

We shall consider throughout this text the definition of the asymmetric Laplace distribution (ALD) of Yu and Zhang (2005). In this way, we must consider that if $Y \sim \text{ALD}(\mu, \sigma, \tau)$, then its distribution function is given by

$$F(y; \mu, \sigma, \tau) = \begin{cases} \tau \exp\left(\frac{1-\tau}{\sigma}(y-\mu)\right), & \text{if } y \leq \mu, \\ 1 - (1-\tau) \exp\left(-\frac{\tau}{\sigma}(y-\mu)\right), & \text{if } y > \mu. \end{cases}$$

where $0 < \tau < 1$ is the skew parameter, $\sigma > 0$ is the scale parameter and $-\infty < \mu < \infty$ is the location parameter.

We will see in the next section the relation between this distribution and the quantile regression framework.

3 Quantile Regression

Since its definition by Koenker and Bassett (1978), quantile regression has been used in several kinds of studies (see, e.g., Yu et al., 2003) as an alternative to the least squares method. First, we should assume the following linear model to describe the relation between $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$

$$Y = \beta_0(\tau) + \beta_1(\tau)x_1 + \cdots + \beta_p(\tau)x_p + \epsilon,$$

where the τ th quantile of ϵ is zero.

Using the asymmetric Laplace distribution, we have that if $\epsilon \sim \text{ALD}(0, \sigma, \tau)$, so its τ th quantile is equal to zero, in agreement with the assumption of the model.

Beyond that, it is known that the quantile regression estimator, $\hat{\beta}(\tau)$, for the parameters of the model above is obtained by finding

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho_{\tau}(y_i - \mathbf{x}'_i \beta),$$

where $\rho_{\tau}(u) = u(\tau - I(u < 0))$.

Nevertheless, considering the asymmetric Laplace distribution for the errors, we have that $\hat{\mu}_i = x_i \hat{\beta}(\tau)$ is the maximum likelihood estimator (MLE) for the conditional location parameter, the τ th conditional quantile of Y , $Q_Y(\tau|x)$, since $\hat{\beta}(\tau)$ is the MLE for $\beta(\tau)$. Therefore, we have a consistent estimator for $Q_Y(\tau|x)$.

4 Quantile Residuals

Following the paper of Dunn and Smyth (1996), we will use the quantile residuals defined as

$$r_{q,i} = \Phi^{-1} \{F(y_i, \hat{\mu}_i, \hat{\sigma}, \tau)\},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal and F is as defined in Section 2. According to the authors, apart from sampling variability in the estimators of μ and σ , the $r_{q,i}$ are exactly normal,

implying that if μ and σ are consistently estimated, then the distribution of $r_{q,i}$ converges to the standard normal distribution. It is important to notice that the above definition is a special case of Cox and Snell (1968) “crude” residuals.

We argued in the last section that μ is consistently estimated by its MLE, with the quantile regression estimator. Using the same idea, it is easy to prove that the MLE for σ is

$$\hat{\sigma} = n^{-1} \sum_{i=1}^n \rho_{\tau}(y_i - \hat{y}_i).$$

Considering these results, we can use the quantile residuals of a quantile regression model to determine if the assumption of the asymmetric Laplace distribution is confirmed after the model fit. For this, we can analyze graphics such as a histogram or a QQ-plot of the quantile residuals.

5 Application

In this part of the article, we will consider data from the National Household Sample Survey, which took place in Brazil, in 2009, to model income as a function of other variables. This type of model is often studied with quantile regression (see Buchinsky, 1994 and Yu et al., 2005).

This survey is done every year by the Brazilian Institute of Geography and Statistics (IBGE). We limited our sample to people who earned at least one third of the minimum wage in 2009, who were between 18 and 80 years old and who worked at least 40 hours/week during the period of the survey. With this filter, we selected 122.727 people.

Our response variable of interest, Y_i , is the real gross monthly income. For the independent variables, we will consider gender, age, age squared, education and a dummy variable indicating whether the person is single or not. We will use the following linear model, and its respective log-linear formulation,

$$y_i = \beta_0(\tau) + \beta_1(\tau)G_i + \beta_2(\tau)A_i + \beta_3(\tau)A_i^2 + \beta_4(\tau)E_i + \beta_5(\tau)S_i + u_i,$$

where G_i is equal to 1 for men and 0 for women, A_i is the age in years, E_i is the years of schooling, and S_i is equal to 1 for single individuals and 0 otherwise. In both cases, we will assume $u_i \sim \text{ALD}(0, \sigma, \tau)$. For the sake of brevity, we will analyze the results only for $\tau = 0.5$, regarding the conditional median of Y .

We refer to Table 1 for the estimates of the fitted models, using the linear formulation and also the log-linear formulation.

Now, we must verify if the assumption of the asymmetric Laplace distribution of the errors is reasonable in each formulation. It is expected, and

TABLE 1. Estimates for the fitted models, $\tau = 0.5$

Variables	Linear Model	Log-linear Model
(Intercept)	-688.44	4.69
Gender	218.00	0.28
Age	27.81	0.04
Age ²	-0.16	-0.00
Education	68.96	0.09
Single	-102.26	-0.14

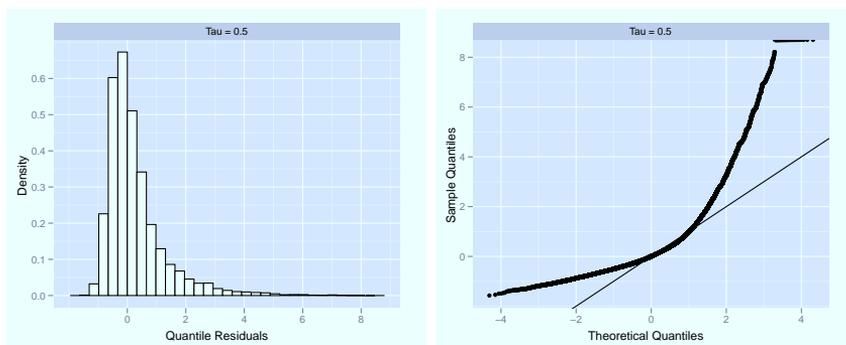


FIGURE 1. Histogram and QQ-plot for the quantile residuals in the linear model.

concluded with the analysis of Figure 1, that the errors, and consequently the response variable, are not distributed according to a symmetric Laplace distribution, which is the case when τ is equal to 0.5. It is well known that income has an asymmetric distribution, with greater concentration in lower incomes. In agreement with this idea, the quantile residuals demonstrate, in both graphics, a bigger concentration in lower values.

On the other hand, with the log transformation of the response variable, as we can visualize with the quantile residuals in Figure 2, we believe that a better approximation is achieved. Using the log-linear model, the quantile residuals show a symmetric behavior, as we can notice in the histogram. In the qq-plot, we can see a bigger difference between the theoretical and the sample quantiles in the left tail than it was expected. Despite this, we believe that the asymmetric Laplace distribution is a good approximation for the error distribution in this example.

Another consideration of these models, that we could address in this example, is the equivariance property that allows us to infer about the conditional median of income with the log-linear model (Koenker, 2005). In this way, we could give the following approach about inference with quantile regression considering this analysis of residuals. One could use monotone transformations until finding the proper asymmetric Laplace distribution

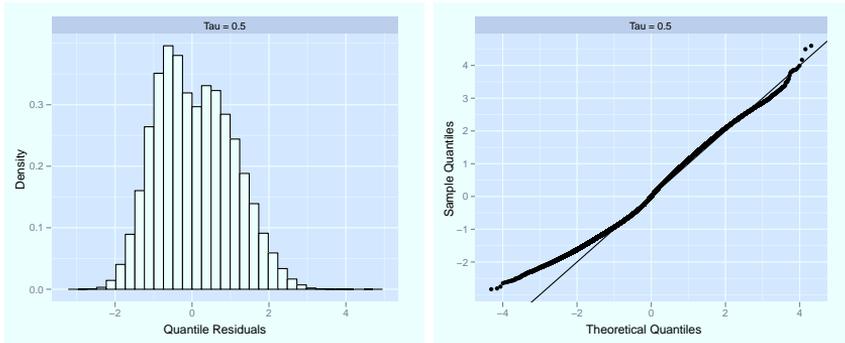


FIGURE 2. Histogram and QQ-plot for the quantile residuals in the log-linear model.

for the error distribution, when this is not obtained in the quantile regression model with the response variable not transformed. This is possible remembering that if $Q_y(\tau|x)$ is the τ th conditional quantile of Y given X and we use $h(\cdot)$, a monotone nondecreasing function, in Y , then by the equivariance property we have that

$$Q_{h(Y)}(\tau|x) = h(Q_Y(\tau|x)).$$

6 Concluding Remarks

Koenker and Machado (1999) and He and Zhu (2003) discuss evaluation methods of goodness of fit and lack-of-fit, respectively, in quantile regression models. We believe that our approach is connected with this kind of analysis, as we verify if τ th quantile of the errors is equal to zero using the asymmetric Laplace distribution, which is a very important assumption of these models. We found that, in the Brazilian example, this distribution was reasonable to explain the error in the log-linear model with income as function of other variables. We used the equivariance property of quantile regression models to note possible inferences about the conditional median using the log-linear model.

The same way as the least squares estimator is closely related to the MLE of regression models with normal distribution for the errors, we described a similar connection between the estimator of quantile regression and MLE with errors distributed according to an asymmetric Laplace distribution. For this reason, we think that the graph analysis proposed in this article could have the same importance as the residuals analysis usually performed in the classical regression analysis.

Although this was not dealt here, these residuals could also be used to verify other assumptions of quantile regression models, such as linearity

and homocedasticity, for example, when plotted together with the fitted values of the model.

Acknowledgments: Special thanks to CAPES for financial support of the first author, during which he was able to complete the Master's Program in Statistics at the University of Sao Paulo, when he studied the main subjects of this article.

References

- Buchinsky, M. (1994). Changes in US Wage Structure 1963-87: An Application of Quantile Regression. *Econometrica*, **62**, 405–458.
- Cox, D. and Snell, E. (1968). A General Definition of Residuals. *Journal of the Royal Statistical Society, Series B*, **30**, 248–275.
- Dunn, P. and Smyth, G. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.
- He, X. and Zhu, L. (2003). A Lack-of-Fit Test for Quantile Regression. *Journal of the American Statistical Association*, **98**, 1013–1022.
- Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33–50.
- Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press.
- Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33–50.
- Koenker, R. and Machado, J. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association*, **94**, 1296–1310.
- Yu, K., Lu, Z. and Stander, J. (2003). Quantile Regression: Application and Current Research Areas. *The Statistician*, **52**, 331–350.
- Yu, K., van Kerm, P. and Zhang, J. (2005). Bayesian Quantile Regression: An Application to the Wage Distribution in 1990s Britain. *Sankhyā - The Indian Journal of Statistics*, **67**, 359–377.
- Yu, K. and Zhang, J. (2005). A Three-Parameter Asymmetric Laplace Distribution and Its Extension. *Communications in Statistics - Theory and Methods*, **34**, 1867–1879.

Effect stars for categorical response models

Gunther Schaubberger¹, Gerhard Tutz¹

¹ Department of Statistics, LMU Munich, Germany

E-mail for correspondence: `gunther.schauberger@stat.uni-muenchen.de`

Abstract: Visualization tools are helpful in all areas of statistics. They can provide an easy and fast access to complex data structures, which is much harder to obtain by looking at numbers only. We propose a visualization tool for regression models with categorical responses focussing on the multinomial logit model. For models with many response categories listings of parameter estimates suffer from the large number of coefficients, which makes interpretation laborious. The proposed method, called effect stars, uses the tool of star plots to visualize the effects of one predictor in one star. The method is illustrated by an application to data from a German election survey.

Keywords: Star plots, parameter visualization, multinomial logit model, effect stars

1 Introduction

Categorical response models like the multinomial logit model or models for ordinal responses are very useful tools for the analysis of categorical data. They are well suited to uncover the relationships between a categorical dependent variable and various explanatory variables. Nevertheless, interpretation is not always intuitive and can be rather tedious. In particular for models with many response categories, the number of coefficients tends to be very large. Moreover, the coefficients have to be interpreted with respect to a non-linear link function, which makes interpretation even less intuitive. In general, it takes practioners some time to extract the relevant information from common software outputs.

Therefore, visualization methods might be very helpful to interpret the fitted models and to compare the influences of different predictors on the response. We propose the tool of effect stars, a modification of the well known tool of star plots, for that purpose. Star plots are a popular instrument for the descriptive analysis of multivariate data sets. Typically, each observation is represented by a star with as many rays as the number of covariates. The ray lengths refer to the values of the corresponding variables. For details on data visualization by star plots see, for example, Chen et al. (2008).

In effect stars, we visualize coefficients of categorical response models by using stars. Each variable, or, more precisely, each column in the design matrix, is represented by a star. A star is built by plotting one ray for every coefficient connected to the variable. The lengths of the stars refer to the exponentials of the estimated coefficients. Therefore, effect stars can visualize all the parameters that are linked to one term within one plot and are a helpful supplement to existing methods for visualizing effects.

Fox and Anderson (2006) and Fox and Hong (2009) proposed effect displays and extended their concept to categorical response models. Effect plots visualize the effects of single covariates by assuming typical values, for example the mean, for the remaining covariates. Especially for more complex models, for example if interactions or non-parametric terms are included, effect displays are an appealing concept. Nevertheless, it suffers from the dependence on the other covariates. This problem is solved by effect stars, which seem to be a good alternative, especially for the most common case of models with linear terms only.

Here we focus on the application of effect stars to the multinomial logit model. First, we shortly present the theory of the underlying model and the interpretation of the estimated coefficients. Then, effect stars are introduced and illustrated by a real data example. In the conclusion, we give a short outlook on possible extensions of the method.

2 Multinomial Logit Models

In the following, we shortly summarize the essential properties of the multinomial logit model, which is the most frequently used model in regression analysis for unordered categorical responses and is extensively treated, for example, in Agresti (2002) or Tutz (2012). For response $Y \in \{1, \dots, k\}$ and the vector of explanatory variables \mathbf{x} it has the form

$$P(Y = r|\mathbf{x}) = \frac{\exp(\beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r)}{\sum_{s=1}^k \exp(\beta_{s0} + \mathbf{x}^T \mathbf{b}_s)},$$

where $\boldsymbol{\beta}_r^T = (\beta_{r1}, \dots, \beta_{rp})$. Since parameters $\beta_{10}, \dots, \beta_{k0}$, $\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_k^T$ are not identifiable, additional constraints are needed. One option is to choose one of the response categories as reference category. For example, by setting $\beta_{k0} = 0$, $\boldsymbol{\beta}_k = \mathbf{0}$, category k is chosen as the reference category and interpretation of all parameters refers to this category. Alternatively one can use the symmetric side constraints $\sum_{s=1}^k \beta_{s0} = 0$, $\sum_{s=1}^k \boldsymbol{\beta}_s^T = (0, \dots, 0)$.

For the interpretation of the parameters it is essential to specify the identifiability constraint that is used. If k is chosen as the reference category one obtains

$$\log \left(\frac{P(Y = r|\mathbf{x})}{P(Y = k|\mathbf{x})} \right) = \beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r, \quad r = 1, \dots, k-1, \quad (1)$$

where the log-odds compare $P(Y = r|\mathbf{x})$ to the probability $P(Y = k|\mathbf{x})$. Then the parameters reflect the effect of predictors on the relation between category r and the reference category k . Symmetric side constraints are less often used although interpretation of parameters is easy. For symmetric side constraints, the interpretation refers to the "mean" response defined by the geometric mean

$$GM(\mathbf{x}) = \sqrt[k]{\prod_{s=1}^k P(Y = s|\mathbf{x})} = \left(\prod_{s=1}^k P(Y = s|\mathbf{x}) \right)^{1/k}.$$

It is easily derived that, in contrast to equation (1), now

$$\log \left(\frac{P(Y = r|\mathbf{x})}{GM(\mathbf{x})} \right) = \beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r, \quad r = 1, \dots, k, \tag{2}$$

holds. Therefore, $\boldsymbol{\beta}_r$ reflects the effects of \mathbf{x} on the logits when $P(Y = r|\mathbf{x})$ is compared to the geometric mean response $GM(\mathbf{x})$.

3 Effect Stars

In a multinomial model the parameters that are linked to the j th predictor are collected in $\boldsymbol{\beta}_r^T = (\beta_{r1}, \dots, \beta_{rp})$. In the star for predictor j the lengths of the rays are proportional to the exponentials of the components, that is, they are given by $\exp(\beta_{r1}), \dots, \exp(\beta_{rp})$. The exponentials of the coefficients are chosen because they are non-negative and have straightforward interpretation. The interpretation is seen from rewriting the model with symmetric side constraints from equation (2), as

$$\begin{aligned} \frac{P(Y = r|\mathbf{x})}{GM(\mathbf{x})} &= \exp(\beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r) \\ &= \exp(\beta_{r0}) \exp(x_1 \beta_{r1}) \dots \exp(x_p \beta_{rp}) \\ &= \exp(\beta_{r0}) \exp(\beta_{r1})^{x_1} \dots \exp(\beta_{rp})^{x_p}. \end{aligned}$$

From

$$\frac{P(Y = r|x_1, \dots, x_j + 1, \dots, x_p)/GM(x_1, \dots, x_j + 1, \dots, x_p)}{P(Y = r|x_1, \dots, x_j, \dots, x_p)/GM(x_1, \dots, x_j, \dots, x_p)} = \exp(\beta_{rj})$$

it is seen that $\exp(\beta_{rj})$ represents the multiplicative effect of variable j on the odds $P(Y = r|\mathbf{x})/GM(\mathbf{x})$ if x_j increases by one unit.

For illustration of effect stars for multinomial logit models, we use election data, which originate from the German Longitudinal Election Study. The response categories refer to the dominant parties in Germany, in particular, the Christian Democratic Union (CDU: 1), the Social Democratic

Party (SPD: 2), the Liberal Party (FDP: 3), the Green Party (4) and the Left Party (5). With the five response categories nine global predictors were collected, age (standardized), political interest (1: less interested, 0: very interested), religion (1: evangelical, 2: catholic, 3: otherwise), regional provenance (west; 1: former West Germany, 0: otherwise), gender (1: male, 0: female), union (1: member of a union, 0: otherwise), satisfaction with the functioning of democracy (1: not satisfied, 0: satisfied), unemployment (1: currently unemployed, 0: otherwise), and high school degree (1: yes, 0: no).

