Proceedings of the

34th International Workshop on Statistical Modelling

Volume I

July 7-12, 2019

Guimarães, Portugal

Title: Proceedings of the 34th International Workshop on Statistical Modelling
Volume I,
Guimarães, July 7-12, 2019,
Luís Meira-Machado, Gustavo Soutinho (editors),
Guimarães, 2019.

Editors:

Luís Meira-Machado,lmachado@math.uminho.pt
University of Minho
Dep. Mathematics and Applications
4810-058 Azurém - Guimarães
Portugal

Gustavo Soutinho, gustavo.soutinho@ispup.up.pt
EPIUnit, ICBADS, University of Porto
Rua das Taipas 135
4050-600 Porto
Portugal

ISBN 978-989-20-9528-8
Printed by Instituto Nacional de Estatística

Preface

We welcome all attendees of the 34th International Workshop on Statistical Modelling (IWSM) in Portugal, and we wish you all a very pleasant stay in Guimarães. Besides being the town hosting the University of Minho, Guimarães is one of the most beautiful towns in the Minho region. Guimarães is a magnificent city of medieval origin, located in the North of Portugal and committed to be awarded the label of European Green City for 2020. The conference will take place at the Vila Flor Cultural Centre, in the city Centre of Guimarães.

For 34 years now, the International Workshop on Statistical Modelling (IWSM) has been a reference for promoting and encouraging statistical modelling in its widest sense. IWSM is now one of the most prestigious world conferences in Statistical Modelling, regularly attracting academic, professional statisticians and data analysts from all parts of the world.

The scientific programme of this year's IWSM follows the well-established traditions of the workshop by having 5 invited lectures. We are glad that renowned experts as Adrian Bowman (University of Glasgow, United Kingdom), Julio Singer (University of São Paulo, Brazil), Maria Antónia Turkman (University of Lisbon, Portugal), Peter Diggle (Lancaster University, United Kingdom) and Philippe Lambert(University of Liege, Belgium) have accepted the invitation to give a one hour presentation each.

In addition to the contributed papers and posters, this year's programme also comprises two special sessions. One session in honor of Professor Murray Aytkin and the other devoted to Statistics Portugal (Instituto Nacional de Estatística - INE). The high standards of the conference and the quality of all presentations were ensured by the scientific committee:

Luís Meira-Machado (University of Minho, Portugal)
Ana Luísa Papoila (NOVA Medical School, Lisbon, Portugal)
Arminda Manuela Gonçalves (University of Minho, Portugal)
Brian Marx (Louisiana State University, USA)
Clarice Demétrio (University of São Paulo, Brazil)
David Conesa (University of Valencia, Spain)
Emilio Porcu (Newcastle University, UK)
Enrico Colosimo (Federal University of Minas Gerais - UFMG, Brazil)
Carlo Giovanni Camarda (French Institute for Demographic Studies - INED, France)
John Hinde (University of Galway, Ireland)
Kenan Matawie (University of Western Sydney, Australia)
Maria Xosé Rodriguez Álvarez (Basque Center for Applied Mathematics, Bilbao, Spain)
Ardo van den Hout (University College London)
Vito Muggeo (University of Palermo, Italy)
Vitor Leiva (Pontifical Catholic University of Valparaiso, Chile)

As usual in the IWSM events there was a large amount of excellent paper submissions, and it was a really stimulating task to select from that big amount 54 abstracts for oral presentations. Each paper was reviewed and scored by three members of the scientific committee. This was a very arduous work, and for their valuable efforts we thank all members of the scientific committee.

It is important to mention that one important characteristic of IWSM workshops is that there are no parallel sessions. We believe that this will provide a stimulating atmosphere, encouraging exchange of ideas and cross-fertilization among different areas of statistics. Also, there will be no oral presentations coinciding with the poster session, so all participants are encouraged to attend and discuss the work being presented.

According to the workshop tradition, in this edition student participation has been strongly encouraged not only to attend the workshop but also to present their work. Students attending the conference had the possibility to compete for three awards: best student paper, best student oral presentation and best student poster. In addition to this, two student travel grants have been provided by the Statistical Modelling Society (SMS). We hope that the Short Course "Statistical modelling with missing data: challenges and practical solutions" given by James Carpenter (London School of Hygiene & Tropical Medicine) was of their interest.

Following the latest tradition of IWSM, two proceeding volumes with manuscripts of presented papers (both oral and poster presentations) are published and distributed at the conference. Part I of proceedings which will be printed will contain all papers being orally presented during IWSM. Part II of proceedings will contain papers corresponding to poster presentations and will be available online.

First of all, thank the Statistical Modelling Society for trusting in our proposal and for giving us this great opportunity to organize IWSM 2019, the first edition of the IWSM to be organized in Portugal. We want to take this opportunity to send out our thanks to all members of my Local Organizing Committee, in particular, Arminda Manuela Gonçalves for her many contributions.

We express our gratitude to all authors and participants for the excellent scientific contributions, and we hope that every participant of IWSM 2019 will have a great and especially research-stimulating week in Guimarães. We are also very grateful to all our sponsors, for their generous support. Without those sponsorships many things could not have been realized. A list of the sponsors of IWSM 2019 can be found on the last pages of this volume. Particular acknowledgement must be given to Statistics Portugal who has contributed to the publication of this Book of Proceedings.

We wish the best success to Maria José Rodriguez Álvarez and Dae-Jin Lee as organizers of the next edition of IWSM, which will be held in Bilbao next summer.

Luís Meira Machado and Gustavo Soutinho

Guimarães, June 2019

Contents

Peter J Diggle, Katharine A Owers, Max Eyre: Longitudinal modelling of leptospirosis prevalence using serial dilution assay data ... 3
Philippe Lambert: Inference based on Laplace approximations in nonparametric additive location-scale model for right- or interval-censored data ... 8
Julio M. Singer, Francisco M.M. Rocha, Antonio Carlos Pedroso-de-Lima, Giovani L. Silva, Giuliana C. Coatti, Mayana Zatz: Random changepoint segmented regression with smooth transition: an example with lateral amyotrophic sclerosis data ... 14
Maria Antónia Amaral Turkman, Kamil Feridun Turkman, Paula Pereira, Soraia Pereira, Miguel de Carvalho, Patrícia de Zea Bermudez: Calibration methods for spatial risk analysis ... 20
Adrian W. Bowman, Stanislav Katina, Liberty Vittert: Statistics with a Human Face ... 29
Murray Aitkin: Statistical modelling for income inequality comparisons ... 37
Inês Rodrigues: Synthetic data as Public Use Files: an application to the Household Budget Survey ... 45
Pedro Campos, Suelma Pina, A. Manuela Gonçalves: Small Area Estimation for Land Use and Land Cover ... 49
Jennifer Pohle, Marius Ötting, Frants Havmand Jensen, Roland Langrock: Modeling interactions between individuals using coupled hidden Markov models ... 57
Trung Dung Tran, Emmanuel Lesaffre, Geert Verbeke, Joke Duyck: Latent Ornstein-Uhlenbeck models for Bayesian analysis of multivariate longitudinal categorical responses ... 62
Andreas Groll, Maike Hohberg: An adaptive lasso Cox frailty model for time-varying covariates based on the full likelihood ... 67
Lucas M. Oliveira, Fernanda De Bastiani, Dimitrios M. Stasinopoulos, Robert A. Rigby: Simultaneous autoregressive models within GAMLSS ... 73
S. Arima, S. Polettini: Variable selection in small area model with measurement error in covariates ... 79
Rosario Barone, Andrea Tancredi: Markov Chain Monte Carlo methods for discretely observed continuous-time multi-state semi-Markov models ... 84
Carlo G. Camarda: Forecasting vital rates from demographic summary measures ... 89

Iain Currie: Invariance and the forecasting of mortality 95
Eduardo Elias Ribeiro Jr', Clarice Garcia Borges Demétrio, John Hinde: COM-Poisson models with varying dispersion 101
Theo Economou, Matthew B. Menary: A latent state model for characterising regime shifts in ocean density 107
Małgorzata Roos, Haakon Bakka, Håvard Rue: Sensitivity and identification quantification by a relative latent model complexity perturbation in the Bayesian meta-analysis 112
Ana C. Guedes, Francisco Cribari-Neto, Patrícia L. Espinheira: Improved likelihood ratio testing inferences for unit gamma regressions 117
M. de Carvalho, R. Rubio, M. Leonelli: Learning and modelling dependence structures with diagonal distributions 121
Marius Ötting, Roland Langrock, Antonello Maruotti: A copula-based multivariate hidden Markov model for modelling momentum in football 125
Paul S. Clarke and Yanchun Bao: Estimating mode effects from a sequential mixed-modes experiment 130
Timo Adam, Roland Langrock, Christian H. Weiss: Nonparametric inference in hidden Markov models for time series of counts 135
Marc Schneble, Göran Kauermann: Estimation of Latent Network Flow in Bike-Sharing Systems from Station Feeds 141
S. Pereira, P. Pereira, M. de Carvalho, P. de Zea Bermudez: Calibration of extreme values of simulated and real data 147
Herwig Friedl, Sanela Omerovic: Mixtures of Generalized Nonlinear Models 151
Nadja Klein, Thorsten Simon and Nikolaus Umlauf: Neural Network Regression with an Application to Leukaemia Survival Data – An Unstructured Distributional Approach 157
Chien-Yu Peng: Degradation Models in Reliability Analysis 161
Ruben Amoros, Ruth King, Hidenori Toyoda, Takashi Kumada, Philip J Johnson, Thomas G Bird: A longitudinal continuous time hidden Markov model on serum biomarkers for the early detection of hepatocellular carcinoma 165
Sina Mews, Roland Langrock, Ruth King, Julia Schemm, Irina Janzen, Nicola Quick: A continuous-time capture-recapture model for annual movements of bottlenose dolphins ... 169
Andrea Sottosanti, Mauro Bernardi, Luis Campos, Aneta Siemiginowska and David van Dyk: Continuous time hidden Markov models for astronomical gamma-ray light curves 175
Fiona Steele, Emily Grundy: Random effects dynamic panel models for unequally-spaced responses 180

Vanda Inácio de Carvalho, María Xosé Rodríguez–Álvarez , Nadja Klein : Density regression via penalised splines dependent Dirichlet process mixture of normals models 184
Wenyu Wang, Ardo van den Hout: Non-parametric Frailty Models for Cardiac Allograft Vasculopathy Data 189
Brian D. Marx, Bin Li: Robust Penalized Signal Regression 194
Wende Clarence Safari, Ignacio López-de-Ullibarri, María Amalia Jácome: Nonparametric cure rate estimation when cure is partially known 200
Fernanda De Bastiani, Robert A. Rigby Dimitrios M. Stasinopoulos and Gillian Z. Heller: A skewness and kurtosis comparison for continuous distributions 204
Jacobo de Uña-Álvarez: The competing risks model with interval sampling 210
Thomas Kneib, Nadja Klein, Nikolaus Umlauf and Stefan Lang: Modular Regression – A Lego System for Building Structured Additive Distributional Regression Models with Tensor Product Interactions 214
Michael Lebacher, Göran Kauermann: Regression-based Network Reconstruction with Covariates and Random Effects 220
Lisa Schlosser, Moritz N. Lang, Torsten Hothorn, Georg J. Mayr, Reto Stauffer, Achim Zeileis: Distributional Trees for Circular Data 226
Nicolò Margaritella, Vanda Inácio De Carvalho, Ruth King: Bayesian functional PCA clustering with applications in neuroscience 232
Mátyás Constans, Attila Lovas, Péter Sótonyi, Brigitta Szilágyi: Non-parametric learning algorithm for evaluating the influence of environmental factors on sudden medical emergencies 236
Andreas Mayr, Leonie Weinhold, Stephanie Titze, Matthias Schmid: Boosting health-related quality of life via distributional beta regression 242
Pedro A. Morettin, Jhames M. Sampaio: Stable Randomized Generalized Autoregressive Conditional Heteroskedastic Models . 248
David Moriña, Juan M. Leyva-Moral, Maria Feijoo-Cid, Pedro Puig: Intervention analysis based on INAR models with applications in public health 254
Emmanuel Lesaffre, Luis Adrian Quintero, Geert Verbeke: Selecting the Number of Factors in Bayesian Factor Analysis 259
Xanthi Pedeli, Dimitris Karlis: Multivariate surveillance using a multivariate integer-valued time series model 263

Ridall, P. G., George, M., Pettitt A. N. : Fast sequential Bayesian analysis of football scores illustrating the evolution of the styles and strengths of each of the British premier league football sides over two decades 267
Jeffrey S. Simonoff, Ningshan Zhang: Joint latent class trees: A tree-based approach to joint modeling of time-to-event and longitudinal data 273
Srimanti Dutta, Arindom Chakraborty: Two-step method for Joint Models of Longitudinal and Time-to-event Data 279
Mauro Bernardi, Daniele Durante, Paola Stolfi: Bayesian Probit Classification Trees 283
de Carvalho, V. I., Lourenço, V. M., de Carvalho, M.: Robust inference for ROC regression 287
Janet van Niekerk, Haakon Bakka and Håvard Rue: A cohesive Bayesian approach to competing risks models 291
Paul Wiemann, Thomas Kneib: Studying the Softplus Function as a Response Function in Regression Models 295
Gianluca Sottile, Vito MR Muggeo: Non-crossing quantile regression via monotone B-spline varying coefficients 301
Santhosh Narayanan, Ioannis Kosmidis, Petros Dellaportas: Flexible multivariate point processes with applications to modelling football matches 306
Inês Sousa, Adriana Vieira, Luis Castro: Longitudinal models with informative time measurements 312
Pedro Puig , David Moriña, Isabel Serra, Álvaro Corral: Estimation of the probability of a giant 'doomsday' solar geomagnetic storm 316
Zhen Chen, Soutik Ghosal: Diagnostic accuracy of ultrasound measures on large for gestational age: a Bayesian regression model for ROC curves with constraints 320
Rafael Pimentel Maia, Clarice Garcia Borges Demétrio, Rodrigo Labouriau: A discrete competing risks mixed model with masked causes: a cow longevity study 324

Part I - Invited Papers

Longitudinal modelling of leptospirosis prevalence using serial dilution assay data

Peter J Diggle[1], Katharine A Owers[2], Max Eyre[1]

[1] Lancaster University, UK
[2] Colorado State University, USA

E-mail for correspondence: p.diggle@lancaster.ac.uk

Abstract: The standard method for identifying sub-clinical leptospirosis, a disease that is transmitted primarily through environmental contact with infected rat urine, is a from Pau da Lima cohort study serial dilution assay. The assay delivers an interval-censored measure of an individual's antibody response. we describe a longitudinal case-study of leptospirosis in a Brazilian favela community, in which the result of a serial dilution assay is the response variable. We distinguish between questions that, in the authors' opinion, can and cannot be answered using standard methods of longitudinal data analysis

Keywords: leptospirosis; longitudinal data; interval censoring.

1 Problem statement

Leptospirosis is a disease of increasing concern in urban slum environments, where its principal mode of transmission is through contact with infected rat urine in the environment. The standard method for identifying sub-clinical infection with leptopsires is a serial dilution assay, which in essence gives an interval-censored measure of an individual's antibody response.
Here, we describe an ongoing longitudinal study of leptospirosis in a Brazilian favela community with a particular focus on the following questions: does a past infection confer partial immunity to future infection?
This extended conference abstract draws heavily on Diggle (2018).

2 Study-design

The microscopic agglutination test is the gold standard serodiagnostic assay for leptospirosis. It is conducted by combining serial dilutions of the sample

of interest with reference strains of Leptospira bacteria, the causative agent of leptospirosis. The mixture is examined under darkfield microscopy to determine if at least 50% of the bacteria are agglutinated by the diluted sample. The result of the assay is given as the highest dilution at which this 50% agglutination threshold was reached.
A prospective cohort study was carried out from February 2013 to August 2017 in Pau da Lima, a vulnerable urban community in Salvador, Brazil. It aimed to detect Leptospira-specific antibodies in individuals living in the community. Individuals living in randomly sampled households were eligible for enrolment in the study if they were aged five years or over and slept for two or more nights in the household each week. A total of 4,441 individuals were enrolled, of which 2,711 were successfully followed up at least once and 447 were available for the entire study period. Follow-up consisted of a serosurvey conducted every six months by team members visiting the households of all study participants. Serological evaluation was then performed using the microscopic agglutination test (MAT) to determine titres of agglutinating antibodies against a a panel of five reference strains and two clinical isolates.

3 Exponential decay of the antibody response absent re-infection

In a serial dilution assay, a blood-sample is tested against a standardised challenge. If the result is negative, the antibody response, W say, is recorded as "below detection limit." If the result is positive, the blood-sample is diluted by a known factor and the test is repeated, using repeated dilutions until a negative result is obtained. This generates an integer response, $K = 0, 1, 2, ...$, the number of dilutions required to return a negative result. The value of K corresponds to an interval-censored version of W, i.e. $K = k$ if and only if $ckd \leq W \leq c(k+1)d$, where c is the detection limit and d the dilution factor. Although this interpretation is rarely used explicitly, it is implicit in the standard practice of declaring the occurrence of an infection event at some time during the follow-up interval in question if either there is an increase of at least two in successive values of K, or a below detection limit result is followed by a positive. This practice converts a series of serial dilution assays at times $t_j : j = 0, 1, ..., n$ to a series of binary responses, $Y_j : j = 1, ..., n$ indicating whether or not the individual concerned has experienced an infection event in the time-interval (t_{j-1}, t_j).
In an unpublished Yale University PhD Thesis, the second author has used data from a single-source outbreak of leptospirosis reported in Lupidi et al (1991) to fit a model in which, absent re-infection, antibody concentrations decay exponentially over time, and estimates the exponential decay rate parameter, ϕ, by maximum likelihood, giving $\hat{\phi} = 0.926$ with 95% confidence interval $(0.918, 0.934)$.

4 Does a past infection confer partial immunity to future infection?

A simple, and superficially attractive way to answer this question is to use a person's antibody response at time t_{j-1} as a covariate of their binary response Y_j at time t_j. Partial immunity would then be indicated by a negative regression coefficient. But the reasoning behind this is flawed. Because a re-infection at some time between consecutive follow-up times is declared when the serial dilution assay response K increases by two or more, the re-infection response at time t_j cannot be independent of the serial dilution assay response at time t_{j-1}.

4.1 Process model

Let $W(t)$ denote the latent antibody response of an individual at time $t \geq 0$; for an infection-naive individual, $W_i(0) = 0$. Assume that $W(t)$ jumps by random iid amounts V_j at infection times t_j and decays exponentially at a rate ϕ between infection events. Finally, assume that infection events follow a Poisson process with intensity

$$\lambda(t) = \exp\{\alpha(t) + W(t)\beta\} \tag{1}$$

The parameter of interest is β; a negative value indicates that infection events confer partial immunity to future infections. Figure 1) shows a realisation of $W(t)$ under model (1), with constant $\alpha(t) = -1$, exponential decay factor 0.1 per month, $\beta = -3$ and iid V_j following exponential distributions with mean 1.

4.2 Data model

The observed response from an individual is the sequence of values of K_j at a set of pre-specified follow-times $t_j : j = 1, ..., n$. Each K_j is an interval-censored version of $\alpha W(t_j)$, where α is an unknown constant of proportionality; an observation $K_j = 0$ corresponds to $\alpha W(t_j) < c$, where c is the detection limit of the assay, whilst $K_j = k > 0$ corresponds to $c \times 2^{k-1} < \alpha W(t_j) < c \times 2^k$.

5 Work-in-progress

At the time of writing, we are searching the literature for other examples of single-point-of-exposure data on the longitudinal progression of leptospirosis antibody response in human blood-samples, to establish the exponential decay assumption holds robustly, to an acceptable level of approximation. We also plan to conduct theoretical studies to understand what sample

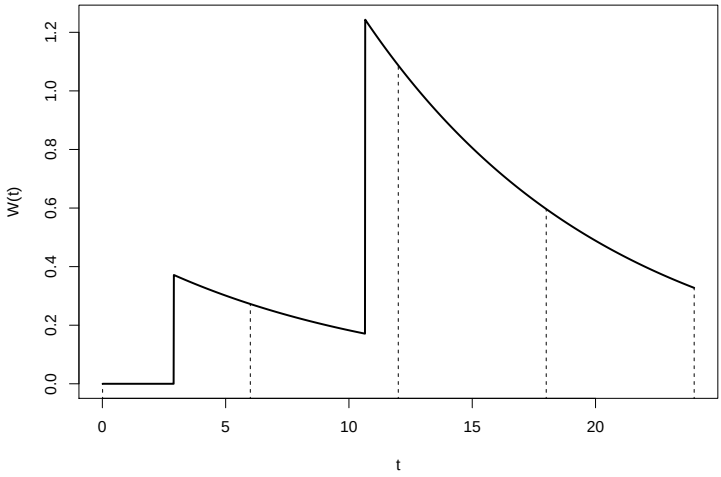


FIGURE 1. A simulated realisation of model (1) for the time-evolution of an individual's antibody response over 24 months, with follow-up times indicated at 0, 6, 12, 18 and 24 months.

size and frequency of follow-up would be needed in order to obtain usefully precise parameter estimates for the model described in Section 4. Finally, and depending on the outcome of these theoretical studies, we will fit the model to an ongoing cohort study with six-month follow-up of approximately NNN inhabitants of SOMEWHERE.

Acknowledgments: We thank the National Institutes of Health, USA, and the Medical Research Council, UK, for funding support.

References

Diggle, P.J. (2018). Analyse problems, not data. *Spatial Statistics*, **28**, 4–7. doi 10.1016/j.spasta.2018.10.003

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

Hagan, J.E., Moraga, P., Costa, F., Capian, N., Ribeiro, G.S., Wunder, E.A., Felzemburgh, R.D.M., Reis, R.B., Nery, N., Santana, F.S., Fraga, D., dos Santos, B.L., Santos, A.C., Queiroz, A.,Tassinari, W., Carvalho, M.A., Reis, M.G., Diggle, P.J. and Ko, A.I. (2016). Spatiotemporal

determinants of urban leptospirosis transmission: Four-year prospective cohort study of slum residents in Brazil. *Public Library of Science: Neglected Tropical Diseases*, **10**, pp.e0004275

Lupidi, R., Cinco, M., Balanzin, D., Delprete, E. and Varaldo, P.E. (1991). Serological follow-up of patients involved in a localized outbreak of leptospirosis. *Journal of Clinical Microbiology*, **29**, 805–809.

Inference based on Laplace approximations in nonparametric additive location-scale model for right- or interval-censored data

Philippe Lambert[1][2]

[1] Institut de Recherches en Sciences Sociales, Université de Liège, Belgium
[2] ISBA, Université catholique de Louvain, Belgium

E-mail for correspondence: p.lambert@uliege.be

Abstract: In a previous publication on Nonparametric additive location-scale models for interval censored data (Lambert 2013), we explained how P-splines could be used in regression models to specify a smooth error density and the joint (possibly) nonlinear effects of covariates on location and dispersion. That methodology extends traditional additive regression models by releasing the parametric constraint on the error distribution and by acknowledging that covariates can affect multiple aspects of the conditional distribution in a non trivial way. These extensions are very attractive and practically useful, but have an important computational cost following from the use of the Metropolis-within-Gibbs algorithm in a richly parameterized model. By extending the results in Gressani & Lambert (2018), we show how Laplace based approximations to the marginal posterior distributions of smoothness parameters can be used to set up a quickly converging iterative algorithm to select penalty parameters and to estimate the spline parameters in the pivotal distribution and in the additive components for location and dispersion. Simulations suggest that the so-obtained estimators have excellent frequentist properties. They can be also be combined in a Bayesian setting to select starting values and proposal distributions in a Metropolis-within-Gibbs algorithm (Gressani & Lambert, 2019).
We illustrate the methodology on various datasets involving different forms of censoring on the response. We also investigate how that strategy can be adapted to analyze survival data with an unknown cured fraction (Lambert & Bremhorst, 2019).

Keywords: Nonparametric additive model ; Location-scale model ; P-splines ; Laplace approximation.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

1 Additive location-scale model

Consider a vector $(Y, \mathbf{z}, \mathbf{x})$ where Y is a univariate continuous response, $\mathbf{z}$ a $p-$vector of categorical covariates, and $\mathbf{x}$ a $J-$vector of quantitative covariates. The response could be right- or interval-censored
Such settings are not only common in survival analysis when studying the time elapsed between a clearly defined time origin and an event of interest, but also in surveys when the respondent reports a quantitive response by pointing one interval or semi-interval in the partition of the variable support. We assume the following location-scale model,

$$Y = \mu(\mathbf{z}, \mathbf{x}) + \sigma(\mathbf{z}, \mathbf{x})\varepsilon$$

where ε is independent of $\mathbf{z}$ and $\mathbf{x}$ with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = 1$. Other constraints based on quantiles are possible, see Lambert (2013).
Assume that independent copies $(y_i, \mathbf{z}_i, \mathbf{x}_i)$ $(i = 1, \ldots, n)$ are observed on n units with the possibility of right– or interval-censoring as described above. We consider additive models for the conditional location and dispersion of the response:

$$\big(\mu(\mathbf{z}_i, \mathbf{x}_i)\big)_{i=1}^n = \mathbf{Z}\beta + \sum_{j=1}^J \mathbf{f}_j^\mu \quad ; \quad \big(\log\sigma(\mathbf{z}_i, \mathbf{x}_i)\big)_{i=1}^n = \mathbf{Z}\delta + \sum_{j=1}^J \mathbf{f}_j^\sigma$$

where $f_j^\mu(\cdot)$ and $f_j^\sigma(\cdot)$ denote smooth additive terms quantifying the effect of the jth quantitative covariate (rescaled and recentered to take values in $(0, 1)$). Consider now a basis of cubic B-splines associated to equally spaced knots on $(0, 1)$ and recentered for identification purposes. Then, the additive terms in the conditional location and dispersion models can be approximated using linear combinations of these (recentered) B-splines,

$$\big(\mu_i = \mu(\mathbf{z}_i, \mathbf{x}_i)\big)_{i=1}^n = \mathcal{X}\boldsymbol{\Psi}^\mu \quad ; \quad \big(\sigma_i = \sigma(\mathbf{z}_i, \mathbf{x}_i)\big)_{i=1}^n = \exp\big(\mathcal{X}\boldsymbol{\Psi}^\sigma\big)$$

with design matrix $\mathcal{X} = [\mathbf{Z}, \mathbf{S}_1, \ldots, \mathbf{S}_J] = [\mathbf{Z}, \mathcal{S}]$; matrices of spline parameters (with one column per additive term) $\boldsymbol{\Theta}^\mu = [\theta_1^\mu, \ldots, \theta_J^\mu]$, $\boldsymbol{\Theta}^\sigma = [\theta_1^\sigma, \ldots, \theta_J^\sigma]$; vectors of (stacked) regression parameters $\boldsymbol{\Psi}^\mu = \big(\beta, \mathrm{Vec}(\boldsymbol{\Theta}^\mu)\big)$, $\boldsymbol{\Psi}^\sigma = \big(\delta, \mathrm{Vec}(\boldsymbol{\Theta}^\sigma)\big)$.

2 Penalized log-likelihood and joint posterior

Estimation of the regression parameters and of the additive terms (for given penalty parameters) can be made from the penalized log-likelihood. Denote standardized residuals by $r_i = (y_i - \mu_i)/\sigma_i$. If $f_\epsilon(\cdot)$ and $S_\epsilon(\cdot)$ denote the density and survival function of the error term ϵ, then the contribution of unit i to the log-likelihood is $\ell_i = -\log\sigma_i + \log f_\epsilon(r_i)$ when uncensored, $\ell_i = \log S_\epsilon(r_i)$ when right-censored, and $\ell_i = \log\big(S(r_i^L) - S(r_i^R)\big)$

when interval-censored. Smoothness of the additive terms can be tuned by penalizing changes in differences of neighbour spline parameters (Eilers, 1996), yielding the penalized log-likelihood

$$\ell_p = \ell(\boldsymbol{\psi}^\mu, \boldsymbol{\psi}^\sigma; \mathcal{D}) - \frac{1}{2}\sum_{j=1}^{J} {\boldsymbol{\theta}_j^\mu}^\top (\lambda_j^\mu \mathbf{P}^\mu) \boldsymbol{\theta}_j^\mu - \frac{1}{2}\sum_{j=1}^{J} {\boldsymbol{\theta}_j^\sigma}^\top (\lambda_j^\sigma \mathbf{P}^\sigma) \boldsymbol{\theta}_j^\sigma .$$

In a Bayesian framework, similar penalties arise through the specification of conditional priors for the spline parameters, yielding for the jth additive terms in the location and dispersion sub-models,

$$p(\boldsymbol{\theta}_j^\mu | \lambda_j^\mu) \propto \exp\left(-\frac{1}{2}\, {\boldsymbol{\theta}_j^\mu}^\top (\lambda_j^\mu \mathbf{P}^\mu) \boldsymbol{\theta}_j^\mu\right) \;;\; p(\boldsymbol{\theta}_j^\sigma | \lambda_j^\sigma) \propto \exp\left(-\frac{1}{2}\, {\boldsymbol{\theta}_j^\sigma}^\top (\lambda_j^\sigma \mathbf{P}^\sigma) \boldsymbol{\theta}_j^\sigma\right).$$

Assuming joint Normal priors for the intercepts and the regression parameters associated to the other covariates $\mathbf{z}$ induce conditional Gaussian Markov random fields (GMRF) (Rue & Held, 2005) for the joint priors for the regression and spline parameters in $\boldsymbol{\psi}^\mu$ and $\boldsymbol{\psi}^\sigma$. Then the joint posterior for the whole set of parameters in the additive location-scale model is

$$p(\boldsymbol{\psi}^\mu, \boldsymbol{\psi}^\sigma, \boldsymbol{\lambda}^\mu, \boldsymbol{\lambda}^\sigma | \mathcal{D}) \propto L(\boldsymbol{\psi}^\mu, \boldsymbol{\psi}^\sigma; \mathcal{D})\; p(\boldsymbol{\psi}^\mu | \boldsymbol{\lambda}^\mu)\; p(\boldsymbol{\psi}^\sigma | \boldsymbol{\lambda}^\sigma)\; p(\boldsymbol{\lambda}^\mu)\; p(\boldsymbol{\lambda}^\sigma)\;. \quad (1)$$

3 Estimation and selection of $(\boldsymbol{\psi^\mu}, \boldsymbol{\lambda^\mu})$ and $(\boldsymbol{\psi^\sigma}, \boldsymbol{\lambda^\sigma})$

Let $\boldsymbol{\psi} = (\boldsymbol{\psi}^\mu, \boldsymbol{\psi}^\sigma)$ and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}^\mu, \boldsymbol{\lambda}^\sigma)$. Starting from the joint posterior for the model parameters, we have the following identity for the marginal posterior of the penalty parameters: $p(\boldsymbol{\lambda}|\mathcal{D}) = p(\boldsymbol{\psi}, \boldsymbol{\lambda}|\mathcal{D})/p(\boldsymbol{\psi}|\boldsymbol{\lambda}, \mathcal{D})$. Given the conditional GMRF prior for $\boldsymbol{\psi}$, we conclude that the conditional posterior in the denominator is approximately Gaussian (Rue & Martino, 2009). Using the Laplace's method, we have $(\boldsymbol{\psi}|\boldsymbol{\lambda}, \mathcal{D}) \;\dot{\sim}\; \mathcal{N}(\hat{\boldsymbol{\psi}}_\lambda, \Sigma_\lambda)$ where $\hat{\boldsymbol{\psi}}_\lambda$ can be obtained using a Newton-Raphson algorithm on (1). Substituting that approximation in the preceding identity, we obtain the following approximation to $p(\boldsymbol{\lambda}|\mathcal{D})$:

$$\tilde{p}(\boldsymbol{\lambda}|\mathcal{D}) \propto p(\hat{\boldsymbol{\psi}}_\lambda, \boldsymbol{\lambda}|\mathcal{D})\; \left|\Sigma_\lambda^{-1}\right|^{-1/2}$$

where an explicit expression for Σ_λ^{-1} is available whatever the censoring status of the data. The maximization of $\tilde{p}(\boldsymbol{\lambda}|\mathcal{D})$ using a Newton-Raphson type algorithm enables a joint selection of the penalty parameters $\hat{\boldsymbol{\lambda}}^\mu$ and $\hat{\boldsymbol{\lambda}}^\sigma$, with as a by-product, an estimation of the regression parameters $\hat{\boldsymbol{\psi}}_\lambda^\mu$ and $\hat{\boldsymbol{\psi}}_\lambda^\sigma$.

4 Nonparametric pivotal density

Parametric or nonparametric choices can be made for the error distribution. Here, we write the underlying hazard $h_\epsilon(\cdot)$ using a linear combination of B-splines, $\log h_\epsilon(r) = \sum_{k=1}^{K} b_k(r)\phi_k$, associated to an equidistant grid of knots on the support of the error distribution. Given the constraints $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{V}(\varepsilon) = 1$, one can practically assume (using Chebyshev's theorem) that (most of) the support is on $(r_{\min}, r_{\max}) = (-6, 6)$, say. Again, a GMRF prior is assumed for $(\boldsymbol{\phi}|\tau)$ with penalty parameter τ to tune the hazard smoothness.
Given (possibly right- or interval-censored) conditionally independent standardized residuals $r_i = \big(y_i - \mu_i(\boldsymbol{\psi}^\mu)\big)/\sigma_i(\boldsymbol{\psi}^\sigma)$, we developed a fast algorithm for the computation of the posterior mode $\hat{\boldsymbol{\phi}}_\tau$ of $(\boldsymbol{\phi}|\tau, \mathcal{D})$ with moment constraints on the underlying distribution of ε. The selection of the penalty parameter also relies on the posterior mode of a Laplace based approximation to $p(\tau|\mathcal{D})$.