Figure 1 shows the effect stars for the multinomial logit model for the election data. The stars are labelled, and, in the case of categorical predictors, by the related predictor category. Every star comes with a circle. The margin of the circle represents the no-effects case where for all parameters corresponding to covariate j , $j \in \{0, \dots, p\}$, $\beta_{1j}, \dots, \beta_{kj} = 0$ holds. Therefore, as the exponentials of the coefficients are plotted, for the radii r_0, \dots, r_p of the circles $r_0, \dots, r_p = \exp(0) = 1$ holds. The different forms of the circles in Figure 1 result from scaling of the stars, which is chosen such that stars are of equal size. Nevertheless, the circles represent the no-effects case for each of the covariates. Values beyond the circle represent positive parameters for the corresponding category and thus represent increasing odds as the covariate is increased. In Figure 1, symmetric side constraints are used. This has to be kept in mind for the interpretation of the stars, especially concerning the no-effects circles.

In addition, we include several p -values. Below the covariates labels, we include p -values referring to likelihood ratio tests of the relevance of the variables. Except for the dummy variable for catholics and the binary variable gender, all the variables are significant referring to significance level of 0.05. Moreover, we include p -values for single coefficients. They refer to Wald tests of the null hypothesis $H_0 : \beta_{rj} = 0$ for $r \in \{1, \dots, k\}$ and $j \in \{0, \dots, p\}$.

Effect stars make the fitting of multinomial logit models easily accessible and interpretation of coefficients becomes much more intuitive. As an example, consider the star for the binary variable high school. The star refers to the category of people whose highest educational level is at least a high school degree, people with no high school degree are the reference group. The strongest positive effect results for the Green Party, indicating strongly increased odds to vote for the Greens for people with high school degree. This confirms the reputation of the Green Party to be a party preferred by the better educated. Also CDU possesses a positive effect for high school, whereas the odds to vote for SPD or the Left Party are decreased for voters with high school degree. Nevertheless, only the Greens and the Left Party have significant effects for this variable.

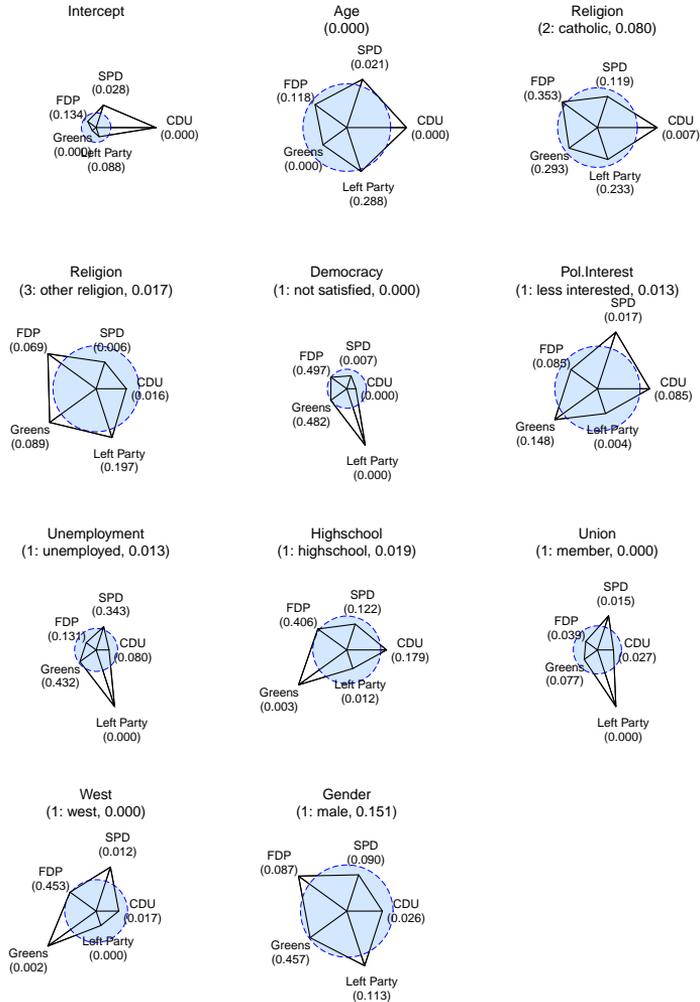


FIGURE 1. Effect Stars for multinomial logit model for election data

4 Concluding Remarks

There is a large repertory of methods to visualize categorical data. The popular tool of mosaic plots, as proposed by Friendly (1994), is especially suited to deal with contingency tables. For visualization of regression models for categorical data, we introduced the tool of effect stars and illustrated the

method by a real data example for the multinomial logit model. Effect stars can be helpful for practitioners to see which effects are relevant. The method can be extended to models for ordinal responses like the sequential logit model or the cumulative logit model. Moreover, additional graphical tools like reliability intervals can be included, see also Tutz and Schauburger (2012). An implementation of effect stars is available on CRAN in the R-package `EffectStars`.

References

- Agresti, A. (2002). *Categorical Data Analysis*. Wiley, New York.
- Chen, C. and Härdle, W. and Unwin, A. (2008). *Handbook of Data Visualization*. Springer Handbooks of Computational Statistics, Springer.
- Fox, J. and Anderson, R. (2006). Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology*, **36**(1), 225–255.
- Fox, J. and Hong, J. (2009). Effect displays in r for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, **32**(1), 1–24.
- Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, **89**, No. 425, 190–200.
- Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press, Cambridge.
- Tutz, G. and Schauburger, G. (2012). Visualization of Categorical Response Models - from Data Glyphs to Parameter Glyphs. *Department of Statistics, LMU Munich*, Technical Report **117**

Comparing the estimation of expectiles and quantiles towards efficiency

Linda Schulze Waltrup¹, Göran Kauermann¹, Fabian Sobotka², Thomas Kneib²

¹ Ludwig-Maximilians-Universität Munich, Germany

² Ernst-August-Universität Göttingen, Germany

E-mail for correspondence: lschulze_waltrup@stat.uni-muenchen.de

Abstract: Quantile regression is a useful tool to model other parts of a response than the ordinary mean. A way to generalize mean regression is by using expectile regression methods. This has the advantage, that the mean regression is a special case of expectile regression. Nevertheless, expectiles lack the intuitive expression which quantiles can offer. We show, how we can use the efficient expectiles to calculate quantiles to gain from both methods: the efficiency of expectiles and the interpretability of quantiles.

Keywords: expectiles; quantiles; regression; least asymmetrically weighted squares; asymmetric regression; loss function

1 Introduction

Modern regression methods offer ways to analyse data beyond mean regression. Quantile regression, for example, does not only model the (conditional) median of the response as a function of various explanatory variables, but also enables to estimate (conditional) quantiles, see Koenker and Bassett (1978). Hence in quantile regression besides the center of the data the tails are modelled as well.

In the following section we briefly describe sample quantiles and expectiles before switching to the regression case. Afterwards we see, how we can convert a set of expectiles into quantiles. Then we exploit the efficiency of the two kind of estimators by a simulation study. The final section contains an outlook on future research.

2 Concepts for Quantiles and Expectiles

In this section we take a closer look at sample quantiles and sample expectiles and at the end of the section we extend the methods to allow for regression. We start with the definition of quantiles: Let Y be a continuous

random variable with distribution function F_Y and density f_Y . Then q_τ is defined implicitly via

$$\tau = \mathbb{P}(Y \leq q_\tau) = F_Y(q_\tau) = \int_{-\infty}^{q_\tau} f_Y(t) dt. \quad (1)$$

For the regression scenario, which will be described later, it is useful to view the estimation of quantiles as an optimization problem. We can calculate quantiles by minimising

$$\sum_{i=1}^n r_\tau(y_i - q_\tau) \quad \text{with} \quad r_\tau(x) = \begin{cases} \tau x & \text{if } x > 0 \\ (1 - \tau)(-x) & \text{if } x \leq 0 \end{cases} \quad (2)$$

as it is described in Koenker and Hallock (2001). Here y_1, \dots, y_n denotes an i.i.d. sample from Y . Function r_τ is a special form of a loss function. For a more detailed description of various loss functions and resulting estimates see Breckling and Chambers (1988).

Similar to the generalization of the median by quantiles, expectiles are a generalization of the mean. A τ -expectile m_τ can be defined implicitly through

$$\tau = \frac{G(m_\tau) - m_\tau F(m_\tau)}{2(G(m_\tau) - m_\tau F(m_\tau)) + m_\tau - m_{0.5}} \quad (3)$$

where $G(x)$ is the partial moment function. Therefore we have $G(\infty) = m_{0.5}$ with $m_{0.5}$ the mean of the distribution and $F(x)$ the distribution function. As seen before with quantiles, the estimation of expectiles can be viewed as an optimization problem as well: We simply have to exchange the loss function in order to obtain expectiles instead of quantiles. We then can calculate expectiles by minimising

$$\sum_{i=1}^n \rho_\tau(y_i - m_\tau) \quad \text{with} \quad \rho_\tau(x) = \begin{cases} \tau x^2 & \text{if } x > 0 \\ (1 - \tau)x^2 & \text{if } x \leq 0 \end{cases}. \quad (4)$$

Function ρ_τ is another special form of a loss function. So instead of quantiles, which are yielded by weighted absolute values, expectiles result from weighted squared values.

It is now straightforward to consider the estimation of regression quantiles and expectiles. Therefore let $\mathbf{x} \in \mathbb{R}^{p+1}$ be a vector of observed covariates with $x_1 = 1$. We simply have to replace q_τ and m_τ by parametric functions $q_\tau(\mathbf{x}, \boldsymbol{\beta})$ and $m_\tau(\mathbf{x}, \boldsymbol{\beta})$ and solve the corresponding minimization problems for $\boldsymbol{\beta}$. In doing so, we obtain optimization problems

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^n r_\tau(y_i - q_\tau(\mathbf{x}, \boldsymbol{\beta})) \quad \text{and} \quad \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^n \rho_\tau(y_i - m_\tau(\mathbf{x}, \boldsymbol{\beta})) \quad (5)$$

which lead to regression quantiles and expectiles, respectively. Schnabel and Eilers (2009) describe a flexible approach of how to compute smooth

expectile curves using P-splines. Note that the second equation in (5) for τ equal to 0.5 leads to the well known least squares regression. In the following section we see, how we can use expectiles to calculate quantiles.

3 Connection between Expectiles and Quantiles

Note that there is a one to one bonding between a distribution function $F_Y(y) = P(Y \leq y)$ and random variable Y . Assuming $F_Y(y)$ to be differentiable this implies that the quantile function $q_\tau = F_Y^{-1}(\tau)$, for $0 < \tau < 1$, also uniquely defines the distribution of Y . Such uniqueness also applies to the expectile function m_τ implicitly defined through equation

$$m_\tau = \frac{(1 - \tau)G(m_\tau) + \tau(m_{0.5} - G(m_\tau))}{(1 - \tau)F(m_\tau) + \tau(1 - F(m_\tau))} \quad (6)$$

where F and G are defined as before.

If we could solve (6) with respect to $F(\cdot)$ we would be able to connect the expectile function m_τ to its corresponding quantile function q_τ . Apparently, an analytic solution looks infeasible but a numerical solution is available and implemented in our R-package `expectreg`.

4 Efficiency of Expectile and Quantile Estimation

The connection between quantiles and expectiles as described in section 3 allows us to compare the estimators by calculating quantiles from expectiles. In our simulations we restrict our attention to estimating population expectiles and quantiles, i.e. we assume a model of the form

$$y = \beta_0 + \epsilon. \quad (7)$$

We suppose, results can be transferred to the regression case. First simulations within a linear regression framework look promising.

Quantiles calculated from (a dense set of) expectiles are denoted by \hat{q}_E . Quantiles are estimated by using function `rq()` from R-package `quantreg` and are denoted by \hat{q}_Q . In the same way root mean square error (RMSE) and mean absolute error (MAE) are indexed by E and Q , respectively. Dependence on τ is suppressed.

We simulated data for different population sizes ($n = 199$ and $n = 499$, 10000 replications) and examined three different kinds of error distributions: $N(0, 1)$, $t(3)$ and $\chi^2(2)$. Selected quantiles were $\tau = 0.01, 0.02, 0.98, 0.99$ and a sequence from 0.05 to 0.95. Results are presented in figures 1 and 2. Figure 1 shows estimated Quantiles, relative RMSE ($\frac{\text{RMSE}_Q}{\text{RMSE}_E}$, relative MAE is defined correspondingly) can be found in 2. It should be mentioned that in two of 10000 replications for $t(3)$ and a sample size of $n = 499$ transformation of expectiles to quantiles failed due to huge outliers.

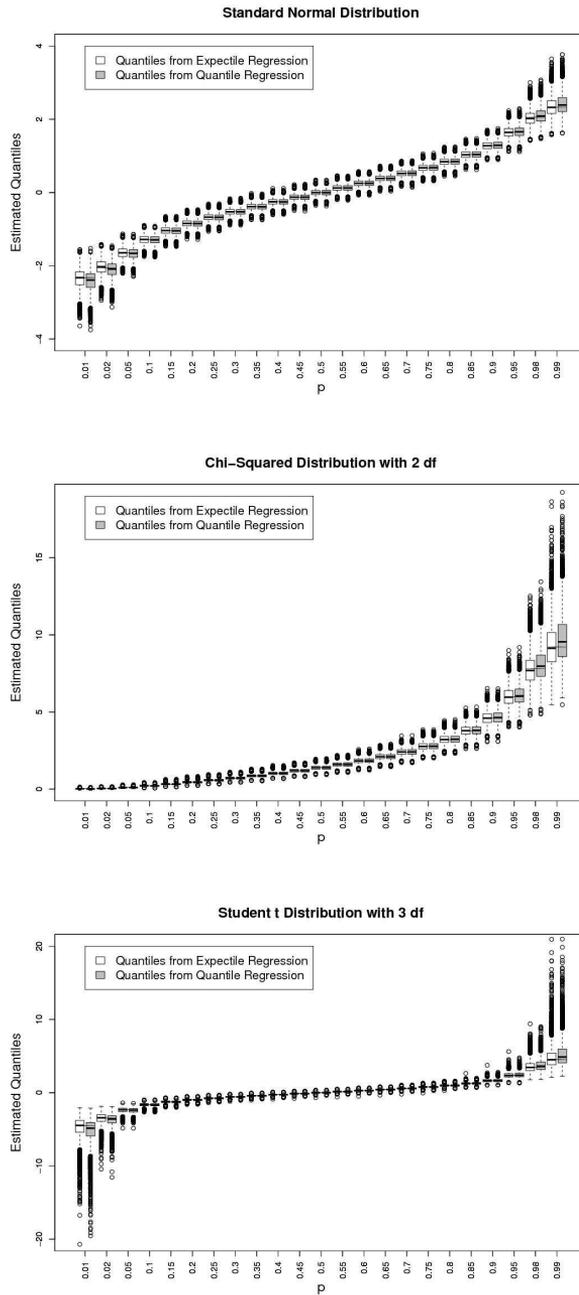


FIGURE 1. Estimated Quantiles for the two methods for $n = 199$. True quantiles are indicated by a thin bar. Results for $n = 499$ look similar.

The boxplots shown in figure 1 indicate that estimated quantiles from expectiles seem to meet true quantiles better than quantiles calculated from $\text{rq}()$. This can also be found in a low RMSE_E and MAE_E (seen figure 2). Note that the relative values are reported. The relative RMSE is given by the proportion $\frac{\text{RMSE}_Q}{\text{RMSE}_E}$ and the relative MAE is defined correspondingly as $\frac{\text{MAE}_Q}{\text{MAE}_E}$. As RMSE_E is the denominator, values greater than one indicate better performance of \hat{q}_E .

Figure 2 shows quite good results for \hat{q}_E over the whole range of quantiles for the two symmetric distributions, i.e. the normal and the student t distribution. Especially extreme quantiles are estimated with high accuracy. In case of the chi-squared distribution we have weaker performance of \hat{q}_E for low quantiles and a good performance of \hat{q}_E for high quantiles. This holds as well in terms of relative RMSE as in terms of relative MAE. As the results for the MAE look nearly the same, we did not include graphics. Thus all in all the simulations above indicate that expectiles may be favourable to quantiles, as the transformed quantiles \hat{q}_E in most of the cases have smaller RMSE and MAE as directly estimated quantiles \hat{q}_Q .

5 Conclusion

As we have seen in the last section, expectiles (at least for the three kinds of distributions we considered) seem to have desirable properties. We showed how one can use expectiles to calculate quantiles which leads to an improvement in terms of efficiency. In future research simulations will include more complex models and first results look promising.

References

- Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, **75**, pp. 761–771.
- Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**, pp. 819–847.
- Koenker, R (2011). quantreg: Quantile Regression. R package version 4.71. *Econometrica*, **46**, 33–50.
- Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33–50.
- Koenker, R. and Hallock, K. F. (2001). Quantile Regression. *Journal of Economic Perspectives*, **15**, 143–156.
- Schnabel, S. K. and Eilers, P. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, **53**, 4168–4177.
- Sobotka, F., Schnabel, S. and Schulze Waltrup, L. (2012). expectreg: Expectile and Quantile Regression. R package version 0.30.