5 Fitting the NP additive location-scale model

We now have all the necessary ingredients for fitting the nonparametric additive location-scale model from possibly right- or even interval-censored data. The algorithm is iterative and alternates the estimation of the error density (Step 1, see Section 4), of the regression and spline parameters in the location (Step 2) and dispersion (Step 3) submodels, selection of the penalty parameters for the additive terms in location (Step 4) and dispersion (Step 5), see Section 3.
Simulations suggest excellent properties of the so-defined estimators. The procedure is extremely fast even with pure R code. Extensions of that model will be presented and discussed during the oral presentation. Several illustrations will also be provided at that occasion.

6 Application

Here, we propose an example with interval- and right-censored responses. The data of interest come from the European Social Survey (ESS) 2016. We focus on the money available per person in Belgian households for respondents aged 25-55 when the main source of income comes from wages or salaries ($n = 756$). Each person reports the net monthly income of the household in one of 10 decile-based intervals: 1: < 1.120 2: $[1.120, 1.400[$, 3: $[1.400, 1.720[$, 4: $[1.720, 2.100[$, 5: $[2.100, 2.520[$, 6: $[2.520, 3.060[$, 7: $[3.060, 3.740[$, 8: $[3.740, 4.530[$, 9: $[4.530, 5.580[$, 10: ≥ 5.580 euros.
We model the relation between the income per person to the availability of (at least) 2 salaries (64.2%) in the household, the age (41.0 ± 8.83 years)

and the number of years of full-time education completed (14.9±3.34 years) by the respondent.

```
Fixed effects for Location:
             est    se   low    up
Intcp      1.580 0.033 1.515 1.645
twoincomes 0.263 0.041 0.182 0.344

Fixed effects for Dispersion:
              est    se    low     up
Intcp      -0.508 0.047 -0.599 -0.417
twoincomes -0.025 0.058 -0.139  0.088
-----------------------------------------------------------------
Effective dimensions for the 2 additive terms in Location:
Age       3.8    Educ      3.9

Effective dimensions for the 2 additive terms in Dispersion:
Age       2.4    Educ      4.3

10 B-splines per additive component in location and dispersion
20 B-splines for the error log-hazard on (-5,10)
-----------------------------------------------------------------
Total sample size: 756 ; Confidence level for CI: 0.95
Uncensored data: 0 (0 percents)
Interval Censored data: 691 (91.4 percents)
Right censored data: 65 (8.6 percents)
Elapsed time: 0.9 seconds  (9 iterations)
```

References

Eilers, P.H.C and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89 – 102.

Gressani, O. and Lambert P. (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. *Computational Statistics and Data Analysis*, **124**, 151 – 167.

Gressani, O. and Lambert P. (2019). Approximate Bayesian inference in generalized additive models with penalized splines. *Working paper (submitted)*.

Lambert, P. (2013). Nonparametric additive location-scale models for interval censored data. *Statistics and Computing*, **23**, 75 – 90.

Lambert, P. and Bremhorst, V. (2019). Estimation and identification issues in the promotion time cure model when the same covariates influence long- and short-term survival. *Biometrical Journal*, **61**: 275 – 279.

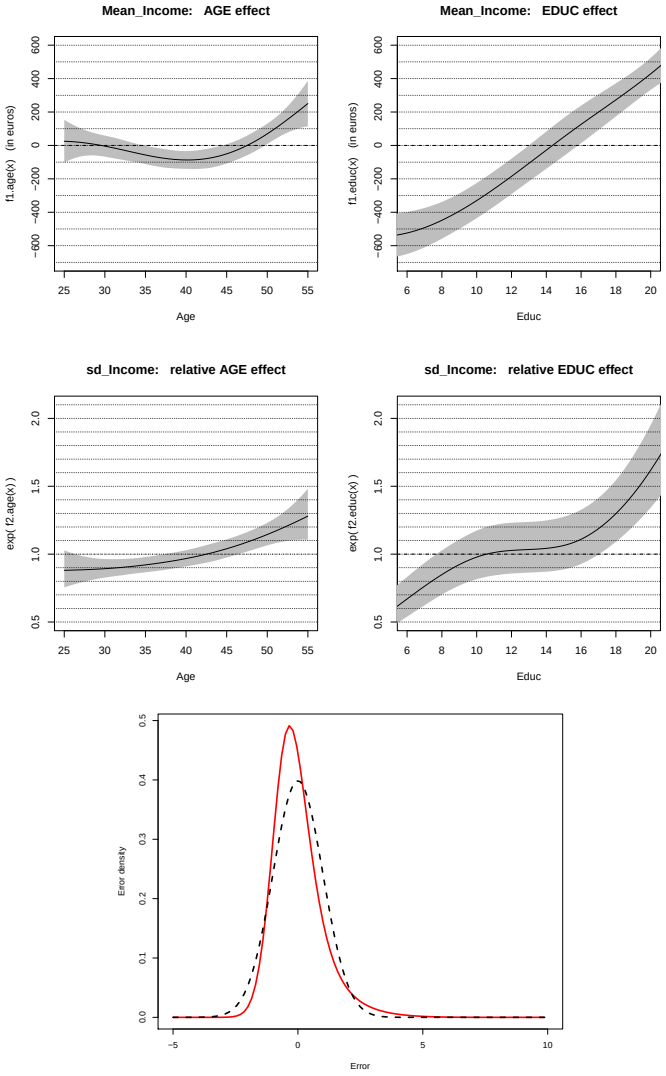


FIGURE 1. NP additive location-scale model fitted on interval- or right-censored income data (ESS 2016, Belgian data).

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall.

Rue H. and Martino, S. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *JRSS-B*, **71**(2), 319 – 392.

Random changepoint segmented regression with smooth transition: an example with lateral amyotrophic sclerosis data

Julio M. Singer[1], Francisco M.M. Rocha[2], Antonio Carlos Pedroso-de-Lima[1], Giovani L. Silva[3], Giuliana C. Coatti[1], Mayana Zatz[1]

[1] Universidade de São Paulo, Brazil
[2] Universidade Federal de São Paulo, Brazil
[3] Universidade de Lisboa, Portugal

E-mail for correspondence: jmsinger@ime.usp.br

Abstract: We consider random changepoint segmented regression models to analyze data from a study conducted to verify whether treatment with stem cells may delay the onset of a symptom of amyotrophic lateral sclerosis in genetically modified mice. The proposed models capture the biological aspects of the data, accommodating a smooth transition between the periods with and without symptoms. An additional changepoint is considered to avoid negative predicted responses. Given the nonlinear nature of the model, we adapt an algorithm proposed by Muggeo et al. (2014) to estimate the fixed parameters and to predict the random effects by fitting linear mixed models at each step. We compare the fixed parameters parameters of the mixed model to averages of parameters obtained by fitting individual models.

Keywords: amyotrophic lateral sclerosis, fitting algorithm, mixed models.

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) is one of the most common adult-onset motor neuron diseases causing a progressive, rapid and irreversible degeneration of motor neurons in the cortex, brain stem and spinal cord. No effective treatment is available and cell therapy clinical trials are currently being tested in ALS affected patients. Mutations in the SOD1 gene represent one of the most frequent causes of ALS.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

Among the different animal models for ALS, SOD1 mice are the most used in preclinical studies. After the initial tremor in the limbs, they develop muscle weakness in early adulthood, become fully paralyzed and die. These mice overexpress the human SOD1 gene bearing the G93A mutation. Interestingly, in this animal model the disease progression is different between the genders as observed in ALS patients. Males have a shorter lifespan and a clinical condition apparently more severe than females and differences in electrophysiological parameters have also been reported.
Treatment of ALS with stem cells is a current research topic. Mesenchymal stromal cells (MSC), specially those derived from adipose tissues, and pericytes have been used in studies that focus on the reduction of the speed of the progression of symptoms of neurodegenerative diseases. In this context we consider a study conducted in the Human Genome and Stem Cell Research Center, at the Biosciences Institute, University of São Paulo, Brazil with the objective of comparing MSC cells and pericytes injected in SOD1-G93A mice with respect to their effects on the evolution of some symptoms of ALS. For details, see Coatti et al. (2017). Our objective is to propose models for the statistical analysis of the data.

2 The study

A set of 34 female and 21 male 8 week old SOD1-G93A mice was divided into 3 groups. Animals in the first group (12 females and 7 males) were submitted to weekly injections of MSC cells, those in second group (11 females and 8 males), to injection with pericytes while animals in the third group (11 females and 6 males) were submitted to the vehicle (*Hank's balanced salt solution* - HBSS). All animals were followed weekly up to their death for clinical analysis of the progression of the disease by means of four variables, the analysis of one of them, *rotarod* is considered in this study. The *rotarod* test was used to evaluate motor coordination and fatigue resistance. For that purpose, the length of time each animal could remain in the rotating cylinder of a *rotarod* apparatus was recorded. The specific objectives of the analysis are:

i) Identification of the moment when animals become symptomatic in each of the six groups defined by the combination of treatment (HBSS, MSC, Pericytes) and sex (male, female).

ii) Estimation of the expected rate of variation in response after symptom onset in each group.

iii) Evaluation of the effects of treatment, sex and their interaction on the expected moment of symptom onset and post onset rate of variation in the expected response.

3 Statistical analysis

Profile plots for the response along with LOESS curves are displayed in Figure 1. An analysis of the behaviour of the response variable corroborates

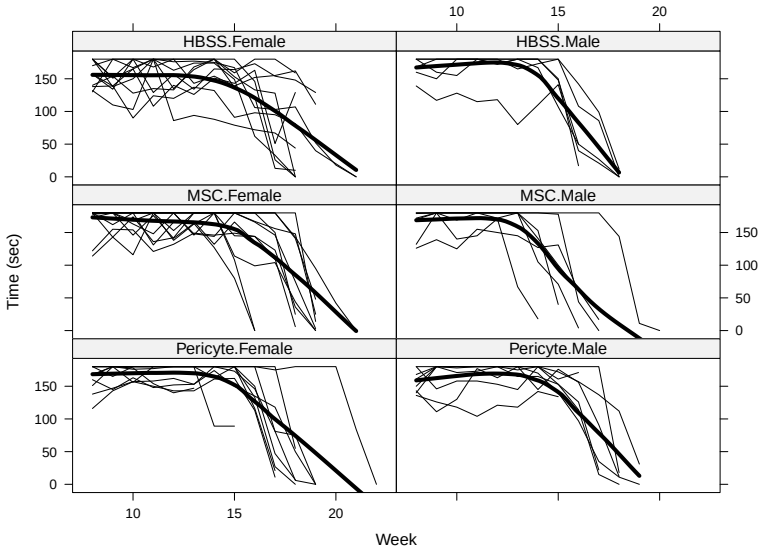


FIGURE 1. Profile plots for the response along with LOESS curves

its expected stable level before the onset of the symptom (a decrease in the length of time during which the animal remains in the rotating cylinder). Furthermore, individual differences in the moment where this occurs as well as differences in the speed with which the intensity of the symptom progresses are also visible. It also seems reasonable to expect a change in the acceleration with which the intensity of the symptom progresses after the disease onset.

Given that such conclusions are in line with the expected biological behaviour, a random changepoint polynomial segmented regression model may be considered for the analysis.

Such models have an attractive practical appeal in many fields and have been the object of statistical research for a long time as detailed in Muggeo et al. (2014). These authors consider a frequentist approach as opposed to the common Bayesian perspective usually employed in the statistical literature.

Keeping in mind the nonnegative nature of the response, we adopt a similar approach and consider an analysis of the ALS data based on the model

$$y_{ijk} = \alpha_{ij} I(t_k < \psi_{2ij}) + \gamma_{ij}[t_k - \psi_{1ij}(\lambda_{ij})]^2 I(\psi_{1ij} \leq t_k < \psi_{2ij}) + e_{ijk} \quad (1)$$

($i = 1, \ldots, 6$, $j = 1, \ldots, n_i$ and $k = 1, \ldots, n_{ij}$) where y_{ijk} denotes the response for the j-th animal observed in the i-th group (defined by the combination of the levels of treatment and sex) at the k-th evaluation instant, α_{ij} is the corresponding stable level of the symptom prior to the first changepoint, γ_{ij} is the coefficient of the quadratic term for the curve that governs the expected response behaviour post changepoint ψ_{1ij}, with

$$\psi_{1ij}(\lambda_{ij}) = [L_1 + L_2 \exp(\lambda_{ij})]/[1 + \exp(\lambda_{ij})]$$

to restrict its value to the interval (L_1, L_2) in which the observations are obtained and ψ_{2ij} denotes the instant where the expected response is null. We assume that $\alpha_{ij} = \alpha_i + a_{ij}$, $\gamma_{ij} = \gamma_i + c_{ij}$, $\lambda_{ij} = \lambda_i + \ell_{ij}$ with $\mathbf{b}_{ij} = (a_{ij}, c_{ij}, \ell_{ij})^\top \sim N(\mathbf{0}, \mathbf{G}_i)$ and $e_{ijk} \sim N(0, \sigma_i^2)$ independent of $\mathbf{b}_{ij}$.
This is an extension of the models proposed by Muggeo et al. (2014) where a smooth transition and a second changepoint are incorporated. For simplicity, we drop the subscript i to specify the the fitting algorithm.
Given that ψ_{2j} corresponds to the instant t_k where $\mathrm{E}(y_{jk}) = 0$, we have $I(t_k < \psi_{2j}) = 1$ and $I(\psi_{1j} \leq t_k < \psi_{2j}) = 1$ and consequently, that $\alpha_j + \{\gamma_j[\psi_{2j} - \psi_{1j}(\lambda_j)]^2\} = 0$, implying that

$$\psi_{2j} = \psi_{2j}(\alpha_j, \gamma_j, \psi_{1j}) = \sqrt{-\alpha_j/\gamma_j} + \psi_{1j}(\lambda_j)$$

Following Muggeo et al. (2014) and Fasola et al. (2018), the model, which is nonlinear, may be approximated by a first order Taylor expansion of

$$f[t_k, \gamma_j, \psi_{1j}(\lambda_j)] = \gamma_j[t_k - \psi_{1j}(\lambda_j)]^2 I(\psi_{1j} \leq t_k < \psi_{2j}).$$

Explicitly,

$$f[t_k, \gamma_j, \psi_{1j}(\lambda_j)] \approx f[t_k, \gamma_j, \psi_{1j}(\widehat{\lambda}_j)] + (\lambda_j - \widehat{\lambda}_j) \frac{\partial f[t_k, \gamma_j, \psi_{1j}]}{\partial \psi_{1j}} \frac{\partial \psi_{1j}(\lambda_j)}{\lambda_j}\Big|_{\lambda_j = \widehat{\lambda}_j}$$

with

$$\frac{\partial f[t_k, \gamma_j, \psi_{1j}]}{\partial \psi_{1j}} = h_j(\lambda_j) = 2\gamma_j[t_k - \psi_{1j}(\lambda_j)] I[\psi_{1j}(\lambda_j) \leq t_k < \psi_{2j}(\lambda_j)]$$

and

$$\frac{\partial \psi_{1j}(\lambda_j)}{\partial \lambda_j} = g_j(\lambda_j) = \frac{(L_2 - L_1)\exp(\lambda_j)}{[1 + \exp(\lambda_j)]^2}.$$

Consequently we may approximate model (1) by

$$y_{jk} \approx \alpha_j I[t_k < \psi_{2j}(\widehat{\lambda}_j)] + f[t_k, \gamma_j, \psi_{1j}(\widehat{\lambda}_j)] - \widehat{\lambda}_j h_j(\widehat{\lambda}_j) g_j(\widehat{\lambda}_j) + \lambda_j h_j(\widehat{\lambda}_j) g_j(\widehat{\lambda}_j) + e_{jk}. \quad (2)$$

Considering the pseudo observations defined by $y^*_{jk} = y_{jk} + \widehat{\lambda}_j h_j(\widehat{\lambda}_j) g_j(\widehat{\lambda}_j)$, the model

$$y^*_{jk} = \alpha_j I[t_k < \psi_{2j}(\widehat{\lambda}_j)] + f[t_k, \gamma_j, \psi_{1j}(\widehat{\lambda}_j)] + \lambda_j h_j(\widehat{\lambda}_j) g_j(\widehat{\lambda}_j) + e_{jk}$$

suggests the following algorithm to fit (1)

1) Let $\psi_{1j}^{(0)} = \psi_1^{(0)}$ and $\psi_{2j}^{(0)} = \psi_2^{(0)}$.

2) Fit model $y_{jk} = \alpha_j I(t_k < \psi_{2j}^0) + \gamma_j (t_k - \psi_{2j}^{(0)})^2 I(\psi_{1j}^{(0)} \leq t_k < \psi_{2j}^{(0)}) + e_{jk}$ to obtain $\alpha^{(0)}$, $a_j^{(0)}$, $\gamma^{(0)}$, $c_j^{(0)}$, $\lambda_j^{(0)} = \log[(\psi_{1j}^{(0)} - L_1)/(L_2 - \psi_{1j}^{(0)})]$ and $\psi_{2j}^{(1)} = \sqrt{-\alpha_j^{(0)}/\gamma_j^{(0)}} + \psi_{1j}^{(0)}$.

3) Let $r = 1$.

4) Compute $y_{jk}^{(r)} = y_{jk} + \lambda_j^{(r-1)} h_j(\lambda_j^{(r-1)}) g_j(\lambda_j^{(r-1)})$.

5) Fit model $y_{jk}^{(r)} = \alpha_j I(t_k < \psi_{2j}^{(r)}) + \gamma_j [t_k - \psi_{1j}^{(r)}]^2 I(\psi_{1j}^{(r)} \leq t_k < \psi_{2j}^{(r)}) + \lambda_j h_j(\lambda_j^{(r-1)}) g_j(\lambda_j^{(r-1)}) + e_{jk}^{(r-1)}$ to obtain $\alpha^{(r)}$, $a_j^{(r)}$, $\gamma^{(r)}$, $c_j^{(r)}$, $\lambda^{(r)}$, $\ell_j^{(r)}$, $\psi_{1j}^{(r)} = [L_1 + L_2 \exp(\lambda_j^{(r)})]/[1 + \exp(\lambda_j^{(r)})]$ and $\psi_{2j}^{(r+1)} = \sqrt{-\alpha_j^{(r)}/\gamma_j^{(r)}} + \psi_{1j}^{(r)}$.

6) Stop if some convergence criterion is satisfied, otherwise, let $r = r + 1$ and repeat steps 4-6.

This algorithm essentially considers iterative fitting of standard linear mixed models by (restricted) maximum likelihood. At convergence, we expect a neglible difference between the third and fourth terms in the right hand side of (2) and as a consequence, that the pseudo observations should well approximate the original ones. Given the linear mixed model nature of the proposed fitting algorithm, we may employ the diagnostic procedures outlined in Singer et al. (2017) to check whether the adopted assumptions for the distribution of the random effects or of the random error are reasonable. The algorithm also provides the elements for the construction of approximate confidence intervals and Wald tests.

4 Results

Fitted profiles for the female mice submitted to the HBSS treatment are depicted in Figure 2.

A significant interaction between treatment and sex with respect to the ψ_1 changepoints ($\chi^2 = 12.96, df = 2, p = 0.002$) may be analysed via multiple comparisons and suggest that the onset of symptoms for the "typical" male in the control group (HBSS) is delayed by 1.9 [CI(95%) = 1.0, 2.9] weeks with respect to the corresponding "typical" female and that treatment with Pericytes (both sexes) or MSC (females) delay the onset of symptoms for the "typical" animals by 1.4 [CI(95%) = 0.6, 2.2] weeks with respect to the HBSS treated "typical" male. The changepoint for the MSC treated "typical" male lies between those for HBSS treated "typical" male and female but the small sample size does not lead to a significant difference in either case.

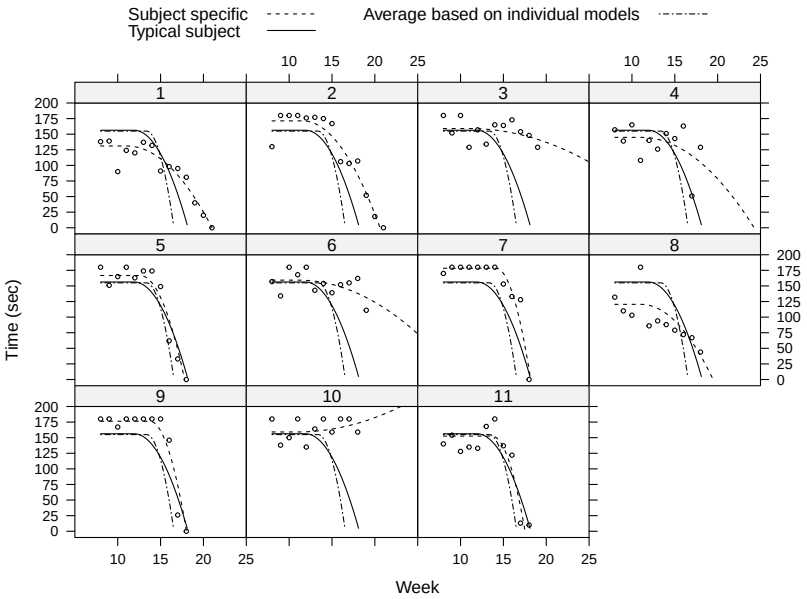


FIGURE 2. Fitted profile plots for HBSS treated females

Acknowledgments: This work received partial financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant 3304126/2015-2) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grant 2013/21728-2), Brazil.

References

Coatti, G.C. et al. (2017). Pericytes extend survival of ALS SOD1 mice and induce expression of antioxidant enzymes in the murine model and in IPSCs derived motor neurons from an ALS patient. *Stem Cell Reviews and Reports*, **13**, 686 – 698.

Fasola, S., Muggeo, V.M.R. and Küchenhoff, H. (2018). A heuristic, iterative algorithm for change-point detection in abrupt change models. *Computational Statistics*, **33**, 997 – 1015.

Muggeo, V.M.R., Atkins, D.C., Gallop, R.J. and Dimidjian, S. (2014). Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Statistical Modelling*, **14**, 293 – 313.

Singer, J.M., Rocha, F.M.M. and Nobre, J.S. (2017). Graphical tools for detecting departures from linear mixed models assumptions and some remedial measures. *International Statistical Review*, **85**, 290 – 324.

Calibration methods for spatial risk analysis

Maria Antónia Amaral Turkman[1], Kamil Feridun Turkman[1], Paula Pereira[13], Soraia Pereira[1], Miguel de Carvalho[12], Patrícia de Zea Bermudez[1]

[1] Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] School of Mathematics, University of Edinburgh, United Kingdom
[3] ESTSetúbal, Instituto Politécnico de Setúbal, Portugal

E-mail for correspondence: maturkman@fc.ul.pt

Abstract: In an environmental framework, extreme values of certain spatio-temporal processes, such as wind speeds, are the main cause of severe damage in property, such as electrical networks, road and agricultural infrastructures. Therefore availability of accurate data on such processes is highly important in risk analysis and in particular in producing probability maps showing the spatial distribution of damage risks. Typically, as is the case of wind speeds, data are available at few stations with many missing observations and consequently simulated data are often used to augment information, since simulated environmental data are typically available at high spatial and temporal resolutions. However, simulated data often mismatch observed data, particulary on tails, therefore calibrating and bringing it in line with observed data may offer practitioners more reliable and richer data sources. Since most damages are caused by extreme winds, it is particulary important to calibrate the right tail of simulated data based on observations. Response relationships between the extremes of simulated and observed data are by nature highly non-linear and non-Gaussian, therefore data fusion techniques available for spatial data may not be adequate for this purpose. Although, our ultimate goal is the development of statistical methods for data fusion and calibration that can extrapolate beyond the range of observed data—into the tails of a distribution—in this talk we will concentrate on calibration methods for the whole range of data. We will also explain how these new data fusion techniques for extremes of simulated and observed data may help in producing more accurate risk analysis in certain environmental problems.

Keywords: Bayesian hierarchical modelling; Calibration; Data fusion; Risk maps; Spatial extremes.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

1 Introduction and motivation

This paper overviews and highlights open challenges on calibration methods in a risk analysis context. A version of Berrocal *et al.* (2012) for non-Gaussian data is devised, and a regression model that extends Naveau *et al.* (2016) is discussed en passant; both methods allow for calibration of observed and simulated data yet via two distinct paradigms.

Our motivation has its roots in a consulting work several members of the research team did for one a major electricity producer and distributor. The electricity grid constantly faces disruptions due to damages in the distribution system, with heavy economic losses. These damages and consequent disruptions occur due to a combination of many factors such as topography and precipitation, however extreme winds and storms are the main cause of such damages. Risk maps that indicate likely places of costly disruptions in electric grids are important decision support tools for administering the power grid and are particularly useful in deciding if costly corrective actions should be taken to improve structures. It is natural that these risk maps should be based primarily on observed wind speeds among other factors. Hence, such risk maps can be interpreted as vulnerability maps of electricity grid to extreme wind speeds, expressed in terms of probability.

In producing such risk maps for damages, it is decided that daily maximum wind speeds should be used as proxy information. However generating such maps depends on reliable wind data at fairly high spatial and temporal resolutions. In Portugal, wind data exist in only 117 stations but missing observations reach to 90% in some stations. Naturally, this reduced number of observation sites does not give sufficient spatial coverage. On the other hand, simulated wind speeds from a simulator (the WRF—Weather Research and Forecast—model, version 3.1.1), obtained at a regular grid of 81ksq grid cell size are available without any missing observations, giving a high resolution spatial coverage. As often is the case, simulated and observed daily-maximum wind speed data, particularly at some stations do not match well. Consequently, corresponding probability maps based on simulated and observed daily-maximum wind speed data may differ. A common practice is to use simulated wind speeds after being 'calibrated', that is, after bringing the simulated wind speeds in line with observed wind speeds. There are many different definitions and consequently methods of calibration, which we briefly describe in Section 2. Although simulated and observed data seem to match reasonably well over the range of data, there are considerable differences on observations coming from the right tail and therefore the adequate calibration method should be particularly adopted to such observations and should be in line with the models and methods suggested by extreme value theory.

The structure of the paper is as follows. In Section 2 we give a brief description of standard data fusion/calibration methods to update simulated data based on the observed data. In Section 3 we describe one specific data

fusion/calibration method and show how our wind speed data can be calibrated using this method; some comments are also made on a model that is developed in a companion paper (Pereira *et al.* 2019). Finally, in Section 4, we briefly explain how calibration can be extended specifically to data coming from the tails of simulated and observed data, using asymptotic models, and methods suggested by extreme value theory.

2 Background on data fusion and calibration

There are many different paradigms for calibration and in this section, we give a brief description of some of the existing methods. Let $Y(s,t)$ and $X(s,t)$ be respectively the observed and simulated wind speeds at location s and time t. Generically we will use Y and X for observed and simulated wind speeds when data are used without any space-time reference.

Quantile matching-based approaches

If we ignore totally space-time variations and dependence structures, then we can define calibrations as simple scaling making use of marginal distributions fitted respectively to X and Y (CDF transform method, Michelangeli *et al*, 2009). According to this idea

$$x_i^* = F_Y^{-1}(F_X(x_i)), \quad i = 1, \ldots, n, \tag{1}$$

can be defined as calibration, without any space-time configuration of the data. Here, x_i^* is the new calibrated (scaled) data, whereas F_Y and F_X are respectively the distribution functions of Y and X. This scaling is justified by the fact that $P(X_i^* \leq z) = F_Y(z)$ so that the calibrated data has the same distribution as the observed data. This idea can be extended to cover space-time non-homogeneity by scaling (calibrating) the data from

$$x(s,t)^* = F_{Y(s,t)}^{-1}(F_{X(s,t)}(x(s,t)),$$

assuming that the marginal distributions of $Y(s,t)$ and $X(s,t)$ are known for every s and t. These distributions can be estimated by fitting these marginal distributions parametrically, whose parameters are smooth function of spatially and temporarily varying covariates as well as space-time latent processes and then extended over space (and time) through the usual space-time smoothing.

Yet this calibration method is not particularly ideal as it only depends on the marginal distributions of Y and Z, consequently does not make use of the expected strong dependence between the two sets of data and do not seem to take care of possible bias in simulated data. Therefore this transformation should be defined as scaling. Ideal calibration should involve joint distributions of Y and X. Therefore, one can suggest calibration based on

$$x_i^* = F_{Y|X}^{-1}(F_X(x_i)), \quad i = 1, \ldots, n,$$

where $F_{Y|X}$ is the conditional distribution function.

Inverse regression

Calibration is commonly defined as a method of adjusting the scale of a measurement instrument on the basis of an informative experiment and therefore can be seen as an inverse regression problem. Aitchison and Dunsmore (1975) approach the problem from a Bayesian perspective by defining the *calibrative distribution*. Under this approach the objective is to obtain the distribution

$$y(s_0) \mid x(s_0), x(s^*), y(s^*)$$

for an unknown $y(s_0)$ based on the observed and simulated data $(x(s^*), y(s^*))$ on N stations and the simulated value $x(s_0)$, with s_0 different from any s^*.

Simulator–emulator-based approaches

Kennedy and O'Hagan (2001) describe calibration as statistical postprocessing of simulator deterministic forecast and assume that detailed information of how emulators work is available in terms of a set of parameters. Sigrist *et al.* (2015) give detailed description of stochastic versions of space-time advection-diffusion PDE's and their solutions as models for emulators and describe a method of postprocessing simulated data.

Data fusion

Data fusion techniques are often seen as possible calibration methods. See for example Zidek *et al.* (2012) and McMillan *et al.* (2010). Two specific data fusion methods are particularly well adopted to calibration namely Fuentes and Raftery, (2005) and Berrocal *et al.* (2012). In both methods, the models for simulated and calibrated data share a latent Gaussian Markov Random Field (GMRF) model, although there are differences in the way these models are built and in the interpretation of the role of the GMRF. In Section 3, we showcase the application of an adequate modification of the Berrocal *et al.* (2012) method to our daily maximum wind speed data. See, e.g, Foley and Fuentes (2008) for an application of Fuentes and Raftery's method to hurricane surface wind prediction.

3 Calibration methods for bulk and tails

3.1 Berrocal–Gelfand–Holland *et al.* (BGH) method

We now devise a version of Berrocal *et al.* (2012) for non-Gaussian data. Let B be the grid cells, s are points in the space, X is the simulated data and Y the observed data. Here, $Y(s)$ is a point referenced process, whereas $X(B)$ is defined over the grid in the sense that for every $s \in B$, $X(s) = X(B)$

is fixed. Also, $Y(s)$ is observed only at fixed and few observation locations s_j, for $j = 1, \ldots, N$. Below, $Z \sim$ GMRF is used to denote that Z follows a Gaussian Markov Random Field. A preliminary data analysis suggests a model based on the following specification:

- *Simulated data.* Let $X(B) \mid Z(B) \sim \text{Gamma}(\alpha_B, \beta_B)$, with

$$\log\left(\frac{\alpha_B}{\beta_B}\right)(B) = \beta_0 + Z(B), \quad Z \sim \text{GMRF}.$$

- *Observed data (given X).* Let $Y(s) \mid X(s), W(s) \sim \text{Gamma}(\alpha_W, \beta_W)$, with

$$\log\left(\frac{\alpha_W}{\beta_W}\right)(s) = \beta_1 + W(s) + \beta_2 X(B), \quad W \sim \text{GMRF}. \tag{2}$$

The specification above is then used to extrapolate Y to grid cell B, given the simulated value (X) at the grid cell B. We call these the *calibrated values*. A smooth version of the model can be obtained by replacing (2) with

$$\log\left(\frac{\alpha_W}{\beta_W}\right)(s) = \beta_1 + W(s) + \beta_3 Z(B),$$

where $W(s)$, $Z(B)$ are independent GMRF defined over an adequate triangulation.

Application to wind speed data

The model is fitted to 51 observation sites using daily maximum wind speed data during two winter months; only days with wind speeds above zero, and without missing observations, are considered corresponding to 58 days. From the simulator we use the daily maximum wind speed simulated data on the same days. We disregard time dependence and hence we assume that for a specific $s \in B$, $(Y(s,t), X(B,t))$, for $t = 1, \ldots, T$ are independent replicates of the same random vector, where T is the number of days under study. Although hourly wind speeds show significant dependence, our daily data do not show such significant dependence. This is a reasonable assumption which brings significant simplifications in the model.

The grid B has 36 columns and 65 rows, so altogether there are 2340 cells; INLA (Rue *et al.*, 2009) and SPDE method of Lindgren *et al.* (2011) are based on triangulation with 3229 vertices (`www.r-inla.org`). The output of interest is the predictive distribution and its expected value at grid cell for the observed daily max wind speed, and we call these the calibrated wind speeds. In Figure 1 we depict a plot of the mean of the observed and simulated values at the 51 sites, together with correspondent 2.5% and 97.5% empirical quantiles, and in Figure 2 we show a map of the observed

mean of the wind speeds at the 51 station sites (left side), and un-smoothed and smoothed maps of the mean of the simulated and calibrated wind speeds for the study period. It is clear from Figure 1 that simulated wind speeds have, in relation to the observed values, a positive bias. This bias is reduced for the calibrated values, as it is clear from the smoothed maps (observe the scales of the simulated and calibrated maps), but the model does not seem to be able to capture large values.

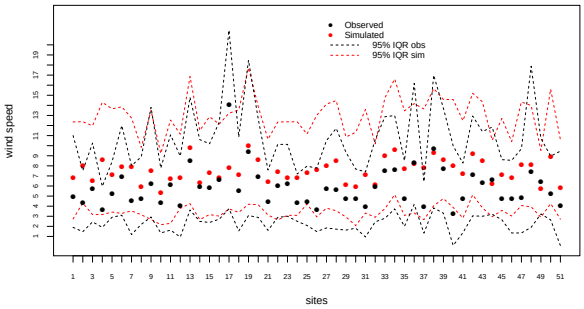


FIGURE 1. Observed (•) and simulated (•) wind speeds and the 95% IQR wind speeds calibrated by station (dashed lines).

3.2 Conditional quantile matching calibration

In a companion paper (Pereira *et al.* 2019) the group develops a covariate-adjusted version of quantile matching-based approach as in (1). To achieve this goal, we have first derived a conditional version of Naveau *et al.* (2016). We briefly discuss the key ingredients of the model below; simulations and further details are available from Pereira *et al.* (2019). To ease notation, we only introduce the model for $F_Y(y \mid \mathbf{x})$, which is given by

$$F_Y(y \mid \mathbf{x}) = G_{\mathbf{x}}\left(H_\xi\left(\frac{y}{\sigma}\right)\right), \tag{3}$$

where $\{G_{\mathbf{x}}\}$ is a family of functions indexed by a covariate, obeying assumptions A, B, and C in Naveau *et al.* (2016), and

$$H_\xi(y) = \begin{cases} 1 - (1 + \xi y)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y), & \xi = 0. \end{cases}$$

Note that (3) is a model for the conditional distribution of y, tailored for both the bulk and tails, and—contrarily to most methods for extremes—it does not require a threshold to be selected.

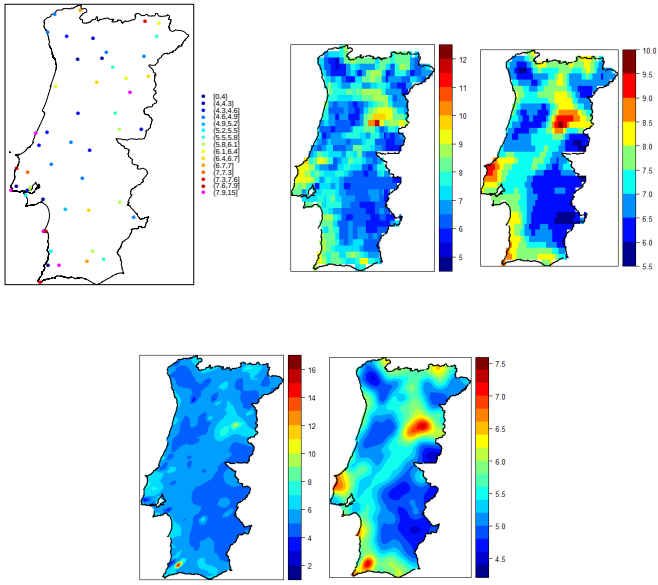


FIGURE 2. Mean wind speeds observed at the 51 sites (top left); mean wind speeds simulated (top right; un-smoothed on the left and smoothed on the righ); mean wind speeds calibrated by BGH method (bottom; un-smoothed on the left and smoothed on the right).

4 Discussion and further extensions

In previous sections we discussed several possible ways of calibrating simulated data based on observations and implemented the modified BGH method as described in Section 3. This method, as expected, does not calibrate well data coming from the right tail.

The fact that damages in electricity grid are basically governed by extreme winds and that primarily simulated-observed data coming from the right tail differ, suggest that adequate calibration methods must be specifically adopted to extreme observations coming the right tails. This in return suggests that methods and models to be used in calibration should ideally be compatible with extreme value theory. A range of approaches for characterising the extremal behaviour of spatial process have been suggested and a brief comparison of these methods can be found in Tawn *et al.* (2018). Let $X_i(s)$, $i = 1, 2, \ldots$ be iid replicates of a spatial process $X(s)$. Essentially, there are 3 different ways of characterising extremal properties of spatial processes and obtaining limiting processes which can be used as models:

1. *Max-stable process:* Limit as $n \to \infty$,

$$M_n(s) = \max_{1 \le i \le n} X_i(s).$$

2. *Pareto process:* Limit as $u \to \infty$,

$$X(s) \mid \max_s X(s) > u.$$

3. *Conditional extremal process:* Limit as $u \to \infty$,

$$X(s) \mid X(\text{fixed site}) > u.$$

Tawn *et al.* (2018) discuss the strong and the weak points of these alternative asymptotic representations. Briefly we mention that max-stable and Pareto processes cannot represent cases when the extremes over space show independence over extended distances, which is the case for extreme wind speeds. Therefore downscaling method described by Towe *et al.* (2017)—which is based on the conditional extremes process—is more suitable, with adequate modifications, to calibrate extreme simulated data based on observed wind speeds. Work on this approach is under progress.

Acknowledgments: The authors acknowledge the financial support received by Fundação para a Ciência e Tecnologia, Portugal, through the projects PTDC/MAT-STA/28649/2017 and UID/MAT/00006/2019

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis.* Cambridge: Cambridge University Press.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2012). Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics*, **68**, 837 – 848.

Foley,K.M. and Fuentes, M. (2008). A Statistical Framework to Combine Multivariate Spatial Data and Physical Models for Hurricane Surface Wind Prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, **13**, 37 – 59.

Fuentes, M. and Raftery, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, **61**, 36 – 45.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society, Series B*, **66**, 497 – 546.

Kennedy, M. and O'Hagan, A. (2001). Bayesian Calibration of Computer Models. *Journal of the Royal Statistical Society, Series B*, **63**, 425 – 464.

Lindgren, F., Rue, H. and Lindstrom, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach (with discussion). *Journal of the Royal Statistical Society, Series B,*, **73**, 423 – 498.

McMillan N. J., Holland, D. M., Morara, M., and Feng J. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics*,**21**, 48 – 65.

Michelangeli, P.-A., M. Vrac, and H. Loukos (2009). Probabilistic downscaling approaches: Application to wind cumulative distribution functions. *Geophys. Res. Lett.*, **36**, 1 – 6.

Pereira, S., Pereira, P. , de Carvalho, M. and de Zea Bermudez, P. (2019). Calibration of extreme values of simulated and real data. *Proceedings of International Workshop on Statistical Modelling 2019.*

Rue H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319 – 392.

Sigrist, F., Künsch, H.R. and Stahel, W.A. (2015). Stochastic partial differential equation based modelling of large space-time data sets. *Journal of the Royal Statistical Society, Series B*, **77**, 3 – 33.

Tawn, J., Shooter, R. , Towe, R., Lamb, R. (2018). Modelling spatial extreme events with environmental applications. *Spatial Statistics.* 10.1016/j.spasta.2018.04.007.

Towe, R.P., Sherlock, E.F. , Tawn, J.A., Jonathan, P. (2017). Statistical downscaling for future extreme wave heights in the North Sea. *Annals of Applied Statistics.* **11**, 2375 – 2403.

Zidek, J.V., Le, N.D. and Liu, Z. (2012). Combining data and simulated data for space-time fields: Application to ozone. *Environmental and Ecological Statistics*, **19**, 37 – 56.

Statistics with a Human Face

Adrian W. Bowman[1], Stanislav Katina[2], Liberty Vittert[3]

[1] School of Mathematics & Statistics, The University of Glasgow, UK
[2] Institute of Mathematics & Statistics, Mazaryk University, Brno, Czech Republic
[3] Department of Mathematics and Statistics, Washington University in St. Louis, USA

E-mail for correspondence: adrian.bowman@glasgow.ac.uk

Abstract: Three-dimensional surface imaging, through laser-scanning or stereo-photogrammetry, provides high-resolution data defining the surface shape of objects. Human faces are of particular interest and there are many biological and anatomical applications, including assessing the success of facial surgery and investigating the possible developmental origins of some adult conditions. An initial challenge is to structure the raw images by identifying features of the face. Ridge and valley curves provide a very good intermediate level at which to approach this, as these provide a good compromise between informative representations of shape and simplicity of structure. Some of the issues involved in analysing data of this type will be discussed and illustrated. Modelling issues include simple comparison of groups, the measurement of asymmetry and longitudinal patterns of shape change. This last topic is relevant at short scale in facial animation, medium scale in individual growth patterns, and very long scale in phylogenetic studies.

Keywords: Shape; Curvature; Visualisation.

1 Introduction

Images which consist of high resolution data on surface shape are becoming increasingly common. Figure 1 shows an example of a human face captured by a stereo-photogrammetric camera system. Laser systems are also available for this kind of imaging. Data of this type can be captured in the context of studies which are based on the usual kinds of scientific questions, involving the comparison of groups, the assessment of the effects of covariates, or the construction of predictions. Appropriate statistical meth-

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

ods are then required to allow these kinds of analysis to take place with data in the form of 3D surfaces.

FIGURE 1. An example of a 3D facial image at different orientations.

2 A model for a face

Faces are surfaces of particular interest for a wide variety of medical, biological, social or security reasons. The large number of 3D points which form the raw image need to be replaced by a systematic description of facial shape that allows analysis based on information that corresponds across subjects and which will then allow meaningful interpretation. There are many ways in which that can be approached. Anatomical landmarks have been the mainstay of shape analysis over many years but with the high resolution images now available more complex descriptions are required. Dryden and Mardia (2016) give a comprehensive description of this general area. A common approach is to 'warp' a template shape onto an individual image using appropriate indicators of local surface characteristics. An alternative approach involves the estimation of well-defined ridge and valley curves which we expect to see on all faces, such as the ridge of the nose, the valley between closed lips etc. This is described in Bowman et al. (2015) and Vittert et al. (2017), while Katina et al. (2015) propose that these curves should be the basis of anatomical definitions. The method involves the characterisation and tracking of local surface properties through the maximal and minimal curvatures present (κ_1, κ_2 respectively). For example, the type of local surface can be helpfully expressed in the 'shape index' defined as $2/\pi \tan^{-1}((\kappa_2 + \kappa_1)/(\kappa_2 - \kappa_1))$; see Koenderink and van Doorn (1992). Values of the shape index are colour coded onto the example face in the left hand image of Figure 2. Once curves have been estimated, the intervening surface patches can be given a simple representation through

interpolated transects. The resulting facial model is displayed in the right hand panel of the Figure.

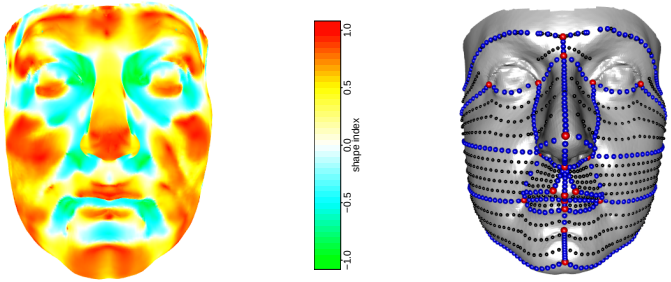


FIGURE 2. An example of a 3D facial image coloured by shape index (left) and with a facial model superimposed (right).

3 The analysis of 3D shape

If a facial model is available for each image then this common shape description can form the basis of analysis. A registration step is required to place the shapes in a common co-ordinate system. Procrustes alignment provides a good solution for this; see Dryden and Mardia (2016) for details. As a simple example of subsequent analysis, Figure 3 shows the results of applying principal components analysis to the aligned nose shapes of 61 males and 69 females, all adults of British ethnic origin. One advantage of the curve-based approach to the construction of a facial model is that particular features such as the nose can then be extracted easily. As usual, principal components analysis provides a means of reducing the very high dimensionality of the space of the model descriptor, providing a much lower dimensional space in which analysis can take place. In this context it is a 'regulariser' of the original space.

Figure 3 plots the scores of the first 10 components, which account for 81% of the variability in the dataset. Confidence intervals (Bonferroni adjusted) are also displayed to allow the evidence of differences in shape between the sexes to be assessed. The nasal images illustrate the nature of each principal component by plotting the shapes which correspond to ± 2 standard deviations from 0 on the scores scale. The images which correspond

to components where there is evidence of differences in shape have been magnified. Component 2 suggests males have narrower and more prominent nose ridges while component 4 indicates a more rounded nasal tip in females. Component 10, which accounts for a very small proportion of variability, may indicate slightly flatter edges to the nasal area in females, particularly close to the eyes.

4 Discussion

This short paper indicates the nature of 3D surface data and describes some initial approaches to modelling. In the presentation associated with the paper a wide variety of other types of analysis will be discussed in the context of several different application areas, including studies of facial surgery, the neuro-developmental origins of adult conditions and the construction of phylogenies. In particular, longitudinal models are relevant at short scale in facial animation, medium scale in individual growth patterns, and very long scale in phylogenetic studies.

References

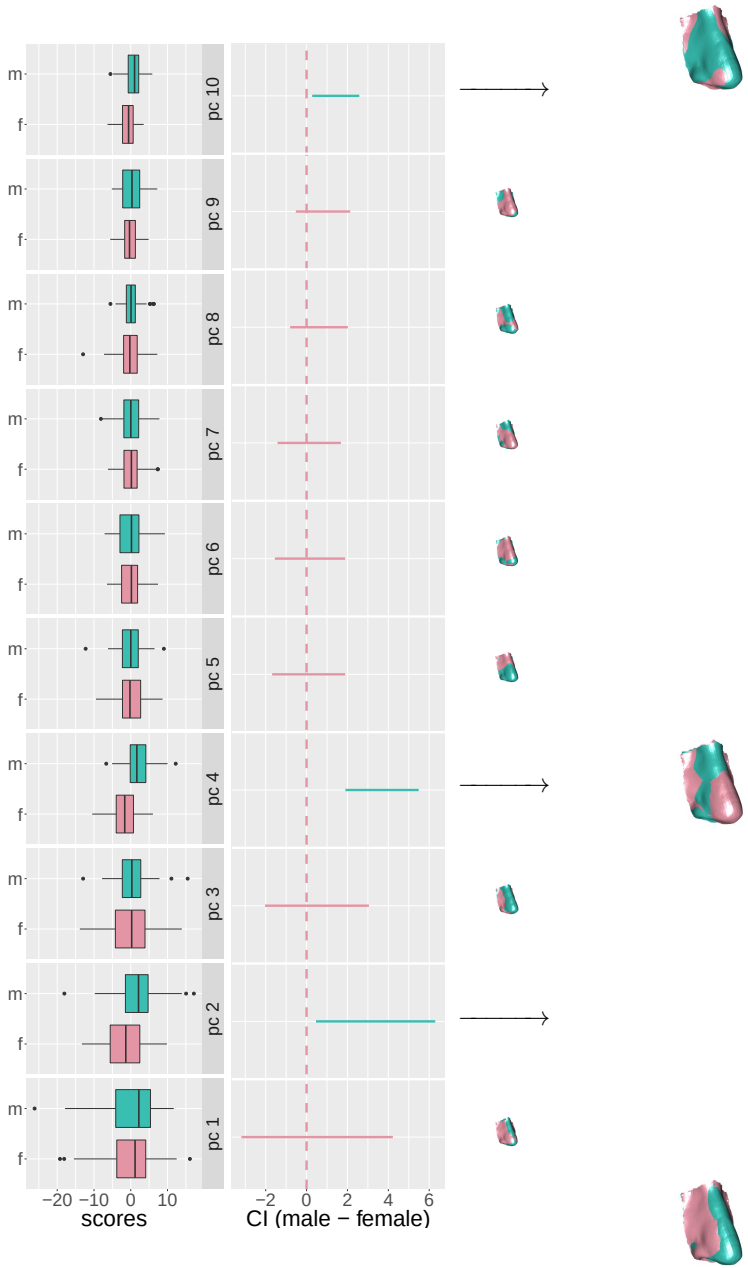
Bowman, A.W., Katina, S., Smith, J., Brown, D. (2015). Anatomical curve identification. *Computational Statistics & Data Analysis*, **86**, 52 – 64.

Dryden, I.L., Mardia, K.V. (2016). *Statistical Shape Analysis, with Applications in R.* 2nd edn. New York: Wiley.

Katina, S., McNeil, K., Ayoub, A., Guilfoyle, B., Khambay, B., Siebert, P. , Sukno, F., Rojas, M., Vittert, L., Waddington, J., Whelan, P.F., A. W. Bowman (2015). The definitions of three-dimensional landmarks on the human face: an interdisciplinary view. *Journal of Anatomy*, **228**, 355 – 365.

Koenderink, J., van Doorn, A. (1992). Surface shape and curvature scales. *Image and Vision Computing*, **10**, 557 – 564.

Vittert, L., Bowman, A., Katina, S. (2017). Statistical models for manifold data with applications to the human face. arXiv preprint 1701.07328.


m
f
pc 10
pc 9
pc 8
pc 7
pc 6
pc 5
pc 4
pc 3
pc 2
pc 1
−20 −10 0 10
scores
−2 0 2 4 6
CI (male − female)

Part II - Special Session in Honour of Prof. Murray Aitkin

Statistical modelling for income inequality comparisons

Murray Aitkin

[1] School of Mathematics and Statistics, University of Melbourne, Australia

E-mail for correspondence: murray.aitkin@unimelb.edu.au

Abstract: This paper is concerned with the measurement of income inequality over successive surveys or censuses of a population. In Australia there has been a recent major argument over the claim by the Australian Bureau of Statistics that inequality had not worsened in Australia over the period 2014-2016.

1 Introduction

> Mr Morrison [the then Australian Treasurer, now Prime Minister] said:
>
> "What I don't accept – that this idea, that people and inequality and incomes have been going in the wrong direction, that's not borne out by the facts.
>
> "The last census showed that on the global measure of inequality, which is the Gini coefficient -– that is the accepted global measure of income inequality around the world -– and that figure shows that it hasn't got worse, inequality, that it's actually got better."
>
> A day earlier Mr Morrison gave a speech to the Australian Industry Group, in which he said:
>
> "Analysis of the more recent census data for the 2016 census shows the Gini coefficient based on gross household income has declined from 0.382 to 0.366 since 2011."
>
> Mr Morrison's figures were derived using gross income data taken from the 2016 census, and were based on internal, unpublished calculations.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

2 Difficulties with the Gini coefficient

There are at least three problems with the Gini index for income inequality comparisons. The principal problem with the index, or any other single number, is that it cannot represent variability in the income distribution. A Gaussian distribution can be summarised by two numbers, but only a single-parameter distribution, like the Poisson or exponential, can be summarised by one number. Recognition of this allows us to develop a statistical modelling approach to changes in income inequality over repeated surveys or censuses, using publicly available data.
The second is that countries (or years within one country) with widely different income distributions may have the same Gini index. This point has been made frequently in criticisms of the Gini coefficient.
The third is that the calculation of the Lorenz curve, from which the Gini index is derived, requires access to individual-level income data, to develop both the percentiles of the individual income distribution and the proportion of national income received by each income percentile group. These data are generally confidential to the national statistical office and are not publicly available. What is publicly available, at least in Australia, is the numbers of households receiving income in ABS-defined income intervals. We give an example of total household income reported in the Australian censuses of 2006, 2011 and 2016.

3 Income distributions by Census year

Table 1 is constructed from the publicly available "Census Table Builder" at the Australian Bureau of Statistics (ABS). It shows the population counts of weekly Total Household Income reported in the 2006, 2011 and 2016 Censuses, in different income intervals in each Census. Households not reporting or reporting partial, zero or negative incomes were excluded. The table gives the "income" variable as the upper end-point of each income interval. The column "median" will be explained below.
The range of incomes reported was extended considerably in the 2016 Census. What can be said about the change in the distribution of Total Household Income over these Census periods? In particular, what proportions of the reporting populations had incomes less than half of the median incomes (a common definition of poverty)?
To *model* these distributions takes us beyond sample variation – these are (or are intended to be) *population* proportions based on very large numbers. However finding a "best-fitting model" representation gives us (by the model smoothing) the medians and other percentiles in the three distributions.

	2006			2011			2016	
median	income	count	median	income	count	median	income	count
105	149	56207	140	199	121661	105	149	34737
220	249	53671	270	299	213918	255	299	55126
320	349	131145	370	399	492804	370	399	51581
455	499	363057	460	599	737671	470	499	178043
605	649	505010	660	799	663125	605	649	184281
755	799	272528	860	999	615387	755	799	445780
949	999	339939	1075	1249	616836	860	999	372972
1140	1199	603408	1325	1499	566837	1075	1249	483418
1260	1399	369503	1650	1999	881783	1325	1499	468132
1490	1699	432424	2150	2499	633195	1575	1749	374240
1790	1999	368798	2650	2999	605381	1825	1999	378560
2150	2499	309149	3150	3499	356245	2165	2549	733187
2800	3499	137629	3650	3999	163860	2685	2999	498209
3650	3999	93702	4300	4999	140282	3150	3499	320392
4150	>4000	90107	5150	>5000	126493	3800	4499	138630
						4650	4999	136917
						5300	5999	122422
						6600	7999	104507
						8300	>8000	11652
Total		4,126,277			6,935,478			5,092,786

TABLE 1. Reported household weekly income in $A

4 Modelling the income distribution

By differencing the cumulative counts at each interval upper endpoint, we obtain a set of interval counts which can be modelled as described by Lindsey (1997). They make up a multinomial distribution of income histogram counts, which can be expressed as a set of conditionally independent Poisson counts, conditional on the total count. The Poisson model for the counts then fits user-specified functions of the income variable as explanatory variables in a log-linear model for the counts. This is widely known as the "Poisson trick". This analysis is a particular case of the *composite link function* (Thompson and Baker 1981), in which the usual link function is applied to a *transformation* – here the successive differences – of the original cumulative observations. The Poisson distribution has been used in several such applications (Eilers 1991, 2017).

The difficulty for the modelling is assigning income locations for the interval counts. The traditional locations are the midpoints of the intervals, used since the invention of the histogram. However there is clearly some bias in this procedure, since all actual values in each interval are moved to the right, and the interval midpoint assumes that these unobserved actual values are symmetrically distributed around the midpoint within each interval. This is clearly not so: a more realistic simple model is that the actual values have a *triangular distribution* in each interval. The interval mass centre would then be at the *median* of the distribution, which would be at 0.7 ($\sqrt{2}/2$), rather than 0.5, of the distance from the lower mass to the higher mass endpoint. These relocated masses can then be analysed with the Poisson

trick. The resulting masspoints (medians) and masses are shown in Table 1.

There is a wide choice of possible continuous distribution models for the linear predictor (regression model). The Gaussian, lognormal and gamma distributions are modelled with the linear and quadratic terms in income, the linear and quadratic terms in log income, and the linear terms in income and log income respectively. None of these distributions gives an adequate fit to the income data in any of the census years.

The substantial skew in all the empirical distributions suggests that the log income scale should be used. We use the *four-moment distribution*, sometimes called the *log-quartic* distribution, with the log density a quartic function of the argument. This ensures that the first four sample moments are reproduced by the fitted model. We fit the Poisson four-moment model in log "median" income to the counts for the three censuses. The fitted cdfs for the three Census years are shown on the probit/log income scale in Figure 1. The skew in the distributions is accounted for by the third and fourth moments.

The fit to the observed data is close for all three censuses. However the income data analysed are expressed in current $A, which need to be adjusted for cost of living increases by the Consumer Price Index, which increased by 15% from 2006 to 2011, and by 25% from 2006 to 2016. The adjustment involves a simple left shift of the two later income distributions, by $\log 1.15 = 0.14$ units for 2011 and $\log 1.25 = 0.22$ units for 2016. Analyses below are expressed in constant 2006 dollars.

TABLE 2. Income medians and quartiles in $A, and poverty percentiles, CPI-adjusted

Year	25%	median	75%	poverty	percentile
2006	635	1031	1451	515	17.6%
2011	462	919	1605	460	24.8%
2016	801	1305	2061	653	18.2%

The definition of (relative) poverty here is family income less than 50% of the median income.

- The median and first quartile *declined* over 2006-11, with more than 50% of families worse off in 2011, and almost 25% being in relative poverty.
- the proportion in poverty *decreased* in 2016 to slightly above the 2006 level.
- This reflects the effect of the 2008/9 GFC, and the improvement since then.

- The 2006 income distribution was changed in 2011 by an increase in the proportion of low-income families,
- but in 2016 the 2006 pattern was repeated, but with a shift to higher incomes.

The increase in the poverty proportion from 2006 to 2011 following the GFC was reversed from 2011 to 2016. The Australian economy improved substantially over this period, partly because of the Australian Government action to stimulate the economy following the GFC, and partly because of the substantial increase in exports to China. The analysis here supports the ABS statement with the Gini index, but is more informative.

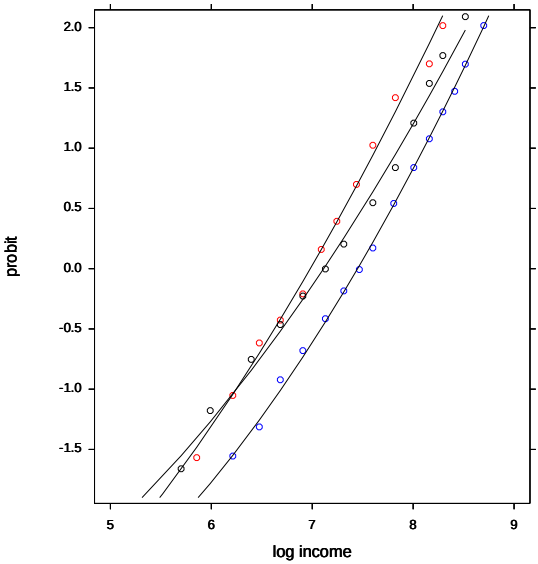


FIGURE 1. Fitted cdfs, income (red 2006, white 2011, blue 2016)

References

Eilers, P.H.C. (1991). Density estimation from coarsely grouped data. *Statistica Neerlandica*, **45**, 255 -– 270.

Eilers, P.H.C (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling* **7**, 239 –2̇54.

Gini, C. (1912). Concentration and dependency ratios (in Italian). English translation in Rivista di Politica Economica, **87** (1997), 769 — 789.

Hodges, J.L., D. Krech and R.S. Crutchfield (1975). *StatLab: An empirical introduction to statistics*. New York: McGraw-Hill.

Lindsey, J.K. (1997). *Applying Generalized Linear Models*. New York, Springer-Verlag.

Lorenz, M.O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* **9**, 209 – 219.

Thompson, R. and Baker, R.J. (1981). Composite link functions in generalized linear models. *Applied Statistics* **30**, 125 – 131.

Part III - Special Session Devoted to Statistics Portugal

Synthetic data as Public Use Files: an application to the Household Budget Survey

Inês Rodrigues[1]

[1] Instituto Nacional de Estatística - Statistics Portugal, Lisboa, Portugal

E-mail for correspondence: ines.rodrigues@ine.pt

Abstract: A methodology for producing Public Use Files (PUF) for the Household Budget Survey by generating synthetic data is presented. Parametric (multinomial logistic and log-linear regressions) and non-parametric methods (classification and regression trees) were used for generating the main identifying variables, as well as income and expenditure totals. The two approaches were compared with a focus on the risk of disclosing confidential information from the PUF.

Keywords: PUF; HBS; Confidentiality; Synthetic data.

1 Introduction

Public Use Files (PUF) include data on individual statistical units and are prepared to be of public access. PUF are intended to be used for education or test purposes - e.g., by researchers when developing their application to access microdata files for research use, the so-called SUF (Scientific Use Files). The aim of this work is to compare a parametric and a non-parametric approach for producing PUF for the Household Budget Survey based on synthetic data, namely regarding the resulting disclosure risk.

2 Producing PUF by generating synthetic data

2.1 A parametric and a non-parametric approach

Let D denote an original microdata set, including a set of k variables represented by Y, whose relationships are intended to be preserved. Following Raghunathan et al. (2001), the joint conditional density of $Y_1, Y_2, \ldots, Y_k$

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

given a set of background variables X, can be factored as:

$$f(Y_1,\ldots,Y_k|X,\theta_1,\ldots,\theta_k) = f_1(Y_1|X,\theta_1)\prod_{v=2}^{k} f_v(Y_v|X,Y_1,\ldots,Y_{v-1},\theta_v)$$

where f_v, $v = 1,\ldots,k$ are the conditional density functions and θ_v is a vector of parameters in the conditional distribution (e.g., regression coefficients). Each conditional distribution is modeled by a given appropriate regression model (e.g., linear, logistic or log-linear regression, if Y_v is a continuous, binary or count variable, respectively).
On the other hand, the CART (classification and regression trees) algorithm provides good results as a non-parametric approach to generate synthetic data (Dreschler and Reiter (2011)). As described by Nowok et al. (2017), it is based on the recursive partition of the original dataset into groups with increasingly homogeneous outcome. Splits are defined based on *yes/no* questions concerning the predictors. In each final group (leaf), values approximate the conditional distribution of the predicted variable for units with predictors meeting the criteria that define that group. Synthetic values are generated by sampling from the appropriate leaf. CART can be used to simulate each variable sequentially, by conditioning on already generated variables, as in the parametric approach.

2.2 Disclosure risk

Following Loong et al. (2013), we assume the user knows which units are included in D and their values regarding m indirect identifiers. Based on this m variables, the user attempts to obtain information on a confidential variable, T, from the synthetic microdata set $D^{'}$. Let n_R and n_S denote the number of units in D and $D^{'}$, respectively. We denote by w_{iq} the value of the indirect identifier q for unit i in D ($i = 1,\ldots,n_R$; $q = 1,\ldots,m$) and by w_{jq} the value of the same identifier for unit j in $D^{'}$. Let $R_{ij} = 1$ ($i = 1,\ldots,n_R$; $j = 1,\ldots,n_S$) if $w_{iq} = w_{jq}\ \forall\ q$ ($q = 1,\ldots,m$) and $R_{ij} = 0$ otherwise. Let also $C_i = \sum_{j=1}^{n_S} R_{ij}$ and $U_{ic} = 1$ ($i = 1,\ldots,n_R$; $c = 1,\ldots,C_i$) if $t_i = t_c$ for categorical T, or $t_c \in [t_i(1-p),\ t_i(1+p)]$ for continuous T and a precision of $p \times 100\%$, and $U_{ic} = 0$ otherwise. We therefore denote by $Z_i = \sum_{c=1}^{C_i} U_{ic}$ the total number of records that are a real match for unit i in D. Let $I_i = 1$ if $Z_i > 0$ and $I_i = 0$ otherwise, and $K_i = 1$ if $C_i = 1 \wedge I_i = 1$ and $K_i = 0$ otherwise. Disclosure risk can therefore be quantified by the following global measures:

- the expected match risk, given by EMR= $\sum_{i=1}^{n_R} \frac{Z_i}{C_i}$
- and the true match risk, given by TMR= $\sum_{i=1}^{n_R} K_i$

EMR reflects the chance of an user randomly establishing a true match for each unit i in D and TMR that of an user correctly and uniquely identifying each unit i in D.

3 Producing PUF for the Household Budget Survey

The Household Budget Survey (HBS) aims at producing data on consumption expenditure; its microdata is composed by records regarding households and household members. We generated as many household records as the number of households in the original sample. We began by simulating region (NUTS II) and household size by sampling from the corresponding estimated multinomial distribution, based on the relative frequency distributions in the original sample - we first simulated region and then household size, given region. For each synthetic household, a real household from the same region and size was randomly selected; the number of members in the synthetic household, as well as their sex and age, was taken to be that from the selected real household. Both approaches presented in 2.1 - Parametric and CART - were then used to generate the main identifying variables (country of birth, country of citizenship, marital status, level of studies completed, status in employment and economic sector in employment), as well as income and expenditure totals. In order to compare both approaches regarding the resulting disclosure risk, we generated 100 synthetic datasets from each approach, considering a random sample of 500 households from the HBS SUF to be our real data.

4 Results and discussion

Figures 1 and 2 illustrate respectively the distributions of two identifying variables and the total expenditure, obtained by generating a single synthetic data following each approach, in comparison with the real data (SUF). Good results were obtained regarding the main statistics computed from HBS data (e.g. the mean consumption expenditures of households (euros) - SUF: 20 391, Parametric: 19 942 and CART: 19 661 - and the at-risk-of-poverty rate (after social transfers) (%) - SUF: 14.8, Parametric: 19.2 and CART: 15.5). However, the additional flexibility from CART results in a slight increase in disclosure risk, as illustrated by figure 3.

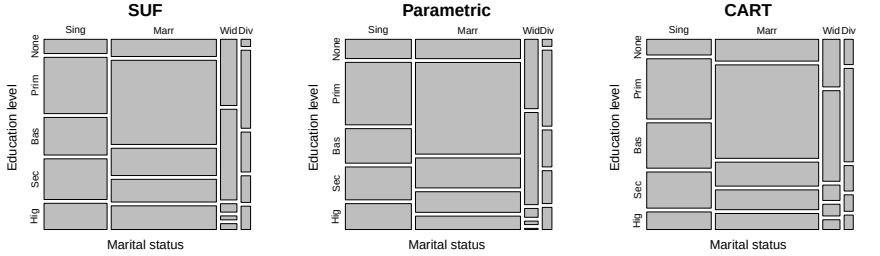


FIGURE 1. Weighted frequency distribution of education level by marital status, in the real (SUF) and synthetic HBS datasets.

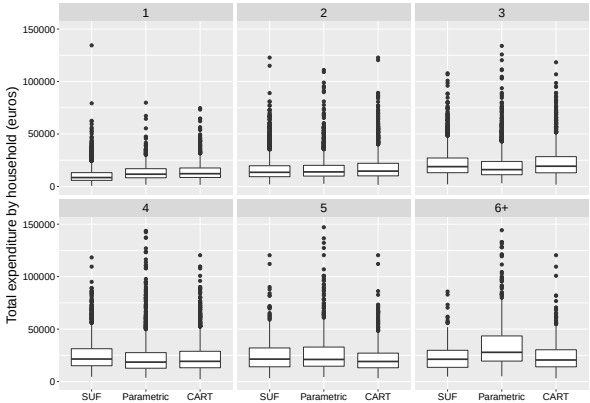


FIGURE 2. Total annual consumption expenditure by household size, in the real (SUF) and synthetic HBS datasets.

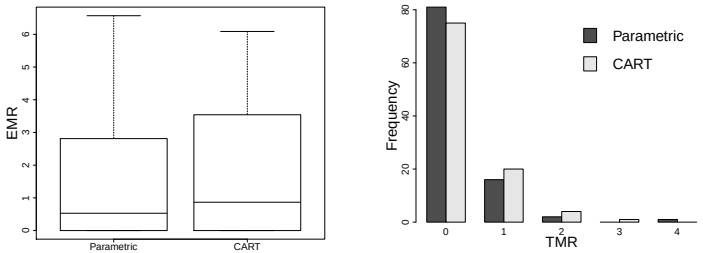


FIGURE 3. EMR and TMR distributions for 100 replications. Following the notation in 2.2, $m = 6$ (sex, age, HH size, marital status, status in employment and country of citizenship), T are income and expenditure totals and $p = 0.05$.