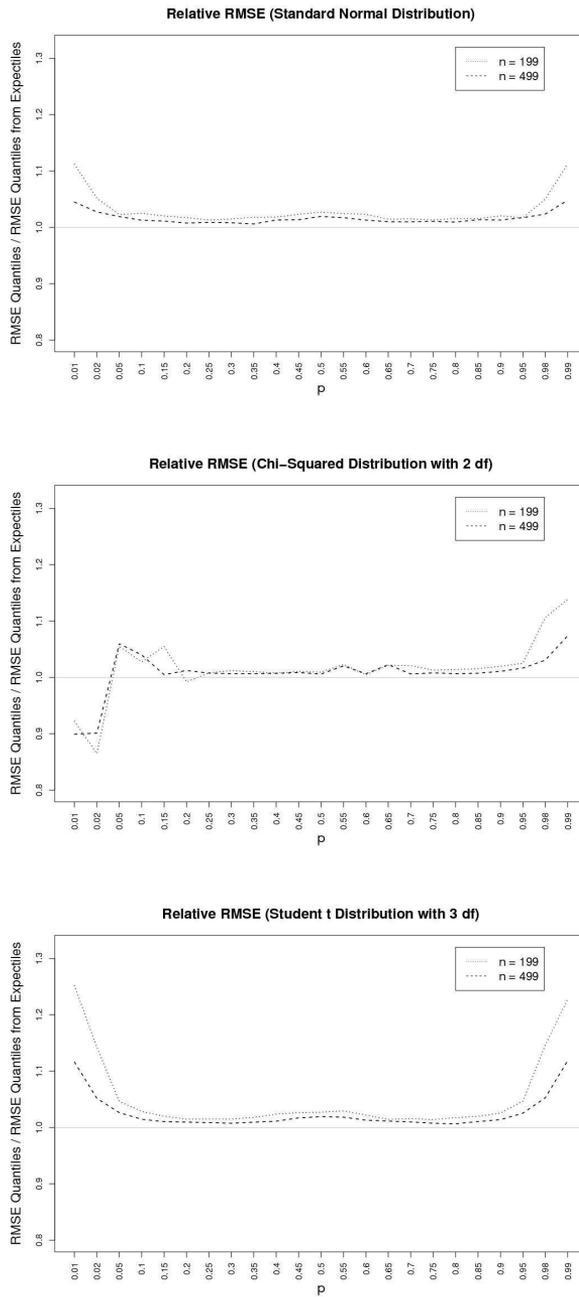


FIGURE 2. Relative root mean squared error (RMSE) for the two methods and sample sizes. Results for MAE look similar.

A likelihood ratio test for detection of single nucleotide polymorphisms (SNPs)

Ali Sheikhi ¹, David Ramsey ¹

¹ Centre of Biostatistics, Department of Mathematics & Statistics, University of Limerick, Limerick, Ireland

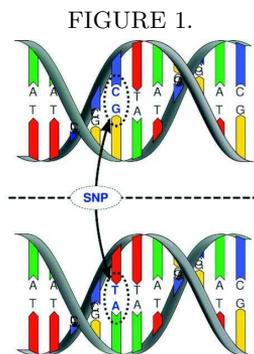
E-mail for correspondence: ali.sheikhi@ul.ie

Abstract: A single nucleotide polymorphism or SNP is a site of the genome where variation occurs within a population. Almost all SNPs have only two alleles (variants). In this work, we consider a statistical method based on a likelihood ratio test to detect these SNPs. We will also present some initial results of the analysis of real genome sequence data.

Keywords: single nucleotide polymorphisms, allele, genome sequencing.

1 Introduction

The biological information contained in a genome of an animal is encoded in its deoxyribonucleic acid (DNA). The genome of a diploid animal can be thought of as a sequence of pairs of nucleotides. There are four nucleotides, denoted by A, C, G and T. DNA contains two strands wrapped around each other in a helix, and these strands are held in place by nucleotides. A single nucleotide polymorphism or SNP is a DNA sequence variation occurring when a single nucleotide - A, T, C or G - in the genome differs within population members. (Figure 1).



In general, there are two variant nucleotides (alleles) at such sites. The least (most) common of these two alleles is called the minor allele (major allele, respectively).

Genome sequencing includes methods and technologies that are used for determining the order of these pairs of nucleotide (bases) along the sequence. Suppose a read of a site is made (a site is a specific position in the sequence of a given individual, i.e. corresponds to a nucleotide pair). It is assumed that one of the two nucleotides in the pair is chosen at random. A read should be understood as the inferred type of the nucleotide chosen (A, C, T or G). Note that multiple reads of a site are possible. An error is made when this inferred type does not correspond to the actual type. An estimate of the probability of error is assigned to each read made. At a large majority of sites, each individual in a population has two copies of the same nucleotide. It should be noted that the genome sequencer only reads the left hand side of the strand. Once the nucleotide on the left hand side is indicated, the other nucleotide on the right hand side is known (if the left is A, the right is T and if the left is C, the right is G and vice versa).

Initially we consider a particular site under a model where there are just two possible alleles. Assume we know from which individual any read comes from. Suppose there are n individuals. As the number of reads varies between different individuals, let m_i be the realisation of the number of reads for individual i , and $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m_i})$ be the vector of reads for individual i . As said before, the genome sequencer can read a base (nucleotide) incorrectly. Let $\hat{\mathbf{p}}_i = (\hat{p}_{i,1}, \hat{p}_{i,2}, \dots, \hat{p}_{i,m_i})$ be the vector of estimates of the probability of error. Note that these estimates of the probability of error are externally calculated (using the software installed in the genome sequencer). It is assumed that the number of reads for a site has a poisson distribution. It should be noted that the likelihood test defined below is based on the conditional likelihood given the number of reads for each individual. Thus, the number of reads takes the form of an ancillary statistics (see Lehmann, 1986, Chapter 10, Section 2).

The major allele at a site is assumed to be the allele with the largest total number of reads. When the minor allele is relatively rare, the major allele will be correctly determined with probability close to 1. On the other hand, when the minor allele is relatively common, any sensible test will detect it with a probability close to 1. Hence, for practical purposes we may ignore the possibility of calling the wrong allele as the major allele. We adapt the likelihood ratio test for detecting SNPs proposed by Futschik & Ramsey (2012).

2 The test

Consider a simplified model in which it is assumed that only two alleles can be read at a site. As said before, the one allele with the largest number of reads is defined to be the major allele. Let γ denote the relative frequency of the minor allele. We first wish to test the following hypothesis for each site,

H_0 : The site is not a SNP, i.e. $\gamma = 0$,

H_A : The site is a SNP, i.e. $\gamma > 0$.

We let $I_{i,j} = 1$ if the j -th read from individual i indicates the major allele and $I_{i,j} = 0$ otherwise. Now we can define $L_i(0)$, the likelihood of the sequence of reads under the null hypothesis, $\gamma = 0$, for individual i by,

$$L_i(0) = \prod_{j:I_{i,j}=1} (1 - \hat{p}_{i,j}) \prod_{j:I_{i,j}=0} \hat{p}_{i,j},$$

where $1 \leq j \leq m_i$. Note that under H_0 a read is an error if and only if it does not indicate the major allele.

Therefore, under H_0 the likelihood for the whole sample is,

$$L(0) = \prod_{i=1}^n L_i(0) = \prod_{i=1}^n \left(\prod_{j:I_{i,j}=1} (1 - \hat{p}_{i,j}) \prod_{j:I_{i,j}=0} \hat{p}_{i,j} \right).$$

If we only have reads of one nucleotide (variant) at a site, obviously we cannot reject H_0 .

Assume that there is a minor allele of relative frequency γ , where $\gamma > 0$. In this case, the probability of a read from an individual indicating a minor allele depends on the genotype of that individual (i.e. on the number of minor alleles that individual i has at the site considered, denoted A_i). Note that A_i has a binomial distribution with parameters 2 and γ . Assume that the genotype is given by MM (homozygote with 2 copies of major allele, $A_i=0$), or Mm (heterozygote with 1 copy of minor allele and 1 copy of major allele, $A_i=1$) or mm (homozygote with 2 copies of minor allele, $A_i=2$). For example, given the genotype is Mm, the minor allele is sampled with probability 0.5. The probability that the minor is sampled and it is read correctly is thus $0.5(1 - \hat{p}_{i,j})$, and the probability that the minor is sampled and it is read wrongly (as a major allele) is $0.5(\hat{p}_{i,j})$.

Let $q_j(a)$ be the probability that the j -th read indicates the prospective minor allele given that it comes from an individual with a minor alleles in their genotype. For small $\hat{p}_{i,j}$, we have $q_j(0) = \hat{p}_{i,j}$, $q_j(1) = 0.5$ and $q_j(2) = 1 - \hat{p}_{i,j}$. Using the law of total probability, the likelihood of the reads from individual i given the minor allele frequency is given by,

$$L_i(\gamma) = \sum_{a_i=0}^2 \left[\binom{2}{a_i} \gamma^{a_i} (1 - \gamma)^{2-a_i} \prod_{j:I_{i,j}=1} [1 - q_j(a_i)] \prod_{j:I_{i,j}=0} q_j(a_i) \right].$$

Multiplying the likelihoods for each individual we obtain,

$$L(\gamma) = \prod_{i=1}^n \left\{ \sum_{a_i=0}^2 \left[\binom{2}{a_i} \gamma^{a_i} (1 - \gamma)^{2-a_i} \prod_{j:I_{i,j}=1} [1 - q_j(a_i)] \prod_{j:I_{i,j}=0} q_j(a_i) \right] \right\}.$$

Let $S = \frac{\max_{0 < \gamma < 0.5} L(\gamma)}{L(0)}$. Since this maximisation is carried out numerically, we may assume that $\gamma \in \{0, \frac{1}{2n}, \frac{2}{2n}, \dots, \frac{n}{2n}\}$. Suppose the maximum is achieved at $\gamma = \hat{p} = \frac{l}{2n}$. The maximum likelihood estimate of the total number of minor alleles in the n genotypes is l . We now consider a method of testing whether a particular site is a SNP.

2.1 The likelihood ratio test

We use the likelihood ratio statistic $T = 2 \ln S$. Using standard asymptotic theory, this statistic will have approximately a chi-square distribution with one degree of freedom. A p -value for this test can thus be calculated under this assumption. However, the minor allele frequency under H_0 is at the border of the state space and so this approximation may well not be appropriate. Simulations indicate that the test described above is conservative and estimation of the critical value via Monte Carlo simulation does not significantly improve the power of the test.

It should be noted that this test is carried out for each site. Hence, we should employ a multiple testing procedure, e.g. the Benjamini-Hochberg (1995) procedure. According to this procedure, we arrange the p -values for each of the n sites in order from the smallest to the largest. Let $p_{(i)}$ denote the i -th smallest p -value. Let the nominal significance level be α . Suppose l is the largest value such that $p_{(l)} \leq \frac{\alpha l}{n}$. We accept that the l sites corresponding to the smallest p -values are SNPs and the remaining sites are not SNPs. Using such a procedure, the false discovery rate (FDR) (i.e. here the expected proportion of detected SNPs which are in fact not SNPs) will be $\leq \alpha$. It should be noted that the actual FDR depends on the true distribution of the number of reads.

2.2 Adaption of the test: four possible variants at a site

Consider the case in which all four possible alleles (variants) may occur. In this case, calculation of L_0 is straightforward because under H_0 (no minor allele) any variant other than the major allele is assumed to be an error. Therefore, once the major allele - the allele with the largest number of reads - is detected, we can calculate L_0 using the same procedure introduced in section 2.

When i non-major alleles are read at a site, then i likelihood ratio statistics may be calculated (one for each non-major allele) using an approach analogous to the one given above. The prospective minor allele is the one for which the greatest likelihood ratio is obtained and the test based on this ratio.

3 The results of the analysis of real genome sequence data

Statistical analysis were performed on a sample of 60000 sites of real genome sequence data. We used a multiple testing procedure (Benjamini-Hochberg) and calculated the p -values for each site. Based on these results we found that out of 60000 sites, 177 sites are SNPs. It means that the proportion of detected SNPs in this sample is 0.00295 or approximately 0.003.

Acknowledgments: Ali Sheikhi and David Ramsey are grateful for the support of Science Foundation Ireland under the BIO-SI project (no. 07MI012)

References

- Balding, D.J., Bishop, M., Cannings, C. (2001). *Handbook of Statistical Genetics*. John Wiley & Sons, LTD.
- Ramsey, D., Futschik, A. (2012). *DNA Pooling and Statistical Tests for the Detection of Single Nucleotide Polymorphisms*. (To appear in *Statistical Applications in Genetics and Molecular Biology*).
- Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Log-Beta modified Weibull regression models in survival analysis

Giovana O. Silva¹, Edwin M. M. Ortega², Gauss M. Cordeiro³

¹ Statistics Department, Federal University of Bahia, Brazil,

² ESALQ, São Paulo of University, Brazil

³ Statistics Department, Federal University of Pernambuco, Brazil

E-mail for correspondence: giovana@ufba.br

Abstract: In this paper, a regression model based in the beta modified Weibull distribution is proposed for modeling data in survival analysis. This model includes as special sub-models the log-modified Weibull, log-generalized modified Weibull and log-Exponentiated Weibull regression models. The parameters are estimated by maximum likelihood method. Besides, we used the sensitivity analysis to detect influential or outlying observations and residual analysis is used to check assumptions in the model such as departures from the error assumptions. The usefulness of the new distribution is illustrated by means of a real data set.

Keywords: beta modified Weibull distribution; censored data; regression models.

1 Introduction

The Weibull distribution, having exponential and Rayleigh as special cases, is a very popular distribution for modeling lifetime data. This standard distribution is suitable only in situations where the failure rate function is constant or monotone. However, this function may frequently present a bathtub-shaped or unimodal form. Many works had introduced new distributions based on modifications of the Weibull distribution to cope with bathtub shaped failure rate.

Numerous parametric regression models have been proposed for lifetime data based in these distribution. For example, log-exponentiated Weibull regression (Cancho, Bolfarine and Achar, 1999), log-generalized modified Weibull (Carrasco, Ortega and Cordeiro, 2009), log-modified Weibull, log-beta Weibull regression models (Ortega, Cordeiro and Hashimoto, 2011). In this paper, we propose a regression model using the beta modified Weibull distribution, referred to as the log-beta modified Weibull regression model, for survival times analysis. Here we note that the various regression models listed above can be embedded in this new regression model. Besides, this model is a feasible alternative for modeling the four existing types of failure rate functions.

We considered a classic analysis for the log-beta modified Weibull regression model. The inferential part was carried out using the asymptotic distribution of the maximum likelihood estimators. First, we give attention to discriminate such regression models through likelihood-ratio test, AIC and BIC. Besides we used several diagnostics measures considering three perturbation schemes in log-beta modified Weibull regression model with censored observations. We present residuals from a fitted model using the Martingale residual proposed by Barlow and Prentice (1988). Finally, the real data set is analyzed.

2 Log-beta modified Weibull regression models

In this paper, we proposed a regression model where the covariate vector, denoted by $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$, is related to the responses $Y = \log(T)$. Hence, it is important to study the distribution of the random variable Y which is denoted by log-beta modified Weibull (LBMW) distribution.

2.1 Regression models

Now, it is also considered that the parameter μ of the LBMW distribution depends on the matrix of explanatory variables X , this is, $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$. We also consider the regression model based on the log-beta modified Weibull relating the response Y and the covariate vector \mathbf{x} , so that the distribution $Y|\mathbf{x}$ can be represents as

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sigma z_i, \quad i = 1, \dots, n, \quad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ and $\sigma > 0$ are unknown parameters, $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ is the explanatory vector.

In this case, the survival function of $Y|\mathbf{x}$ is given by

$$S(y|\mathbf{x}) = 1 - I_{G(y)}(a, b) = 1 - \frac{1}{B(a, b)} \int_0^{G(y)} w^{a-1} (1-w)^{b-1} dw$$

where $G(y) = 1 - \exp \left\{ - \exp \left(\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} \right) \exp[\lambda \exp(y)] \right\}$

2.2 Special submodels

The log-beta modified Weibull (1) opens new possibilities for fitted many different types of data. It contains as special submodels the following well-known regression models and new regression models, for example, log-Weibull (LW) or extreme value regression model, log-modified Weibull regression model, log-generalized modified Weibull regression model, log-Exponentiated Weibull regression model and the new log-beta Weibull regression model.

2.3 Estimation by maximum likelihood

The values corresponding to the sample $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \dots, (y_n, \mathbf{x}_n)$ of n observations, where $y_i = \min [\log(T_i), \log(C_i)]$, T_i is the lifetime for the i -th individual, C_i is the censoring for the i -th individual, $i = 1, \dots, n$ and \mathbf{x}_i is the explanatory variables vector associated with the i -th individual. We assume no informative censoring and independence of the observed lifetime and censoring time. The log-likelihood function of the model given in (1) for parameter $\boldsymbol{\theta} = (a, b, \lambda, \sigma, \boldsymbol{\beta}^T)^T$ takes the form

$$l(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i \in F} l_1(\boldsymbol{\theta}, z_i) + \sum_{i \in C} l_2(\boldsymbol{\theta}, z_i), \tag{2}$$

where \mathcal{D} is the set of observed data, F denotes the set of uncensored observations; C denotes the set of censored observations, $z_i = (y_i - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma$ and

$$\begin{aligned} l_1(\boldsymbol{\theta}, z_i) &= z_i + \log\left(\frac{1}{\sigma} + u_i\right) + u_i - \log(B(a, b)) \\ &\quad + (a - 1) \log(1 - \exp(-\exp(z_i + u_i))) - b \exp(z_i + u_i) \end{aligned}$$

$$l_2(\boldsymbol{\theta}, z_i) = \log(1 - I_{(1 - \exp(-\exp(z_i + u_i)))})(a, b),$$

where $u_i = \lambda \exp(\mathbf{x}_i^T \boldsymbol{\beta} + \sigma z_i)$. Maximum likelihood estimates for parameter vector $\boldsymbol{\theta}$ can be obtained by maximizing the likelihood function. In this paper, we used the matrix programming language Ox (MaxBFGS subroutine) (see Doornik, 2007) to compute maximum likelihood estimates (MLE).

3 Sensitivity analysis

As a tool for sensitivity analysis the local influence method will now be described for log-beta modified Weibull regression model with censored data. Local influence calculation can be carried out in the model. If the likelihood displacement $LD(\boldsymbol{\omega}) = 2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\}$ is used, where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ denotes the MLE under the perturbed model, the normal curvature for $\boldsymbol{\theta}$ at the direction $\mathbf{d}, \|\mathbf{d}\| = 1$, is given by $C_{\mathbf{d}}(\boldsymbol{\theta}) = 2|\mathbf{d}^T \boldsymbol{\Delta}^T \ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1} \boldsymbol{\Delta} \mathbf{d}|$, where $\boldsymbol{\Delta}$ is a $(p + 4)n$ matrix that depends on the perturbation scheme and whose elements are given by $\Delta_{ji} = \partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial \theta_j \partial \omega_i, i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p + 2$ evaluated at $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega}_0$, where $\boldsymbol{\omega}_0$ is the no perturbation vector (see Cook, 1986).