References

Drechsler, J. and Reiter, J.P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, **55**, 3232 – 3243.

Loong, B., Zaslavsky, A.M., He, Y. and Harrington, D.P. (2013). Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. *Statistics in Medicine*, **32**, 4139 – 4161.

Nowok, B., Raab, G.M. and Dibben, C. (2017). Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. *Statistical Journal of the IAOS*, **33**, 785 – 796.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, **27**, 85 – 95.

Small Area Estimation for Land Use and Land Cover

Pedro Campos[1], Suelma Pina[2], A. Manuela Gonçalves[2]

[1] Statistics Portugal, Portugal
[2] University of Minho, Portugal

E-mail for correspondence: pedro.campos@ine.pt

Abstract: Small Area Estimation (SAE) is a part of statistical science that combines survey sampling and inference of finite populations with statistical modelling. The main objective of this paper is to analyze and test the implementation of different types of estimators of small domains in order to improve the quality of the estimates produced within the framework of the Farm Structure Survey (FSS) at NUTS III level. Under the EUROSTAT Land Use and Cover Area Statistical Survey (LUCAS) project, this is a fundamental tool for environmental studies, forestry and agricultural resource planning.

Keywords: Small Area Estimation; Regression Estimator; EBLUP; SEBLUP; Farm Structure Survey

1 Introduction

Nowadays, public and private institutions are increasingly seeking more detailed information to aid their decision-making process, and the National Statistical Offices do fall into this new paradigm. The need to produce reliable estimates for the total of variables of interest in small domains is fundamental. However, estimates cannot always be obtained through direct estimators (that use only the observations of the variable of interest belonging to the domain for the time period under analysis), because often there are no samples for these domains, or they are too small to obtain sufficient quality estimates. In order to solve this problem, several types of estimators for small domains have been proposed: some of them combine auxiliary information of the variable of interest in the domain and in different periods of time, or even consider variable sources of other domains (the so-called indirect estimators). The main objective of this paper is to develop, analyze and test the implementation of different types of small area

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

estimators in order to improve the quality of the estimates produced within the framework of the Farm Structure Survey (FSS) at regional (NUTS III) level. Currently, Statistics Portugal publishes these estimates at National (NUTS I) and Regional (NUTS II) levels. Under the EUROSTAT Land Use and Cover Area Statistical Survey (LUCAS) project, Statistics Portugal intends to use this information to detail the agriculture class, thus providing information on agricultural land use up to the third level of patent nomenclature in the Land Use and Land Cover Mapping (LULC), a fundamental tool for environmental studies, forestry and agricultural resource planning (EUROSTAT,2013). In this work, five different estimators (direct, modified and combined) are used to estimate 44 variables by NUTS III in mainland Portugal: the direct estimator (1 and 2), the estimator modified by the Regression, the EBLUP estimator using the Fay-Herriot method and the EBLUP estimator by the spatial level of the area (SEBLUP). Based on the results, we may conclude that when auxiliary variables are available, the estimator modified by the Regression performs better when compared to other estimators.

2 Small Area Estimators

In this section we introduce Small Area Estimation (SAE) and shortly describe the main estimators used in this work. In a stratified random sampling design, let U be a finite population of N distinct elements, $U = \{1, ..., N\}$, the subpopulations (in this case, strata), U_h, with $U_h \subset U, h = \{1..., H\}$, for which certain parameters have to be estimated according to the domain d. (see Figure 1).

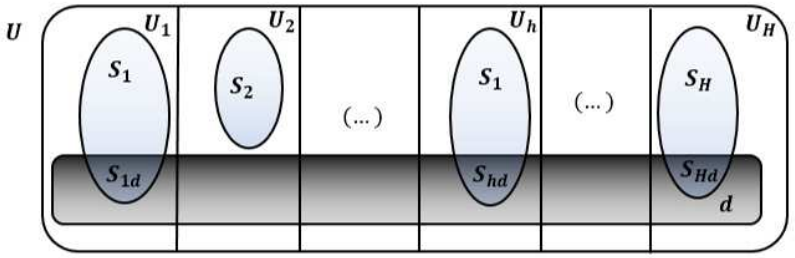


FIGURE 1. Representation of domains, under the SAE perspective

The population dimension of each stratum U_h is denoted by N_h with $h = \{1, ..., H\}$, where $N = \sum_{h=1}^{H} N_h$, and the subpopulation dimension in U_{hd} is denoted by N_{hd}, where $N_d = \sum_{h=1}^{H} N_{hd}$; we consider s as a sample

of size n collected from U that may be decomposed in $s = \sum_{h=1}^{H} s_h$ and $s_d = \sum_{h=1}^{H} s_{hd}$, which are sampling units of size n_d and n_{hd} randomly selected, where $n = \sum_{h=1}^{H} n_h$ and $n_d = \sum_{h=1}^{H} n_{hd}$.
We usually denote population U as being composed by two quantities, Y (the explained variable, or variable of interest) and $X = (X_1, ..., X_j) \in \mathbb{R}^j$, the values of the covariates or auxiliary variables. Auxiliary variables are always assumed to be known, whereas the variable of interest may be unknown for some areas if individuals in these areas are not sampled. Assuming that we want to obtain estimates of the total, τ_d the total of the variable of interest for the population of the domain of interest d is given by: $\tau_d = \sum_{i \in U_d} Y_i$.
In general, SAE models can be categorized in direct and indirect estimators. Direct estimators only consider the observations of the variable of interest belonging to the study domain for the time period under analysis, whereas indirect estimators take observations of the variable of interest as well as auxiliary sources outside the study domain for the considered period of time. The Model-based approach belongs to the class of indirect estimators and regression models are used here between data from the sample and auxiliary variables from other data sources, such as census and administrative records to "lend" information from similar areas (Rao and Molina, 2015). Indirect estimators can also be divided in synthetic and combined estimators which can be derived under a design-based approach or taking into account the fact that an explicit area level or unit level model exists. Combined estimators are basically weighted averages of a direct estimator and an indirect estimator (Rao and Molina, 2015, Pfeffermann, 2013).

2.1 Direct Estimators (D_1 and D_2)

We start with the fundamental Horvitz-Thompson estimator, defined in Rao and Molina (2015):

$$D_1 = \hat{\tau}_{D_1} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_{hd}} y_i$$

$$Var(\hat{\tau}_{D_1}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} (s_{hd}^2 + (1 - \frac{N_h d}{N_h}) \bar{y}_{hd}^2)$$

A second estimator is used, where we assume to know the dimension of each population defined by the intersection of NUTS III with the strata defined a priori in the sampling plan: (N_{hd} e n_{hd}):

$$D_2 = \hat{\tau}_{D_2} = \sum_{h=1}^{H} \frac{N_{hd}}{n_{hd}} \sum_{i \in s_{hd}} y_i$$

$$Var(\hat{\tau}_{D_2}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} s_{hd}^2$$

Where,s^2_{hd} is the sampling variance in the subsample defined by the intersection of stratum h with domain d.

2.2 Direct Estimator modified by Regression (Reg)

For the application of this estimator, it is necessary to know the values of the auxiliary variables for all units of the population at individual level, the vector of the totals of the auxiliary variables in domain τ_{xd} and their observed values in the sample units of the subpopulation g,x_i, $i \in s_g$. The regression estimator for the total estimate is given by:

$$\hat{\tau}_{d,reg} = \hat{\tau}_d + (\tau_{xd} - \hat{\tau}_{xd})' \hat{\beta}_g$$

where $\hat{\beta}_g$ is the estimator of regression parameters $\beta_g = (\beta_{g1}, \beta_{g2}, ..., \beta_{gp})'$. In this case there is an implicit link model: $y_i = x'\beta_g + \epsilon_i$, with $i \in U_g$

2.3 EBLUP and SEBLUP

The EBLUP is a combined estimator. Considering a finite population divided into D small domains, the Fay-Herriot base model (Rao and Molina, 2015) linearly relates the value of the d-th domain of the variable of interest θ_d to a vector of p auxiliary variables aggregated at the x_d area level and includes an associated random v_d effect. The model is given by $\theta = x'_d\beta + v_d, d = 1, ...D$; where β is a vector of regression parameters; v_d are the random effects. Then, the combined estimator SEBLUP, $\hat{\theta}_{SEBLUP}$ of parameter θ_d may be written as:

$$\hat{\theta}_{SEBLUP} = x'_d\beta + v_d + e_d = x'_d\beta + (I_D - \rho W)^{-1}u + e_d$$

The SEBLUP estimator considers a spatial component. The main difference between the two models (EBLUP and SEBLUP) lies in the fact that SEBLUP uses the information of the distances between the domains through a proximity matrix (Pfeffermann, 2013).

3 Data, Software, and Results

3.1 Data and Software

The Farm Structure Survey (FSS), also known as the Survey on the structure of agricultural holdings, is carried out by all European Union (EU) Member States and provides comparable statistics across countries and time, at regional levels (down to NUTS 3 level). The edition of 2013 considers more than 650 variables. In this study several strata has been considered, based on size class, area status, legal status of the holding, objective zone and farm type (INE, 2013). Therefore, the population has been divided in 765 strata, (h=1,. . . ,765) and 23 domains or small areas, corresponding

to NUTS III, (d=1,...,23). The overall population size (N) is 236696 agricultural holdings and the sample size (n) is 23108, representing about 9,76 % of the population. Algorithms to calculate the estimates, with the exception of the EBLUP estimator, were all programmed in R by the authors. The SEBLUP algorithm was obtained through the eblupSFH function of the R package sae (Molina and Marhuenda, 2013). In order to measure and compare the quality of the estimators, the coefficients of variation (CV) are computed and shown in percentage. To see if the spatial information introduced by the SEBLUP provided some improvement in the CV estimates, in the analysis of the results we also consider the results of the EBLUP estimator computed through the Fay-Herriot method ($EBLUP_{FH}$).

3.2 Results

Results of the coefficient of variation (CV) of the five estimators are presented in Table 1.

TABLE 1. Results of the coefficient of variation (CV) of the five estimators

Estimator	CV range (%)	1st Quartile	Median	Mean	3rd Quartile	Quartile
$\hat{\tau}_{D_1}(Direct_1 or D_1)$	1.63-41.21	2.99	3.99	7.14	5.83	9.32
$\hat{\tau}_{D_2}(Direct_2 or D_2)$	1.29-18-82	2.12	2.57	3.72	3.84	3.61
$\hat{\tau}_{d,reg}(Reg)$	0.93-24.00	2.23	3.64	4.87	4.88	4.93
$\hat{\theta}_{SEBLUP}$	1.64-44.09	3.04	3.99	7.33	5.89	9.86
$\hat{\theta}_{EBLUP_{FH}}$	1.63-39.37	2.86	3.93	6.83	5.84	8.66

The wide variation of the CV range is due to the fact that different small areas (the NUTS III regions) differ much in terms of sample sizes. We can see (see Figure 2) that lowest values of CV were provided by Reg (the Direct Estimator modified by Regression) , although Direct 2 (the Direct Estimator 2) also performed well.

4 Conclusions

With regard to modified and indirect estimators Reg, SEBLUP and EBLUP, we found out that they present greater gains in precision when the sample size is larger and when the correlation between the dependent and independent variables is greater. When analyzing the CV estimates of the different estimators studied by NUTS III for one of the most important variables, UAA (Utilized Agricultural Area), the regions of Baixo Alentejo (184) and Alentejo Central (187) are the ones with the highest CV values when compared with those of the other NUTS III regions. This result ends

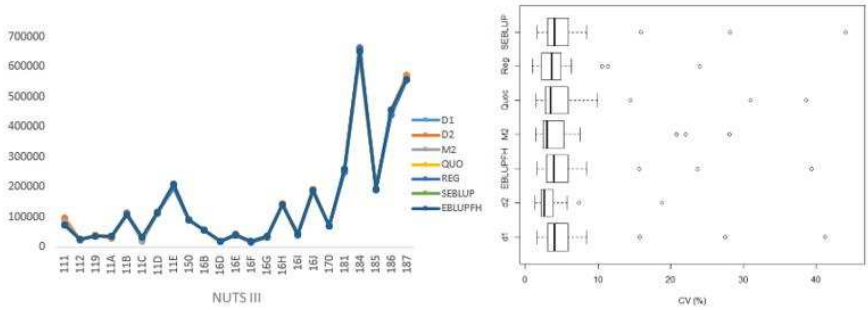


FIGURE 2. Graphical comparison of the estimates and boxplots of CV for the five estimators under analysis. (Note: we introduced two extra estimators: M2, the modified estimator and Quo, the Quotient estimator).

up harming the interpretation of the mean CV values of the estimators, since in general the CV estimates for the other regions are much lower.

References

EUROSTAT (2005). LUCAS 2009 (Land Use / Cover Area Frame Survey), *Quality report*. Luxembourg: Eurostat.

Instituto Nacional de Estatística (INE) (2013). Inquerito a Estrutura das Exploracoes Agricolas *Documento Metodologico*. Lisboa: INE.

Molina, I., Marhuenda, Y. (2013). sae: An R package for Small Area Estimation In: *The R Journal*, 7 ,1 .

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation . *Statistical Science*, **28, 1, 40**, 40 – 68.

Rao, J.N.K., Molina, I. (2015). Small Area Estimation, 2nd Edition *Wiley Series in Survey Methodology.*, John Wiley and Sons, Inc., Hoboken, New Jersey

Part IV - Contributed Papers

Modeling interactions between individuals using coupled hidden Markov models

Jennifer Pohle[1], Marius Ötting[1], Frants Havmand Jensen[2], Roland Langrock[1]

[1] Bielefeld University, Germany
[2] Woods Hole Oceanographic Institution, Massachusetts, USA

E-mail for correspondence: jennifer.pohle@uni-bielefeld.de

Abstract: Hidden Markov models are popular tools for modeling time series that are driven by latent state processes. When multiple such time series, associated with different individuals, are observed simultaneously, then independence is commonly assumed. Here we discuss how coupled hidden Markov models, where potential dependence between individuals is explicitly addressed, can be used to model interaction between individuals. We provide two case studies to demonstrate the potential of this class of models. First, we apply the model to animal movement data of a dolphin mother and its calf. Second, we analyse the performance of a football team and how it depends on the performance of the opposing team.

Keywords: time series; latent variables; Markov chains

1 Introduction

The question of how individuals interact and affect each other is of much interest in disciplines such as the social sciences, epidemiology, and ecology. Structural econometric models (Hartmann et al., 2008) or network analysis (Jacoby and Freeman, 2016) are two example methods used to investigate these interdependencies. In this paper, however, we focus on situations with the additional complexity that each individual's observations are driven by an underlying latent process, and the dependence between individuals manifests itself in their underlying unobserved processes. For instance, in the context of animal movement, the observed movement patterns of an animal depend on its unobserved behavioral modes, like foraging or resting, which in turn will often depend on the behaviors exhibited by conspecifics.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

Similarly, in sport competitions, the observed performance of a team depends both on its own but also its opponent's current form. Coupled hidden Markov models (CHMMs) are flexible time series models for multiple time series which assume the observations to depend on underlying interacting state sequences. Hence, they provide an intuitive and convenient framework for the situations considered. After introducing the basic model formulation, we illustrate its use by applying it to two different data sets, first to animal movement data, and second to football data of the German Bundesliga. Our preliminary results are presented in the third section.

2 Coupled hidden Markov models

Let $\{Y_{i,t}\}_{t=1}^{T}$ denote the observed time series of length T belonging to individual $i = 1, \ldots, I$. A CHMM assumes each of the I time series to depend on an underlying state sequence $\{S_{i,t}\}_{t=1}^{T}$ with a finite number of states, i.e. $S_{i,t} \in \{1, \ldots, N\}$. At each time point, the current state $S_{i,t}$ completely determines the distribution of $Y_{i,t}$. Hence, given the state sequences, the observations are conditionally independent of each other: $\Pr(Y_{i,t}|Y_{i,t-1}, \ldots, Y_{i,1}, S_{i,t}, \ldots, S_{i,1}) = \Pr(Y_{i,t}|S_{i,t})$. To account for interactions between the individuals, however, the future state $S_{i,t+1}$ depends not only on its current state $S_{i,t}$ — as would be the case if we were to consider I separate hidden Markov models (HMMs) — but on the current states of all I state sequences. Thus, summarising the states in the vector $\mathbf{S}_t = (S_{1,t}, \ldots, S_{I,t})$,

$$\Pr(S_{i,t+1}|\mathbf{S}_1, \ldots, \mathbf{S}_t) = \Pr(S_{i,t+1}|\mathbf{S}_t) \neq \Pr(S_{i,t+1}|S_{i,t}).$$

This dependence structure is displayed in Figure 1.

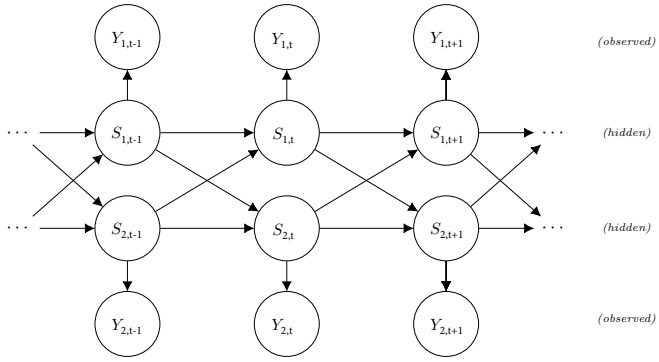


FIGURE 1. Dependence structure of a CHMM for $I = 2$ individuals.

The CHMM can be regarded as an HMM with an extended state space of dimension N^I, in which each state represents an I–tuple corresponding

to the possible states of $\mathbf{S}_t$. Hence, the corresponding transition probability matrix (t.p.m.) is of dimension $N^I \times N^I$. This model formulation has the advantage that the standard HMM machinery can be used for parameter estimation and inference, for instance numerical maximization of the likelihood based on the forward algorithm or the use of the Viterbi algorithm for state decoding (see, for example, Zucchini et al., 2016). The main disadvantage of this model formulation is that the number of parameters grows exponentially with the number of states and individuals. The t.p.m. of a CHMM can also be parameterized more parsimoniously, for instance as a probability product (Brand, 1997) or as a mixture distribution (Saul and Jordan, 1999). Nevertheless, these approaches are less flexible and correspond to more restrictive dependence assumptions regarding the interaction between individuals.

3 Case studies

3.1 Dolphin movement data

In our first case study, we model the tortuosity of a dolphin mother and its calf, calculated across 10–second intervals, with a total sample size of $T = 6546$. The tortuosity values are bounded between 0 and 1 and provide a measure of how tortuous the track of an animal is, with 1 meaning that the animal goes straight without any turnings. To model the different movement patterns of the interacting dolphins, we fit a CHMM, with $N = 3$ (i.e. N^I states in the HMM formulation of the CHMM) and beta state–dependent distributions, using numerical maximization of the likelihood. Figure 2 displays the estimated state–dependent distributions for the dolphin mother and calf, respectively. For both animals, their 3 different states correspond to low (1), medium (2) and high (3) tortuosity levels.
From Table 1 it can be seen that while mother and calf show a high level of synchrony in their behavior, there are also occasional deviations: For 4.6% of the observed time points, the two individuals' behavioral states are classified differently. These results could be used as a starting point for further studies of the dolphin behavior. For instance, it might be interesting to investigate why the dolphin movement differs at these time points.

TABLE 1. Dolphin CHMM: Number of occasions on which states were active according to the Viterbi-decoded sequence.

(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
2330	85	23	46	2145	108	2	39	1768

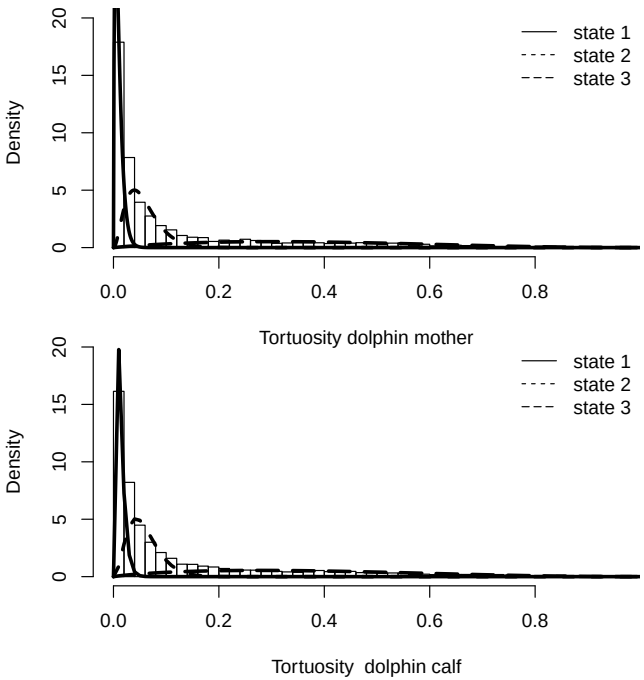


FIGURE 2. Histograms and estimated state—dependent distributions for the tortuosity of the dolphin mother and calf, respectively.

3.2 Football performance data

In our second case study, we are interested in the dynamics in football matches and how the performances of the two teams involved evolve over time within a match. For our case study, we focus on the German football team Bayer Leverkusen and its matches in the German Bundesliga (season 2017/2018). We assume that the performance is driven by the current (latent) form (or momentum) of the team, which can vary throughout the match and which may also be influenced by the form (momentum) of the opposing team. As a performance measure we use the number of passes at one-minute intervals. We fit a 2-state CHMM with Conway-Maxwell-Poisson state-dependent distributions to account for possible over- and underdispersion. The resulting state-dependent mean number of passes are 3.14 and 9.02 for Bayer Leverkusen, and 3.12 and 8.62 for the opposing teams, respectively. Under the fitted model, the stationary distribution, rounded to 2 decimal places, is $\boldsymbol{\delta} = (0.26, 0.29, 0.45, 0.00)$ for states $(1,1)$,

$(1,2)$, $(2,1)$ and $(2,2)$, respectively. As expected, it hardly ever happens that both teams are in the state corresponding to frequent passing. The model indicates that Bayer Leverkusen was the dominant team most of the time, represented by the third state (2,1).

4 Discussion

Our preliminary results suggest that CHMMs are promising and convenient tools for modeling interacting individuals, which can be used in various areas of empirical research. The major caveat of these models is that the number of parameters increases rapidly, such that long time series may be needed for stable estimation. The incorporation of covariates in both the observed and the hidden processes are straightforward and could offer insights into how the interactions are affected by external variables.

References

Brand, M. (1997). Coupled hidden Markov models for modeling interacting processes. *Technical Report 405*, Massachusetts Institute of Technology Media Laboratory, Cambridge.

Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., Hosangar, K. and Tucker, C. (2008). Modeling social interactions: Identification, empirical methods and policy implications. *Marketing Letters*, **19**, 287—-304.

Jacoby, D. M. P. and Freeman, R. (2016). Emerging network-based tools in movement ecology. *Trends in Ecology & Evolution*, **31**, 301—314.

Saul, L. K. and Jordan, M. I. (1999). Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simpler ones. *Machine Learning*, **37**, 75—-87.

Zucchini, W., MacDonald, I. L. and Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R*, 2nd Edition. Boca Raton: Chapman & Hall/CRC.

Latent Ornstein-Uhlenbeck models for Bayesian analysis of multivariate longitudinal categorical responses

Trung Dung Tran[1], Emmanuel Lesaffre[1], Geert Verbeke[1], Joke Duyck[2]

[1] Public Health and Primary Care, KU Leuven, 3000 Leuven, Belgium
[2] Department of Oral Health Sciences, KU Leuven, 3000 Leuven, Belgium

E-mail for correspondence: trungdung.tran@kuleuven.be

Abstract: We propose a Bayesian latent Ornstein-Uhlenbeck model to analyze unbalanced longitudinal data of binary and ordinal variables, which are manifestations of fewer continuous latent variables. Existing approaches are limited to data collected at regular time intervals. Our proposal makes use of an Ornstein-Uhlenbeck (OU) process for the latent variables to overcome this limitation. It also allows for both non-oscillating and oscillating processes. We illustrate our proposed model with two motivating datasets. The BelRAI dataset was obtained from a registry on the elderly population in Belgium. We were interested in predictive relationships between oral health and general health status. The ALS dataset contains patients with amyotrophic lateral sclerosis disease. We were interested in how bulbar, cervical, and lumbar functions evolve over time.

Keywords: Bayesian analysis; Eigenvalues; Latent variables; Ornstein-Uhlenbeck processes.

1 Introduction

Frequently in longitudinal data, subjects are measured at irregular time points, resulting in unbalanced data. In many cases, the outcomes are manifestations of one or more underlying latent characteristics. For example, in amyotrophic lateral sclerosis (ALS) disease, ten indicators are used to represent three latent functions: bulbar, cervical, and lumbar. We were interested in how the latent variables evolve over time.
The research question led to a joint framework consisting of an item response theory model linking the responses to the latent variables and a

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

model describing continuous latent variables over time. For balanced data, a vector autoregressive process can be fitted (e.g. Tran et al., 2019). For unbalanced data, an Ornstein-Uhlenbeck (OU) process is a better choice (e.g. Oravecz et al., 2016). However, this model has not been applied to a latent structure. Moreover, current approaches are not suitable here because of assuming unrealistic constraints. We show that our approach is able to fit the data appropriately.

2 Proposed model

2.1 Model specification

Denote Y_{ijk} the observed response for the k^{th} item of the i^{th} individual at time t_{ij} where $i = 1, ..., N$, $j = 1, ..., n_i$, $k = 1, ..., K$, n_i is the number of occasions for individual i, K is number of observed responses and N is number of individuals. The observed items are assumed to represent R latent variables, $\boldsymbol{\xi}_{ij} = (\xi_{ij1}, ..., \xi_{ijr}, ..., \xi_{ijR})^T$, as specified by:

$$h(P(Y_{ijk} \leq m)) = \theta_{km} + \boldsymbol{\beta}_k^T \boldsymbol{x}_{ij} + \boldsymbol{\lambda}_k^T \boldsymbol{\xi}_{ij} + b_{ik},$$

where $h(.)$ is a link function (typically a logit or probit function) and m $(0 \leq m \leq c_k - 2)$ is some score of item k with c_k the number of categories. The parameters θ_{km} and $\boldsymbol{\lambda}_k$ are item-specific location and discrimination (factor loading) parameters, respectively. Furthermore, $\boldsymbol{\beta}_k$ is a $p \times 1$ vector of regression parameters and $\boldsymbol{x}_{ij}$ is a $p \times 1$ vector of covariates for individual i at time t_{ij}. Finally, $b_{ik} \sim N(0, \sigma_{bk}^2)$, the random effect for item k of individual i, is incorporated to take local dependence into account (Tran et al., 2019). The model for the latent variables is specified as follows (e.g. Blackwell, 2003):

$$\begin{aligned} \boldsymbol{\xi}_{ij} &\sim N(\boldsymbol{\mu} + e^{-\Gamma d_{ij}} \left(\boldsymbol{\xi}_{i,j-1} - \boldsymbol{\mu}\right), \Omega - e^{-\Gamma d_{ij}} \Omega e^{-\Gamma^T d_{ij}}), \\ \boldsymbol{\xi}_{i1} &\sim N(\boldsymbol{\mu}, \Omega), \end{aligned}$$

where $d_{ij} = t_{ij} - t_{i,j-1}$, and $\boldsymbol{\mu}, \Omega, \Gamma$ satisfy the following conditions:

$$\begin{aligned} &\text{The real part of each eigenvalue of } \Gamma \text{ is positive,} \\ &\Gamma\Omega + \Omega\Gamma^T \text{ is a covariance matrix,} \\ &\Omega \text{ is a covariance matrix,} \end{aligned} \tag{1}$$

where $e^M = I + \sum_{j=1}^{+\infty} \frac{M^j}{j!}$ for a square matrix M with $M^j = M \times ... \times M$ (j times). We fixed $\boldsymbol{\mu}$=$\mathbf{0}$ and Ω is a correlation matrix for model identification.

2.2 Eigenvalues of the drift matrix Γ

Although constraint (1) specifies that the real part of each eigenvalue of Γ is positive, a number of proposals are limited to real eigenvalues (e.g.

Blackwell, 2003; Oravecz et al., 2016). Assuming real eigenvalues might facilitate computation but this assumption is unrealistic as seen in the ALS application. In contrast, our proposal allows for both real and complex eigenvalues. Specifically, we applied the original constraint (1) and solved the mathematical conditions so that Γ satisfies this constraint. In short, when $R = 2$, constraint (1) is replaced by the following

$$\begin{cases} \gamma_{11} + \gamma_{22} > 0 \\ \gamma_{11}\gamma_{22} - \gamma_{12}\gamma_{21} > 0 \end{cases},$$

whereas the following was used to replace constraint (1) in case $R = 3$:

$$\begin{cases} -\gamma_{33} - \gamma_{22} - \gamma_{11} < 0 \\ -\gamma_{31}\gamma_{13} - \gamma_{32}\gamma_{23} + \gamma_{33}\gamma_{22} + \gamma_{33}\gamma_{11} - \gamma_{21}\gamma_{12} + \gamma_{22}\gamma_{11} > 0 \\ -\gamma_{31}\gamma_{12}\gamma_{23} - \gamma_{32}\gamma_{21}\gamma_{13} + \gamma_{31}\gamma_{13}\gamma_{22} + \gamma_{32}\gamma_{23}\gamma_{11} + \gamma_{33}\gamma_{21}\gamma_{12} - \gamma_{33}\gamma_{22}\gamma_{11} < 0 \end{cases},$$

where γ_{ij} denotes the (i, j) element of Γ.

3 Application to the BelRAI and ALS dataset

3.1 BelRAI

Three binary oral health (OH) indicators: non-intact teeth, chewing difficulty, and dry mouth, and four ordinal general health (GH) scales: Activities of Daily Living, Cognitive Performance Scale, Depression Rating Scale, and Changes in Health, End-Stage Disease, Signs, and Symptoms Scale, represent OH and GH status, respectively. It is of interest to describe the development of OH and GH status over time and to assess the importance of the cross-lagged effects, i.e. the additional information that the current OH (resp. GH) status provides on the future GH (resp. OH) status.

The results in Figure 1 indicate that the current OH (resp. GH) status provides additional information in predicting the future value of GH (resp. OH) status given their current status. The cross-lagged effects from OH to GH in the BelRAI population suggest that the presence of OH problem can be considered as a symptom of GH problem in the future.

3.2 ALS

Amyotrophic lateral sclerosis is a progressive neurological disease that causes a gradual degeneration and death of motor nerve cells. ALS Functional Rating Scale was developed to monitor disease progression by measuring clinical features. It contains ten items falling into three functions: bulbar

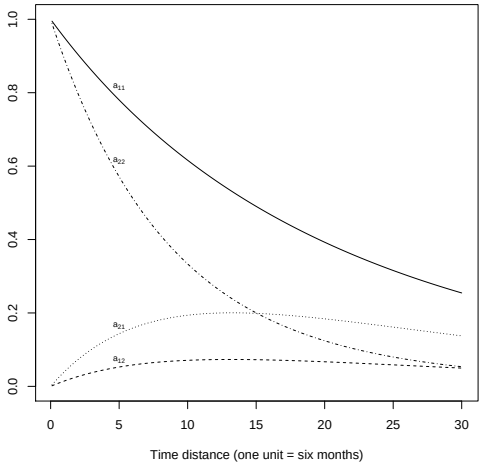


FIGURE 1. BelRAI: Estimated $e^{-\Gamma\Delta t}$ where a_{11}, a_{22}, and a_{12}, a_{21} are the autoregressive and cross-lagged parameters, respectively.

(speech, salivation, swallowing, breathing), fine motor (or cervical) (handwriting, cutting, and dressing), and gross motor (or lumbar) (turning, walking, and climbing) function. Predictive relationships between these functions were of interest. From the PRO-ACT database (Atassi et al., 2014), a random subset containing 300 subject with 2911 observations were taken. Figure 2 indicates predictive relationships between the latent functions. In addition, when time distance changes, the orders and signs of the parameters in the transition matrix $e^{-\Gamma\Delta t}$ also change. It is because every line in Figure 2 oscillates. The reason is that two out of three eigenvalues of the drift matrix are not real and therefore the process of three latent neurological functions is oscillating. In this case, assuming only real eigenvalues is not appropriate because it eliminates the class of oscillating processes (Kuiper and Ryan, 2018).

4 Discussion

We have introduced the multivariate OU process for analyzing the latent continuous variables, allowing a continuous time analysis at the latent level. Our simulation study (not shown here) and the ALS application show that assuming real eigenvalues for the drift matrix of the OU process can lead to biased estimates and/or misleading inference when the true process is oscillating. Our proposal allows real and complex eigenvalues, making it available for analyzing both non-oscillating and oscillating processes.

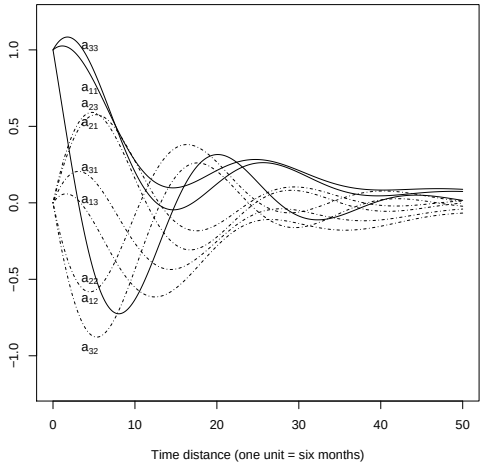


FIGURE 2. Estimated transition matrix $e^{-\Gamma\Delta t}$ as functions of time distance (Δt) where a_{11}, a_{22}, and a_{33} are the autoregressive parameters, and the others are the cross-lagged parameters

Acknowledgments: We thank the funding C24/15/034, KU Leuven, Belgium for financial support, and Prize4Life Israel for using the ALS dataset.

References

Atassi, N., Berry, J., Shui, A., et al. (2014). The PRO-ACT database. *Neurology*, **83**, 1719 – 1725.

Blackwell, P. G. (2003). Bayesian inference for Markov processes with diffusion and discrete components. *Biometrika*, **90**, 613 – 627.

Kuiper, R. M. and Ryan, O. (2018). Drawing conclusions from cross-lagged relationships: Re-considering the role of the time-interval. *Structural Equation Modeling: A Multidisciplinary Journal*, **25**, 809 – 823.

Oravecz, Z., Tuerlinckx, F., and Vandekerckhove, J. (2016). Bayesian data analysis with the bivariate hierarchical Ornstein-Uhlenbeck process model. *Multivariate Behavioral Research*, **51**, 106 – 119.

Tran, T. D., Lesaffre, E., Verbeke, G., and Duyck, J. (2019). Modeling local dependence in latent vector autoregressive models. *Re-submitted to Biostatistics*.

An adaptive lasso Cox frailty model for time-varying covariates based on the full likelihood

Andreas Groll[1], Maike Hohberg[2]

[1] Faculty of Statistics, TU Dortmund University, Germany
[2] Chair of Statistics, University of Goettingen, Germany

E-mail for correspondence: groll@statistik.tu-dortmund.de

Abstract: This paper proposes a method to regularize Cox frailty models that accommodates time-varying covariates and is based on the full likelihood. A particular advantage of this framework is the explicit modeling of the baseline hazard in a non-linear way, e.g. via P-splines. Additionally, adaptive weights are included to stabilize the estimation. The method is implemented in R in the function coxlasso and will be compared to other packages for regularized Cox regression.

Keywords: Cox frailty model; lasso; full likelihood; time-varying covariates; penalization.

1 Introduction

Cox's well known proportional hazards model (Cox, 1972) assumes the semi-parametric hazard

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t)\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \tag{1}$$

where $\lambda(t|\mathbf{x}_i)$ denotes the hazard function for individual i at time t conditional on covariates $\mathbf{x}_i$. The shared baseline hazard $\lambda_0(t)$ is usually not further specified and $\boldsymbol{\beta}$ is a vector for p fixed effects. Estimation of the model is typically based on maximizing the partial likelihood which has the advantage of removing $\lambda_0(t)$ from the estimation of $\boldsymbol{\beta}$. For the case of a large number of predictors, the lasso penalty was incorporated into the Cox model to enable variable selection and shrinkage (Tibshirani, 1997). Different algorithms to fit the penalized model have been proposed by e.g. Gui and Li (2005) using least-angle regression (LARS), Simon et al. (2011) via a

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

coordinate descent algorithm, and Goeman (2010) who combines gradient ascent optimization with the Newton Raphson algorithm.
These procedures are based on the partial likelihood which can lead to a loss in efficiency and precision in small and moderate samples (see, e.g., Cox and Oakes, 1984). Considering the popularity of regularized Cox models and given that in many medical applications the sample size is often rather small, it seems surprising that there is, to the best of our knowledge, currently no available implementation that uses a lasso penalty within the Cox full likelihood model. Especially for datasets with a small or moderate number of observations, using the full likelihood does not drastically increase computing time.
Despite the predominance of the partial likelihood in existing `R` routines, there are (at least) two major advantages when using the full likelihood:

1. the baseline hazard can be modeled explicitly, e.g., using a basis function approach such as P-splines (see, e.g., Eilers and Marx, 1996),

2. the full likelihood model can easily be extended by a wide class of frailty distributions including random intercepts and random slopes.

Our approach is implemented in an `R` function called `coxlasso` that includes a (adaptive) lasso penalization and can easily accommodate changing covariates, frailties, and time-varying coefficients. Currently, a working version is directly available from the authors upon request, which will be incorporated in the `R` package `PenCoxFrail` (Groll, 2016) soon.
Besides a small sample size, the full likelihood might become relevant when covariates change frequently. In survival analysis we typically deal with data consisting of a tuple (T_i, d_i) with d_i indicating whether an event happened, i.e $d_i = 1$ if the survival time is completely observed, whereas $d_i = 0$ if this observation is right censored. The random variable T_i can be described by event time $\tilde{T}_i$ and censoring time C_i via $T_i = \min(\tilde{T}_i, C_i)$. Since $f(t) = \lambda(t)S(t)$, we get

$$
\begin{aligned}
f(T_i, d_i) &= \left(f(\tilde{T}_i)P(\tilde{T}_i < C_i)\right)^{d_i} \cdot \left(g(C_i)P(\tilde{T}_i \geq C_i)\right)^{1-d_i} \propto f(T_i)^{d_i} \cdot S(T_i)^{(1-d_i)} \\
&= \left(\lambda_0(T_i)\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{d_i} \exp\left(-\int_0^{T_i} \lambda_0(s)\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})ds\right).
\end{aligned}
$$

A suitable expansion of the full likelihood over all individuals thus yields

$$
\begin{aligned}
L(\lambda_0(t), \boldsymbol{\beta}) &= \prod_{i=1}^{n} \left(\lambda_0(t_i)\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{d_i} \underbrace{\exp\left(-\int_0^{t_i} \lambda_0(s)\right) ds}_{S_0(t_i)} \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \\
&= \prod_{i=1}^{k} \frac{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\sum_{j \in R(t_{(i)})} \exp(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta})} \cdot \left(\sum_{j \in R(t_{(i)})} \lambda_0(t_{(i)})\exp(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}) \cdot \prod_{i=1}^{n} S_0(t_i)\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right), \quad (2)
\end{aligned}
$$

where $t_{(1)} < \cdots < t_{(i)} < \cdots < t_{(k)}$ are the event times of the uncensored individuals, and $R(t)$ is the set of individuals under risk at time t. The first factor of equation (2), which is the ratio of the event probabilities of all individuals that died at time t and the event probabilities of all individuals at risk at that time, corresponds to the *partial likelihood*. Besides removing the baseline hazard from the inference process of $\boldsymbol{\beta}$, the partial likelihood is attractive since covariate information can be easily included and it is not affected by the censoring pattern (Efron, 1977). However, it actually is not a real likelihood as it ignores the integral part of the full likelihood and, hence, certain covariate information from non-failure intervals. It is thus not based on all observations. Since the partial likelihood nearly contains all of the information about $\boldsymbol{\beta}$, the estimate $\hat{\boldsymbol{\beta}}$ is still asymptotically efficient. Ignoring the second factor of equation (2), might not give satisfying estimates if there is a lot of of information in non-failure intervals that influence the survival outcome. In particular, time-varying covariates result in splits of the data and could create several new, censored observations. The more often covariates change, the more splits get neglected since the partial likelihood only considers the status of the covariates at the event times but not in between.

The literature on the full likelihood for the regularized Cox model is rather scarce and only a few approaches exist for the standard $n > p$ case. However, none of these analyze the case of changing covariates nor provide a readily implemented `R` packages. For this reason, alongside with the `coxlasso` function we also provide the function `coxFL`, which implements the Cox full likelihood approach and allows for changing covariates, frailties, and time-varying coefficients.

2 Methodology

A Cox frailty model accounts for heterogeneity in the population and is given by

$$\lambda_{ij}(t|\mathbf{x}_{ij}, b_j) = b_j \lambda_0(t) \exp(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}), \qquad i = 1, \ldots, n;\ j = 1, \ldots, N, \tag{3}$$

where individual i belongs to cluster j resulting in frailty component b_j for that particular cluster. Due to its mathematical convenience, these frailties are often assumed to follow a gamma distribution but to allow for a more flexible predictor structure, assuming log-normally distributed frailties is more appropriate. Hence, we specify $\mathbf{b}_j \sim N(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta}))$ with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{Q}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. Replacing b_j in the conditional hazard function in (3) with multiplicative frailties following a multivariate log-normal distribution possibly also containing random slopes, yields:

$$\lambda(t|\mathbf{x}_{ij}, \mathbf{u}_{ij}, \mathbf{b}_j) = \lambda_0(t) \exp(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{u}_{i,j}^{\mathrm{T}} \mathbf{b}_j)\,,$$

where $\mathbf{u}_{ij}$ is a vector of covariates associated with frailties $\mathbf{b}_j$. An extension to the simple predictor $\eta_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{u}_{ij}^{\mathrm{T}}\mathbf{b}_j$ are time-varying covariate effects $\gamma_k(t)$. These effects can be estimated using a B-spline representation, i.e.

$$\gamma_k(t) = \sum_{m=1}^{M} \alpha_{k,m} B_m(t, d)\,, \tag{4}$$

with $\alpha_{k,m} = 1, \ldots, M$ denoting unknown spline coefficients associated with the m-th B-spline basis function $B_m(t, d)$ of degree d. Equivalently, the baseline hazard can also be modeled using B-splines. In this way, it is shifted into the predictor $\eta_{ij}(t)$ using a $\log(\cdot)$ transformation, $\gamma_0(t) := \log(\lambda_0(t))$, where $\gamma_0(t)$ is again expanded in B-splines, i.e. $\gamma_0(t) = \sum_{m=1}^{M} \alpha_{0,m} B_m(t, d)$. The conditional hazard function can thus be written as

$$\lambda(t|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{u}_{ij}, \mathbf{b}_j) = \exp(\eta_{ij}(t))\,, \tag{5}$$

with corresponding predictor

$$\eta_{ij}(t) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=0}^{r} \nu_{ijk}^{\mathrm{T}}\boldsymbol{\alpha}_k + \mathbf{u}_{ij}^{\mathrm{T}} b_j\,, \tag{6}$$

where $\nu_{ijk} = z_{ijk} \cdot \mathbf{B}(t)$ and $\mathbf{z}_{ij}^{\mathrm{T}} = (1, z_{ij1}, \ldots, z_{ijr})$ is a covariate vector belonging to baseline hazard and time-varying coefficients $\gamma_k(t)$. Furthermore, $\boldsymbol{\alpha}_k^{\mathrm{T}} = (a_{k,1}, \ldots, \alpha_{k,M})$ are corresponding spline coefficients, where $k = 0, \ldots, r$ indexes the baseline hazard or the k-th time varying effect. The matrix $\mathbf{B}(t)^{\mathrm{T}} = (B_1(t, d), \ldots, B_M(t, d))$ captures the M basis functions evaluated at time points t.
Let now $\boldsymbol{\alpha}^{\mathrm{T}} = (\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_r^{\mathrm{T}})$ collect all spline coefficients corresponding to the baseline hazard and time-varying effects in case they are included. Analogously to equation (2), estimation of (5) can be based on the full likelihood which is given for a single cluster j by

$$L_j = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{b}_j) = \prod_{j=1}^{N_i} \exp(\eta_{ij}(t_{ij}))^{d_{ij}} \exp\left(-\int_0^{t_{ij}} \exp(\eta_{ij}(s))ds\right)\,. \tag{7}$$

The corresponding log-likelihood can be maximized using a penalized quasi-likelihood approach proposed by Breslow and Clayton (1993), that involves the marginal log-likelihood given by

$$\ell^{mar}(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{j=1}^{N} \log\left(\int L_j(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{b}_j) p(\mathbf{b}_j|\boldsymbol{\theta}) d\mathbf{b}_j\right)\,, \tag{8}$$

depending on parameter vector $\boldsymbol{\delta}^{\mathrm{T}} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \mathbf{b}^{\mathrm{T}})$ and on $\boldsymbol{\theta}$, the parameters of the covariance structure of random effects $\mathbf{b}_j$ as specified before. The density of the random effects is given by $p(\mathbf{b}_j|\boldsymbol{\theta})$. Following Breslow and

Clayton (1993) and applying Laplace approximation, leads to a penalty term $\mathbf{b}^\top \mathbf{Q}^{-1}(\boldsymbol{\theta})\mathbf{b}$ that is deducted from the likelihood contribution of each cluster, i.e. the approximated log-likelihood is given by

$$\ell^{app}(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log L_j(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{b}_j) - \frac{1}{2}\mathbf{b}^\top \mathbf{Q}^{-1}(\boldsymbol{\theta})\mathbf{b}\,. \tag{9}$$

In order to perform variable selection and shrinkage, a lasso-type penalty is applied to linear effects while a second penalty controls the wiggliness of the smooth baseline hazard (and of additional time-varying coefficients, if present). Including the penalties, the log-likelihood can be written as

$$\ell^{pen}(\boldsymbol{\delta}, \boldsymbol{\theta}) = \ell^{app}(\boldsymbol{\delta}, \boldsymbol{\theta}) - J_\beta(\boldsymbol{\beta}) - J_\zeta(\boldsymbol{\alpha}), \tag{10}$$

where $J_\beta(\boldsymbol{\beta}) = \xi \sum_{l=1}^{p} w_l \, |\beta_l|$ is a lasso penalty that shrinks less important (time-constant) fixed effects $\beta_l, l = 1, \ldots, p$, towards zero and is able to exclude them from the predictor. Furthermore, $\xi \geq 0$ is a tuning parameter controlling the strength of the penalization that needs to be chosen by an appropriate technique, e.g., K-fold cross-validation. Additionally, we incorporate adaptive weights $w_l := 1/\,|\hat{\beta}_l^{(ML)}|$ given by the inverse of the corresponding (unpenalized) maximum likelihood (ML) estimator.
If categorical variables are present, the lasso penalty can be combined with a group lasso penalty (see Meier et al., 2008). In this case, the categorical variable is dummy encoded forming a group of dummies and $\boldsymbol{\beta}_l$ collects the corresponding coefficients of the particular group. The L_2 norm of vector $\boldsymbol{\beta}_l$ is penalized yielding penalty

$$J_\beta(\boldsymbol{\beta}) = \xi \sum_{l=1}^{p} w_l \sqrt{df_l} \, ||\boldsymbol{\beta}_l||_2 \,,$$

where df_l is the number of dummies in group l and is used to rescale the penalty according to the dimensionality of $\boldsymbol{\beta}_l$. In this case the corresponding weights have the general form $w_l := 1/||\hat{\boldsymbol{\beta}}_l^{(ML)}||_2$.
Since both baseline hazard and time-varying coefficients are expanded in B-spline basis functions, second order differences of adjacent spline coefficients $\alpha_{k,m}$ are penalized in $J_\zeta(\boldsymbol{\alpha})$ to control the roughness of the smooth functions. To determine the optimal amount of smoothing, we suggest a mixed-model representation of the penalized spline approach allowing data driven, fast selection. In this view, the regression spline coefficients $\boldsymbol{\alpha}_k$ that are subject to penalization are taken to be random with corresponding random effect distributions $N \sim (\mathbf{0}, \sigma^2_{\boldsymbol{\alpha}_k}\mathbf{I})$. The reciprocal of $\sigma^2_{\boldsymbol{\alpha}_k}$ can then be used as an optimal smoothing parameter (see, e.g., Ruppert et al., 2003).
The estimation of the penalized likelihood in (10) is based on a Newton-Raphson algorithm and makes use of local quadratic approximations of the penalty terms following Oelker and Tutz (2017). Its performance will be investigated both in simulations and a real world data application

References

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association*, **88**, 9 – 25.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, bf 34(2), 187 – 220.

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman & Halls.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**(359), 557 – 565.

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

Goeman, J.J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, **52**, 70 – 84.

Groll, A. (2016). *PenCoxFrail: Regularization in Cox Frailty Models.* R package version 1.0.1.

Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**(13), 3001 – 3008.

Meier, L., Van de Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, **70**(1), 53 – 71.

Oelker, M.-R. and Tutz, G. (2017). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classiffcation*, **11**(1), 97 – 120.

Ruppert, D., Wand, M.P., and Carroll,R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, **39**(5), 1 – 13.

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, **16**, 385 – 395.

Simultaneous autoregressive models within GAMLSS

Lucas M. Oliveira[1], Fernanda De Bastiani[1], Dimitrios M. Stasinopoulos[2], Robert A. Rigby[2]

[1] Universidade Federal de Pernambuco, Brazil
[2] London Metropolitan University, UK

E-mail for correspondence: ldmo1@de.ufpe.br

Abstract: The main goal of this paper is to introduce the class of simultaneous autoregressive models within the GAMLSS framework. This implementation allows any or all the parameters of the distribution to be modelled as function of the explanatory variables, while the distribution do not have to belong to the exponential family. The methodology is applied to study income inequality in the State of Pernambuco, Brazil, considering the spatial structure of cities.

Keywords: areal data, conditional autoregressive, spatial effect

1 Introduction

In spatial data analysis, the data is indexed to a set of locations in space. According to the form that this set of locations is defined, there are three major fields of spatial statistics: *geostatistical data*, *areal data*, and *point pattern data* (Banerjee *et al.* 2014). This paper focus on the second approach.We assume that the data is a realization of a stochastic process where the space of variation is discrete. Each element is associated to a geographic region (unit area). A well-known model in the field is the simultaneous autoregressive model (SAR), which has applications in many fields sucha as ecological data, texture analysis and spatial econometrics.

1.1 GAMLSS

The generalized additive models for location, scale and shape (GAMLSS) were proposed by Rigby and Stasinopolous (2005). The observations of the response variable y_i are conditional independent with probability (density)

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

function $f(y_i|\boldsymbol{\theta}^i)$, conditioned on the vector $\boldsymbol{\theta}^{i\top} = (\theta_{i1}, \theta_{i2}, ..., \theta_{ip})$ of p unknown parameters. Each parameter can be modeled through different explanatory variables or/and random effects. For $k = 1, 2, 3, 4$, let $g_k(.)$ be a known monotone link function that associates $\boldsymbol{\theta}_k$ with independent variables. The random effect formulation of the model is given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{1}$$

where the vector $\boldsymbol{\eta}_k$ is the linear predictor and has length n. Similarly, $\boldsymbol{\theta}_k^\top$=$(\theta_{1k}, \theta_{2k}, ..., \theta_{nk})$ has the same length. The vector of the parameters $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J'_k k})$ has dimension J'_k, and the matrices of covariates $\boldsymbol{X}_k$ and $\boldsymbol{Z}_{jk}$ are of orders $n \times J'_k$ and $n \times q_{jk}$. The vector γ_{jk} has length J'_k and follows a normal distribution with $\boldsymbol{\gamma}_{jk} \sim N_{qjk}(\mathbf{0}, \mathbf{G}_{\boldsymbol{jk}}^-)$. The variance-covariance matrix $\mathbf{G}_{jk}^-$ and the precision matrix $q_{jk} \times q_{jk}$, $\mathbf{G}_{jk}$, is a function of a vector of hyperparameters $\boldsymbol{\lambda}_{jk}$.

2 Gaussian Markov random fields

As shown by Rue and Held (2005), consider $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top$ a random vector normally distributed with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. Let *G=(V,E)* be an non-directed graph, with $V = \{1, \ldots, q\}$, the set of vertices or nodes representing the q-area units and E is the set of edges that connect these areas. Hence, define that $\boldsymbol{\gamma} \in \mathbb{R}^n$ will be a Gaussian Markov random field (GMRF) with respect to the graph G if its density function is given by

$$\pi(\boldsymbol{\gamma}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-1/2} \exp(-\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma}^{-1})(\boldsymbol{\gamma} - \boldsymbol{\mu})), \tag{2}$$

and $\Sigma_{ij}^{-1} \neq 0$ if and only if $\{i, j\} \in E$ for all $i \neq j$. Hence, the symmetric precision matrix $\boldsymbol{\Sigma}^{-1}$ informs which areas are neighbors, given a criterion. For $\Sigma_{ij}^{-1} = 0$, we state that i and j are conditional independent, by the Markov property. The SAR Models, introduced by Whittle (1954), are GMRF models with a density function given by equation 2. They defined a spatial process simultaneously in $\mathbb{R}^2$ on a countable grid. This model with zero mean is given as follows:

$$\gamma_i = \sum_{j=1}^{q} b_{ij}\gamma_j + \varepsilon_i, \quad i = 1, \ldots, q,$$

which we can written in matrix form by

$$(\mathbf{I} - \mathbf{B})\boldsymbol{\gamma} = \boldsymbol{\varepsilon},$$

where $\mathbf{I}$ is a $q \times q$ identity matrix, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Lambda})$, $\mathbf{B}$ is a spatial dependence matrix with elements b_{ij}, denoting the dependence between the area

units, and $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B}^{\top})^{-1})$. This can be implemented within GAMLSS through random effects form given in (1). The equivalence between inverse penalty matrix of the GAMLSS and the precision matrix of SAR models is: $\mathbf{G}^{-1} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B}^{\top})^{-1} = \mathbf{K}^{-1}$. Thus, $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \lambda\mathbf{K}^{-1})$.
This is possible by writing the covariance matrix of the SAR model as covariance matrix of conditional autoregressive models (CAR) (Besag, 1974), as described in Hoef *et al.* (2018). The equivalence relation between these two occurs when the matrices of variances of these models are equal, and a first order SAR model is equivalent to a third order CAR model. For details about the implementation of CAR models in GAMLSS, see De Bastiani *et al.* (2018). Thus, $\boldsymbol{\gamma}$ is a *intrinsic* GMRF, and the penalty matrix $\mathbf{K}$ has dimension $q \times q$ and has elements:

$$K_{u,v} = \begin{cases} 0, \text{if } u \text{ and } v \text{ if they are not neighbors,} \\ -1, \text{if } i \text{ and } j \text{ are neighbors ,} \\ n_u, \text{ the number of neighbors of } u, \forall u = v. \end{cases}$$

The penalty matrix represents the pseudo-inverse of the covariance matrix of the CAR model. The structure of the penalty matrix for the two models is the same. The difference is in the neighborhood order for these two models. In the SAR model, the penalty matrix will be less sparse than in the CAR model.

3 Modeling Gini index with spatial dependence

The income inequality as measured by the Gini index in 2010 in the State of Pernambuco in Brazil is analysed. The variables considered are:

`Gini`: the index of gini each city of Pernambuco;

`Pibcap`: the gross domestic product per capita municipality;

`TX_DESEMP`: the proportion of unemployed people;

`IDoJOV`: the ratio between the elderly population and young people;

`TX_ANALF`: the proportion of illiterate people;

`PBF`: a benefit received by poor families;

`BPC`: a retirement benefit for poor people.

We use the **gamlss** package in `R` to select variables and the response distribution based on the Generalized Akaike Information Criterion (GAIC). The

distribution chosen was the zero-inflated beta (`BEZI`) distribution (available in package **gamlss.dist**) and the final chosen model is given by

$$
\begin{aligned}
Y &\sim \text{BEZI}(\hat{\mu}, \hat{\sigma}, \hat{\nu}),\\
\log(\frac{\hat{\mu}}{1-\hat{\mu}}) &= 0.088 - h_{11}(\texttt{Pibcap}) + h_{21}(\texttt{PBF}) + s(\texttt{city})\\
\log(\hat{\sigma}) &= 1.241 + 0.2433\texttt{Pibcap} + 7.93\texttt{IDoJOV} + h_{12}\texttt{BPC}\\
\log(\frac{\hat{\nu}}{1-\hat{\nu}}) &= -28.53.
\end{aligned}
$$

In the above equations h denotes *penalized splines* and s is a spatial SAR smoothing function. The variables `TX_ANALF` and `TX_DESEMP` were not statistically significant in the fitting. In relation to the modeling of the dispersion parameter, the increase of `Pibcap` in 1 unit, with everything more constant, increases the dispersion in 1.275 $[= e^{0.2433}]$. If the number of older people in relation to that of young people more dispersed is the distribution of the coefficient of `Gini`. The variable `PBC` has a positive impact on $\log(\hat{\sigma})$. The above model was selected with GAIC equal to -738.1156.
Figure 1 shows the city effect on $\log(\frac{\hat{\mu}}{1-\hat{\mu}})$, highlighting higher values for the economically richest cities in the state. Figure 2 shows the worm plot,

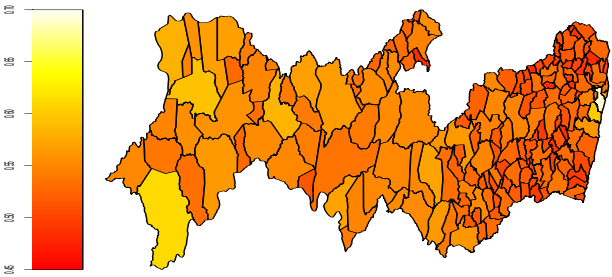


FIGURE 1. The fitted spatial effect for μ for the chosen model

introduced by Buuren and Fredriks (2001), of the residuals of the estimated model. We verified in the figure that the ordered residuals are close to their expected values due to their proximity to the horizontal line.

4 Conclusion

We show in this work an innovative approach to modeling areal data that is configured with SAR covariance structure within the GAMLSS framework. Also, we show an important way of employing this approach in the field of spatial econometrics and in the Gini coefficient modelling.

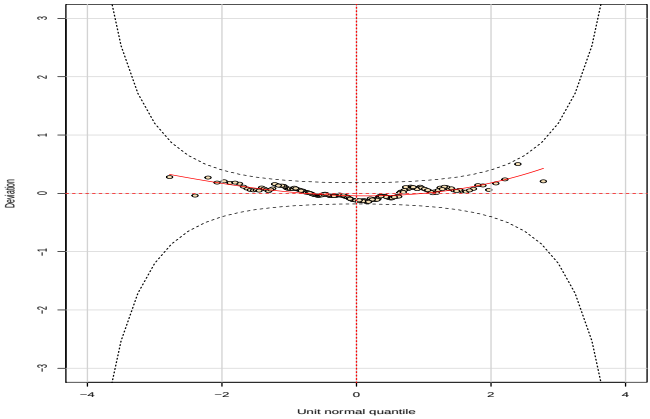


FIGURE 2. Worm plot from the of the fitted BEZI model.

5 Acknowledgments

The authors thank Coordination for the Improvement of Higher Education Personnel (CAPES) and Propesq/UFPE for the financial support. And they also thank the National Supercomputing Center at Federal University of Rio Grande do Sul (CESUP/UFRGS) for the computational support.

References

Banerjee, S. Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. London: Chapman & Hall.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*. Series B (Methodological). **36(2)**, 192 – 236.

Buuren, S. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, **20(1)**, 1259 – 1277.

De Bastiani, F., Rigby, R.A., Stasinopoulous, D.M., Cysneiros, A.H.M.A., and Uribe-Opazo, M.A. (2018). Gaussian Markov random field spatial models in GAMLSS. *Journal of Applied Statistics*, **45(1)**, 168 – 186.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54(3)**, 507 – 554.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. London: Chapman & Hall.

Hoef, J. M., Hanks, E. M., and Hooten, M. B. (2018). *On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models*. Spatial Statistics, **25**, 68 – 85.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, **41**, 434 – 449.

Variable selection in small area model with measurement error in covariates

S. Arima[1], S. Polettini[2]

[1] Dep. of Methods and Models for Economy, Territory and Finance, Sapienza University of Rome, Rome, Italy
[2] Dep. of Social and Economic Sciences, Sapienza University of Rome, Italy

E-mail for correspondence: serena.arima@uniroma1.it

Abstract: Model based small area estimation relies on mixed effects regression models that link the small areas and borrow strength from similar domains. The variability of the random effects, while accounting for lack of fit, affects uncertainty of both point and interval estimators of small area means. Random effects models play an important role in model-based small area estimation. Indeed, random effects account for any lack of fit of a regression model for the population of small areas on a set of explanatory variables. In the presence of good covariates, small variation of the random small area effects is expected, but when measurement error is present it has been proved that parameter estimates may be dramatically biased and the variability of the random effects and, consequently, of the small area means significantly increases. While the random effect may improve the adaptivity and flexibility of the Fay-Herriot model, it also increases the uncertainty of both point and interval estimators of small area means. Because of that, several tests and variable selection procedures have been developed in order to verify the presence or not of the random effects in such models. Adopting a fully Bayesian approach, we model the measurement error through a mixture that allows us, using spike and slab priors, to infer the presence or not of measurement error in the covariates. We empirically evaluate the accuracy of the estimates in different simulation scenario. We also apply the proposed procedure to the well known Battese data and to data from the 2010 Italian household budget survey (Banca d'Italia, Indagine sui bilanci delle famiglie italiane).

Keywords: Small area models; measurement error; spike and slab prior.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

1 Small area estimation and measurement error models

In the analysis of survey data, growing interest in specific subpopulations, obtained by breakdown of the population according to socio-demographic or geographic variables introduces unplanned domains (small areas), characterized by small -or even zero- sample sizes. Design-based estimators may have unacceptably large variances for such domains; small area estimation (Rao and Molina, 2015) defines estimation procedures that increase the effective sample size using sample information observed in other areas or previous periods. Mixed effects regression models are often introduced to this aim; random effects allow to capture the variation of the small area means not accounted for by the covariates. Area level small area models relate direct estimates to suitable auxiliary variables that are available from other surveys and administrative records; unit-level models relate the unit values of the study variable to suitable auxiliary variables with known area means. To obtain reliable model-based small area estimates, the availability of good covariates, implying small variation of the random small area effects, is crucial. However one may define a good model with poor covariates because they are affected by measurement error, an ubiquitous problem (Carroll *et al.*, 2006) also studied within the small area literature (see Ghosh *et al.*, 2006, Arima *et al.*, 2015 and references therein).
Denoting by uppercase letters the variables observed with error, and by lowercase letters the corresponding latent values, the measurement error model assumes that the covariate x_i $(i = 1, ..., n)$ is not available and that we observe $r \geq 1$ replicates $X_{ij} = x_i + \eta_{ij} \quad j = 1, ..., r$, with η_{ij}'s independent and identically distributed errors with zero mean.
Suppose there are m areas and let N_i be the known population size of area i. We define Y_{ij} $(i = 1, ..., m$ and $j = 1, ..., n_i)$ the response variable collected for the $j-$th individual in the $i-$th area and X_{ij} an auxiliary variable measured with error. Goal is to predict the small area means $\Gamma_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}$ given the available information.
Ghosh et al. (2006) consider the following model

$$
\begin{aligned}
Y_{ij} = \quad & \theta_i + \epsilon_{ij}, && \epsilon_{ij} \sim N(0, \sigma_e^2) && (1)\\
\theta_i = \quad & \beta_0 + \beta_1 x_i + v_i, && v_i \sim N(0, \sigma_v^2) && (2)\\
& X_{ij} \sim N(x_i, \sigma_\eta^2), && x_i \sim N(\mu_x, \sigma_x^2) && (3)
\end{aligned}
$$

Under (1)–(3) the expected value of the variable of interest given (β, u_i, x_i) is θ_i; assuming no selection bias and that the auxiliary information is available for each area, prediction of Γ_i can be based on prediction of θ_i.
Ghosh *et al.* (2006) derived a predictor for Γ_i based on the expected response on the $N_i - n_i$ unsampled units, conditional on the unknown parameters and the observed sample, denoted as $Y^{(1)}$; the shrinkage factors are

$B_i = \sigma_e^2/[\sigma_e^2 + n_i(\sigma_u^2 + \beta^2\sigma_x^2)]$. The empirical Bayes predictor is obtained by replacing the unknown model parameters with their estimators.
In the presence of poor covariates, the variability of the random effects increase, with shrinkage to the sampling component of the small area estimator. Correcting for measurement error reduces the random effects variability and improves the resulting estimates, provided the induced uncertainty is small compared to the sampling variation.
Datta and Mandal (2015) propose an area-level mixture model where the inclusion of the random component depends on the area-specific lack of fit of the model; this is obtained through a 'spike and slab" distribution assigned to the random small area effects, that are given probability p to be present in the model for the i-th area.

2 Proposed model and application

Working on the the Ghosh et al. (2006) model just described, we propose a unit-level small area model with measurement error in auxiliary variables that, borrowing from Datta and Mandal (2015), includes a "spike and slab" distribution for modelling the inclusion of the covariate measured with error. For greater generality, we also include a set of area-level covariates W_i measured without error. The hierarchical Bayes representation of our model is given as follows:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 z_i + \beta_2 W_i + v_i + \epsilon_{ij} \qquad v_i \sim N(0, \sigma_v^2), \epsilon_{ij} \sim N(0, \sigma_e^2) \\ z_i &= (1-\delta)\bar{X}_i + \delta x_i \end{aligned}$$

where $x_i \sim N(\mu_x, \sigma_x^2)$, $X_{ij} \sim N(x_i, \sigma_\eta^2)$ and $P(\delta = 1) = p = 1 - P(\delta = 0)$. In other words, conditional on δ, the auxiliary variable X is modelled with or without error. Indeed, if $\delta = 0$, then z_i equals the area-level mean of the observed covariate $\bar{X}_i$; on the other hand, when $\delta = 1$, the observed covariate is measured with error as $X_{ij} \sim N(x_i, \sigma_\eta^2)$, where σ_η^2 defines the variability of the measurement error and $x_i \sim N(\mu_x, \sigma_x^2)$ is the true latent value. Indeed, the prior distribution on z_i assigns a positive mass $(1-p)$ at $\bar{X}_i$ and spreads the remaining mass p according to a normal distribution centred at true unknown value of the covariate X, x_i.
Prior distributions for all unknown parameters are specified as in Ghosh et al. (2006): independent flat normal distributions have been specified ($\mu_{\beta_k} = 0$ and $\sigma_{\beta_k} = 10^5$, $k = 0, 1, 2$) for regression parameters, and independent flat inverse gamma distributions have been specified for all variance parameters with shape and scale parameters both equal to 0.001. With respect to p, we propose to specify a Beta prior. In many small area problems, it is likely to have some information about the measurement error mechanism that can be easily elicited in the specification of the parameters of the Beta distribution.