4 Residual analysis

In order to study departures from the error assumption as well as presence of outliers, there are various residuals proposed in the literature (see Collett, 2003). We will consider the Martingale-type residual.

4.1 Martingale-type residual

The martingale residual was introduced in the counting process, and more details can be found in Fleming and Harrington (1991). Therneau et. al (1990) introduced the deviance residuals, which, for the Cox model with no time-dependent explanatory variables, can be described as

$$r_{D_i} = \text{sign}(r_{M_i}) \left[-2 \left\{ r_{M_i} + \delta_i \log(\delta_i - r_{M_i}) \right\} \right]^{\frac{1}{2}}, \quad (3)$$

where r_{M_i} is the martingale residual. In the log-beta modified Weibull regression model, the residual given in (3) is not a component of the deviance, but we will use it in order to have a transformation of the martingale residual.

5 Application

As an illustration, we consider the data from the Veterans Administration lung cancer trial given in Prentice (1973) and reported in Kalbfleisch and Prentice (2002). This data set considered males with advanced inoperable lung cancer that received chemotherapy. The survival time is the time from start of treatment. The main purpose of the study was to compare the effects of two chemotherapy treatments in prolonging survival time.

The explanatory variables involved in the study were: performance status at diagnosis (x_5), a measure of general fitness on a scale from 0 to 100, the age in years of the patient (x_6), the number of months from diagnosis of cancer (x_7) and priori therapy (x_8), 0: no priori therapy and 10: priori therapy. In addition, each patient was assigned one of two chemotherapy treatments (standard or test) and the tumors were classified into four types: large, adeno, small and squamous. The data contains $n = 137$ observations of which 9 were censored.

As did in Lawless (2003), we centered only the explanatory variables x_1, x_2 and x_3 and work with the following model

$$y_i = \beta_0 + \sum_{p=1}^4 \beta_p x_{i_p} + \beta_5(x_{i5} - \bar{x}_5) + \beta_6(x_{i6} - \bar{x}_6) + \beta_7(x_{i7} - \bar{x}_7) + \beta_8 x_{i8} + \sigma z_i, \quad i = 1, \dots, 137, \quad (4)$$

where y_i is the logarithm of the survival time t_i , the random error z_i and $x_{i1} = 0$ if treatment is test, 1 otherwise; $x_{i2} = 1$ if tumor type is squamous, 0 otherwise; $x_{i3} = 1$ if tumor type is small, 0 otherwise; $x_{i4} = 1$ if tumor type is adeno, 0 otherwise; $x_{i8} = 10$ if patient priori therapy, 0 otherwise. We fitted the log-beta modified Weibull and log-modified Weibull regression models and some submodels to these data. The required numerical evaluations were implemented by using an Ox program (sub-routine

MaxBFGS)(see, Doornik, 2007). Then, we select the best model based on the values of the statistics AIC and BIC and likelihood ratio test.

In addition, the sensitivity and residual analysis indicate that the log-beta modified Weibull Weibull regression model do not seem to be unsuitable to fit the data. In summary, we recommend using the LBMW regression model based on the analysis above.

Thus, the results suggest that X_3 , X_4 and X_5 are significant, and we interpret the estimated coefficients of the model as the following: the expected survival time should increase approximately 4% [$e^{0.0390} \times 100\%$] as the center *performance status* increases one unit, keeping the other variables fixed. In addition, the treatment does not appear to have sizeable effects, but the adeno tumor type is important. As the values of the β_3 and β_4 are negative, the patients whose tumor type is adeno or small present survival smaller probabilities than do the patients with large tumor types.

6 Concluding Remarks

In this paper, a log-beta modified Weibull regression model with the presence of censored data is proposed as an alternative to model lifetime in survival analysis. We used the Quasi-Newton algorithm to obtain the estimators of maximum likelihood and an asymptotic test was performed for the parameters based on the asymptotic distribution of the maximum likelihood estimators. The local influence theory was discussed in this study. So, we performed a general model by checking the analysis which makes this model a very attractive option for modelling censored and uncensored lifetime data. In the application within a real data set we showed good adjustment of the log-beta modified Weibull regression model through residual and sensitivity analysis. We hope this generalization may attract wider applications in survival analysis. In addition, the new model can be modified to cope with possible long-term survivors in data.

Acknowledgments: Special Thanks to Fundação de Amparo Pesquisa da Bahia - FAPESB.

References

- Cancho, V. G., Bolfarine, H., Achcar, J. A. (1999). A Bayesian analysis for the exponentiated-Weibull distribution. *Journal of Applied Statistical Science*, **8**, 227-242.
- Collet, D., 2003. *Modelling Survival data in medical research*. London: Chapman and Hall.
- Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, **48**, 133-169.

- Doornik, J., 2007. *Ox 5: Object-oriented matrix programming language*. London: Timberlake Consultants.
- Fleming, T. R. and Harrington, D. P. (1991). *Counting Process and Survival Analysis*. Wiley: New York.
- Kalbfleisch, J. D., Prentice, R. L. (2002). *The statistical analysis of failure time data*. New York: John Wiley.
- Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. New York: Wiley.
- Ortega, E. M. M., Cordeiro, G. M., Carrasco, J. M. F. (2011a). The log-generalized modified Weibull regression model. *Brazilian Journal of Probability and Statistics*, **25**, 64-89.
- Ortega, E. M. M., Cordeiro, G. M., Hashimoto, M. E. (2011). A log-linear regression model for the beta-Weibull Distribution. Submitted.
- Prentice, R. L. (1973). Exponential survival with censoring and explanatory variables. *Biometrika*, **60**, 279-288.
- Silva, G.O.; Ortega, E.M.M., Cordeiro, G.M. (2010). The Beta Modified Weibull Distribution. *Lifetime Data Analysis*, 16, 409-430.
- Therneau, T.M., Grambsch, P.M and Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147-60.

Factor analytic mixed models with inclusion of pedigrees in the analysis of plant breeding trials

Katia Stefanova¹

¹ University of Western Australia, Crawley WA 6009, Australia

E-mail for correspondence: katia.stefanova@uwa.edu.au

Abstract: Factor analytic mixed model are successfully used for the analysis of multi-environment trial data. A further step in the modelling process is the inclusion of pedigree information. This allows decomposition of the total genetic effects into additive and non-additive components, which provides essential feedback for selection of the best parents and the best test lines (varieties). These techniques are illustrated on data set of Lupin *Angustifolius* (narrowleaf lupin) breeding trials in Western Australia. Models in which the pedigree information was included demonstrated significantly better fit.

Keywords: Factor analytic models; Multi-environment trials; Pedigrees.

1 Introduction

Multiplicative mixed models, particularly factor analytic (FA) model, have been routinely used for the analysis of multi-environment trials (METs), Smith *et al* (2001). In this approach, the genetic variances are assumed independent and the parental covariances arising from common ancestors are ignored. Oakey *et al* (2007) suggested decomposition of the total genetic effects into additive, dominance and residual non-additive components in the context of factor analytic model applied to MET data. Recent papers by Beeck *et al* (2010), Cullis *et al* (2010) and Piepho *et al* (2008) discuss the inclusion of pedigree information for METs when analyzing traits as yield and oil content.

The aim of this paper is to present the techniques and models including pedigree information, and to illustrate it on MET yield data from an Australian lupin breeding program, while addressing the issues of selection of the best varieties and best parents using the pedigree information.

TABLE 1. Lupin trials summary.

Site	Columns	Rows	Yield Mean (kg/ha)	Varieties
08BAWH	6	28	1700	79
08BBWH	6	27	1486	75
08BCWH	6	32	1342	89
09BALV	6	42	1325	79
09BAVR	6	42	2719	79
09BAWH	6	42	1023	79

2 Motivating Example

2.1 Phenotypic Data

The MET data set included 6 lupin angustifolius breeding trials conducted at 3 environments for two sequent years. The trials were laid out as rectangular row-column arrays with 6 columns and varying number of rows. Table 1 presents a summary of the trials, including trial layout, total number of varieties and mean yield. The trials were designed using spatially optimized row column designs and generated by DiGGeR (Coombes, 2009). The variety concurrence between trials were high only for 09BALV, 09BAVR and 09BAWH. For the rest of the sites it varied between 5 and 42.

2.2 Pedigree Information

Pedigree information on total of 273 varieties was available and used in the METs analysis. Of those 13 varieties were not present in the trials. Also there were 11 varieties, 5 of which commercial, with unknown pedigrees and regarded as founders. The degree of inbreeding varied from 0 to nearly fully inbred for the majority of the varieties with inbreeding coefficient greater than 0.99.

3 Statistical Methods

Denote the vector of plot yields $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \dots, \mathbf{y}_t^T)^T$, t is the number of trials. The traditional model for \mathbf{y} is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\boldsymbol{\tau}$ is the vector of fixed effects with design matrix \mathbf{X} , \mathbf{u} is the vector of random effects with design matrix \mathbf{Z} and \mathbf{e} is the vector of plot errors. We extend the model by partitioning the vector of random effects to random genetic and random non-genetic effects. Furthermore, the vector of genetic effects is partitioned into additive and non-additive genetic effects.

The final model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_a + \mathbf{Z}_g\mathbf{u}_{\bar{a}} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e}$$

where $\mathbf{u}_g = \mathbf{u}_a + \mathbf{u}_{\bar{a}}$ and $\mathbf{u}_g, \mathbf{u}_a, \mathbf{u}_{\bar{a}}$ are respectively the vectors of genetic, additive genetic and non-additive genetic effects.

Note that the vector of genetic effects \mathbf{u}_g consists of two sub-vectors: vector of genetic effects for varieties in the pedigree present in the MET data set (\mathbf{u}_{go}) and vector of genetic effects for varieties in the pedigree not present in the MET data set (\mathbf{u}_{gp}), i.e. $\mathbf{u}_g = (\mathbf{u}_{gp}^T, \mathbf{u}_{go}^T)^T$.

Assume that \mathbf{u}_a and $\mathbf{u}_{\bar{a}}$ are independent and their joint distribution is Gaussian, with zero mean and variance matrices, $var(\mathbf{u}_a) = \mathbf{A} \otimes \mathbf{G}_{ea}$, $var(\mathbf{u}_{\bar{a}}) = \mathbf{I}_m \otimes \mathbf{G}_{e\bar{a}}$, and $\mathbf{G}_{es} = \Lambda_{es}^{t \times k_s} \Lambda_{es}^T + \Psi_{es}$, $\mathbf{s} = \mathbf{a}, \bar{\mathbf{a}}$.

The information on pedigree enters the model through the relationship matrix $\mathbf{A} = \{\mathbf{a}_{ij}\}$ given by $a_{ii} = 1 + F_i$ and $a_{ij} = 2f_{ij}$, where F_i is the inbreeding coefficient of entry i and f_{ij} is the coefficient of parentage between entries i and j . The matrix \mathbf{A}^{-1} and the models are computed using ASREML-R (Butler *et al*, 2009). For more details on matrix \mathbf{A} see Mrode & Thompson, 2005.

4 Model Selection

Model selection is a crucial step in the analysis of MET data sets. It includes selection of the model for the variance structure of the additive and non-additive effects and for the residuals. The latter assumes inclusion of terms to account for the randomization process (reflecting the experimental design) and terms to account for the spatial heterogeneity. The pedigree information is not included in the model till finalizing the spatial model and obtaining the total genetic effects. At this stage formal and informal diagnostics are used, e.g. sample variogram with simulation based coverage intervals (Stefanova *et al*, 2009 and Gilmour *et al*, 1997), REML log-likelihood ratio test. The next phase in the modelling process is partitioning of the the total genetic effects into additive and non-additive effects and selection of the final model.

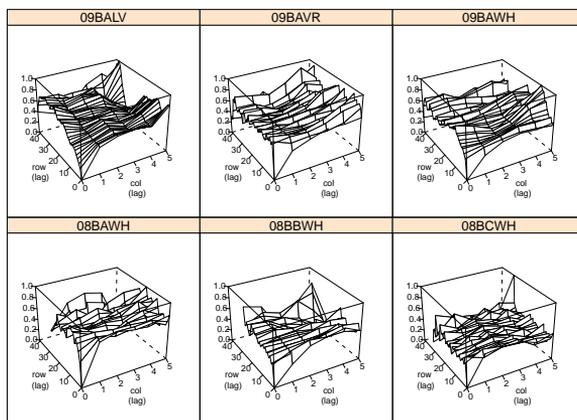
5 Results and Discussion

Yield data tend to exhibit spatial variability. Random column effects were present for all 2009 trials and random row effect for trial 08BCWH. Similarly, linear row effect was significant for site 09BAVR and column effect for sites 08BAWH and 09BALV. The local trend (AR1 for columns x AR1 for rows) reached 0.61 for columns and 0.35 for rows. There was some manifestation of aphids for all 2008 trials. A covariate for aphids was included in the models, consisting of scores on a scale 0-9, where 0 is used for no manifestation and 9 for the highest possible manifestation. Overview of the models fitted for each environment is presented in Table 2. Yield sample variograms for each trial after accounting for the spatial trends (see Figure 1) demonstrate good spatial adjustments for all sites.

TABLE 2. Models fitted for *yield*.

Site	Fixed effect (<i>p</i> value)†				Random effects†			AR1xAR1		Resid σ^2	
	lin(row)	lin(col)	hdir	cone	aphids	col	row	rep	ρ_r		ρ_c
08BAWH		.040			< .001			.000	.04	.33	.022
08BBWH					< .001			.000	.35	.24	.035
08BCWH					< .001		.008	.008	.03	.02	.033
09BALV		.044		< .001		.005		.000	.11	.61	.024
09BAVR	.003		.004			.028		.000	.23	.03	.033
09BAWH			.002			.060		.000	.31	.31	.027

† $\text{lin}(\text{row})$ and $\text{lin}(\text{col})$ represent the linear regression of yield on the row and column index, respectively, which is included in \mathbf{X} ; hdir and cone is included in \mathbf{X} , col and row represent factors based on the column and row indices and are included in \mathbf{Z} .

FIGURE 1. Variograms for *yield*.

Prior to the inclusion of pedigrees in the model, additional exploration of the data was done by grouping the pedigrees. Figure 2 shows yield level for each one of the 22 pedigree groups at each site. Pedigree groups 9, 17 and 18 show consistent pattern of high yield along all sites. The varieties from these groups were further followed up after the inclusion of the pedigree information in the model.

Table 3 presents the REML estimates of the additive and non-additive components of the genetic variance. The proportion of additive and non-additive variance depends on the trait and on the environment. The results here confirm that yield is less inheritable trait and is usually more affected by the environment in comparison to other traits.

Presented example reinforce the need for inclusion of pedigree information in the analysis of plant breeding trials. Models with the pedigree included

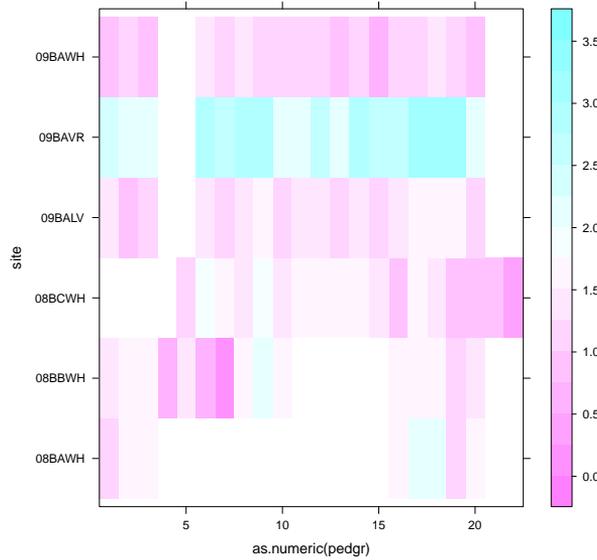


FIGURE 2. Yield levels for Pedigree Groups.

TABLE 3. REML Estimates of Additive and Non-Additive Genetic Variance. Site Yield

	Additive	Non-Additive
08BAWH	0.0000	0.9818
08BBWH	0.5612	0.0000
08BCWH	0.3563	0.0000
09BALV	0.3966	0.2568
09BAVR	1.5277	0.4068
09BAWH	0.4625	0.0817

provided significantly better fit in comparison to the models with pedigree excluded. Also partitioning of the genetic variance into additive and non-additive allows selection of potential parents, based on the additive component only.

The MET data set discussed here also illustrated the importance of complete pedigree information. The pedigree information included 273 varieties, with 11 entries, 5 of which commercial, with unknown pedigrees and regarded as founders. The inclusion of pedigrees for the commercial varieties is expected to improve the model and present better partitioning of the genetic variance.

Future work aims: (i) considering other traits, in particular protein; (ii)

getting more insight on yield and protein relationship: (iii) better understanding of the nature of the additive and non-additive genetic variance.

Acknowledgments: The author thanks Bevan Buirchell for providing the data sets and Brian Cullis for useful comments. The financial support of GRDC is gratefully acknowledged.

References

- Beeck, C.P., Cowling, W.A., Smith, A.B., and Cullis, B.R. (2010) Analysis of yield and oil from a series of canola breeding trials. Part I: Fitting factor analytic models with pedigree information. *Genome* **53**: 992-1001.
- Butler, D.G., Cullis, B.R., Gilmour, A.R., and Gogel, B.J. (2009) *ASReml-R Reference Manual. Release 3*. Technical Report. QDPI.
- Coombes, N.E. (2009). *DiGGeR, a spatial design program*. Biometrics Bulletin, NSW DPI.
- 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.
- Gilmour, A.R., Cullis, B.R. and Verbyla, A.P. (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* **2**: 269-273.
- Mrode, R.A., Thompson, R. (2005) *Linear Models for the Prediction of Animal Breeding Values*, 2nd Edition. CABI Publishing.
- Oakey, H.A., Verbyla, A., Cullis, B.R., Wei, X., Pitchford, W. (2007) Joint modelling of additive and non-additive genetic line effects in multi-environment trials. *Theoretical and Applied Genetics* **117**: 1319-1332.
- Piepho, H.P., Möring, J., Melchinger, A.E., and Büchse, A. (2008) BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* **161**: 209-228.
- Smith, A.B., Cullis, B.R., and Thompson, R. (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* **57**: 1138-1147.
- Stefanova, K.T., Smith, A.B. and Cullis, B.R. (2009) Enhanced Diagnostics for the Spatial Analysis of Field Trials. *Journal of Agricultural, Biological, and Environmental Statistics* **14**: 392-410.