According to the model specification in (1), (2) and (3), after integrating out $v_1, ..., v_m$ and δ from the joint density of $(Y_{ij}, v_i, \delta, X_{ij})$, we get, conditional on the unknown parameters, that $Y_{i1}, ..., Y_{in_i}$ are independently distributed as a two-component mixture of normal distributions, where a positive mass $(1-p)$ is assigned to the model with no measurement error and the remaining mass p is spread to the model involving the measurement error.
While small area means will be estimated via Gibbs sampling, we study the conditional small area mean $\tilde{\theta}_i$, given the model parameters and the observed data $Y^{(1)}$. By direct calculation,

$$\tilde{\theta}_i(\beta, \sigma_v^2, \sigma_e^2, Y^{(1)}) = (1-\pi)(B(\beta_0+\beta_1\bar{X}_i)+(1-B)\bar{y}_i)+\pi(\tilde{B}(\beta_0+\beta_1\mu_x)+(1-\tilde{B})\bar{y}_i)$$

where $B = \frac{\sigma_e^2}{\sigma_e^2+n_i\sigma_u^2}$, $\tilde{B} = \frac{\sigma_e^2}{\sigma_e^2+n_i(\sigma_u^2+\beta_1^2\sigma_x^2)}$ and π is the conditional probability that $\delta = 1$ (that is, measurement error is present), $\pi = \frac{p}{p+(1-p)A}$,

$$A = \gamma^{-\frac{n}{2}} e^{-\frac{1}{2}\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\beta_0-\beta_1\bar{X}_i-\beta_2 W_i)^2}{\sigma_e^2+\sigma_u^2} - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\beta_0-\beta_1\mu_x-\beta_2 W_i)^2}{\sigma_e^2+n_i(\sigma_u^2+\beta_1^2\sigma_x^2)}\right)}$$

with $\gamma = \frac{\sigma_e^2+n_i(\sigma_u^2+\beta_1^2\sigma_x^2)}{\sigma_e^2+\sigma_u^2}$.
From the expression above, it can be easily grasped that, if the residual sum of squares (SSE) of the measurement error model increases, then the probability of belonging to the model with no measurement error increases. Moreover, the weight of the SSE of the measurement error model depends on the variances σ_e^2 and σ_v^2 but also on $\beta_1^2\sigma_x^2$, that is on the size of the effect of the covariate measured with error penalized by its variability.
We perform an extensive simulation study where we compare the performance of the proposed model with the model ignoring the measurement error and the model accounting for the measurement error, computing the mean squared error of the estimated small area means. The proposed model performs very similarly to the model ignoring the measurement error when the measurement error is very small and the probability p is coherently estimated very close to 0. When the measurement error is present, the proposed model performs very similarly in terms of parameter estimates to the model accounting for measurement error: however, small area predictions obtained with the proposed model show smaller variability.
A real data application has also been performed using data from the 2010 Italian household budget survey (Banca d'Italia, Indagine sui bilanci delle famiglie italiane). Also in this case, we conclude that small area predictions obtained with the proposed model gain is smaller variability with respect to the model accounting for measurement error.

References

Arima, S., Datta, G. S., and Liseo, B. (2015). Bayesian estimators for small area models when auxiliary information is measured with error. *Scandinavian Journal of Statistics,* **42**, 518 –529

Carroll, R.J., Ruppert, D., Stefanski, L. and Crainiceanu, C. (2006) *Measurement error in nonlinear models: a modern perspective.* 2nd edn. Chapman & Hall, CRC.

Datta, G.S. and Mandal, A. (2015). Small area estimation with uncertain random effects *Journal of the American Statistical Association*, **110 (512)**, 1735 – 1744.

Ghosh, M., Sinha, K. and Kim, D. (2006). Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error model, *Scandinavian Journal of Statistics*, **33**, 591 – 608.

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation.* Wiley.

Markov Chain Monte Carlo methods for discretely observed continuous-time multi-state semi-Markov models

Rosario Barone[1], Andrea Tancredi[1]

[1] Sapienza University of Rome, Italy

E-mail for correspondence: andrea.tancredi@uniroma1.it

Abstract: Inference for continuous time multi-state models presents considerable computational difficulties when the process is only observed at discrete time points with no additional information about the state transitions. In particular, when transitions between states may depend on the time since entry into the current state, and semi-Markov models should be fitted to the data, the likelihood function is neither available in closed form. In this paper we propose a Markov Chain Monte Carlo algorithm to simulate the posterior distribution of the model parameters.

Keywords: Markov models; Metropolis-Hastings; Weibull distribution.

1 Introduction

Let $\{X(t), t \geq 0\}$ be a continuous time multi-state process with state space $\mathcal{S} = \{1, 2, \ldots, S\}$. Models for continuous time multi-state process $X(t)$ can be defined via the transition intensity functions

$$q_{rs}(t, \mathcal{F}_t) = \lim_{\delta t \to 0} \frac{P\{X(t + \delta t) = s | X(t) = r, \mathcal{F}_t\}}{\delta t}$$

representing the instantaneous probability of a transition from state r to state s at time t when $\mathcal{F}_t$ is the past history up to time t. Considering

$$P\{X(t + \delta t) = s | X(t) = r, \mathcal{F}_t\} = \begin{cases} \gamma_{rs}\delta t + o(\delta t) & s \neq r \\ 1 + \gamma_{rr}\delta t + o(\delta t) & s = r \end{cases} \quad (1)$$

where $\gamma_{rs} \geq 0$ and $\gamma_{rr} = -\sum_{s \neq r} \gamma_{rs} = -\gamma_r$, we have Markov continuous time model. In the semi-Markov models, the transition intensity functions

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

also depend on the time spent in the current state, that is

$$q_{rs}(t, \mathcal{F}_t) = \lim_{\delta t \to 0} \frac{P\{X(t+\delta t) = s | X(t) = r, T^* = t - u\}}{\delta t}$$

where T^* denotes the entry time in the last state assumed before time t. Setting

$$P\{X(t+\delta t) = s | X(t) = r, T^* = t-u\} = \begin{cases} q_{rs}(u)\delta t + o(\delta t) & s \neq r \\ 1 - \sum_{l \neq r} q_{rl}(u)\delta t + o(\delta t) & s = r \end{cases}$$

where the expressions $q_{rs}(u)$ represent the cause-specific hazard functions, we describe the whole process $X(t)$. In fact, let $F_r(u)$ be the distribution with hazard function $F_r'(u)/(1 - F_r(u)) = \sum_{l \neq r} q_{rl}(u)$. Consider $p_{rs} = \int_0^\infty q_{rs}(u)(1 - F_r(u))du$ and $F_{rs}(u) = \frac{1}{p_{rs}} \int_0^u q_{rs}(v)(1 - F_r(v))dv$, for $s \neq r$. Then, $X(t)$ is the result of the state sequence generated by the Markov chain with transition probabilities p_{rs} and sojourn times depending on the departure and arrival states generated independently with distributions F_{rs}. To specify the functions $q_{rs}(u)$ we can also proceed directly by fixing the transition probabilities p_{rs} and the conditional sojourn distributions F_{rs}.

In this paper we assume that the sojourn time is Weibull distributed with density $f(u; \gamma_r, \alpha_r) = \gamma_r \alpha_r (\gamma_r u)^{\alpha_r - 1} e^{-(\gamma_r u)^{\alpha_r}}$ and does not depend on the exit stage. The model parameters are then $\theta = (p, \gamma, \alpha)$ with $p = (p_1, \ldots, p_S)$ comprising the transition probabilities $p_r = (p_{r1}, \ldots, p_{r\,r-1}, p_{r\,r+1}, \ldots, p_{rS})$ with $\sum_{s \neq r} p_{rs} = 1$, $\gamma = (\gamma_1, \ldots, \gamma_S)$ representing the rate parameters and $\alpha = (\alpha_1, \ldots, \alpha_S)$ representing the shape parameters.

2 Inference for semi-Markov models

Inference for the model parameters is straightforward when the whole process trajectory $x(t)$ is observed on the interval $[0, T]$. Let $s = (s_0, s_1, \ldots, s_\ell)$ be the state sequence and let $w = (w_0, w_1, \ldots, w_\ell)$ be the times in which the state transitions occur. Moreover let $n_{rs} = \sum_{j=0}^{\ell-1} I(s_j = r, s_{j+1} = s)$ be the transition counts, $n_r = \sum_s n_{rs}$ be the total number of visits of the state r and let $d_r = (d_{r1}, \ldots, d_{rn_r})$ be the set of sojourn times $w_{j+1} - w_j$ into the state r with $d_{s_\ell n_{s_\ell}} = T - w_\ell$. Then the density of (s, w) is

$$p(s, w|\theta) = \prod_{rs} p_{rs}^{n_{rs}} \prod_r \alpha_r^{n_r} \gamma_r^{\alpha_r n_r} \left(\prod_{j=1}^{n_r} d_{rj} \right)^{\alpha_r} e^{-\gamma_r^{\alpha_r} \sum_{j=1}^{n_r} d_{rj}^{\alpha_r}} \times \frac{1}{\alpha_{s_\ell} \gamma_{s_\ell}^{\alpha_{s_\ell}} d_{s_\ell n_{s_\ell}}^{\alpha_{s_\ell}}}$$

where the last factor is due to the truncation of the sojourn time for the last visited state.

Now suppose to observe the trajectory $x(t)$ only at fixed points so that the state sequence s and the transition times w are not available. Let

$y = (y_0, y_1, \ldots, y_m)$ be the observed states at times $0 = t_0 < t_1 < \ldots < t_m$. Note that the observation times can be irregularly spaced. Suppose also that they are non-informative for the underlying process. In this case, apart from the specific case of the phase-type distributions, e.g. Titman and Sharples (2010), the likelihood function for the semi-Markov processes is not analytically available and to make inference only numerical or approximate solutions have been proposed (Titman, 2014, Tancredi 2019).
For discretely observed data, Hobolth and Stone (2009) showed how to simulate the whole trajectory $x(t)$ conditional on the observed states y from a Markov process, that is when $\alpha_r = 1\ \forall r$. This way they were able to recover the vectors s and w. Their approach was based on the *uniformization* technique. To perform Bayesian inference, here we propose to embed their approach in a Metropolis-Hastings step to simulate the posterior distribution $\pi(s, w|y)$ under the semi-Markov assumption.
In fact suppose to draw (s, w) from the conditional distribution of $(s, w|y)$ assuming that $x(t)$ is a realization from a Markov process with rates $\tilde{\gamma}$. The corresponding proposal density is

$$q_M(s, w|y) = \frac{\pi_M(s, w)}{\pi_M(y)} \propto \prod_{rs} \tilde{p}_{rs}^{n_{rs}} \prod_r e^{-\tilde{\gamma}_r \sum_{j=1}^{n_r} d_{rj}} \times \frac{1}{\tilde{\gamma}_{s_\ell}}$$

where $\tilde{p}_{rs} = \tilde{\gamma}_{rs}/\tilde{\gamma}_r$. Employing an independent Metropolis-Hastings algorithm, the proposal (s', w') is accepted with probability

$$\min\left\{1, \frac{\pi(s', w'|y) q_M(s, w|y)}{\pi(s, w|y) q_M(s', w'|y)}\right\}$$

as a new value of the chain, where

$$\frac{\pi(s', w'|y)}{\pi(s, w|y)} = \frac{\prod_{rs} p_{rs}^{n'_{rs}} \prod_r {\alpha_r}^{n'_r} {\gamma_r}^{\alpha_r n'_r} \left(\prod_{j=1}^{n'_r} d'_{rj}\right)^{\alpha_r - 1} e^{-\gamma_r^{\alpha_r} \sum_{j=1}^{n'_r} {d'}_{rj}^{\alpha_r}} \alpha_{s_\ell} \gamma_{s_\ell}^{\alpha_{s_\ell}} d_{s_\ell}^{\alpha_{s_\ell} n_{s_\ell}}}{\prod_{rs} p_{rs}^{n_{rs}} \prod_r {\alpha_r}^{n_r} {\gamma_r}^{\alpha_r n_r} \left(\prod_{j=1}^{n_r} d_{rj}\right)^{\alpha_r - 1} e^{-\gamma_r^{\alpha_r} \sum_{j=1}^{n_r} d_{rj}^{\alpha_r}} \alpha_{s'_{\ell'}} \gamma_{s'_{\ell'}}^{\alpha_{s'_{\ell'}}} {d'}_{s'_{\ell'}}^{\alpha_{s'_{\ell'}} n'_{s'_{\ell'}}}}$$

and

$$\frac{q_M(s, w)}{q_M(s', w')} = \frac{\prod_{rs} \tilde{p}_{rs}^{n_{rs}} \prod_r e^{-\tilde{\gamma}_r \sum_{j=1}^{n_r} d_{rj}} \gamma_{s'_{\ell'}}}{\prod_{rs} \tilde{p}_{rs}^{n'_{rs}} \prod_r e^{-\tilde{\gamma}_r \sum_{j=1}^{n'_r} d'_{rj}} \tilde{\gamma}_{s_\ell}}.$$

When the parameters θ are unknown we can simulate the posterior distribution $\pi(\theta, s, w|y)$ by alternating the simulation of $(s, w|y, \theta)$ via the Metropolis-Hastings step described above and the simulation of $(\theta|s, w, y)$. The latter can be obtained by Gibbs and Metropolis steps accordingly to the prior distribution for the model parameters and their parameterization. Note also that the proposed algorithm can be easily generalized to handle models with absorbing states and panel data configurations where a set of observed states $y_i = (y_{i0}, y_{i1}, \ldots, y_{im_i})$ at the follow-up times $(t_{i0}, t_{i1}, \ldots, t_{im_i})$ is available for $i = 1, \ldots, n$, i.e. for each sample unit.

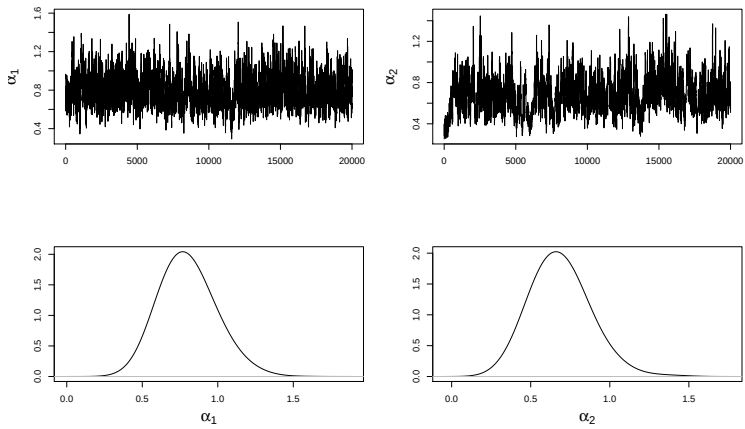


FIGURE 1. Breast cancer data: traces and posterior distributions for the parameters α_1 and α_2.

3 Application

We consider a panel data set comprising 37 women with breast cancer treated for spinal metastases; see De Stavola (1988), Davison (2003) where these data have been used to fit Markov models. The ambulatory status of the women, defined as ability to walk unaided or not, was recorded when the treatment began and then 3, 6, 12, 24, and 60 months after treatment. The three states are: able to walk unaided (1) unable to walk unaided (2) and dead (3).

We fitted the semi-Markov Weibull model with death as an absorbing state. The model parameters are $\theta = (p_{12}, p_{13}, p_{21}, p_{23}, \gamma_1, \gamma_2, \alpha_1, \alpha_2)$. Figure 1 shows the traces of the Metropolis-Hastings algorithm and posterior distributions for the shape parameters α_1 and α_2 under a vague prior distribution for θ. Posterior mean and standard deviations for all the model parameters are reported in Table 1. The results are similar to those reported in Tancredi (2019) where an approximate Bayesian computation (ABC) algorithm have been proposed to perform Bayesian inference. In particular note that both the posterior means of the shape parameters are less than 1 suggesting that the hazard of a transition is decreasing with time for both the states. Anyway the corresponding 95% credible intervals equal to $[0.51, 1.17]$ and $[0.38, 1.07]$ do not provide substantial evidence for this pattern. An extended version of this paper will provide more details about the MCMC algorithm and the real data application where the results of the MCMC algorithm will be compared with those of the ABC approach.

TABLE 1. Breast cancer data: posterior mean and standard deviation for the model parameters.

	p_{12}	p_{13}	p_{21}	p_{23}	γ_1	γ_2	α_1	α_2
$E(\cdot\|y)$	0.87	0.13	0.20	0.80	0.14	0.36	0.80	0.68
$SD(\cdot\|y)$	0.11	0.11	0.10	0.10	0.05	0.21	0.17	0.18

References

Davison, A.C. (2003). *Statistical Models*. Cambridge University Press.

De Stavola, B. L. (1988). Testing departures from time homogeneity in multistate Markov processes. *Journal of the Royal Statistical Society: Series C.* **37** 242 – 250.

Hobolth, A., Stone, E. A. Simulation from end-point conditioned, continuous time Markov chains on a finite state space, with application to molecular evolution. *The Annals of Applied Statistics.* **9** 1204 – 1231

Tancredi, A. (2019). Approximate Bayesian inference for discretely observed continuous time multi-state models. *Biometrics*

Titman, A. C. (2014). Estimating parametric semi-Markov models from panel data using phase-type approximations. *Statistics and Computing* **24**, 155 – 164.

Titman, A. C., Sharples, L. D. (2010). Semi-Markov models with phase-type sojourn distributions. *Biometrics* **66** , 742 – 752.

Forecasting vital rates from demographic summary measures

Carlo G. Camarda[1]

[1] Institut national d'études démographiques (INED), Paris, France

E-mail for correspondence: carlo-giovanni.camarda@ined.fr

Abstract: In population and actuarial sciences, time-trends of summary measures (such as life expectancy or the average number of children per woman) are easy to interpret and predict. Most summary measures are nonlinear functions of the vital rates, the key variable we usually want to estimate and forecast. Furthermore smooth outcomes of future age-specific vital rates are desirable. Therefore, optimization with nonlinear constraints in a smoothing setting is necessary. We propose a methodology that combines Sequential Quadratic Programming and a P-spline approach, allowing to forecast age-specific vital rates when future values of demographic summary measures are provided. We provide an application of the model on Italian mortality and Spanish fertility data.

Keywords: Vital rates forecast; Smoothing; Constrained nonlinear optimization; Summary measures.

1 Introduction

Future mortality and fertility levels can be predicted either by modelling and extrapolating rates over age and time, or by forecasting summary measures, later converted into age-specific rates. The latter approach takes advantage of the prior knowledge that demographers and actuaries have on possible future values of measures such as life expectancy at birth and total fertility rate. Among others, this methodology has been lately adopted by the United Nations (Ševčíková et al., 2016). In this paper, we propose a model to derive future mortality and fertility age-patterns complying with projected summary measures. Unlike comparable approaches, we assume only smoothness of future vital rates, which is achieved by a two-dimensional P-spline approach as in Currie et al. (2004), and we allow constraints to multiple series of summary measures. Since these measures are

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

commonly nonlinear functions of the estimated penalized coefficients, Lagrangian multipliers cannot be directly implemented. We hence opted for a Sequential Quadratic Programming (SQP) procedure (Nocedal & and Wright, 2006) to perform the associated constrained nonlinear optimization. We illustrate our approach with two data sets. We forecast mortality of Italian females, based on future life expectancy predicted by UN World Population Prospects (2017) and a future trend of a lifespan disparity measure obtained by time-series analysis. We also forecast Spanish fertility constrained to future values of total fertility rates, mean and variance of age at childbearing, derived by time-series analysis.

2 Model on Italian mortality data

For ease of presentation, we formulate the model on mortality data. We suppose that we have deaths, and exposures to risk, arranged in two matrices, $\boldsymbol{Y} = (y_{ij})$ and $\boldsymbol{E} = (e_{ij})$, each $m \times n_1$, whose rows and columns are classified by age at death, $\boldsymbol{a}$, $m \times 1$, and year of death, $\boldsymbol{t}_1$, $n_1 \times 1$, respectively. We assume that the number of deaths y_{ij} at age i in year j is Poisson distributed with mean $\mu_{ij}\, e_{ij}$. Forecasting aims to reconstruct trends in μ_{ij} for n_2 future years, $\boldsymbol{y}_2, n_2 \times 1$.
It is common practice to summarize mortality age-patterns by computing measures such as life expectancy at birth (e_0) and lifespan disparity measures. Time-trends of these summary measures are often regular and well-understood. Forecasting these time-series is therefore an easier task. Figure 1 (top-left panel) presents observed e_0 for Italian females from 1960 to 2016 along with the medium variant up to 2050 as computed by the UN. A second constraint is given by future values of $e^\dagger$, a lifespan disparity measure defined as the average years of life lost in a population attributable to death (Vaupel & Canudas Romo, 2003). Future values of this measure are obtained by conventional time-series models and portrayed in the top-right panel of Figure 1. Future mortality patterns, both by age and over time, must adhere to these predicted trends.
We arrange data as a column vector, that is, $\boldsymbol{y} = \mathtt{vec}(\boldsymbol{Y})$ and $\boldsymbol{e} = \mathtt{vec}(\boldsymbol{E})$ and we model our Poisson death counts as follows: $\ln(E(\boldsymbol{y})) = \ln(\boldsymbol{e}) + \boldsymbol{\eta} = \ln(\boldsymbol{e}) + \boldsymbol{B}\,\boldsymbol{\alpha}$, where $\boldsymbol{B}$ is the regression matrix over the two dimensions: $\boldsymbol{B} = \boldsymbol{I}_{n_1} \otimes \boldsymbol{B}_a$, with $\boldsymbol{B}_a \in \mathbb{R}^{m \times k_a}$. Over time, we employ an identity matrix of dimension n_1 because we will incorporate a constraint for each year. Over age, $\boldsymbol{B}_a$ includes a specialized coefficient for dealing with mortality at age 0. In order to forecast, data and bases are augmented as follows:

$$\breve{\boldsymbol{E}} = [\boldsymbol{E} : \boldsymbol{E}_2]\,, \qquad \breve{\boldsymbol{Y}} = [\boldsymbol{Y} : \boldsymbol{Y}_2]\,, \qquad \breve{\boldsymbol{B}} = \boldsymbol{I}_{n_1+n_2} \otimes \boldsymbol{B}_a\,, \tag{1}$$

where $\boldsymbol{E}_2$ and $\boldsymbol{Y}_2$ are filled with arbitrary future values. If we define a weight matrix $\boldsymbol{V} = \mathrm{diag}(\mathtt{vec}(\boldsymbol{1}_{m \times n_1} : \boldsymbol{0}_{m \times n_2}))$, the coefficients vector $\boldsymbol{\alpha}$

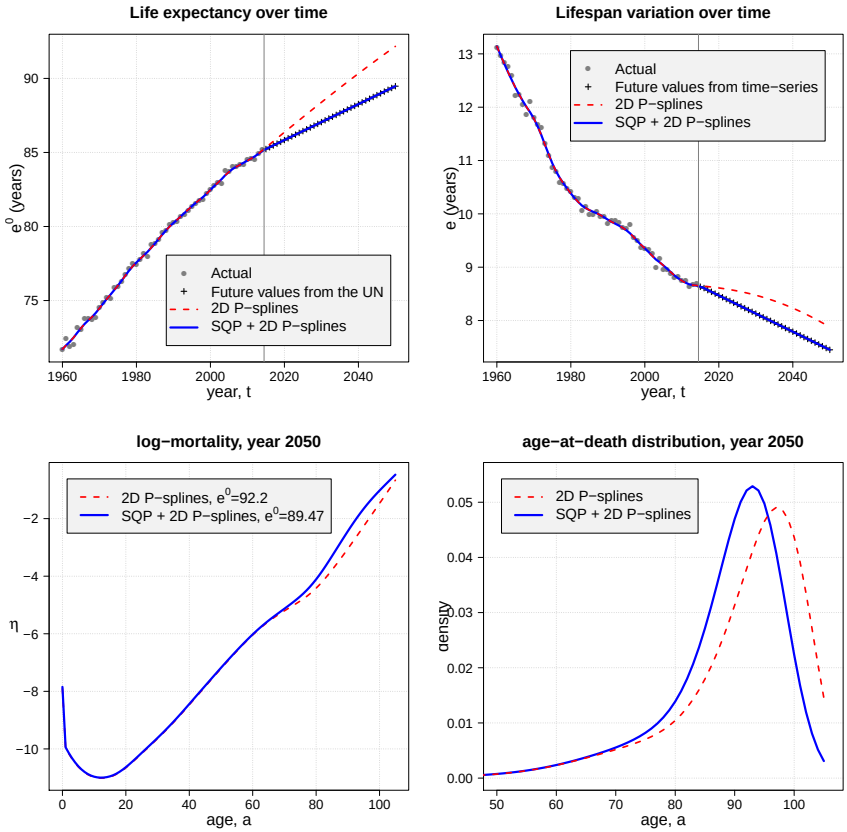


FIGURE 1. Top panels: Actual, estimated and forecast life expectancy at birth and lifespan disparity measure by United Nations and time-series, 2D P-splines and the SQP+2D P-splines. Bottom panels: Mortality in 2050 described by log–hazards and associated densities (ages 50+) by 2D P-splines and the SQP+2D P-splines. Italian females, ages 0-105, years 1960-2014, forecast up to 2050.

can be estimated by a penalised version of the iteratively reweighted least squares algorithm:

$$(\breve{\boldsymbol{B}}^T\boldsymbol{V}\tilde{\boldsymbol{W}}\breve{\boldsymbol{B}}+\boldsymbol{P})\tilde{\boldsymbol{\alpha}}=\breve{\boldsymbol{B}}^T\boldsymbol{V}\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}\,, \tag{2}$$

where a difference penalty $\boldsymbol{P}$ enforces smoothness behaviour of mortality both over age and time. Outcomes from this approach in terms of life expectancy and $e^\dagger$ are depicted with a dashed line in Figure 1 (top panels), and departures from the UN and time-series projected values are evident. Both life expectancy and average years of life lost are nonlinear function of the coefficients vector $\boldsymbol{\alpha}$. For a year j and associated k_a coefficients $\boldsymbol{\alpha}_j$, we denote mortality by $\boldsymbol{\mu}_j=\exp(\boldsymbol{B}_a\boldsymbol{\alpha}_j)$. We can write our summary measures

as follows

$$\begin{aligned} e^0(\boldsymbol{\alpha}_j) &= \mathbf{1}_m^{\mathrm{T}} \exp[\boldsymbol{C}\,\boldsymbol{\mu}_j] + 0.5 \\ e^\dagger(\boldsymbol{\alpha}_j) &= -\exp[\boldsymbol{C}\,\boldsymbol{\mu}_j]^T\,\boldsymbol{C}\boldsymbol{\mu}_j \end{aligned} \tag{3}$$

where $\boldsymbol{C}$ is a $(m \times m)$ lower triangular matrix filled only with -1.
Constrained nonlinear optimization is therefore necessary and a SQP approach is implemented. Let denote with $\boldsymbol{N}^0$ and $\boldsymbol{N}^\dagger$ the $(k_a n_2 \times n_2)$ matrices with block-diagonal structures containing derivatives of (3) with respect to $\boldsymbol{\alpha}_j$ for $j = n_1 + 1, \ldots n_1 + n_2$:

$$\begin{aligned} \frac{\partial e^0(\boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}_j} &= \mathbf{1}_m^{\mathrm{T}} \mathtt{diag}[\exp(\boldsymbol{C}\boldsymbol{\mu}_j)]\,\boldsymbol{C}\,\mathtt{diag}(\boldsymbol{\mu}_j)\boldsymbol{B}_a \\ \frac{\partial e^\dagger(\boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}_j} &= -\boldsymbol{B}_a^{\mathrm{T}}\left\{\boldsymbol{C}^{\mathrm{T}}[\boldsymbol{C}\boldsymbol{\mu}_j \circ \exp(\boldsymbol{C}\boldsymbol{\mu}_j)] \circ \boldsymbol{\mu}_j\right\} + \\ &\quad -\boldsymbol{B}_a^{\mathrm{T}}\left\{[\boldsymbol{C}^{\mathrm{T}} \exp(\boldsymbol{C}\boldsymbol{\mu}_j)] \circ \boldsymbol{\mu}_j\right\}, \end{aligned} \tag{4}$$

where $\circ$ represents element-wise multiplication. Target life expectancy and lifespan disparity for future years are given by n_2-vectors $\boldsymbol{e}^0_{\mathrm{T}}$ and $\boldsymbol{e}^\dagger_{\mathrm{T}}$.
Solution of the associated system of equations at the step $\nu + 1$ is given by

$$\begin{bmatrix} \boldsymbol{\alpha}_{\nu+1} \\ \boldsymbol{\omega}_{\nu+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_\nu & : & \boldsymbol{H}^0_\nu & : & \boldsymbol{H}^\dagger_\nu \\ \boldsymbol{H}^{0T}_\nu & : & \mathbf{0}_{n_2 \times n_2} & : & \mathbf{0}_{n_2 \times n_2} \\ \boldsymbol{H}^{\dagger T}_\nu & : & \mathbf{0}_{n_2 \times n_2} & : & \mathbf{0}_{n_2 \times n_2} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{r}_\nu - \boldsymbol{L}_\nu \boldsymbol{\alpha}_\nu \\ \boldsymbol{e}^0_{\mathrm{T}} - \boldsymbol{e}^0(\boldsymbol{\alpha}_\nu) \\ \boldsymbol{e}^\dagger_{\mathrm{T}} - \boldsymbol{e}^\dagger(\boldsymbol{\alpha}_\nu) \end{bmatrix}, \tag{5}$$

where $\boldsymbol{L}$ and $\boldsymbol{r}$ are left- and right-hand-side of the system in (2), and matrices $\boldsymbol{H}^0 = \left[\mathbf{0}_{k_a n_1 \times n_2} : \boldsymbol{N}^0\right]^T$ and $\boldsymbol{H}^\dagger = \left[\mathbf{0}_{k_a n_1 \times n_2} : \boldsymbol{N}^\dagger\right]^T$. Vector of $\boldsymbol{\omega}$ denotes the current solution of the associated Lagrangian multipliers for both set of constraints.
Future values for e^0 and $e^\dagger$ forecast by the proposed method are exactly equal to the UN and time-series values (Figure 1, top panels). The bottom panels show the forecast mortality age-pattern in 2050: the shape obtained by the suggested approach is not a simple linear function of the plain P-splines outcome, and differences are evident by looking at the associated age-at-death distributions.

3 Spanish Fertility Data

We forecast Spanish fertility using three commonly-used summary measures: Total Fertility Rate describing average number of children per women in a given year, and mean and variance of childbearing age which measure fertility shape over age. In formulas:

$$\begin{aligned} TFR(\boldsymbol{\alpha}_j) &= \mathbf{1}_m^{\mathrm{T}}\,\boldsymbol{\mu}_j \\ MAB(\boldsymbol{\alpha}_j) &= \boldsymbol{\mu}_j^T\,(\boldsymbol{a} + 0.5)\,/\,TFR(\boldsymbol{\alpha}_j) \\ VAB(\boldsymbol{\alpha}_j) &= \boldsymbol{\mu}_j^T\,(\boldsymbol{a} + 0.5)^2\,/\,TFR(\boldsymbol{\alpha}_j) - MAB(\boldsymbol{\alpha}_j)^2\,. \end{aligned} \tag{6}$$

We forecast trends of these measures by time-series analysis. We then smooth and constrain future fertility age-patterns to comply forecast values of (6) as in (5). Summary measures as well as fertility rates in 2050 are presented in Figure 2. Differences between proposed approach and plain 2D P-splines are clear. Whereas P-splines blindly extrapolate previous trends mainly accounting for the last observed years, the proposed approach enforces future age-patterns to adhere combinations of summary measures, guiding future fertility toward demographic meaningful trends.

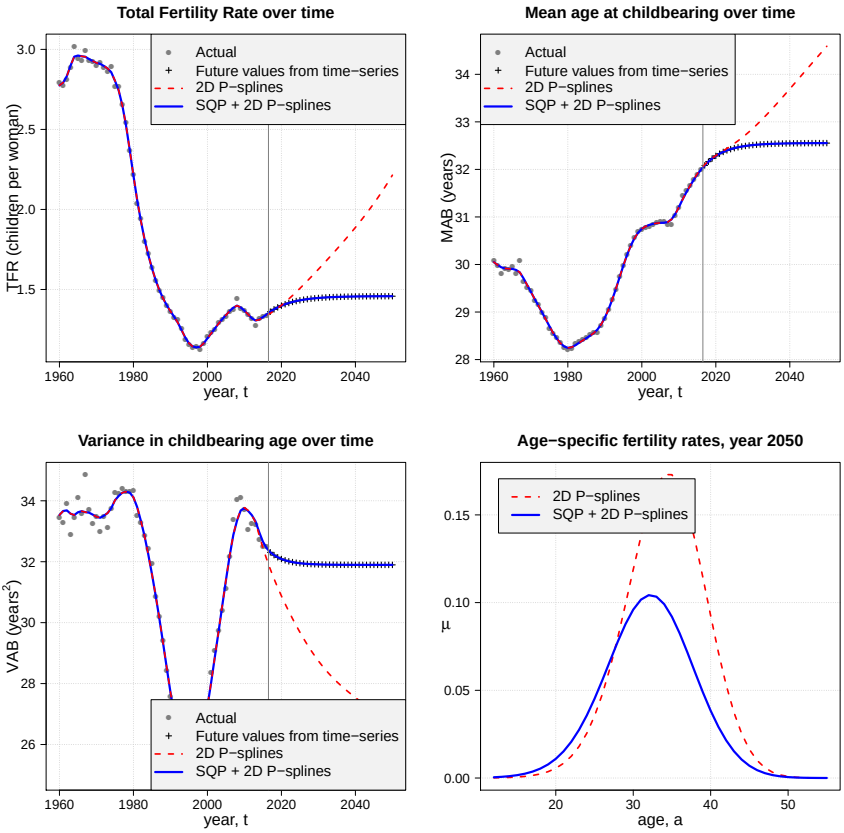


FIGURE 2. Top and left-bottom panels: Actual, estimated and forecast Total Fertility Rate, Mean and Variance in childbearing age by time-series analysis, 2D P-splines and the SQP+2D P-splines. Right-bottom panel: Age-specific fertility rate in 2050 by 2D P-splines and the SQP+2D P-splines. Spain, ages 12-55, years 1960-2016, forecast up to 2050.

4 Concluding remarks

In this paper, we combine smoothing models (P-splines) and optimization with nonlinear constraints (Sequential Quadratic Programming) to forecast vital rates when future values of demographic summary measures are provided.
We envisage further applications. Forecast of vital rates for partially completed cohorts is often relevant in population studies. For instance, final fertility history of a given cohort may be hypothesized though age-pattern is not yet observed and its estimation will be necessary. We also plan to adopt our approach to reconstruct demographic scenarios which are conventionally based on summary measures.
From a methodological perspective, future work will be realized to incorporate uncertainty and to objectively select the amount of smoothness in future mortality and fertility age-patterns.

References

Currie, I. D. et al. (2004). Smoothing and Forecasting Mortality Rates. *Statistical Modelling*, **4**, 279-298.

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization*. Springer.

Ševčíková, H. et al. (2016). Age-Specific Mortality and Fertility Rates for Probabilistic Population Projections. In R. Schoen (Ed.), *Dynamic demographic analysis*, 285 – 310. Springer.

United Nations, Population Division (2017). *World Population Prospects: The 2017 Revision, Volume II*. ST/ESA/SER.A/400.

Vaupel & Canudas Romo (2003). Decomposing change in life expectancy: A bouquet of formulas in honor of Nathan Keyfitz's 90th birthday. *Demography*, **40**, 201-216.

Invariance and the forecasting of mortality

Iain Currie[1]

[1] Heriot-Watt University, UK

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

Abstract: The forecasting of human mortality is an import topic for providers of pensions and care of the elderly. Many models of mortality are not identifiable so parameter constraints are used to obtain parameter estimates that can be used for forecasting. We show that when an ARIMA model is used to forecast parameter estimates the resulting forecasts of mortality are invariant with respect to the choice of constraints. These results remain true when some model terms are smoothed. We illustrate our results with Portuguese data.

Keywords: Forecasting; Identifiability; Invariance; Mortality.

1 Introduction

The mortality of an individual depends on their current age, the current year and their year of birth (among other risk factors). These determinants are generally known as the *age effect*, the *period effect* and the *cohort effect*. A problem of major interest to the providers of pensions and care of the elderly is the forecasting of mortality. In the financial world the usual approach is to construct a model in terms of the age, period and cohort effects. The model is fitted to appropriate data, the period and cohort effects are forecast, and a forecast of mortality is obtained.

However, most such models are not identifiable and it is not clear what exactly is being forecast. For example, Clayton and Schifflers (1987) in a carefully argued paper "doubt the wisdom" of such forecasting. Nevertheless the method does give plausible answers that are consistent across a range of models. Our purpose here is to try to explain why this is so.

We use Portuguese mortality data downloaded from the Human Mortality Database on December 18, 2018. We have the number of deaths $d_{x,y}$ and the corresponding central exposed to risk $e_{x,y}$ for ages 50 to 90 and years 1970 to 2015. For simplicity we will index the ages by $\boldsymbol{x}_a = (1, \ldots, n_a)^{\mathrm{T}}$, the

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

years by $\boldsymbol{x}_y = (1, \ldots, n_y)^{\mathrm{T}}$ and the years of birth by $\boldsymbol{x}_c = (1, \ldots, n_c)^{\mathrm{T}}$ where $n_c = n_a + n_y - 1$ is the number of distinct cohorts. With this convention age x has index $i = x - \min(x) + 1$ and year y has index $j = y - \min(y) + 1$. The oldest cohort in the first year is indexed one; thus, the cohort index for age x in year y is $n_a - i + j$. We suppose that the number of deaths at age x in year y follows a Poisson distribution $\mathcal{P}(e_{x,y}\lambda_{x,y})$ where $\lambda_{x,y}$ is the force of mortality or hazard of death at age x in year y. With our data we have $n_a = 41$, $n_y = 46$ and $n_c = 86$.
The plan of the paper is: section 2 describes our method, section 3 gives an example and section 4 contains some concluding remarks.

2 Method

We consider a generalized linear model or GLM with model matrix $\boldsymbol{X}$, $n \times p$, $n > p$ and rank $p - q$ where $q \geq 1$. We denote the vector of parameters by $\boldsymbol{\theta}$. Since $\boldsymbol{X}$ is not of full rank $\boldsymbol{\theta}$ is not identifiable. However, there exists a matrix $\boldsymbol{H}$, $q \times p$, with rank q such that $\boldsymbol{H\theta} = \boldsymbol{0}$. Now, subject to the condition that $\boldsymbol{H\theta} = \boldsymbol{0}$, we do have a unique estimate of $\boldsymbol{\theta}$. We refer to $\boldsymbol{H}$ as a *constraints matrix* and we note that $\boldsymbol{H}$ is not unique.
We will use the P-spline system of smoothing when we wish to smooth certain parameters in our models; see Eilers and Marx (1996) for a general introduction and Currie et al. (2004) for details in our present application. We denote the penalty matrix of the P-spline system by $\boldsymbol{P}$.
Currie (2013) generalized the Nelder and Wedderburn (1972) algorithm for estimation of $\boldsymbol{\theta}$ in a GLM. The following is a scoring algorithm for estimation of $\boldsymbol{\theta}$ subject to (a) the constraint $\boldsymbol{H\theta} = \boldsymbol{0}$ and (b) smoothing via the penalty matrix $\boldsymbol{P}$

$$\begin{pmatrix} \boldsymbol{X}^{\mathrm{T}}\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P} & : & \boldsymbol{H}^{\mathrm{T}} \\ \boldsymbol{H} & : & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\omega}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^{\mathrm{T}}\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \\ \boldsymbol{0} \end{pmatrix}; \tag{1}$$

here $\tilde{\boldsymbol{W}}$ is the diagonal matrix of weights, $\tilde{\boldsymbol{z}}$ is the so-called working variable and $\hat{\boldsymbol{\omega}}$ is an auxiliary variable. If a canonical link is used then (1) is a Newton-Raphson scheme.
Let $\hat{\boldsymbol{\theta}}_i$ be the maximum likelihood estimate of $\boldsymbol{\theta}$ under the constraint $\boldsymbol{H}_i\boldsymbol{\theta} = \boldsymbol{0}$, $i = 1, 2$. The fitted values are invariant with respect to the choice of constraints so $\boldsymbol{X}\hat{\boldsymbol{\theta}}_1 = \boldsymbol{X}\hat{\boldsymbol{\theta}}_2$. We are interested in the relationship between $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$. This is characterized by the *null space* of $\boldsymbol{X}$ which we define as

$$\mathcal{N}(\boldsymbol{X}) = \{\boldsymbol{v} : \boldsymbol{Xv} = \boldsymbol{0}\}. \tag{2}$$

With this notation, $\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2 \in \mathcal{N}(\boldsymbol{X})$.
The idea is to use two constraint systems: the first will be a system used in the literature, a *standard* system, and the second will be a *random* system. We will then use the null space of $\boldsymbol{X}$ to show that forecasts of mortality under these two systems are also invariant when an ARIMA model is used to forecast.

3 Example

One popular model for forecasting mortality in the financial world is the age-period-cohort model. We have

$$\log \lambda_{i,j} = \alpha_i + \kappa_j + \gamma_{n_a - i + j},\ i = 1, \ldots, n_a,\ j = 1, \ldots, n_y, \tag{3}$$

where i is the age of death, j is the year of death and $n_a - i + j$ is the year of birth. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n_a})^{\mathrm{T}}$, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{n_y})^{\mathrm{T}}$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{n_c})^{\mathrm{T}}$; let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\kappa}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. Model (3) has $n_a + n_y + n_c$ parameters but the rank of its model matrix is $n_a + n_y + n_c - 3$ so three constraints are required to give a unique estimate of $\boldsymbol{\theta}$. One common or *standard* set of constraints is

$$\sum_1^{n_y} \kappa_j = \sum_1^{n_c} \gamma_c = \sum_1^{n_c} c\gamma_c = 0; \tag{4}$$

see Cairns et al. (2009) for example. We also consider a set of *random* constraints

$$\sum u_{i,j}\theta_j,\ i = 1, 2, 3,\ j = 1, \ldots, n_a + n_y + n_c \tag{5}$$

where the $u_{i,j}$ are independent uniform variables, $\mathcal{U}(0, 1)$. We denote the estimates under the two constraint systems by $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_r$ respectively. Figure 1 shows the estimates of $\boldsymbol{\alpha}, \boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$ under the two systems for our Portuguese data and one set of random constraints. The estimates are strikingly different in both shape and scale, yet the invariance of the fitted values guarantees that the fitted values $\log \hat{\lambda}$ are equal, as in the bottom right panel.

A basis for the null space, $\mathcal{N}(\boldsymbol{X})$, of the model matrix, $\boldsymbol{X}$, is

$$\left\{ \begin{pmatrix} \mathbf{1}_{n_a} \\ -\mathbf{1}_{n_y} \\ \mathbf{0}_{n_c} \end{pmatrix}, \begin{pmatrix} \mathbf{1}_{n_a} \\ \mathbf{0}_{n_y} \\ -\mathbf{1}_{n_c} \end{pmatrix}, \begin{pmatrix} \boldsymbol{x}_a \\ -\boldsymbol{x}_y \\ \boldsymbol{x}_c - n_a \mathbf{1}_{n_c} \end{pmatrix} \right\} \tag{6}$$

where $\mathbf{1}$ and $\mathbf{0}$ are vectors of 1s and 0s respectively of the indicated lengths. The estimates $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_r$ are intimately related; their difference $\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_r$ lies in $\mathcal{N}(\boldsymbol{X})$. We define $\Delta\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_r$, $\Delta\hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r$ and $\Delta\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r$. Then, equating coefficients in (6), we find

$$\Delta\hat{\boldsymbol{\alpha}} = (A + B)\mathbf{1}_{n_a} + C\boldsymbol{x}_a \tag{7}$$

$$\Delta\hat{\boldsymbol{\kappa}} = -A\mathbf{1}_{n_y} - C\boldsymbol{x}_y \tag{8}$$

$$\Delta\hat{\boldsymbol{\gamma}} = -(B + n_a C)\mathbf{1}_{n_c} + C\boldsymbol{x}_c \tag{9}$$

for some constants A, B and C; in our example, we found $A = -4.07$, $B = 3.90$ and $C = -0.0818$. Clayton and Schifflers (1987) among others have observed that the age, period and cohort parameters are only estimable up

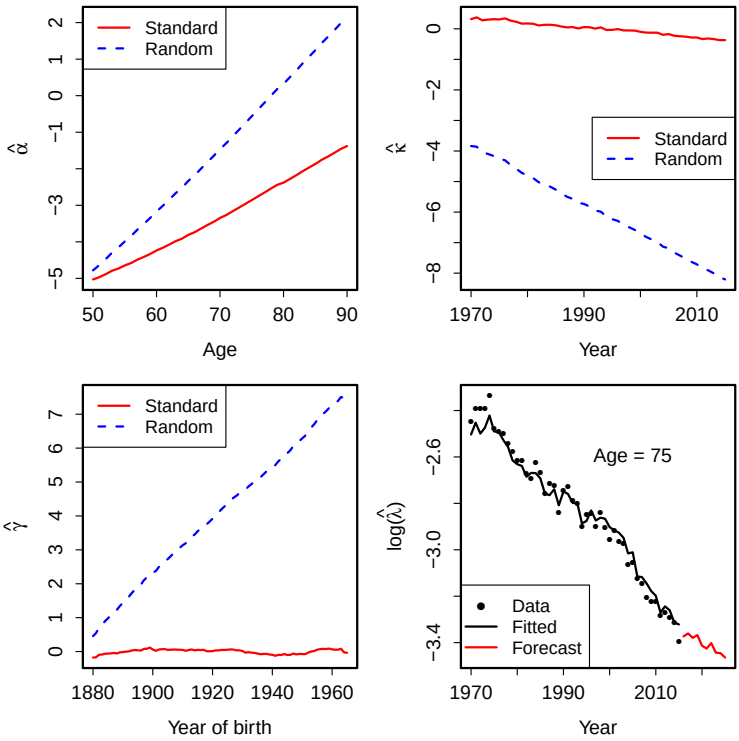


FIGURE 1. Parameter estimates in the age-period-cohort model under standard and random constraints; observed and the invariant fitted and forecast $\log(\hat{\lambda})$.

to a linear function. Equations (7), (8) and (9) make this precise; the left panel of Figure 2 illustrates these relations.

We turn now to forecasting. We forecast the period and cohort terms with an ARIMA model and then, with $\boldsymbol{\alpha}$ fixed at its estimated value, use (3) to forecast the log λ. Despite the difference between $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$, and between $\hat{\boldsymbol{\gamma}}_s$ and $\hat{\boldsymbol{\gamma}}_r$, the forecast values of log λ are invariant with respect to the choice of constraints, just like the fitted values. This invariance is illustrated in the bottom right panel of Figure 1; it can also be proved with (7), (8) and (9).

We turn briefly to the effect of smoothing. Regular forecasts of log λ are desirable for the pricing and reserving of many financial products. In the age-period-cohort model this regularity can be achieved by smoothing the age term $\boldsymbol{\alpha}$. We use the P-spline system with cubic B-splines, a second order penalty and a knot spacing of $\delta_a = 5$. Let $\boldsymbol{B}_a$, $n_a \times c_a$, be the resulting regression matrix along age and set $\boldsymbol{\alpha} = \boldsymbol{B}_a \boldsymbol{a}$. We refer to this model as the smooth age-period-cohort model. The regression coefficients

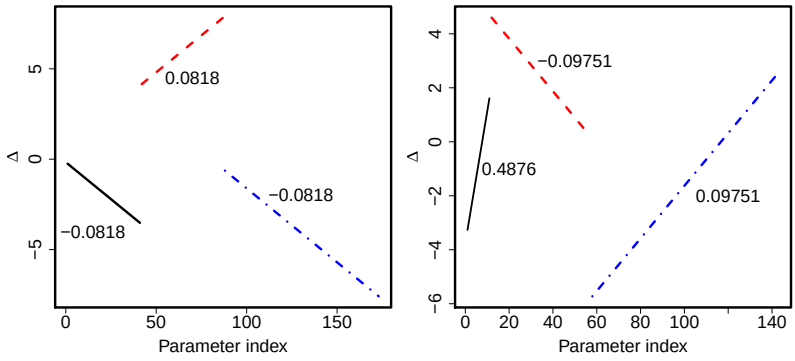


FIGURE 2. $\Delta\hat{\boldsymbol{\alpha}}$ (——), $\Delta\hat{\boldsymbol{\kappa}}$ (– – –) and $\Delta\hat{\boldsymbol{\gamma}}$ (– · –). Left: age-period-cohort model and right: age-period-cohort model with smooth $\boldsymbol{\alpha}$.

are $\boldsymbol{\theta} = (\boldsymbol{a}^{\mathrm{T}}, \boldsymbol{\kappa}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ with length $c_a + n_y + n_c$. We use the same approach as in the unsmoothed model. We use the constraints (4) and the random constraints (5) but with $j = 1, \ldots, c_a + n_y + n_c$. A basis for the null space of the model is

$$\left\{ \begin{pmatrix} \mathbf{1}_{c_a} \\ -\mathbf{1}_{n_y} \\ \mathbf{0}_{n_c} \end{pmatrix}, \begin{pmatrix} \mathbf{1}_{c_a} \\ \mathbf{0}_{n_y} \\ -\mathbf{1}_{n_c} \end{pmatrix}, \begin{pmatrix} \delta_a \boldsymbol{x}_{c_a} \\ -\boldsymbol{x}_y \\ \boldsymbol{x}_c - \omega_c \mathbf{1}_{n_c} \end{pmatrix} \right\}; \tag{10}$$

here $\boldsymbol{x}_{c_a} = (1, 2, \ldots, c_a)^{\mathrm{T}}$ and $\omega_c = n_a + 2\delta_a - 1$. The right panel of Figure 2 illustrates the relations; we note that $0.4876 = \delta_a 0.09751$, in agreement with the basis in (10). Again we see that the age, period and cohort effects are estimable only up to a linear function. It follows in a similar fashion to the unsmoothed case that not only are fitted values invariant with respect to the choice of constraints but so also are their forecast values.

4 Conclusions

The rates of mortality in models of mortality are uniquely estimable, despite the fact that the component terms, ie, the age, period and cohort terms, are not uniquely estimable. This follows from the invariance of fitted values in a GLM. This gives rise to the following paradox: the period and cohort terms are not estimable and so neither are their forecast values. Why is it then that in practice the forecast rates of mortality seem plausible across a range of mortality models? In this short paper we provide a partial resolution of this paradox, namely that the forecast values of $\log \lambda$ are also invariant with respect to the choice of constraints used to estimate the age, period and

cohort effects. For a fuller discussion of these ideas and further examples see Currie (in preparation).

References

Cairns, A.J.G., Blake, D., Dowd, K. et al. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, **13**, 1 – 35.

Clayton, D. and Schifflers, E. (1987). Models for temporal variation in cancer rates. II: Age-period-cohort models. *Statistics in Medicine*, **6**, 469 – 481.

Currie, I.D. (in preparation). Constraints, the identifiability problem and the forecasting of mortality.

Currie, I.D. (2013). Smoothing constrained generalized linear models with an application to the Lee-Carter model. *Statistical Modelling*, **13**, 69 – 93.

Currie, I.D., Durbán, M. and Eilers, P.H.C. (2004). Smoothing and forecasting mortality rates, *Statistical Modelling*, **4**, 279 – 298.

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

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A*, **135**, 370 – 384.

COM-Poisson models with varying dispersion

Eduardo Elias Ribeiro Jr[1,2], Clarice Garcia Borges Demétrio[1], John Hinde[3]

[1] Department of Exact Sciences, USP-ESALQ, Piracicaba, SP, Brazil
[2] USP-IME, São Paulo, SP, Brazil
[3] School of Mathematics, Statistics and Applied Mathematics, NUI Galway, Galway, Ireland

E-mail for correspondence: clarice.demetrio@usp.br

Abstract: We propose an extension of the COM-Poisson model to jointly model the mean and the dispersion as functions of covariates taking into account, possibly, under- and overdispersion in the same count data set. Estimation and inference are based on the likelihood paradigm. Results from a simulation study show that the maximum likelihood estimators are consistent and unbiased for both mean and dispersion parameters. The methodology is illustrated with the analysis of a data set. The R codes and data set are available online.

Keywords: COM-Poisson, Double generalized linear models, Varying dispersion.

1 Introduction

Standard Gaussian linear models are based on the assumption of variance homogeneity. Generalized linear models relax this assumption by assuming the observations come from some distribution in the exponential family. A key feature of exponential family distribution is the so-called mean-variance relationship, $\text{Var}(Y) = \phi V(\mu)$. The main examples are $V(\mu) = \mu(1-\mu)$ for the binomial, $V(\mu) = \mu$ for the Poisson, $V(\mu) = \mu^2$ for the gamma, and $V(\mu) = \mu^3$ for the inverse-Gaussian distributions (McCullagh & Nelder, 1989). However, once the mean-variance relationship is specified, the variance is assumed to be known up to a constant of proportionality, the dispersion parameter ϕ. To provide more flexibility in the analysis of heterogeneous count data, we explore methods for modelling dispersion as a function of covariates.

Modelling dispersion with covariates in the analysis of count data has received little attention in the literature. The class of double generalized

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

linear models (Smyth 1988, McCullagh & Nelder 1989, Smyth & Verbyla 1999) provide a possible approach. This class has been widely explored for continuous data. Another approach that has gained momentum in the last decade is the generalized additive models for location, shape, and scale (GAMLSS) (Rigby & Stasinopoulos,2005).
In this paper, we propose to jointly model the mean and dispersion based on the COM-Poisson distribution. This approach is very similar to that of GAMLSS, however, we develop and explore our own estimation methods. This approach allows modelling of data that exhibit both under- and overdispersion.

2 Toxicity of nitrofen in aquatic systems

Nitrofen is a herbicide that was used extensively for the control of broad-leaved and grass weeds in cereals and rice. Although it is relatively non-toxic to adult mammals, nitrofen is a significant tetragen and mutagen. This data set comes from an experiment to measure the reproductive toxicity of the herbicide nitrofen on a species of zooplankton (*Ceriodaphnia dubia*). Fifty animals were randomized into batches of ten and each batch was placed in a solution with a measured concentration of nitrofen $(0, 0.8, 1.6, 2.35$ and 3.10 $\mu g/10^2$litre) (`dose`). Subsequently, the number of live offspring was recorded.
Figure 1 shows the data and summary statistics for each batch. It is clear that the number of live offspring decreases as the nitrofen dose increases. However, it seems that the dispersion is also influenced by the nitrofen concentration level, with underdispersion for low doses and overdispersion for high doses.

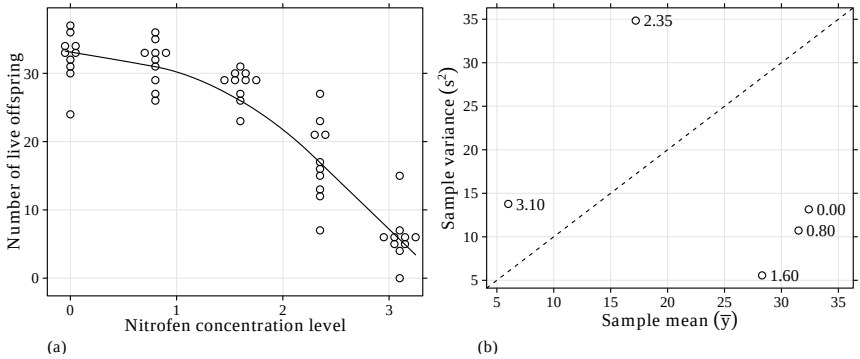


FIGURE 1. (a) Number of live offspring observed for each nitrofen concentration level (solid lines represent loess curve) and (b) sample variance against sample mean for each concentration level (dotted line is the identity line).

3 COM-Poisson models with varying dispersion

The COM-Poisson distribution is a two-parameter generalization of the Poisson distribution that can handle under-, over- and equidispersion (Shmueli et al. 2005). The probability mass function of the COM-Poisson distribution is

$$\Pr(Y = y) = \frac{\lambda^y}{(y!)^\nu \mathrm{Z}(\lambda, \nu)}, \quad y = 0, 1, 2, \ldots; \quad \mathrm{Z}(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}, \quad (1)$$

where $\lambda > 0$ and $\nu \geq 0$. The $\mathrm{Z}(\lambda, \nu)$ is a normalizing constant that cannot be expressed in closed form, except for special cases.
The moments for the COM-Poisson distribution also cannot be obtained in closed forms. Shmueli et al. (2005) showed that the expectation of the COM-Poisson distribution can be approximated by

$$\mathrm{E}(Y) = \frac{\mathrm{d}\{\log[\mathrm{Z}(\lambda, \nu)]\}}{\mathrm{d}\lambda} \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu}.$$

The parameter ν is the dispersion parameter and has a clear interpretation. When $\nu = 1$, the Poisson distribution results as a special (equidispersion) case, while for $0 < \nu < 1$ we have overdispersion and for $\nu > 1$ underdispersion. On the other hand, the parameter λ has no clear interpretation, except for $\nu = 1$ when it is a rate parameter and the Poisson mean, and in general it is strongly related to ν. To circumvent this dependency, Ribeiro Jr et al. (2018) proposed a reparameterization of the COM-Poisson distribution to provide an approximate mean parameter. Replacing λ by the new parameter $\mu > 0$,

$$\mu = \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \quad \Rightarrow \quad \lambda = \left(\mu + \frac{(\nu - 1)}{2\nu}\right)^\nu,$$

the authors showed that the new parameterization has good properties for estimation and inference, with approximate orthogonality of μ and ν. They proposed the use of a regression model for this approximate mean, rather than for λ as in Sellers & Shmueli (2010), and here we extend this to allow both μ and ν to depend on covariates.
Let y_i, $i = 1, 2, \ldots, n$, be independent realizations of Y_i from COM-Poisson distributions with parameters μ_i and ν_i. The proposed COM-Poisson varying dispersion model assumes

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} \quad \text{and} \quad \xi_i = h(\nu_i) = \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\gamma},$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)^{\mathrm{T}}$ are the parameters to be estimated, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{np})^{\mathrm{T}}$ and $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{nq})^{\mathrm{T}}$ are vectors of known covariates, and $g(.)$ and $h(.)$ are suitable link functions, such as the log.

4 Estimation and inference

To fit COM-Poisson models with varying dispersion, we use maximum likelihood and inferences are based on the standard asymptotic likelihood theory. The log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ is

$$\ell = \ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \nu_i \log\left(\mu_i + \frac{\nu_i - 1}{2\nu_i}\right) - \nu_i \log(y_i) - \log[Z(\mu_i, \nu_i)] \right\}, \tag{2}$$

where $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, $\nu_i = \exp(\boldsymbol{z}_i^\top \boldsymbol{\gamma})$, and $Z(\mu_i, \nu_i)$ is the normalizing constant computed for the parameters μ_i and ν_i.

Parameter estimation requires the numerical maximization of (2). Since the derivatives of ℓ cannot be obtained in closed forms, we compute them by central finite differences using the Richardson method from the `R` package `numDeriv` (Gilbert & Varadhan, 2016).

Standard errors are obtained from the observed information matrix and hence the variance-covariance matrix of the maximum likelihood estimators is

$$V_{\boldsymbol{\theta}} = \begin{pmatrix} -\partial\ell^2/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top & -\partial\ell^2/\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^\top \\ -\partial\ell^2/\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}^\top & -\partial\ell^2/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top \end{pmatrix}^{-1} = \begin{pmatrix} V_{\boldsymbol{\beta}} & V_{\boldsymbol{\beta\gamma}} \\ V_{\boldsymbol{\gamma\beta}} & V_{\boldsymbol{\gamma}} \end{pmatrix}.$$

Variances for $\hat{\eta}_i$ and $\hat{\xi}_i$ can be obtained using the delta method, $\mathrm{Var}(\hat{\eta}_i) = \boldsymbol{x}_i^\top V_{\boldsymbol{\beta}|\boldsymbol{\gamma}} \boldsymbol{x}_i$ and $\mathrm{Var}(\hat{\xi}_i) = \boldsymbol{z}_i^\top V_{\boldsymbol{\gamma}|\boldsymbol{\beta}} \boldsymbol{z}_i$, where $V_{\boldsymbol{\beta}|\boldsymbol{\gamma}} = V_{\boldsymbol{\beta}} - V_{\boldsymbol{\beta\gamma}} V_{\boldsymbol{\gamma}}^{-1} V_{\boldsymbol{\gamma\beta}}$ and $V_{\boldsymbol{\gamma}|\boldsymbol{\beta}} = V_{\boldsymbol{\gamma}} - V_{\boldsymbol{\gamma\beta}} V_{\boldsymbol{\beta}}^{-1} V_{\boldsymbol{\beta\gamma}}$. Since $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are nearly orthogonal, $V_{\boldsymbol{\beta\gamma}} = V_{\boldsymbol{\gamma\beta}}^\top \approx 0$, hence $V_{\boldsymbol{\beta}|\boldsymbol{\gamma}} \approx V_{\boldsymbol{\beta}}$ and $V_{\boldsymbol{\gamma}|\boldsymbol{\beta}} \approx V_{\boldsymbol{\gamma}}$, which implies that inferences based on the conditional log-likelihood and the marginal log-likelihood are the same. Confidence intervals for μ_i and ν_i are obtained by back-transforming the confidence intervals for η_i and ξ_i. Maximum likelihood estimation for fitting COM-Poisson models and methods for computing the associated confidence intervals are implemented in the `R` package `cmpreg` (`https://github.com/jreduardo/cmpreg`).

5 Data analysis of nitrofen experiment

To analyse the number of live offspring (Y_{ij}) for ith nitrofen dose and jth repetition we use a cubic polynomial in dose for both mean and dispersion

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \mathtt{d}_i + \beta_2 \mathtt{d}_i^2 + \beta_3 \mathtt{d}_i^3 \quad \text{and} \quad \log(\nu_{ij}) = \gamma_0 + \gamma_1 \mathtt{x}_{1i} + \gamma_2 \mathtt{x}_{2i} + \gamma_3 \mathtt{x}_{3i},$$

where $\mathtt{x}_{qi}$ is the orthogonal polynomial of degree q evaluated at dose $\mathtt{d}_i$. For the dispersion we also consider nested submodels.

Table 1 shows clear evidence that the linear predictor for the dispersion is at least linearly dependent on the nitrofen concentration level. However, there is no strong evidence to favour the quadratic, or cubic, models over the linear model for the dispersion.

TABLE 1. Nitrofen data: goodness-of-fit measures (deviance and AIC) and model comparisons (based on deviance differences, Δ-Dev) of the dispersion models.

	df	Deviance	AIC	Δ-Dev	$\Pr(>\chi^2)$
Constant	45	288.13	298.13		
Linear	44	274.11	286.11	14.02	0.0002
Quadratic	43	270.49	284.49	3.62	0.0572
Cubic	42	269.50	285.50	0.99	0.3198

TABLE 2. Nitrofen data: Parameter estimates and standard errors for the fitted COM-Poisson models.

Par	Const	Linear	Quad	Cubic
		Mean		
β_0	3.48 (0.05)*	3.48 (0.03)*	3.48 (0.04)*	3.48 (0.04)*
β_1	−0.09 (0.20)	−0.11 (0.14)	−0.12 (0.13)	−0.12 (0.13)
β_2	0.16 (0.17)	0.17 (0.15)	0.19 (0.14)	0.19 (0.13)
β_3	−0.10 (0.04)*	−0.10 (0.04)*	−0.11 (0.04)*	−0.11 (0.04)*
		Dispersion		
γ_0	0.05 (0.20)	0.29 (0.21)	0.24 (0.26)	0.35 (0.23)
γ_1	–	−5.24 (1.36)*	−7.00 (2.30)*	−5.73 (1.84)*
γ_2	–	–	−3.98 (2.44)	−2.92 (1.90)
γ_3	–	–	–	1.52 (1.41)

Est (SE)* indicates |Est/SE| > 1.96.

Parameter estimates, standard errors and significance (based on Wald tests) are given in Table 2. For the dispersion structure, there is no evidence to keep the quadratic term, (p-value = 0.10), but the mean model standard errors do decrease once the linear term is included in the dispersion.
Figure 2 shows the fitted values with confidence bands for the mean model and linear and quadratic dispersion models. For constant dispersion, the fitted model corresponds to equidispersion ($\nu = 1$), with $\exp(\hat{\gamma}_0) = 1.05$. However, there is some evidence that the dispersion changes across nitrofen levels. In particular, all models show that at around $2\mu g/10^2$ litre, the numbers of live offspring change from under- to over-dispersed. The variances for nitrofen doses obtained from the fitted models with linear and quadratic models for the dispersion are also shown.

Acknowledgments: This work was partially supported by CNPq and by FAPESP, Brazilian science funding agencies.

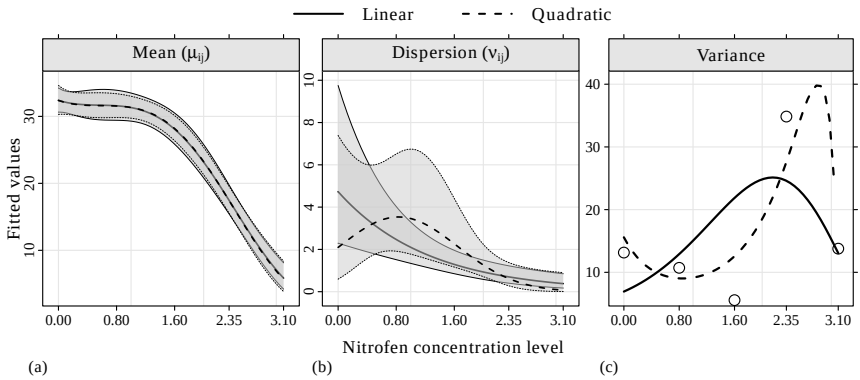


FIGURE 2. (a-b) Fitted values for mean and dispersion parameters with 95% confidence bands and (c) variances obtained from the fitted models.

References

Gilbert, P. and Varadhan, R. (2016). *numDeriv: Accurate Numerical Derivatives*, R package version 2016.8-1.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Monographs on Statistics and Applied Probability, 2nd edition, Chapman & Hall, London.

Ribeiro Jr, E.E., Zeviani, W.M., Bonat, W.H., Demétrio, C.G.B. and Hinde, J. (2018). Reparametrization of COM-Poisson regression models with applications in the analysis of experimental data. Preprint *arXiv*.

Rigby, R. A. and Stasinopoulos, D. M . (2005). Generalized additive models for location, scale and shape (with discussion), *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **54**, 507 – 554.

Sellers, K. F. and Shmueli, G. (2010). A flexible regression model for count data, *Annals of Applied Statistics*, **4**, 943 – 961.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P., (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **54**, 127 – 142.

Smyth, G.K. (1988). Generalized Linear Models with Varying Dispersion. *Journal of the Royal Statistical Society, Series B*, **51**, 47 – 60.

Smyth, G. K. and Verbyla, A. P. (1999). Adjusted likelihood methods for modelling dispersion in generalized linear models, *Environmetrics*, **10**, 695 – 709.

A latent state model for characterising regime shifts in ocean density

Theo Economou[1], Matthew B. Menary[2]

[1] Department of Mathematics, University of Exeter, UK
[2] LOCEAN/IPSL, Sorbonne Universités (UPMC)-CNRS-IRD-MNHN, Paris, France

E-mail for correspondence: t.economou@exeter.ac.uk

Abstract: Decadal predictions of temperature and precipitation over Europe are largely affected by variability in the North Atlantic Ocean. Within this region, the Labrador Sea is of particular importance due to its link between surface-driven density variability and the Atlantic Meridional Overturning Circulation (AMOC). Using physical justifications, we propose a statistical model to describe the temporal variability in ocean density, in terms of salinity and temperature. This is a hidden semi-Markov model that allows for alternating temperature- and salinity-driven ocean density. The model is Bayesian, and a reversible jump MCMC algorithm is proposed to deal with a single-regime scenario. The model is applied to an observations-based data set as well as to data from 43 climate models. Estimates of the mean holding time for each regime are used to establish a link between regime behaviour and the AMOC.

Keywords: Reversible jump MCMC; Bayesian; HMM; Forward algorithm; Adaptive Metropolis.