Estimation of the conditional distribution of two censored gap times based on a nonparametric approach

Ewa Strzalkowska-Kominiak¹, Elisa M. Molanes-López²,
Emilio Letón³

¹ Department of Mathematics, Universidade da Coruña, A Coruña, 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: `strzalkowska@udc.es`

Abstract: In many survival studies, recurrent or consecutive events may occur during the follow-up study of the individuals. This situation can be found, for example, in transplant studies, where there are two consecutive events of interest and therefore two consecutive gap times subject to a common censoring time. In this work, we incorporate the information of covariates into the bivariate distribution of the two gap times of interest and propose two non-parametric methods to cope with it. These two methods will be referred to as Beran based estimator and Kaplan-Meier based estimator. We perform a simulation study to see the performance of both approaches and illustrate their use with a well-known real data example.

Keywords: bivariate distribution; copula function; covariates; serial dependence.

1 Introduction

In many biomedical studies, recurrent or consecutive events may occur during the follow-up study of the individuals. In this setting, it is of interest to study the time between consecutive events subject to a common censoring. In the literature, these consecutive times are known as gap times. Examples of this situation can be found in the recurrence of breast cancer, bleeding episodes in patients with liver cirrhosis or transplantation in heart studies (see, for example, Meira-Machado and Roca-Pardiñas, 2011).

In Section 2, we introduce the model and propose two estimators of the conditional distribution function of two consecutive times. Then, we check their behavior through a simulation study in Section 3. Finally, in Section 4 we illustrate the use of this two new approaches with a real data example.

2 Model description and estimation

Let T_1 and T_2 be two consecutive times subject to a common censoring variable C . Define $T = T_1 + T_2$, $\delta_1 = 1_{\{T_1 \leq C\}}$, $\delta_2 = 1_{\{T_2 \leq (C - T_1)1_{\{T_1 \leq C\}}\}}$ and $\delta = 1_{\{T \leq C\}}$. Denote by \tilde{T}_1 and \tilde{T}_2 the observed times, that is, $\tilde{T}_1 = \min(T_1, C)$, $\tilde{T}_2 = \min(T_2, (C - T_1)1_{\{T_1 \leq C\}})$ and define $\tilde{T} = \tilde{T}_1 + \tilde{T}_2$. The following three situations may occur:

- a) $T \leq C \Rightarrow T_1$ and T_2 are observed, that is, $\tilde{T}_i = T_i$, $\delta_i = 1$, for $i = 1, 2$.
- b) $T_1 \leq C < T \Rightarrow \tilde{T}_1 = T_1$, $\delta_1 = 1$, $\tilde{T}_2 = C - T_1$ and $\delta_2 = 0$.
- c) $C < T_1 \Rightarrow \tilde{T}_1 = C$, $\delta_1 = 0$, $\tilde{T}_2 = 0$ and $\delta_2 = 0$.

The joint cumulative distribution function (cdf) of the pair of consecutive survival data has been recently studied by Strzalkowska-Kominiak and Stute (2010) and de Uña-Álvarez and Amorim (2011), among others. In this work, we consider extra information given by a one-dimensional covariate, let say X . Our goal is to estimate the bivariate cdf of (T_1, T_2) given that $X = x$, that is,

$$\mathbf{F}(y_1, y_2|x) = \mathbb{P}(T_1 \leq y_1, T_2 \leq y_2|X = x), \tag{1}$$

under the assumptions that (T_1, T_2) is independent of C , (T_1, T_2) is independent of C given X and T_2 is independent of $(C - T_1)1_{\{T_1 \leq C\}}$ given T_1 and X . In the following subsections, we propose two estimators of (1) using two different approaches that we will refer to as Beran based estimator and Kaplan-Meier based estimator.

2.1 Beran based estimator

This estimator is based on the fact that

$$\mathbf{F}(y_1, y_2|x) = \int_0^{y_1} F_{21}(y_2|t_1, x)F_1(dt_1|x),$$

where

$$F_{21}(y_2|t_1, x) = \mathbb{P}(T_2 \leq y_2|T_1 = t_1, \delta_1 = 1, X = x)$$

and

$$F_1(y_1|x) = \mathbb{P}(T_1 \leq y_1|X = x).$$

Following the same ideas as in Van Keilegom (2004), we propose to estimate \mathbf{F} by

$$\mathbf{F}_n(y_1, y_2|x) = \int_0^{y_1} F_{21n}(y_2|t_1, x)F_{1n}(dt_1|x), \tag{2}$$

where $F_{1n}(t_1|x)$ is the Beran estimator of the conditional cdf of T_1 given $X = x$, and $F_{21n}(t_2|t_1, x)$ is the Beran estimator of the conditional cdf of T_2 given $(T_1, \delta_1, X) = (t_1, 1, x)$.

2.2 Kaplan-Meier based estimator

Following the same ideas as in de Uña-Álvarez and Meira-Machado (2008), we propose to estimate \mathbf{F} by

$$\tilde{\mathbf{F}}_n(y_1, y_2|x) = \frac{\frac{1}{h_1} \sum_{i=1}^n W_i^{KM} 1_{\{\tilde{T}_{1i} \leq y_1, \tilde{T}_{2i} \leq y_2\}} K\left(\frac{x-X_i}{h_1}\right)}{\frac{1}{h_1} \sum_{i=1}^n W_i^{KM} K\left(\frac{x-X_i}{h_1}\right)}, \tag{3}$$

where h_1 is a bandwidth parameter and

$$W_i^{KM} = \hat{F}_T(\tilde{T}_i) - \hat{F}_T(\tilde{T}_i -),$$

$$1 - \hat{F}_T(t) = \prod_{i=1}^n \left[1 - \frac{\delta_i}{\sum_{j=1}^n 1_{\{\tilde{T}_j \geq \tilde{T}_i\}}} \right]^{1_{\{\tilde{T}_i \leq t\}}}.$$

3 Simulation study

A simulation study is carried out here to check the performance of the two new estimators, previously proposed in equations (2) and (3). The scenarios that we consider are based on Huang et al. (2011). More specifically, we generate n i.i.d. observations $\{(T_{1i}^0, T_{2i}^0), i = 1, \dots, n\}$, from two correlated exponential random variates, (T_1^0, T_2^0) , with unit means, where $T_k^0 = -\ln(1 - \Phi(A + B_k))$, with Φ referring to the cdf of a standard normal variate and A and B_k denoting zero-mean normal random variates with variances ρ and $1 - \rho$, respectively, for $k = 1, 2$.

Since $(A + B_1, A + B_2)'$ follows a standardized bidimensional normal variate with correlation ρ , it is easy to prove that

$$\begin{aligned} \mathbf{F}^0(t_1, t_2) &= \mathbb{P}(T_1^0 \leq t_1, T_2^0 \leq t_2) \\ &= \mathbb{P}((A + B_1) \leq \Phi^{-1}(1 - e^{-t_1}), (A + B_2) \leq \Phi^{-1}(1 - e^{-t_2})) \\ &= \Phi_2(\Phi^{-1}(1 - e^{-t_1}), \Phi^{-1}(1 - e^{-t_2})), \end{aligned}$$

where Φ_2 denotes the cdf of $(A + B_1, A + B_2)'$. Taking into account Sklar's theorem, it is easy to see that $\mathbf{F}^0(t_1, t_2) = \mathbf{C}(F_1^0(t_1), F_2^0(t_2))$, where \mathbf{C} refers to the bivariate Gaussian copula with parameter ρ and F_k^0 denotes the cdf of T_k^0 , for $k = 1, 2$.

To incorporate the effect of the covariate X in the gap times (T_1, T_2) , we generate them as follows, $T_k = (T_k^0/a(X))^{1/\kappa}$, for $k = 1, 2$, where $a(X) = \exp(\beta X)$ and β refers to the parameters of the Cox model given by $h(t|x) = h_0(t) \exp(\beta x)$. It is easy to check that, conditionally on $X = x$, T_k , for $k = 1, 2$, is Weibull distributed random variate with conditional cdf $F_k(t|x) = 1 - e^{-(t/\lambda)^\kappa}$ for $t \geq 0$, where the scale parameter equals $\lambda = (1/a(x))^{1/\kappa}$ and the shape parameter equals κ . Therefore, it follows that $h(t|x) = a\kappa t^{\kappa-1}$, with $a = \exp(\beta x)$ and $h_0(t) = \kappa t^{\kappa-1}$.

In order to get a copula representation of the conditional cdf of (T_1, T_2) given $X = x$, we use the fact that

$$\begin{aligned} \mathbf{F}(t_1, t_2|x) &= \mathbb{P}(T_1^0 \leq a(X)t_1^\kappa, T_2^0 \leq a(X)t_2^\kappa | X = x) \\ &= \mathbf{F}^0(a(x)t_1^\kappa, a(x)t_2^\kappa) \\ &= \Phi_2(\Phi^{-1}(1 - e^{-a(x)t_1^\kappa}), \Phi^{-1}(1 - e^{-a(x)t_2^\kappa})) \\ &= \Phi_2(\Phi^{-1}(F_1(t_1|x)), \Phi^{-1}(F_2(t_2|x))), \end{aligned}$$

where $F_k(t|x)$ denotes the cdf of T_k given $X = x$, for $k = 1, 2$. Therefore, this procedure leads us to two Weibull variates, T_1 and T_2 , linked through the bivariate Gaussian copula. In order to get more flexibility, other copula functions could be used to link T_1 and T_2 . In that case, the conditional cdf of (T_1, T_2) given $X = x$ would be as follows $\mathbf{F}(t_1, t_2|x) = \mathbf{C}(F_1(t_1|x), F_2(t_2|x))$, where \mathbf{C} denotes a bivariate copula function. For the sake of brevity, we only consider the Gaussian copula with $\rho = 0.5$. Regarding the marginals, we consider $\kappa = 2$, $\beta = 0.3$ and $X \sim U(0, 10)$. Furthermore, $C \sim U(0, \tau_c)$, where we choose two values for τ_c , $\tau_c = 4.6$ and $\tau_c = 3.1$, such that the proportion of subjects with zero events is approximately 10% and 15%, that is, there are 10% n and 15% n subjects with $C_i < T_{1i} \Rightarrow \tilde{T}_{1i} = C_i, \delta_{1i} = 0, \tilde{T}_{2i} = 0$ and $\delta_{2i} = 0$. Additionally, for $\tau_c = 4.6$ and $\tau_c = 3.1$ we have, respectively, 20% and 30% events for which $\delta_{1i} = 1$ but $\delta_{2i} = 0$.

TABLE 1. Bias, variance and MSE in (t_1, t_2, x) when $\rho = 0.5$ and $n = 100$

(t_1, t_2, x)		$(0.46, 0.46, 5)$			$(1, 1, 5)$		
τ_c	Method	Bias	Variance	MSE	Bias	Variance	MSE
4.6	Beran	-0.0010	0.0064	0.0064	-0.0198	0.0009	0.0013
	KM	0.0137	0.0070	0.0072	-0.0180	0.0012	0.0015
3.1	Beran	-0.0138	0.0060	0.0062	-0.0229	0.0010	0.0015
	KM	0.0019	0.0074	0.0074	-0.0232	0.0016	0.0021

This simulation study is carried out in R and compares the two approaches given in the former section in terms of bias, variance and mean squared error (MSE). These estimators are computed in the middle point $(t_1, t_2, x) = (0.46, 0.46, 5)$, where $F(0.46, 0.46|5) \approx 0.45$, and in the right side of the support $(1, 1, 5)$, where $F(1, 1|5) \approx 0.98$, using 200 trials (see Table 1). As can be seen in Table 1, the Beran estimator gives better results than the Kaplan-Meier estimator. Nevertheless, the Beran estimator, because of its complexity, is much more time consuming. Moreover, both estimators have in the middle point smaller bias and larger variance than in the right side of the support.

Additionally, we estimate the conditional cdf of T_2 given $X = x$ based on Beran and Kaplan-Meier estimators by using $F_{2n}(t|x) = \mathbf{F}_n(\infty, t|x)$ and

$\tilde{F}_{2n}(t|x) = \tilde{\mathbf{F}}_n(\infty, t|x)$, respectively. In order to measure the performance of these two conditional estimators, we consider the Kolmogorov-Smirnov (KS) distance, that is,

$$KS^B(x) = \sup_t |F_{2n}(t|x) - F_2(t|x)|,$$

$$KS^{KM}(x) = \sup_t |\tilde{F}_{2n}(t|x) - F_2(t|x)|.$$

For $x = 5$ we obtain $KS^B(5) = 0.1456$ and $KS^{KM}(5) = 0.1473$, for $\tau_c = 4.6$, and $KS^B(5) = 0.1468$ and $KS^{KM}(5) = 0.1503$, for $\tau_c = 3.1$.

4 Example

In this section, we analyze the example of Stanford Heart Transplant data, described in Meira-Machado and Roca-Pardiñas (2011), among others. In this dataset, there are 103 individuals, for each one there are two times of interest, $T_1 =$ time from acceptance into the transplantation program to transplant (in days/30), and $T_2 =$ time from transplant to death (in days/30). Additionally, there are other variables (*delta* and *status*) that specify if the individual has received a transplant (or not), $\delta_1 = 1$ ($\delta_1 = 0$), and if he/she has died (or not), $\delta_2 = 1$ ($\delta_2 = 0$). Those individuals with both times observed correspond with those that have $\delta_1 = 1$ and $\delta_2 = 1$, in the case that the first time is observed but not the second time the individuals have $\delta_1 = 1$ and $\delta_2 = 0$, for the rest (with $\delta_1 = 0$ and $\delta_2 = 0, 1$) none of the two times are observed. Note that in our notation, $\delta_1 = \delta_1$ and $\delta_2 = \delta_1 \times \delta_2$. Besides, there is one covariate available, $X = \text{age}$.

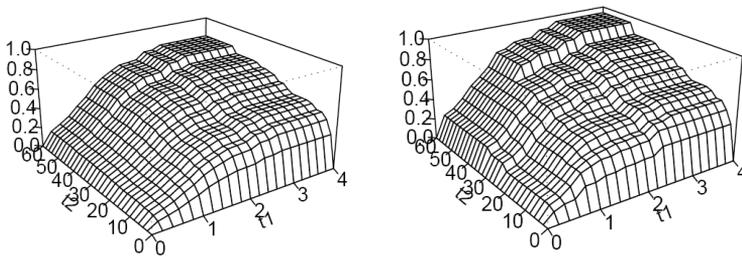


FIGURE 1. Estimated $\mathbf{F}(t_1, t_2|x)$ for $x = 55$ with the Beran method (left-hand panel) and the Kaplan-Meier method (right-hand panel)

A graphical representation of \mathbf{F}_n and $\tilde{\mathbf{F}}_n$ is given in Figure 1. In Table 2, we collect the estimated $\mathbf{F}(t_1, t_2|x)$ for a fixed value of x and $(t_1, t_2) \in (1, 2, 3, 4) \times (12, 24, 36, 48, 60)$.

TABLE 2. Estimated $\mathbf{F}(t_1, t_2|x)$ for $x = 55$

Method	$t_1 \backslash t_2$	12	24	36	48	60
	Beran	1	0.2837	0.3176	0.3776	0.4111
KM		0.3972	0.3972	0.5245	0.5245	0.5245
Beran	2	0.4705	0.5277	0.6247	0.6805	0.6805
KM		0.5013	0.5834	0.7108	0.8188	0.8188
Beran	3	0.5537	0.6215	0.7335	0.7994	0.7994
KM		0.5967	0.6788	0.8306	0.9386	0.9386
Beran	4	0.5537	0.6215	0.7335	0.7994	0.7994
KM		0.5967	0.6788	0.8306	0.9386	0.9386

Acknowledgments: This research has been supported by Grants from the Spanish Ministry of Science and Innovation. E. Strzalkowska-Kominiak acknowledges support from Juan de la Cierva scholarship. E.M. Molanes-López acknowledges support to MTM2010-09213-E, ECO2011-25706 and MTM2011-28285-C02-02. E. Letón acknowledges support to SEJ2007-64500, TIN2009-09158 and MTM2010-09213-E.

References

- de Uña-Álvarez, J. and Amorim, A.P. (2011) A semiparametric estimator of the bivariate distribution function for censored gap times. *Biometrical Journal* **53**, 113-127.
- de Uña-Álvarez, J., Meira-Machado, L.F. (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters* **78**, 2440-2445.
- Huang, C.-Y., Luo, X. and Follmann, D.A. (2011) A model checking method for the proportional hazards model with recurrent gap time data. *Biostatistics* **12**, 535-547.
- Meira-Machado, L. and Roca-Pardiñas, J. (2011). p3state.msm: Analyzing survival data from an illness-death model. *Journal of Statistical Software* **38**, Issue 3.
- Strzalkowska-Kominiak, E. and Stute, W. (2010) The statistical analysis of consecutive survival data under serial dependence. *Journal of Non-parametric Statistics* **22**, 585-597.
- Van Keilegom, I. (2004). A note on the nonparametric estimation of the bivariate distribution under dependent censoring. *Journal of Non-parametric Statistics* **16**, 659-670.

Robust estimation of large data sets by piecewise-linear interpolating estimator

Anatoli Torokhti¹, Stanley Miklavcic²

¹ School of Mathematics & Statistics, University of South Australia, Australia

² Phenomics and Bioinformatics Research Centre, School of Mathematics & Statistics, University of South Australia, Australia

E-mail for correspondence: anatoli.torokhti@unisa.edu.au

Abstract: Let K_Y and K_X be large sets of observed and reference random vectors, respectively, each containing N vectors. Is it possible to construct an estimator $\mathcal{F} : K_Y \rightarrow K_X$ that requires a priori information only on *few random vectors*, $p \ll N$, from K_X but *performs better* than the known estimators based on a priori information on *every* reference random vector from K_X ? It is shown that the positive answer is achievable under quite unrestrictive assumptions. The device behind the proposed method is based on a special extension of the piecewise linear interpolation technique to the case of random vector sets. The estimator is determined in terms of pseudo-inverse matrices so that it always exists.

Keywords: large data sets; random vector; estimation.

1 Introduction

A purpose of the proposed new estimating methodology is to provide an effective way to process large data sets. The device behind the proposed method is based on a special extension of the piecewise linear interpolation technique to the case of random signal sets.

1.1 Motivations

The problem under consideration is motivated by the following observations.

1.1.1 Estimating large sets of data; less initial information for better estimating. Suppose we need to transform a data set K_Y to another data set K_X . The sets consist of random vectors. A major difficulty and inconvenience common to many known estimating methodologies (see, for example, Chen *et al* (2006), Torokhti and Howlett (2007), Torokhti and Manton (2009), Torokhti and Miklavcic (2009), (2010), (2011)) is that they require a priori information on *each reference random vector* to be estimated. In particular, the known estimators are based on the use of either

the reference signal $\mathbf{x} \in K_x$ itself or its estimate. The Wiener estimating approach assumes that a covariance matrix formed from a reference data, $\mathbf{x} \in K_x$, and an observed data, $\mathbf{y} \in K_y$, is known or can be estimated. The latter can be done, for instance, from samples of \mathbf{x} and \mathbf{y} . In particular, this means that \mathbf{x} can be measured.

In the case of processing large data sets, such restrictions become much more inconvenient.

The major motivating question for this work is as follows. Let $\mathcal{F} : K_y \rightarrow K_x$ denote an estimator that estimates a large set of reference data, K_x , from a large set of observed data, K_y . Each set contains N data-vectors. Is it possible to construct an estimator \mathcal{F} that requires a priori information only on *few vectors*, $p \ll N$, from K_x but *performs better* than the known estimators based on a prior information on *every* reference data-vector from K_x ? We denote such an estimator by $\mathcal{F}^{(p-1)}$.

It is shown that the positive answer is achievable under quite unrestrictive assumptions. The required features of filter $\mathcal{F}^{(p-1)}$ are satisfied due to its special structure.

1.1.2 Estimating based on idea of piecewise function interpolation. The specific structure of the proposed filter follows from the extension of piecewise function interpolation. This is because the technique of piecewise function interpolation has significant advantages over the methods of linear and polynomial approximation used in known estimating techniques.

1.1.3 Exploiting pseudo-inverse matrices in the filter model. Most of the known estimating techniques are based on exploiting inverse matrices in their mathematical models. In the cases of grossly corrupted signals or erroneous measurements those inverse matrices may not exist and, thus, those filters cannot be applied.

The estimator proposed here avoids this drawback since its model is based on exploiting pseudo-inverse matrices. As a result, the proposed estimator always exist. That is, it processes any kind of noisy data.

1.1.4 Computational work. For the estimator $\mathcal{F}^{(p-1)}$ to be introduced below, the associated computational work is substantially less than that for the known estimators. This is because $\mathcal{F}^{(p-1)}$ requires the computation of only p pseudo-inverse matrices associated with p selected signals in K_x , where p is much less than the number of signals in K_x .

1.1.5 Difficulties associated with the known estimating techniques. Basic difficulties associated with applying the known estimating techniques to the case under consideration (i.e. to processing of large data sets, K_x and K_y) are that:

- (i) they require an information on *each* reference data-vector (in the form of a sample, for example),
- (ii) matrices used in the known estimators can be not invertible and then the estimator does not exist, and

(ii) the associated computation work may require a very long time.

1.1.6 Differences from the known estimating techniques. The differences from the known estimating techniques discussed above are as follows.

- (i) We consider a single estimator that processes *arbitrarily large* sets of stochastic data-vectors. The known estimators have been developed for the processing of an individual signal-vector only. In the case of their application to arbitrarily large signal sets, they imply difficulties described above.
- (ii) As a result, our piecewise linear estimator model, the statement of the problem and consequently, the device of its solution are different from the known ones.

1.1.7 Contribution. In general, for the processing of large data sets, the proposed estimator allows us to achieve better results in the comparison with the known techniques. In particular, it allows us to

- (i) achieve a desired accuracy in data-vector estimation (in practice, of course, the accuracy is increased to a prescribed reasonable level),
- (ii) exploit a priori information only on *few reference data-vectors*, p , from the set K_x that contains $N \gg p$ data-vectors or even infinite number of data-vectors,
- (iii) determine the estimator in terms of pseudo-inverse matrices so that the estimator always exists, and
- (iv) decrease the computational load compared to the known techniques.

2 Some preliminaries

2.1 Notation. The data sets we consider are, in fact, special representations of time series. Let (Ω, Σ, μ) be a probability space, and K_x and K_y be arbitrarily large data sets such that

$$K_x = \{\mathbf{x}(t, \cdot) \in L^2(\Omega, \mathbb{R}^m) \mid t \in T\} \text{ and } K_y = \{\mathbf{y}(t, \cdot) \in L^2(\Omega, \mathbb{R}^n) \mid t \in T\}$$

where $T := [a, b] \subseteq \mathbb{R}$. We interpret $\mathbf{x}(t, \cdot)$ as a reference data-vector and $\mathbf{y}(t, \cdot)$ as an observable data-vector, an input to the estimator \mathcal{F} studied below. The variable $t \in T \subseteq \mathbb{R}$ represents time. Then, for example, the random data-vector $\mathbf{x}(t, \cdot)$ can be interpreted as an arbitrary stationary time series.

Let $\{t_k\}_1^p \subset T$ be a sequence of fixed time-points such that $a = t_1 < \dots < t_p = b$. Because of this partition, the sets K_y and K_x are divided in ‘smaller’ subsets $K_{X,1}, \dots, K_{X,p-1}$ and $K_{Y,1}, \dots, K_{Y,p-1}$, respectively, so that, for each $j = 1, \dots, p$,

$$K_{X,j} = \{\mathbf{x}(t, \cdot) \mid t_j \leq t \leq t_{j+1}\} \quad \text{and} \quad K_{Y,j} = \{\mathbf{y}(t, \cdot) \mid t_j \leq t \leq t_{j+1}\}.$$

Therefore, K_y and K_x can now be represented as $K_x = \bigcup_{j=1}^{p-1} K_{X,j}$ and $K_y = \bigcup_{j=1}^{p-1} K_{Y,j}$.

2.2 Brief description of the problem. Given two *arbitrarily large* sets of random data-vectors, K_Y and K_X , find a single estimator $\mathcal{F} : K_Y \rightarrow K_X$ that estimates the data-vector $\mathbf{x} \in K_X$ with a controlled, associated error. Note that in our formulation the set K_Y can be finite or infinite.

2.3 Brief description of the method. The solution of the above problem is based on the representation of the proposed estimator in the form of a sum with $p - 1$ terms $\mathcal{F}_1, \dots, \mathcal{F}_{p-1}$ where each term, \mathcal{F}_j , is interpreted as a particular sub-estimator (see (1) below). Such an estimator is denoted by $\mathcal{F}^{(p-1)} : K_Y \rightarrow K_X$.

The sub-estimator \mathcal{F}_j transforms data-vectors that belong to ‘piece’ $K_{Y,j}$ of set K_Y to data-vectors in ‘piece’ $K_{X,j}$ of K_X , i.e. $\mathcal{F}_j : K_{Y,j} \rightarrow K_{X,j}$. Each sub-estimator \mathcal{F}_j depends on two parameters, α_j and \mathcal{B}_j .

The prime idea is to determine \mathcal{F}_j (i.e. α_j and \mathcal{B}_j) separately, for each $j = 1, \dots, p - 1$. The required α_j and \mathcal{B}_j follow from the solutions of the linear equation and an associated minimization problem (see below). This procedure adjusts \mathcal{F}_j so that the error associated with the estimation of $\mathbf{x}(t, \cdot) \in K_{X,j}$ is minimal.

A motivation for such a structure of the estimator $\mathcal{F}^{(p-1)}$ is as follows. The method of determining α_j and \mathcal{B}_j provides an estimate $\mathcal{F}_j[\mathbf{y}(t, \cdot)]$ that interpolates $\mathbf{x}(t, \cdot) \in K_{X,j}$ at $t = t_j$ and $t = t_{j+1}$. Due to this way of determining \mathcal{F}_j , it is natural to expect that the processing of a ‘smaller’ data-vector set, $K_{Y,j}$, may lead to a *smaller associated error* than that for the processing of the whole set K_Y by an estimator which is not specifically adjusted to each particular piece $K_{Y,j}$.

As a result, $\mathcal{F}^{(p-1)}[\mathbf{y}(t, \cdot)]$ represents a special *piecewise interpolation procedure* and, thus, should be attributed with the associated advantages such as, for example, the high accuracy of estimation.

3 Main results

3.1 Piecewise linear estimator model. Let $\mathcal{F}^{(p-1)} : K_Y \rightarrow K_X$ be an estimator such that, for each $t \in T$,

$$\mathcal{F}^{(p-1)}[\mathbf{y}(t, \cdot)] = \sum_{j=1}^{p-1} \delta_j \mathcal{F}_j[\mathbf{y}(t, \cdot)], \tag{1}$$

where $\mathcal{F}_j[\mathbf{y}(t, \cdot)] = \alpha_j + \mathcal{B}_j[\mathbf{y}(t, \cdot)]$ and $\delta_j = \begin{cases} 1, & \text{if } t_j \leq t \leq t_{j+1}, \\ 0, & \text{otherwise.} \end{cases}$

Here, \mathcal{F}_j is a sub-estimator defined for $t_j \leq t \leq t_{j+1}$, $\alpha_j = [\alpha_j^{(1)}, \dots, \alpha_j^{(m)}]^T \in \mathbb{R}^m$ and $\mathcal{B}_j : L^2(\Omega, \mathbb{R}^n) \rightarrow L^2(\Omega, \mathbb{R}^m)$ is a linear operator given by a matrix $B_j \in \mathbb{R}^{m \times n}$, so that $[\mathcal{B}_j(\mathbf{y})](t, \omega) = B_j[\mathbf{y}(t, \omega)]$. Thus, \mathcal{F}_j is defined by a matrix $F_j \in \mathbb{R}^{m \times n}$ such that

$$F_j[\mathbf{y}(t, \omega)] = \alpha_j + B_j[\mathbf{y}(t, \omega)]. \tag{2}$$

Estimator $\mathcal{F}^{(p-1)}$ defined by (1)–(2) is called the *piecewise* estimator. Parameters of the estimator $\mathcal{F}^{(p-1)}$, i.e. vector α_j and matrix B_j , for $j = 1, \dots, p - 1$, are unknown. Therefore, the problem is to determine α_j and B_j , for $j = 1, \dots, p - 1$.

Interpolation conditions that lead to a determination of α_j and B_j are as follows.

3.2 Interpolation conditions. Let us denote $\|\mathbf{x}(t_j, \cdot)\|_\Omega^2 = \int_\Omega \|\mathbf{x}(t_j, \omega)\|_2^2 \times d\mu(\omega)$ where $\|\mathbf{x}(t_j, \omega)\|_2$ is the Euclidean norm of $\mathbf{x}(t_j, \omega) \in \mathbb{R}^m$. Let $\widehat{\mathbf{x}}(t_{k-1}, \cdot)$ be an estimate of $\mathbf{x}(t_{k-1}, \cdot)$ defined by the preceding steps as

$$\widehat{\mathbf{x}}(t_{k-1}, \cdot) = \mathcal{F}_{k-2}[\mathbf{y}(t_{k-1}, \cdot)]. \tag{3}$$

Then sub-estimator \mathcal{F}_{k-1} is defined as follows. For $j = k - 1$, α_{k-1} and B_{k-1} solve

$$\widehat{\mathbf{x}}(t_{k-1}, \cdot) = \alpha_{k-1} + B_{k-1}[\mathbf{y}(t_{k-1}, \cdot)], \quad \min_{B_{k-1}} \|\alpha_{k-1} + B_{k-1}[\mathbf{y}(t_k, \cdot)] - \widehat{\mathbf{x}}(t_{k-1}, \cdot)\|_\Omega^2,$$

respectively. Then an estimate of $\mathbf{x}(t, \cdot)$, $\widehat{\mathbf{x}}(t, \cdot)$, for $t \in [t_{k-1}, t_k]$, is determined as

$$\widehat{\mathbf{x}}(t, \cdot) = \mathcal{F}_{k-1}[\mathbf{y}(t, \cdot)] = \widehat{\mathbf{x}}(t_{k-1}, \cdot) + B_1[\mathbf{y}(t, \cdot) - \mathbf{y}(t_{k-1}, \cdot)]. \tag{4}$$

These conditions are motivated by the device of piecewise function interpolation and associated advantages.

Estimator $\mathcal{F}^{(p-1)}$ of the above form is called the *piecewise linear interpolation* filter. The pair of data-vectors $\{\mathbf{x}(t_k, \cdot), \mathbf{y}(t_k, \cdot)\}$ associated with time t_k is called the *interpolation pair*.

3.3 Determination of piecewise linear interpolation filter. Let us denote $\mathbf{z}(t_j, t_{j+1}, \cdot) = \mathbf{x}(t_{j+1}, \cdot) - \widehat{\mathbf{x}}(t_j, \cdot)$ and $\mathbf{w}(t_j, t_{j+1}, \cdot) = \mathbf{y}(t_{j+1}, \cdot) - \mathbf{y}(t_j, \cdot)$.

We need to represent $\mathbf{z}(t_j, t_{j+1}, \cdot)$ and $\mathbf{w}(t_j, t_{j+1}, \cdot)$ in terms of their components as follows:

$$\mathbf{z}(t_j, t_{j+1}, \cdot) = [\mathbf{z}^{(1)}(t_j, t_{j+1}, \cdot), \dots, \mathbf{z}^{(m)}(t_j, t_{j+1}, \cdot)]^T \text{ and}$$

$$\mathbf{w}(t_j, t_{j+1}, \cdot) = [\mathbf{w}^{(1)}(t_j, t_{j+1}, \cdot), \dots, \mathbf{w}^{(n)}(t_j, t_{j+1}, \cdot)]^T,$$

where $\mathbf{z}^{(j)}(t_j, t_{j+1}, \cdot) \in L^2(\Omega, \mathbb{R})$ and $\mathbf{w}^{(i)}(t_j, t_{j+1}, \cdot) \in L^2(\Omega, \mathbb{R})$ are random variables, for all $j = 1, \dots, m$.

Then for

$$\langle \mathbf{z}^{(i)}(t_j, t_{j+1}, \cdot), \mathbf{w}^{(k)}(t_j, t_{j+1}, \cdot) \rangle = \int_\Omega \mathbf{z}^{(i)}(t_j, t_{j+1}, \omega) \mathbf{w}^{(k)}(t_j, t_{j+1}, \omega) d\mu(\omega),$$

we can introduce the covariance matrix

$$E_{z_j w_j} = \left\{ \left\langle \mathbf{z}^{(i)}(t_j, t_{j+1}, \cdot), \mathbf{w}^{(k)}(t_j, t_{j+1}, \cdot) \right\rangle \right\}_{i,k=1}^{m,n}. \tag{5}$$

Below, M^\dagger is the Moor-Penrose generalized inverse of a matrix M .

Theorem 1 For any $t \in [a, b]$, the proposed piecewise linear interpolation filter $\mathcal{F}^{(p-1)} : L^2(\Omega, \mathbb{R}^n) \rightarrow L^2(\Omega, \mathbb{R}^m)$ transforming any signal $\mathbf{y}(t, \cdot) \in L^2(\Omega, \mathbb{R}^m)$ to an estimate of $\mathbf{x}(t, \cdot)$, $\hat{\mathbf{x}}(t, \cdot)$, is given by

$$\hat{\mathbf{x}}(t, \cdot) = \mathcal{F}^{(p-1)}[\mathbf{y}(t, \cdot)] = \sum_{j=1}^{p-1} \delta_j \mathcal{F}_j[\mathbf{y}(t, \cdot)]$$

where $\mathcal{F}_j[\mathbf{y}(t, \cdot)] = \hat{\mathbf{x}}(t_j, \cdot) + \mathcal{B}_j[\mathbf{y}(t, \cdot) - \mathbf{y}(t_j, \cdot)]$,
 $\hat{\mathbf{x}}(t_j, \cdot) = \mathcal{F}_{j-1}[\mathbf{y}(t_j, \cdot)]$ (for $j = 2, \dots, p-1$),

$$\mathcal{B}_j = E_{z_j w_j} E_{w_j w_j}^\dagger + M_{B_j} [I_n - E_{w_j w_j} E_{w_j w_j}^\dagger],$$

and where I_n is the $n \times n$ identity matrix and M_{B_j} is an $m \times n$ arbitrary matrix.

4 Simulations

Simulations have been done with data sets arising in a DNA analysis and with digitized images of moving objects. The results of the simulations confirm theoretical results and they will be presented in the final manuscript.

References

- Chen, J., Benesty, J., Huang, Y., and Doclo, S. (2006). New Insights Into the Noise Reduction Wiener Filter, *IEEE Trans. on Audio, Speech, and Language Processing*, **14**, No. 4, pp. 1218–1234.
- Torokhti, A. and Howlett, P. (2007). *Computational Methods for Modelling of Nonlinear Systems*. Elsevier.
- Torokhti, A. and Miklavcic, S. (2009). Data compression and de-compression by causal filters with variable finite memory. In: *Advanced Technologies*, INTECH, pp. 631–654.
- Torokhti, A. and Manton, J. (2009). Generic Weighted Filtering of Stochastic Signals. *IEEE Trans. on Signal Processing*, **57**, issue 12, pp. 4675–4685.
- Torokhti, A. and Miklavcic, S. (2010). Data Compression under Constraints of Causality and Variable Finite Memory. *Signal Processing*, **90**, Issue 10, 2822–2834.
- Torokhti, A. and Miklavcic, S. (2011). Filtering of infinite sets of stochastic signals: an approach based on interpolation techniques. *Signal Processing*, **91**, Issue 11, pp. 2556–2566.

Age and sex specific social contact patterns stratified by location and day of the week

Jan van de Kastelee¹, Jan van Eijkeren¹, Jacco Wallinga¹

¹ National Institute for Public Health and the Environment - RIVM, the Netherlands

E-mail for correspondence: jan.van.de.kastelee@rivm.nl

Abstract: A statistical methodology using Gaussian Markov Random Fields is presented for determining age and sex specific social contact patterns stratified by location and day of the week. Contact patterns are crucial information in mathematical models for infectious diseases. We show that, depending on the setting and age, contact rates differ between men and women

Keywords: Social contact patterns; Infectious Diseases; Gaussian Markov Random Fields; Smoothing

1 Introduction

The incidence and severity of most human respiratory infections such as influenza, tuberculosis, measles, rubella and mumps depend on age and sex. For the planning and evaluating of vaccination programmes against these infectious diseases it is necessary to know how an infection spreads among age groups and among women and men.

Mathematical modeling of infectious diseases transmitted by the respiratory or close-contact route is increasingly used to determine the impact of possible interventions. The contact patterns, which are crucial determinants for model outcome, are often limited to coarse age-specific contact rates. Stratification by sex, location of contact or weekday is usually absent, even though it can be expected that contact patterns differ between males and females, and depend very much on the location and day of the week. There is a need for more detailed empirical data on contact behavior by sex and by setting.

Large data sets on the contact behavior of a representative population have been collected and made publicly available, Mossong et al. (2008). The estimation of contact rates by both age and sex from these data sets has proven to be statistically challenging. The stratification by age and sex of both the study participants and their contacts leads to a very large number of contact rates to be estimated. For example, if we would choose to estimate the contacts by men and women in 81 age cohorts to cover the age

range in the general population this would require estimating $(2 \times 81)^2 = 26,244$ contact rates. This number will increase further if we stratify by the settings where they occur (home, work, school, leisure, transport, other) and by day of the week at which they occur (weekday or weekend).

There are two ways to handle such a large number of contact rates. The first invokes the reciprocal nature of contact that induces symmetry in contact rates between age groups and sexes. This symmetry removes the need to estimate nearly half the number of the contact rates. The second relies on the similarity in contact patterns for individuals of similar age. This allows us to use statistical smoothing techniques to increase the precision of each estimate. We combine these two approaches by imposing smoothness in contact rates while accounting for the reciprocity of contacts between age groups and sexes, using Gaussian Markov Random Fields (GMRF). This allows us to extract contact rates from the existing data sets at an unprecedented level of detail.

2 Data and Methods

2.1 Population based survey

For respiratory infections, the number of different conversational partners provides a good proxy. We explore data from a population-based prospective survey of contact patterns in eight European countries. 7,466 participants recorded characteristics of 98,309 contacts with different individuals during one day, age and sex, at six different settings at different days in the week.

2.2 Hierarchical Bayesian model

Our aim is to describe contact rates among 81 year age groups and the two sexes, given a specific combination of setting of contact and day of the week. We infer contact rates using a hierarchical Bayesian model. There are three levels.

The first level, the observation level, refers to the total number of contacts of any age and sex specific combination (i, j) , location l and day of the week d in the dataset, simply found by cross-tabulation. At this observation level, the total number of contacts y_{ijld} is described by a negative binomial distribution with mean μ_{ijld} and dispersion parameter θ_{ld} :

$$y_{ijld} \mid \mu_{ijld}, \theta_{ld} \sim \text{NegBin}(\mu_{ijld}, \theta_{ld}).$$

At the second level, the mean of the total number of contacts is described by the product of an age, sex and day of the week specific offset E_{ijd} and age, sex, location and day of the week specific contact rate c_{ijld} :

$$\mu_{ijld} = E_{ijd} \exp(\beta_{ld} + x_{ijld}).$$

The offset is the product of the total number of participants of a certain age and sex and the total population number of the contacted age and sex. We can interpret β_{ld} as an intercept and x_{ijld} as deviations from an average contact pattern. These deviations have a smooth and symmetric structure. Our approach is to model \mathbf{x}_{ld} as a zero mean two-dimensional Gaussian Markov Random Field (GMRF) over the ages i and j .

A GMRF is a random field following a multivariate normal (Gaussian) distribution with conditional (Markov) independence assumptions. This conditional independence is defined by a precision matrix $\mathbf{Q}_{ld} = \tau_{ld}\mathbf{R}$, where \mathbf{R} is a structure matrix and where τ_{ld} is the precision parameter that controls the smoothness of the deviations \mathbf{x}_{ld} :

$$\mathbf{x}_{ld} | \tau_{ld} \sim \text{Normal}(\mathbf{0}, \mathbf{Q}_{ld}).$$

At the third level uninformative hyperpriors are specified for the overdispersion parameter of the negative binomial model and the precision (smoothing) parameter of the GMRF.

$$\begin{aligned} \tau_{ld} &\sim \text{Gamma}(1, 0.01), \\ \log(\theta_{ld}) &\sim \text{Normal}(0, 0.1). \end{aligned}$$

We estimate the rates and parameters for such highly structured hierarchical Bayesian models by the recently established Integrated Nested Laplace Approximations technique (INLA).

2.3 Tailor-made structure matrix

The contact rates are symmetric, such that $c_{ijld}^{MM} = c_{jild}^{MM}$, $c_{ijld}^{FF} = c_{jild}^{FF}$, and $c_{ijld}^{MF} = c_{jild}^{FM}$. These symmetries carry over to the deviations x_{ijld} . The symmetry is illustrated in Figure 1 for the two sexes and $n = 5$ age groups. There are $(2 \times 5)^2 = 100$ nodes, each corresponding to one data record. A node corresponds to a value x_{ijld} . By forcing identical node values in the lower and upper triangular parts of the matrix block, symmetry is guaranteed. Thus, 55 unique node values are inferred from the 100 data records. For illustrative purposes, identical nodes are indicated by identical colors.

Smoothing is achieved by imposing the condition that neighboring node values of \mathbf{x}_{ld} should be similar. The neighborhood structure is defined by the entries of the structure matrix \mathbf{R} and is illustrated in Figure 1 by the edges. Only the nodes in the lower triangular parts of the matrix blocks need to be linked, the nodes in the upper triangular part follow because of the imposed symmetry.

We use the second order random walk model (RW2), which reflects the prior beliefs that the gradient of x_{ld} varies smoothly and that sudden jumps between neighboring values of the gradient of \mathbf{x}_{ld} are unlikely. In one dimension smoothness is achieved by placing a normal (Gaussian) prior on

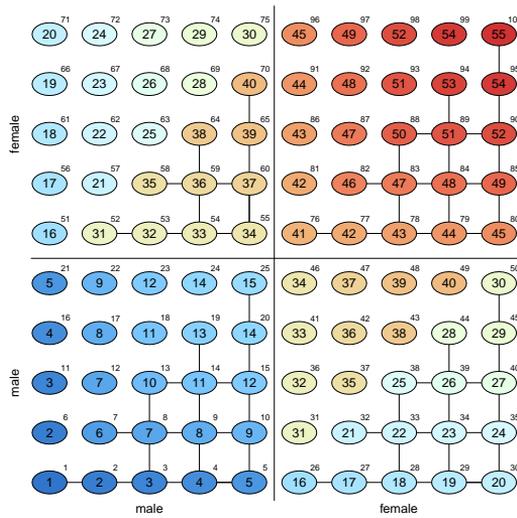


FIGURE 1. Graphical representation of a contact matrix stratified by sex ($n = 5$ example). The matrix consists of four blocks: MM (lower left), FM (lower right), MF (upper left) and FF (upper right) contacts. Each node represents an age combination between participants (horizontal axis) and contacts (vertical axis). The numbering of the nodes is such that symmetry between ages and sexes is achieved. The coloring is for illustrative purpose. The edges denote the dependencies between triplets of nodes (the lower triangular parts only). The superscripts represent record id's in the dataset.

the second order differences of \mathbf{x}_{ld} :

$$\Delta^2 x_{m,ld} = x_{m-1,ld} - 2x_{m,ld} + x_{m+1,ld} \sim \text{Normal}(0, \tau_{ld})$$

In two dimensions smoothness is achieved by placing a RW2 prior on the columns and rows of \mathbf{x}_{ld} using kronecker products.

3 Results

The settings with the highest contact rates are home, school and work. In general the contacts tend to be assortative with respect to sex for most ages in most settings. At home (weekend) adult men make most contacts with adult women and vice versa, whereas children make contacts with either sex (Figure 2). At work (weekday), adult men have overall higher contact rates than women; remarkably, older men and younger women have higher contact rates at work than younger males and older females. At school (weekday), boys contact boys rather than girls and vice versa. The contact rates differ slightly between boys and girls. The contact rate of children with adult women is slightly higher than their contact rate with adult men.

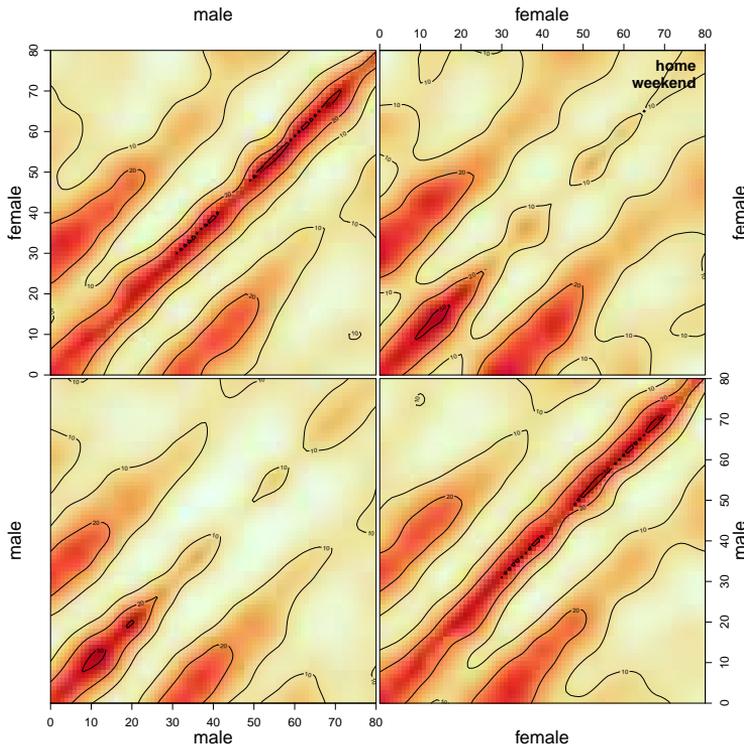


FIGURE 2. Age and sex specific contact rates at home during a weekend. See Figure 1 for explanation of the four blocks. The color scale indicates relative values of the contact rates. Absolute values of contact rates are indicated by contour lines.

The relevance of the difference between sexes becomes clear when we simulate the spread of infection in a completely susceptible population. We calculate the distribution of the (relative) incidence of new infections for both sexes by age and by setting in which infections would occur (Figure 3). We find that the highest incidence of new infections occurs among the 16 year old, and the lowest incidence occurs among the 0 year olds and the 80 year olds. Boys and girls have a similar age-specific risk of infection for ages 0 up to 20 years. Women have a substantially higher the risk of infection from the age of 20 up to 50 years, men have a substantially higher risk of infection from the age of 50 to 80 years.

4 Conclusions

The statistical model enables us to quickly estimate smooth and symmetric contact rate matrices. We have highlighted the relevance of sex-specific dif-

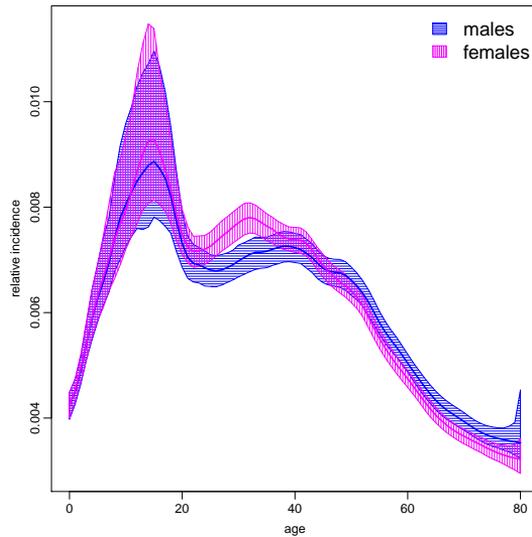


FIGURE 3. Relative incidence of infection when a new emerging infection spreads among men and women in a completely susceptible population. The shaded region give the 95% credible interval of the estimates. The incidence is scaled such that the area under the curves adds to one.

ferences in these patterns for respiratory infections. Contacts tend to be assortative with respect to sex for most ages in most settings. Two exceptions: at home men contact more men than women and v.v.; at school children contact more adult women than adult men. The observations strongly suggest that exposure to infection is substantially differs between sexes, where the magnitude and direction of the difference depends on age.

References

- Mossong, J., *et al.* (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, **5**(3):e74, 381–391.
- Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields - Theory and Applications*. Chapman & Hall.
- Rue, H., Martino, S., and Chopin N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, **71**(2), 319–392.

Modelling interactions of multiple events with application to human papillomavirus virus infections

Zihua Yang¹, Jack Cuzick¹

¹ Wolfson Institute for Preventive Medicine, Queen Mary University of London.

E-mail for correspondence: z.h.yang@qmul.ac.uk

Abstract: Excess risk in the probability of concurrent events is often observed in multiple infections of human papillomavirus (HPV). We model the excess risk for each multiplicity of concurrences using a single parameter across all combination of HPV types, this approach yields simple expressions for outcome probabilities. The model is fitted to a dataset of infections of 35 HPV types in 11,155 New Mexico women with abnormal cervical cytology smears and age of less and equal to 30.

Keywords: categorical data, multiple events, human papillomavirus virus.

1 Introduction

We explore a dataset of multiple infections of 35 human papillomavirus (HPV) types on 11,155 women (with abnormal cytology and age of less and equal to 30) who had a cervical smear in the state of New Mexico between December 2007 to April 2009. As observed in previous studies, the observed concurrences of different types is significantly larger than that would be expected under independence. We believe that this excess risk may reflect a population level heterogeneity in susceptibility to the infections - this is supported by existing literature on cohort and case-control studies of HPV infections which suggest no evidence for specific synergistic interactions between multiple types (Kong *et al.* (2010), Maucort *et al.* (2010), Plummer *et al.* (2008), Trottier *et al.* (2006), Vaccarella *et al.* (2010) and Chaturvedi *et al.* (2011)). Previous studies look mostly at second-order interactions and do not provide overall models for the whole event space.

2 The Model

Let δ_i^n be the indicator of individual n being infected with type i ($i = 1, \dots, K, n = 1, \dots, N$). For a given set of type indices $\{i_1, \dots, i_K\} \in \{1, \dots, K\}$, let us denote the probability of having types i_1, \dots, i_k by $p_{i_1, \dots, i_k}^n = pr(\delta_{i_1}^n =$

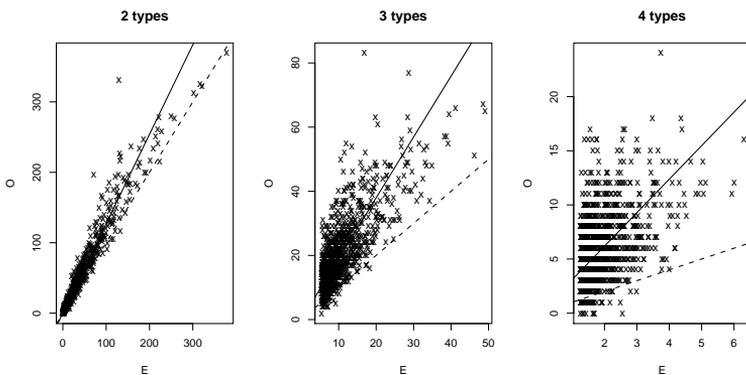


FIGURE 1. Observed numbers versus expected numbers of 2-types (595 pairs), 3-types (6,545 triplets) and 4-types (52,360 4-tuples) interactions for the 35 HPV types, sample restricted to women younger than 30 with abnormal cytology (N=11,155). Solid lines denote simple linear fits and dashed lines give the diagonal.

$1, \dots, \delta_{i_k}^n = 1) = E(\prod_{j=i_1}^{i_k} \delta_j^n)$, $k \leq K$. We drop superscript n in probabilities as we assume the events over individuals are i.i.d. Under independence, p_{i_1, \dots, i_k} reduces to $\prod_{j=i_1}^{i_k} p_j$. For our dataset on multiple HPV infections (for women younger than 30 with abnormal cytology), we observe a significant excess risk in being infected with multiple types - Figure 1 illustrates the 2, 3, 4-way fit of observed concurrences O_{i_1, \dots, i_k} versus expected concurrence E_{i_1, \dots, i_k} under independence where $O_{i_1, \dots, i_k} = \sum_n \prod_{j=i_1}^{i_k} \delta_j^n$ and $E_{i_1, \dots, i_k} = N \prod_{j=i_1}^{i_k} \hat{p}_j$ (using marginal prevalence estimates $\hat{p}_i = N^{-1} \sum_{n=1}^N \delta_i^n$). We consider a model in which

$$p_{i_1, \dots, i_k} = \theta_k \prod_{j=i_1}^{i_k} p_j; \quad \theta_0 = \theta_1 = 1$$

where θ_k is an excess risk parameters. It then follows that $pr(\delta_{i_1} = 1 | \delta_{i_2} = 1, \dots, \delta_{i_k} = 1) = (\theta_k / \theta_{k-1}) p_{i_1}$. We could impose further constraint on the excess risk θ_k by modelling it as a population heterogeneity effect or ‘frailty’. For example, the simplest case would be a two-point frailty model with $\theta_k = \theta^{k-1}$. Detailed results on fitting a frailty model will appear elsewhere.

Using the inclusion/exclusion principle, the likelihood for any individual can be written as

$$pr(\delta_1^n, \delta_2^n, \dots, \delta_K^n) = \left(\prod_{i=1}^K p_i^{\delta_i^n} \right) \left\{ \sum_{m=0}^K (-1)^m \theta_{M+m} P_{m, \delta}^{(K)} \right\}$$

where $M \equiv \sum_{i=1}^K \delta_i^n$ and $P_{m, \delta}^{(K)} \equiv \sum_{i_1 < i_2 < \dots < i_m} \prod_{j=i_1}^{i_m} p_j (1 - \delta_j^n)$.

2.1 Estimation

The marginal prevalence for type i can be estimated using $\hat{p}_i = N^{-1} \sum_{n=1}^N \delta_i^n$ which is clearly unbiased, $N^{1/2}$ consistent and asymptotically normal. θ_k is estimated by

$$\hat{\theta}_k = \sum_{i_1 < \dots < i_k}^K O_{i_1, \dots, i_k} \left\{ \sum_{i_1 < \dots < i_k}^K E_{i_1, \dots, i_k} \right\}^{-1} \tag{1}$$

which can be shown to be asymptotically normal with variance

$$\text{var}(\hat{\theta}_k) = \theta_m N P_k^{-2} \sum_{i_1 < \dots < i_k} \sum_{j_1 < \dots < j_k} p_{l_1} \dots p_{l_m} - N^{-1} \theta_k^2$$

where $\{l_1, \dots, l_m\}$ is the unique set of indices $\{i_1, \dots, i_{k_1}\}$ and $\{j_1, \dots, j_{k_2}\}$ and $P_k \equiv \sum_{i_1 < i_2 < \dots < i_k}^K p_{i_1} \dots p_{i_k}$. Similarly,

$$\text{cov}(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}) = \theta_m (N P_{k_1} P_{k_2})^{-1} \sum_{i_1 < \dots < i_{k_1}} \sum_{j_1 < \dots < j_{k_2}} p_{l_1} \dots p_{l_m} - N^{-1} \theta_{k_1} \theta_{k_2} \tag{2}$$

3 HPV Genotype Interactions

We fit the model to the 35 HPV infections of the 11,155 New Mexico women with abnormal cytology and age less and equal to 30 years. The marginal prevalences vary from 0.2% (type 69) to 22% (type 16).

The results for multiple event measures are summarised in Table 1. For $k = 2, 3, 4$, estimates of the overall interaction coefficient θ_k are significantly larger than unity. The theoretical standard deviation of $\hat{\theta}_k$, $s.d.(\hat{\theta}_k)$ ($k = 2, 3, 4$) uses $\hat{\theta}_k$ up to $k = 8$ and we found it to match closely with bootstrap estimates of the variance.

We tested the goodness-of-fit of the model using the test statistic

$$T_k = \sum_{i_1 < \dots < i_k} (O_{i_1, \dots, i_k} - \theta_k E_{i_1, \dots, i_k})^2 / \max(1, \theta_k E_{i_1, \dots, i_k})$$

where the maximum in the denominator is used to downweight the less prevalent concurrences. We refer the value of the observed $T_k(\hat{\theta}_k)$ to 1000 simulations of T_k using $\hat{\theta}_k$ for a given k . Overall, the model did not provide a good fit to the 35 types (goodness of fit p -value < 0.001). The index of dispersion, defined as

$$D(\theta_k) = \sum_{i_1, \dots, i_k} (O_{i_1, \dots, i_k} - \theta_k E_{i_1, \dots, i_k})^2 / \sum_{i_1, \dots, i_k} \theta_k E_{i_1, \dots, i_k}$$

was calculated for the three multiplicities under the fitted model and the independence model. While the D values were substantially smaller using

TABLE 1. Multiple HPV events for the 35 types, $O_k \equiv \sum_{i_1 < \dots < i_k} O_{i_1, \dots, i_k}$.

	O_k	$\hat{\theta}_k$	$s.d.(\hat{\theta}_k)$	D_{model}	D_{indep}
	35 types				
$k = 2$	33470	1.27	0.022	5.308	10.868
$k = 3$	35638	2.015	0.065	2.484	11.560
$k = 4$	18193	3.736	0.220	1.409	9.609

$\hat{\theta}_k$ than for independence, it is clear that there remains clustering within the concurrences that has not been captured by the model.

Plummer *et al.* (2008) found the excess risk in second order interactions within the $\alpha 9$ species (6 HPV types in our dataset) and $\alpha 7$ species (6 HPV types in our dataset) to disappear once they included a population-level random effect into their model. For our dataset, the model using second order excess risk estimates gives a good fit to concurrences belonging to $\alpha 9$ species ($\hat{\theta}_2^{(\alpha 9)} = 1.05$, goodness of fit p -value = 0.686) and $\alpha 7$ species ($\hat{\theta}_2^{(\alpha 7)} = 1.29$, goodness of fit p -value = 0.417).

References

- Chaturvedi, A.K., Katki, H.A., Hildesheim, A., Rodríguez, A.C., Quint, W., Schiffman, M., Van Doorn, L.J., Porras, C. and Wacholder, S. and Gonzalez, P. and others (2011). Human papillomavirus infection with multiple types: pattern of coinfection and risk of cervical disease *Journal of Infectious Disease*, **7**, 910–920.
- Kong, X., Gray, R.H., Moulton, L.H., Wawer, M. and Wang, M.C. (2010). A modeling framework for the analysis of HPV incidence and persistence: A semi-parametric approach for clustered binary longitudinal data analysis. *Statistics in Medicine*, **29**, 2880–2889.
- Maucort-Boulch, D., Plummer, M., Castle, P.E., Demuth, F., Safaeian, M., Wheeler, C.M. and Schiffman, M. (2010). Hierarchical generalized linear models. *International Journal of Cancer*, **126**, 684–691.
- Plummer, M., Schiffman, M., Castle, P.E., Maucort-Boulch, D. and Wheeler, C.M. (2008). A 2-Year prospective study of human papillomavirus persistence among women with a cytological diagnosis of atypical squamous cells of undetermined significance or low-grade squamous intraepithelial lesion. *Journal of Lower Genital Tract Disease*, **12**, 1582–1589.
- Trottier, H., Mahmud, S., Costa, M.C., Sobrinho, J.P., Duarte-Franco, E., Rohan, T.E., Ferenczy, A., Villa, L.L. and Franco, E.L. (2006). Human papillomavirus infections with multiple types and risk of cervical

neoplasia *Cancer Epidemiology Biomarkers & Prevention*, **15**, 1274–1280.

Vaccarella, S., Franceschi, S., Snijders, P.J.F., Herrero, R., Meijer, C.J.L.M. and Plummer, M. (2010). Concurrent infection with multiple human papillomavirus types: pooled analysis of the IARC HPV Prevalence Surveys *Cancer Epidemiology Biomarkers & Prevention*, **19**, 503–510.

Spatial model for multivariate traffic accident count data

Zamira Zamzuri¹, Ross Sparks², Graham Wood¹, Gillian Z. Heller¹

¹ Department of Statistics, Macquarie University, Australia

² Centre of Mathematical and Information Sciences, CSIRO, Australia

E-mail for correspondence: zamira.zamzuri@students.mq.edu.au

Abstract: This paper introduces a new spatial model to accommodate multivariate accident count data. Available models in the accident literature to date can either only cope with the spatial component, or were developed under a univariate framework. Based on the multivariate Poisson lognormal model, we further develop this model by introducing linear combinations of random impulses to capture spatial correlation. The estimation of this model is then carried out using the Markov Chain Monte Carlo simulation method. Simulated data sets were used in assessing the performance of this model. An advantage of this new model is that it not only copes with between and within location variations, but also provides information on spatial dependency that has been often ignored in past models.

Keywords: Spatial model; Multivariate count model; Traffic accidents; Markov Chain Monte Carlo.

1 Introduction

Generalized linear model has been used to model the relationship between the number of accidents and explanatory variables. Most common choices of generalized linear model applied to traffic accident counts are Poisson regression model, negative binomial regression model and zero inflated models (Miao, 1994; Shankar et al. 1997; Oh et al. 2006). In these models, accident counts happen at different locations were assumed to be independent from each other. This fails to make the best use of the spatial association of accident counts. Spatial correlation may arise from two or more neighbouring areas. The area can be countries, regions, suburbs or road intersections. In this paper, the spatial correlation is focused on neighbouring intersections. Neighbouring intersections share many unmeasured features, such as weather conditions, and this contributes to the spatial correlation between accident counts. Several models have been suggested in the traffic accident literature that exploited the correlation between different intersections (MacNab, 2004; Song et al., 2006; Quddus, 2008). Meanwhile Park and

Lord (2006) and Ma et al. (2008) look at different type of correlation, correlation that exists between different levels of severity. Since severity within the same intersections are expected to be associated, their correlation is modelled using a random effect. The use of Multivariate Poisson Lognormal (MPL) model (Chib and Winkelmann, 2001) to capture this type of correlation become our fundamental base to further extending this model to capture both spatial (between intersection) and the within intersection (between levels of severity) correlation. The spatial correlation was introduced to the model through a linear combination of random impulses as suggested by Wolpert and Ickstadt (1998). The specification of this model is presented in the next section followed by the estimation procedure and simulation study results.

2 The model

Let i represent intersection, t the day and j the level of severity, and let y_i be a column vector of counts for the i th intersection at t th day, $y_{it} = (y_{i1t}, y_{i2t}, \dots, y_{iJt})'$. We have N intersections, J levels of severity and T days. Parameter β_j is a vector of regression coefficients for the j th level without the constant coefficient, while X_{ijt} is the vector of covariates for i th intersection, j th level of severity and t th day. Assume that observations between different days are independent. Parameter D is an unrestricted covariance matrix used in the generation of vector b_i while c_i is a linear combination of g_i , the random impulses. Let $g = (g_1, g_2, \dots, g_N)'$, β_j be a vector of regression coefficients for j th level of severity and A is the adjacency matrix that carry spatial correlation between intersection. The model specification is given as

$$\begin{aligned} Y_{ijt} | b_{ijt}, c_{it} &\sim \text{Poisson}(\mu_{ijt}) \\ \mu_{ijt} &= \exp(X'_{ijt}\beta_j + b_{ij} + c_i) \\ b_i &\sim N_J(0, D) \\ c_i &= Ag \\ g_i &\sim N(\xi, v) \end{aligned}$$

2.1 Estimation procedure

The likelihood

The likelihood is given by the product of the probability function from $i=1$ to N and from $t=1$ to T (N is the number of intersections while T is the

number of days).

$$\begin{aligned}
 L(y|\beta, D, \xi, v) &= \prod_{t=1}^T \prod_{i=1}^N \int \int f(y_{it}, b_{it}, g_{it}|\beta, D, \xi, v) db dg \\
 &= \prod_{t=1}^T \prod_{i=1}^N \int \int \prod_{j=1}^J (f(y_{ij}|\beta, b_{ij}, g_{ij}) \phi_J(b_{ij}|D) \phi(g_{ij}|\xi, v)) db dg \\
 &= \prod_{t=1}^T \prod_{i=1}^N \int \phi_J(b_{it}|D) \phi(g_{it}|\xi, v) \int \prod_{j=1}^J f(y_{ij}|\beta_j, b_{ij}, c_i) db dg
 \end{aligned}$$

The likelihood is a complex function involving high-dimensional integration. To work in a Bayesian framework, we need to define the priors and then calculate the posterior distribution to perform the MCMC simulation.

The priors

The prior distributions for parameters β, D^{-1}, ξ and v are given as follows

$$\phi_K(\beta|\beta_0, B_0^{-1}) \quad f_w(D^{-1} | v_0, R_0) \quad \phi(\xi|\xi_0, \tau_0) \quad f_{IG}(v|\alpha_0, \theta_0)$$

where ϕ_K is a K – variate normal distribution (K is the length β), f_w is a Wishart distribution, ϕ is a normal distribution, f_{IG} is an inverse gamma distribution and $\beta_0, B_0, v_0, R_0, \xi_0, \tau_0, \alpha_0$ and θ_0 are hyperparameters.

The posterior

Using Bayes theorem, the posterior distribution of parameters given information from the observations is proportional to the product of priors and likelihood, as given below

$$\begin{aligned}
 &\phi_K(\beta|\beta_0, B_0^{-1}) f_w(D^{-1}|v_0, R_0) \phi(\xi|\xi_0, \tau_0) f_{IG}(v|\alpha_0, \theta_0) \\
 &\quad \prod_{t=1}^T \prod_{i=1}^N f(y_{it}, b_{it}, g_{it}|\beta, D, \xi, v)
 \end{aligned}$$

Next, we will look at a Markov chain sampling procedure for this model. The Markov chain is constructed using full conditional distributions of parameters,

$$[b|y, \beta, D, c] \quad [\beta|y, b, c] \quad [D^{-1}|b] \quad [g|y, \xi, v, \beta, b] \quad [\xi|g, v] \quad [v|g, \xi]$$

There is no sampling stage for c_i because it has been specified as a linear combination of g . Using updated values of g in each MCMC iteration, the value of c_i is then computed. For each of the sampling stage, we take a sample from the corresponding conditional posterior density. Two different sampling procedure were used here; the Metropolis-Hastings and Gibbs

sampling algorithm. The first algorithm was used in the case when the conditional posterior density is an unrecognized distribution while the latter is for the opposite case. For the Metropolis-Hastings algorithm, parameters of the proposal density were first estimated through maximizing the log of the conditional posterior density.

2.2 Spatial correlation through the adjacency matrix

Let $c = (c_1, c_2, \dots, c_n)$. Then c can be computed as a product of the adjacency matrix and the random impulses g , where $g = (g_1, g_2, \dots, g_n)'$. Consider the case of three intersections, let A be the adjacency matrix, representing which intersection are adjacent to which other intersections, and assume equal weights are applied to neighbouring g_i , then $A = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \end{pmatrix}$ implying

$$\begin{aligned} c &= Ag \\ &= \left(\frac{g_1}{2} + \frac{g_2}{2}, \frac{g_1}{3} + \frac{g_2}{3} + \frac{g_3}{3}, \frac{g_2}{2} + \frac{g_3}{2} \right) \end{aligned}$$

Then the computed values of c is used in sampling β and b . As discussed in Wolpert & Ickstadt (1998), overlapping random impulses create the spatial correlation between locations.

3 Simulation study

A simulation study has been conducted to validate the accuracy of the model estimation. Four different simulated data sets were generated with different setting of number of intersections (N), days (T) and correlation (ρ). Results are given in the following table.

From Table 1, it can be seen that estimations of all parameter are reasonably well. It also can be seen that the estimations were improved when we have more information from the data, through the increment of a number of intersections (N) or a longer time period (T).

4 Summary

In this paper, an extended Multivariate Poisson Lognormal model was presented. The additional component in this model is intend to capture spatial correlation between intersections through linear combinations of random impulses. Estimations of parameters in this model were derived from MCMC procedure based on Bayesian framework. Simulation study shown that the parameters of the model were estimated satisfactorily.

TABLE 1. Results of simulation study.

Data	Parameter	True	Estimate
Data 1	$\beta_{1,j=1}$	0.260	0.234 (0.020)
$N = 9$	$\beta_{2,j=1}$	0.176	0.162 (0.022)
$T = 100$	$\beta_{3,j=1}$	0.100	0.147 (0.033)
	$\beta_{1,j=2}$	0.260	0.262 (0.021)
	$\beta_{2,j=2}$	0.176	0.159 (0.020)
	$\beta_{3,j=2}$	0.100	0.071 (0.032)
	D_{11}	0.099	0.151 (0.065)
	D_{12}	0.023	0.018 (0.029)
	D_{22}	0.049	0.027 (0.017)
	ρ	0.330	0.282
	ξ	-2.078	-1.991 (0.200)
	v	0.179	0.097 (0.052)
Data 2	$\beta_{1,j=1}$	0.260	0.268 (0.009)
$N = 9$	$\beta_{2,j=1}$	0.176	0.177 (0.009)
$T = 300$	$\beta_{3,j=1}$	0.100	0.123 (0.017)
	$\beta_{1,j=2}$	0.260	0.259 (0.010)
	$\beta_{2,j=2}$	0.176	0.170 (0.010)
	$\beta_{3,j=2}$	0.100	0.088 (0.018)
	D_{11}	0.099	0.105 (0.052)
	D_{12}	0.067	0.066 (0.069)
	D_{22}	0.145	0.141 (0.099)
	ρ	0.559	0.542
	ξ	-1.902	-2.030 (0.160)
	v	0.264	0.265 (0.237)
Data 3	$\beta_{1,j=1}$	0.260	0.250 (0.006)
$N = 25$	$\beta_{2,j=1}$	0.176	0.183 (0.006)
$T = 300$	$\beta_{3,j=1}$	0.100	0.095 (0.010)
	$\beta_{1,j=2}$	0.260	0.254 (0.006)
	$\beta_{2,j=2}$	0.176	0.176 (0.006)
	$\beta_{3,j=2}$	0.100	0.099 (0.011)
	D_{11}	0.082	0.069 (0.022)
	D_{12}	0.064	0.052 (0.022)
	D_{22}	0.107	0.097 (0.029)
	ρ	0.683	0.636
	ξ	-2.059	-2.062 (0.082)
	v	0.199	0.227 (0.251)
Data 4	$\beta_{1,j=1}$	0.260	0.258 (0.059)
$N = 9$	$\beta_{2,j=1}$	0.176	0.192 (0.070)
$T = 300$	$\beta_{3,j=1}$	0.100	0.060 (0.040)
	$\beta_{1,j=2}$	0.460	0.462 (0.026)
	$\beta_{2,j=2}$	0.276	0.278 (0.020)
	$\beta_{3,j=2}$	0.210	0.232 (0.053)
	D_{11}	0.097	0.073 (0.052)
	D_{12}	0.049	0.033 (0.014)
	D_{22}	0.076	0.059 (0.016)
	ρ	0.571	0.533
	ξ	-6.542	-6.593 (0.233)
	v	0.415	0.468 (0.170)

References

- Miao, S.P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis and Prevention*,**26**, 471–482
- Shankar, V., Milton, J., and Mannering, F.L. (1997). Modelling accident frequency as zero altered probability process: an empirical enquiry. *Accident Analysis and Prevention*,**29**, 829–837
- Wolpert, R.L., and Ickstadt, K. (1998). Poisson gamma random field model for spatial statistics. *Biometrika*,**85**, 251–267
- Chib, S., and Winkelmann, R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data. *American Statistical Association Journal of Business & Economic Statistics*,**19**, 428–435
- MacNab Y.C. (2004). Bayesian spatial and ecological models for small-area accident and injury analysis. *Accident Analysis and Prevention*,**36**, 1019–1028
- Song J.J, Ghosh M., Miao S., and Mallick B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*,**97**, 246–273
- Oh, J., Washington, S.P., and Nam, D. (2006). Accident prediction model for railway highway interfaces. *Accident Analysis and Prevention*,**38**, 346–356
- Park, E.S, and Lord, D. (2006). Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. In: *Proceedings of the 86th Annual Meeting of the Transportation Research Board*, Washington, D.C.
- Ma, J., Kockelman, K.M., and Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention*,**40**, 964–975
- Quddus M.A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. *Accident Analysis and Prevention*,**40**, 1486–1497