1 Introduction

Skillful decadal predictions of changes in temperature and precipitation over Europe are extremely important. Low frequency variability, e.g. the Atlantic Meridional Overturning Circulation (AMOC), in the North Atlantic ocean is a key component of any skilful prediction (Collins, 2002). This variability has been linked to the Labrador Sea region, and variability in seawater density therein. Understanding the nature of seawater density changes in this region is therefore valuable in improving the skill of predictions.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

The question here is whether density changes in the Labrador Sea are driven by sea temperature T or salinity S changes and whether the driver is stationary, or whether there are regime shifts between salinity-driven and temperature-driven density over time. The relatively short time span of instrumental observations makes it necessary to also rely on coupled general circulation climate models (CGCMs). The goal is to develop a statistical model for quantifying variability in the Labrador Sea density in terms of salinity and temperature regime shifts, and apply it to available observations as well as to data from 43 free-running climate models. We use the results to investigate whether CGCMs are able to simulate regime shifts and thus obtain a better understanding the temporal structure of these shifts.

2 Model formulation

Seawater density ρ, can be described as a function of T and S such that:

$$\rho = f_T(T) + f_S(S) + f_{S,T}(S,T).$$

This equation is non-linear in temperature and salinity over the full, observed temperature/salinity space, however using approximations as described in Menary et al. (2015), we can describe the density anomaly (i.e. mean-centred) ρ using two possible equations:

$$\rho = \beta_S \rho_S + \epsilon_S \qquad \text{(salinity driven density)} \tag{1}$$

$$\rho = \beta_T \rho_T + \epsilon_T \qquad \text{(temperature driven density)} \tag{2}$$

where ρ_S and ρ_T are components of ρ that are driven solely due to S and T respectively, where the ϵ terms capture residual variation. As such, a model M_S or M_T for density being solely driven by salinity or temperature across all time is:

$$M_S: \quad \rho(t) \sim N(\beta_S \rho_S(t), \sigma_S^2) \tag{3}$$

$$M_T: \quad \rho(t) \sim N(\beta_T \rho_T(t), \sigma_T^2) \tag{4}$$

Previous analyses (Menary et al., 2016) provide evidence that in any given point in time, ocean density is described by either an S-driven regime (1) or a T-driven regime (2). A natural modelling framework for describing underlying regime changes in a random variable, is the hidden Markov model or HMM. These, model latent regimes or states over time as a Markov chain, where the holding time for each state is implicitly Geometric. A generalisation of HMMs, are hidden semi-Markov models (HSMMs), that allow for explicit modelling of the holding times. Both (3) and (4) above assume no regime shifts, therefore we consider a third model, M_{ST} where a regime switching mechanism is described by a latent semi-Markov chain

$C(t)$ with two states: $C(t) \in \{S, T\}$. The model, depicted in Figure 1, is given by:

$$\rho(t)|C(t) = \beta_{C(t)}\rho_{C(t)}(t) + \epsilon_{C(t)}(t) \quad (5)$$

$$\epsilon_{C(t)}(t) \sim N(0, \sigma^2_{C(t)}). \quad (6)$$

This is a model that jumps between M_S and M_T. The semi-Markov chain is defined by two Poisson holding time distributions with means ϕ_S and ϕ_T respectively. Self-transitions are not allowed implying that neither M_S nor M_T are special cases of M_{ST}, so in fitting the models in the Bayesian framework (Economou et al., 2014), we use reversible jump MCMC to decide which of the three models best describes ocean variability.

3 Model application

Fitting the model to $t = 1, \ldots, 115$ years of reanalysis data (which is as close as we can get to observations), indicates that model M_{ST} is chosen with probability 0.999. Figure 2 plots the estimated probability of being in the S regime. This indicates that regime S is much more persistent (mean holding time of 11 years) than the regime T, which does occur albeit in short bursts of 1-2 years.

The model was also applied to 43 pre-industrial control simulations from CGCMs. These simulations aim to recreate an equilibrium climate (prior to the secular trend that is now evident) using interannualy invariant external forcings (e.g. greenhouse gas) appropriate for pre-industrial times. Each control simulation was at least 200 years in length. They represent different approaches to simulating this pre-industrial climate and by comparing them it is possible to investigate the strength of internal variability in the climate system. Most (but not all) CGCMs are able to simulate a regime changing ocean density in line with the reanalysis data, albeit with varying degrees of temporal persistence of the regimes. Overall, regime S has larger mean holding times across CGCMs, much like the reanalysis. Some CGCMs however have regime T that lasts longer on average.

Estimates of parameters ϕ_S and ϕ_T provide a way of quantifying the temporal regime behaviour of each CGCM. An important question is whether the strength of the AMOC in CGCMs is systematically related to the preference for one density regime or the other. We consider the measure $\log(\phi_T/\phi_S)$, where larger (smaller) values suggest a more temperature (salinity) dominated density regime. Figure 3, shows that AMOC strength in complex CGCMs is indeed linked to their preference for one density regime over another. For robustness, we use two different definitions of the AMOC hence the two plots, that show a strong linear relationship. CGCMs that have increasingly T-driven density tend to have a stronger AMOC in the mean, with correlations of 0.79 and 0.66 for each plot.

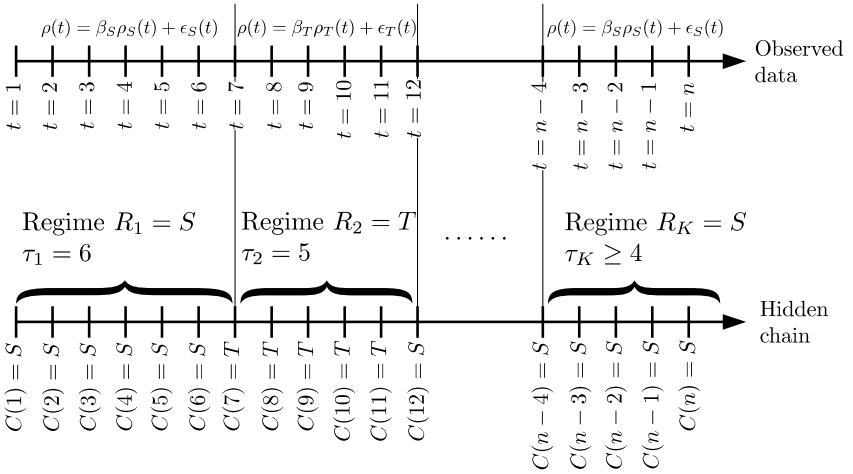


FIGURE 1. Schematic showing a particular realisation of the hidden semi-Markov model for ocean density given by (5)–(6).

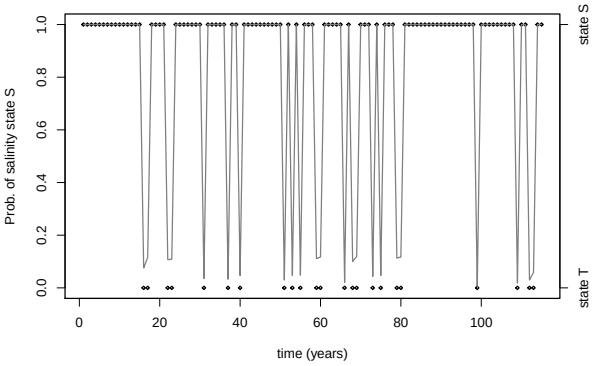


FIGURE 2. Plot of the probability of being in the salinity regime S (as opposed to temperature T) against time t. The most likely state (S or T) is decided by whether this probability is above or below 0.5, and is depicted using a right hand y-axis and diamond symbols.

References

Economou, T., Bailey, T., and Kapelan, Z. (2014). MCMC implementation for Bayesian hidden semi-Markov models with illustrative applications. *Statistics and Computing*. **24**(5), 739 – 752.

Collins, M. (2002). Climate predictability on interannual to decadal time scales: The initial value problem. *Climate Dynamics*, **19**(8), 671 – 692.

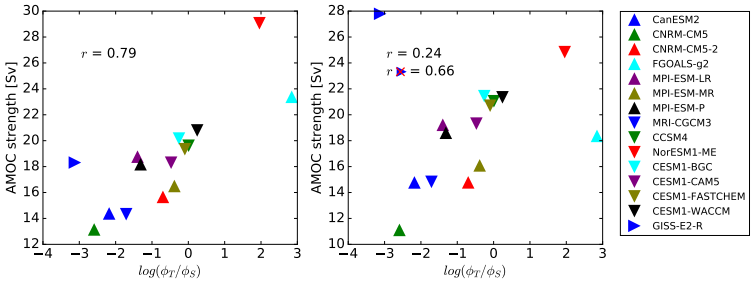


FIGURE 3. Relationship between relative regime persistence and the mean AMOC strength at 26.5°N (left) and 45°N (right) in CGCMs. AMOC strength is measured in Sverdrups [1 Sv = $10^6 m^3/s$].

Menary, M., Hodson, D.., Robson, J., Sutton, R., Wood, R., and Hunt, J. (2015). Exploring the impact of CMIP5 model biases on the simulation of North Atlantic decadal variability. *Geophysical Research Letters*. **42**(14), 5926 – 5934.

Menary, M., Hermanson, L., and Dunstone, N. (2016). The impact of Labrador sea temperature and salinity variability on density and the subpolar AMOC in a decadal prediction system. *Geophysical Research Letters*. **43**(23), 12217 – 12227.

Sensitivity and identification quantification by a relative latent model complexity perturbation in the Bayesian meta-analysis

Małgorzata Roos[1], Haakon Bakka[2], Håvard Rue[2]

[1] University of Zurich, Switzerland
[2] King Abdullah University of Science and Technology, Saudi Arabia

E-mail for correspondence: malgorzata.roos@uzh.ch

Abstract:
Meta-analysis is a well established statistical methodology, strongly recommended for synthesizing scientific knowledge. Frequently, however, only a very small number of studies is available for meta-analysis, and thus the results can be misleading due to unobserved heterogeneity. The Bayesian methodology mitigates this problem by incorporating priors on heterogeneity standard deviation in a normal-normal hierarchical model (NNHM). Nonetheless, in NNHM two relevant concerns remain: parameter identification and sensitivity of the posterior inference with respect to the heterogeneity prior. We still lack a systematic account of how these two concerns affect the posterior inference. Here, we develop a novel two-dimensional sensitivity-identification measure based on numerical derivatives of the Bhattacharyya coefficient with respect to relative latent model complexity perturbations. Our results show that the proposed two-dimensional approach accurately assesses sensitivity and identification. It also explicitly reveals the inherent amount of smoothing in Bayesian meta-analysis applications.

Keywords: Bayesian Meta-Analysis; Normal-Normal Hierarchical Model; Formal Sensitivity And Identification Diagnostics.

1 Introduction

In light of the recent widespread crisis of replicability, decision making under uncertainty has become a very challenging task. The Cochrane, which promotes evidence-based medicine, strongly recommends meta-analysis to aggregate scientific knowledge. Meta-analysis is a formal methodology which allows for a combined evaluation of evidence. It shifts the focus away from

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

evidence in a single study and towards a combined view of evidence provided by several studies. Meta-analysis takes also into account between-study heterogeneity. When the between-study heterogeneity is ignored, the standard errors of estimates are over-optimistically small and the significance of the effect may actually be invalid. Frequently, however, only a very small number of studies is available for meta-analysis, and thus the results can be misleading due to unobserved heterogeneity. The Bayesian methodology mitigates this problem by incorporating priors on heterogeneity standard deviation in a NNHM.

Formally, a Bayesian NNHM consists of three parts: the sampling model (likelihood), the random-effects model (latent field) and priors. The sampling model for a number of I studies assumes (iid) normally distributed outcome Y_i with a fixed within-study standard deviation σ_i, which arise around the latent random-effect parameter θ_i

$$Y_i|\theta_i, \sigma_i \ \sim \ \mathrm{N}(\theta_i, \sigma_i^2). \tag{1}$$

In addition, the exchangeability of latent parameters θ_i is imposed by assuming that the parameters follow a normal distribution with mean μ and a heterogeneity (between-study) standard deviation τ

$$\theta_i|\mu, \tau \ \sim \ \mathrm{N}(\mu, \tau^2), \tag{2}$$

for $i = 1, \ldots, I$. Finally, the priors for $\mu \ \sim \ \pi(\mu)$ and $\tau \ \sim \ \pi(\tau)$ are assumed. In particular, $\pi(\mu)$ is set to $\mathrm{N}(0, 4^2)$ and $\pi(\tau)$ to a Half-Normal (HN) prior as discussed by Röver (2018) and Friede et al. (2017).

The Bayesian meta-analysis gives rise to two concerns: sensitivity (Roos et al. (2015)) and identification (Gelfand and Sahu (1999)). Whereas sensitivity quantifies the heterogeneity prior impact, identification is concerned with the data impact on the marginal posterior inference. Both concerns are rarely addressed, if at all, and they have not yet been addressed simultaneously. Therefore, there is a need to unify the approach and to develop a combined two-dimensional sensitivity-identification (S-I) measure.

2 Data

One typical data set for a medical Bayesian meta-analysis considered by Friede et al. (2017) is shown in Table 1.

These data originated in a systematic review of two randomized controlled trials providing evidence of the efficacy and safety of immunosuppresive therapy with interleukin-2 receptor antibodies following liver transplantation in children. Based on observations from both studies $y_i = \log(OR_i)$ and $\sigma_i = \mathrm{SE}(\log(OR_i))$ were computed and supplied for a Bayesian meta-analysis.

Note that all meta-analyses presented by Friede et al. (2017) have the following three properties in common: first, they incorporate a small number

TABLE 1. Data for the Acute Graft Rejection discussed by Friede et al. (2017).

	experimental		control		y	σ
study	events	total	events	total	$\log(OR)$	$\text{SE}(\log(OR))$
1	14	61	15	20	-2.31	0.60
2	4	36	11	36	-1.26	0.64

of studies. Second, the individual studies have small sample sizes, which renders the within-study standard deviation σ_i estimates uncertain. Third, they assume approximate normality by imposing a NNHM.

3 Methods

We quantify the impact of a formal perturbation from a base (b) to an altered (a) model on marginal posterior distributions by a symmetric measure of affinity: the Bhattacharyya coefficient (BC) (Roos et al. (2015))

$$\text{BC}(\pi_b(\psi|y), \pi_a(\psi|y)) = \int_{-\infty}^{\infty} \sqrt{\pi_b(\psi|y)\pi_a(\psi|y)}d\psi, \tag{3}$$

with $\psi \in \{\mu, \log(\tau), \theta_1, \ldots, \theta_I\}$. We utilize the model complexity p_D (Spiegelhalter et al. (2002)) and the reference standard deviation σ_{ref} (Sørbye and Rue (2014)) to define the relative latent model complexity (RLMC) in a NNHM

$$\text{RLMC} = p_D/I = \frac{\tau^2}{\tau^2 + \sigma_{ref}^2}. \tag{4}$$

RLMC can be thought of as the amount of smoothing inherent to the NNHM with a value 0 indicating perfect and 1 no smoothing. Lower values of RLMC can be obtained by either perturbing the heterogeneity prior (P), to put more weight on values close to 0 with fixed data, or perturbing data to get a less pronounced likelihood with a fixed heterogeneity prior. A derivative of BC (dBC) with respect to RLMC perturbations quantifies how quickly the marginal posterior changes with changing RLMC. The unified two-dimensional S-I measure is defined as a ratio of two derivatives

$$S^{\phi}(\psi) = \frac{d\text{BC}_P(\psi)}{d\text{BC}_P(\phi)} \quad \text{and} \quad I^{\phi}(\psi) = \frac{d\text{BC}_L(\psi)}{d\text{BC}_L(\phi)} \tag{5}$$

with $\psi \in \{\mu, \log(\tau), \theta_1, \ldots, \theta_I\}$ and $\phi \in \{\mu, \log(\tau)\}$. Note that the S-I measure is invariant to the size and direction of RLMC perturbations.

4 Results

Table 2 shows the two-dimensional S-I measure for a NNHM applied to the Acute Graft Rejection data in Table 1 with a HN heterogeneity prior

TABLE 2. Sensitivity and identification estimates for the NNHM applied to the Acute Graft Rejection data with RLMC = 0.25 and a HN heterogeneity prior.

parameter	$d\text{BC}_P$	S^μ	S^τ	$d\text{BC}_L$	I^μ	I^τ
μ	0.078	1	0.386	0.196	1	13.067
$\log(\tau)$	0.202	2.589	1	0.015	0.077	1
θ_1	0.031	0.397	0.153	0.459	2.342	30.600
θ_2	0.039	0.500	0.193	0.385	1.964	25.667

and RLMC fixed at 0.25. Whereas a high $S^\mu(\log(\tau))$ value indicates that τ is 2.6 times more sensitive to the heterogeneity prior than μ, a high $I^\tau(\mu)$ value shows that μ is 13 times more informed by the data than τ.

5 Discussion

The novel, unified methodology for a combined, two-dimensional S-I analysis accurately assesses sensitivity and identification. It also involves an explicit specification of the inherent amount of smoothing in the Bayesian meta-analysis applications. Developments in Sørbye and Rue (2014) indicate that our approach has the potential to be extended to complex Bayesian hierarchical models in the context of Latent Gaussian Models.

Acknowledgments: This research was funded by the Swiss National Science Foundation (Project Number 175933).

References

Gelfand, A.H. and Sahu, S.K. (1999). Identifiability, improper priors, and Gibbs sampling for Generalized Linear Models. *Journal of the American Statistical Association*, **94**, 247 – 253.

Friede, T., Röver, C., Wandel, S., and Neuenschwander, B. (2017). Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometrical Journal*, **59**, 658 – 671.

Roos, M., Martins, T.G., Held, L., and Rue, H. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis*, **10**, 321 – 349.

Röver, C. (2018). Bayesian random-effects meta-analysis using the bayesmeta R package. *Journal of Statistical Software*, (in press).

Sørbye, S.H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, **8**, 39 – 51.

Spiegelhalter, D.J., Best, N.G., Carlin, B.R., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B.*, **64**, 583 – 616.

Improved likelihood ratio testing inferences for unit gamma regressions

Ana C. Guedes[1], Francisco Cribari-Neto[1], Patrícia L. Espinheira[1]

[1] Departamento de Estatística, Universidade Federal de Pernambuco, Brazil

E-mail for correspondence: cribari@gmail.com

Abstract: We derive two modified likelihood ratio test statistics for use with unit gamma regressions. The tests bases on such statistics are expected to deliver more accurate inferences in small samples relative to the standard likelihood ratio test.

Keywords: Doubly limited variables; Likelihood ratio test; Unit gamma distribution; Unit gamma regression.

1 Introduction

Regression analysis is commonly used to explain the behavior of doubly limited continuous dependent variables (DBCDVs) that assume values in (a, b), where a and b are known and $-\infty < a < b < \infty$. The beta regression model (Ferrari and Cribari-Neto, 2004) is the most well known and the most widely used model with DBCDVs. Alternative models have, nonetheless, been introduced in the literature since it is useful for practitioners to have more than a single model at disposal. Alternative models have, nonetheless, been introduced in the literature. For example, Mousa et al. (2016) proposed the unit gamma regression model.

The likelihood ratio (LR) test is the most commonly used test in regression analysis in general and also in regression models for DBCDVs. A shortcoming of such as test is that it relies on an asymptotic approximation and can be considerably size-distorted in small samples. In this paper we shall focus on the the unit gamma regression model. We shall consider mean effects modeling and also joint mean and precision effects modeling. Our interest is on hypothesis testing inferences performed with samples of small sizes. In particular, our chief goal is to derive two modified LR test statistics that can be used to perform reliable testing inferences in unit gamma regressions

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

when the sample size is small. The Monte Carlo evidence that we present shows that testing inferences based on the two modified test statistics can be considerably more accurate than that based on the standard likelihood test statistic in small samples. The latter can be considerably oversized. Tests with improved finite sample behavior are obtained for use with beta regressions; e.g., Ferrari and Pinheiro (2011). To the best our knowledge, however, no similar results have been obtained for the unit gamma regression model.

2 The unit gamma regression model

Let $y_1, \ldots, y_n$ be independent random variables, where each $y_i \sim ug(\mu_i, \phi_i)$, $i = 1, \ldots, n$. That is, each y_i is unit gamma-distributed with mean μ_i and precision ϕ_i. In the unit gamma regression model proposed by Mousa et al. (2016), the ith mean response and the ith precision can be written as

$$g_1(\mu_i) = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij} \quad \text{and} \quad g_2(\phi_i) = \zeta_i = \sum_{j=1}^{q} \delta_j h_{ij}, \tag{1}$$

respectively. Since $\mathrm{var}(y_i) = \mu_i\{[1/(2-\mu_i^{1/\phi_i})^{\phi_i}] - \mu_i\}$, where $\mu_i = g_1^{-1}(\eta_i)$, the regression model is heteroskedastic. The fixed dispersion unit gamma regression model is obtained by setting $g_2(\phi_i) = g_2(\phi) = \delta_0$.

3 Two improved likelihood ratio tests

Consider the model in (1) and also the corresponding log-likelihood function (ℓ), where $\theta = (\beta^T, \delta^T)^T$ is the model k-dimensional parameter vector, β being a p-vector and δ being a q-vector ($p + q = k$). In what follows, $\kappa = (\kappa_1, \ldots, \kappa_l)^T$ is the parameter of interest and $\psi = (\psi_1, \ldots, \psi_s)^T$ is the nuisance parameter. (Note that $l + s = p + q$). We wish to test $\mathcal{H}_0 : \kappa = \kappa^0$ vs. $\mathcal{H}_0 : \kappa \neq \kappa^0$, where κ^0 is a fixed l-vector. The LR test statistic is $w = 2[\ell(\hat{\kappa}, \hat{\psi}) - \ell(\kappa^0, \tilde{\psi})]$, where $(\kappa^{0T}, \tilde{\psi}^T)$ and $(\hat{\kappa}^T, \hat{\psi}^T)$ are, respectively, the restricted and unrestricted maximum likelihood estimators of (κ^T, ψ^T). Under $\mathcal{H}_0$, w is asymptotically distributed as χ^2_l. The null hypothesis is rejected at the α significance level ($0 < \alpha < 1$) if $w > \chi^2_{1-\alpha,l}$, where $\chi^2_{1-\alpha,l}$ is the $1-\alpha$ upper χ^2_l quantile. When n is small, the approximation used in the LR test may not be accurate, and as a result size distortions may take place.

We follow an approach developed by Skovgaard (2001), who proposed the following modified LR test statistic: $w^* = w - 2\log\xi$. Here,

$$\xi = \frac{\{|\tilde{I}||\hat{I}||\tilde{J}_{\psi\psi}|\}^{1/2}}{|\bar{\Upsilon}||\{\tilde{I}\bar{\Upsilon}^{-1}\hat{J}\hat{I}^{-1}\bar{\Upsilon}\}_{\psi\psi}|^{1/2}} \frac{\{\tilde{U}^T\bar{\Upsilon}^{-1}\hat{I}\hat{J}^{-1}\bar{\Upsilon}\tilde{I}^{-1}\tilde{U}\}^{l/2}}{w^{l/2-1}\tilde{U}^T\bar{\Upsilon}^{-1}\bar{q}},$$

where U is the score vector and $J_{\psi\psi}$ is the Hessian matrix. When the relevant regularity conditions are satisfied, $-J_{\psi\psi}$ is the observed information matrix relative to ψ. Note that $\bar{q}$ is a vector of dimension $l+s$ and $\bar{\Upsilon}$ is a matrix of dimension $(l+s) \times (l+s)$. Under $\mathcal{H}_0$, w^* is asymptotically distributed as χ^2_l. The quantities $\bar{q}$ and $\bar{\Upsilon}$ come from $q = \mathbb{E}[U(\theta_1)(\ell(\theta_1) - \ell(\theta))]$ and $\Upsilon = \mathbb{E}[U(\theta_1)U^T(\theta)]$ by replacing θ_1 with $\hat{\theta}$ and θ with $\tilde{\theta}$ after the expected values are computed. Here, $\hat{\theta}$ and $\tilde{\theta}$ denote, respectively, the unrestricted and restricted maximum likelihood estimators of θ.
An asymptotically equivalent test statistic is $w^{**} = w\left(1 - w^{-1}\log\xi\right)^2$. A clear advantage of w^{**} is that it is always non-negative.
We derived closed form expressions for $\bar{q}$ and $\bar{\Upsilon}$ in the class of unit gamma regression models. After some algebra, we arrived at

$$\bar{q} = \begin{bmatrix} X^T\hat{T}_1\hat{\Phi}^{-1}\hat{\mathcal{M}}^{-1}\hat{D}(\mathcal{I}+\hat{D})\{\hat{V}^*(\hat{D}-\tilde{D}) + \hat{C}(\hat{\Phi}-\tilde{\Phi})\}\iota \\ H^T\hat{T}_2\{(\hat{P}\hat{V}^*ˆ + \hat{C})(\hat{D}-\tilde{D}) + (\hat{P}\hat{C} + \hat{V}^\dagger)(\hat{\Phi}-\tilde{\Phi})\}\iota \end{bmatrix}$$

and

$$\bar{\Upsilon} = \begin{bmatrix} X^T\hat{T}_1\hat{\Phi}^{-1}\hat{\mathcal{M}}^{-1}\hat{D}(\mathcal{I}+\hat{D})\hat{V}^* & X^T\hat{T}_1\hat{\Phi}^{-1}\hat{\mathcal{M}}^{-1}\hat{D}(\mathcal{I}+\hat{D}) \\ \times(\mathcal{I}+\tilde{D})\tilde{D}\tilde{\mathcal{M}}^{-1}\tilde{\Phi}^{-1}\tilde{T}_1X & \times\{\hat{V}^*\tilde{P} + \hat{C}\}\tilde{T}_2H \\ H^T\hat{T}_2\{\hat{P}\hat{V}^* + \hat{C}\}(\mathcal{I}+\tilde{D})\tilde{D}\tilde{\mathcal{M}}^{-1}\tilde{\Phi}^{-1}\tilde{T}_1X & H^T\hat{T}_2\{\hat{P}\hat{V}^*\tilde{P} + (\hat{P}+\tilde{P})\tilde{C} + \tilde{V}^\dagger\}\tilde{T}_2H \end{bmatrix},$$

where V^*, $V^\dagger$ and C are diagonal matrices properly defined. The other quantities are also defined in terms elements of the unit gamma regression model.

4 Numerical evidence

We shall report the results of Monte Carlo simulations. The number of Monte Carlo replications is 10,000. All simulations were performed using the Ox matrix programming language.
We consider the varying precision unit gamma regression model given by $\log(\mu_i/(1-\mu_i)) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$ and $\log(\phi_i) = \delta_1 + \delta_2 h_{i2} + \delta_3 h_{i3} + \delta_4 h_{i4}$, $i = 1, \ldots, n$. The covariates valyes are obtained as random standard uniform draws. We test $\mathcal{H}_0 : \beta_3 = \beta_4 = 0, \delta_2 = \delta_3 = \delta_4 = 0$ ($l = 5$). Data generation was carried out using $\beta_1 = 1.5$, $\beta_2 = 1.5$, $\beta_3 = 0$, $\beta_4 = 0$, $\delta_1 = (\log(30), \log(10), \log(5))$, $\delta_2 = \delta_3 = \delta_4 = 0$. The sample sizes are $n = 20, 40, 60$. Figure 1 contains QQ plots constructed using the test statistics values. The null distribution of w is poorly approximated by the limiting χ^2 distribution. Such an approximation is much more precise when used with the two corrected test statistics derived in this paper, especially when $n \geq 40$. Notice that the case in which there are only twenty observations in the sample is quite challenging since there are five restrictions under test. Even with $n = 20$, however, the χ^2 approximation is somewhat precise when used with w^* and w^{**}, except in the distribution upper tail.

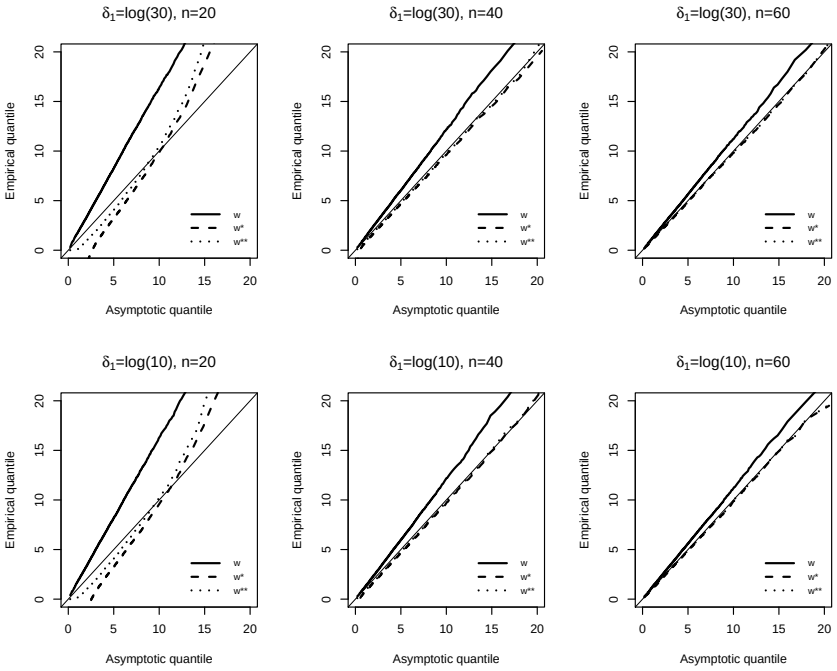


FIGURE 1. Quantile-quantile plots.

5 Concluding remarks

We obtained two modified likelihood ratio test statistics that can be reliably used to perform testing inferences on the parameters that index the unit gamma regression model when the number of observations is small.

References

Ferrari, S.L.P., and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799 – 815.

Ferrari, S.L.P., and Pinheiro, E.C. Improved likelihood inference in beta regression. *Journal of Statistical Computation and Simulation*, **81**, 431 - 443.

Mousa, A.M., El-Sheikh, A.A., and Abdel-Fattah, M.A. (2016). A gamma regression for bounded continuous variables. *Advances and Applications in Statistics*, **49**, 305 – 326.

Skovgaard, I.M. (2001). Likelihood asymptotics. *Scandinavian Journal of Statistics*, **28**, 3 – 32.

Learning and modelling dependence structures with diagonal distributions

M. de Carvalho[1], R. Rubio[2], M. Leonelli[3]

[1] School of Mathematics, University of Edinburgh, UK
[2] Department of Mathematics, Pontificia Universidad Católica de Chile, Chile
[3] School of Mathematics & Statistics, University of Glasgow, UK

E-mail for correspondence: Miguel.deCarvalho@ed.ac.uk

Abstract: This paper introduces *diagonal distributions* as a copula-based method that extends the concept of marginal distributions. The main diagonal is studied in detail, which consists of a mean-constrained univariate distribution function on the unit interval that summarizes main aspects on the dependence structure of a random vector, and whose variance has links with Spearman's rho. An application is given illustrating how diagonal densities can be used so to contrast the diversification of a portfolio based on FAANG stocks against one based on crypto-assets.

Keywords: Copula; Dependence modeling; Marginal distribution; Mean constrained inference.

1 Marginals and diagonals

1.1 Diagonals

We start by laying the groundwork. The dependence within a random vector can be fully characterized by means of a copula, that provides the formal link between the joint distribution F and the marginal distributions F_1, F_2. Formally, the copula is the distribution function $C : [0,1]^2 \to [0,1]$, with uniform marginals, obeying

$$C(F_1(y_1), F_2(y_2)) = F(\mathbf{y}), \quad \mathbf{y} = (y_1, y_2).$$

Let $(U, V) \sim C(u, v)$ with $U = F_1(Y_1)$ and $V = F_2(Y_2)$. Given the definition of copula, if we project (U, V) onto the u or v axis the resulting projections [namely $(U, 0)$ and $(0, V)$] are tantamount to a Unif$(0, 1)$ distribution, and

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

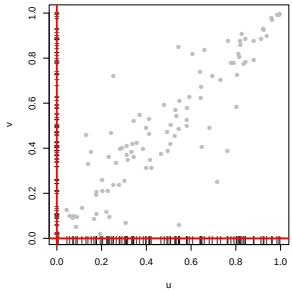


FIGURE 1. Projected pseudo-observations from a Gumbel copula on the u and v axes.

thus only contain information about marginal distributions; see Figure 1. Yet, as we observe below, if we project the random vector (U, V) over any straight line that passes through the origin, the corresponding projection will preserve key features of the dependence structure; of particular interest here will be the distribution of (U, V) over the line $u = v$.

1.2 Main diagonals

It can be shown that the orthogonal projection of $\mathbf{r} = (U, V)$ over the line $u = v$ is $\mathbf{p}(\mathbf{r}) = (Z, Z)$, where $Z \equiv (U+V)/2$. As we discuss below, the law of Z provides information on the dependence between X and Y. Indeed, for $z \in [0, 1]$,

$$F_Z(z) = P(Z \leq z) = P\{(U+V)/2 \leq z\} = \int\int_{\{u+v \leq 2z\}} c(u, v)\, \mathrm{d}u\, \mathrm{d}v,$$

where c is the copula density. We refer to $F_Z(z)$ as the (main) *diagonal distribution function*, and if F_Z is absolutely continuous we refer to $f_Z = \mathrm{d}F_Z/\mathrm{d}z$ as the (main) *diagonal density*. If $U = V$, then $Z \sim \mathrm{Unif}(0, 1)$. In the case of independence the diagonal density can is the symmetric triangular distribution on $[0, 1]$. The case of perfect positive dependence leads to Z is degenerated at $1/2$. Note further that

$$\mathrm{E}(Z) = \frac{1}{2}, \quad \mathrm{var}(Z) = \frac{1}{24} + \frac{1}{2}\,\mathrm{cov}(U, V) = \frac{1}{24}(1 + S), \tag{1}$$

where $S = 12\,\mathrm{cov}(U, V)$ denotes the Spearman's rho. Figure 2 depicts the Normal diagonal density $f(z; \rho)$ with shape parameter $\rho \in [-1, 1]$. If $\rho = 0$, the main diagonal distribution of Z is distributed according to the symmetric triangular distribution, whereas as we increase correlation it gets closer to the uniform distribution.

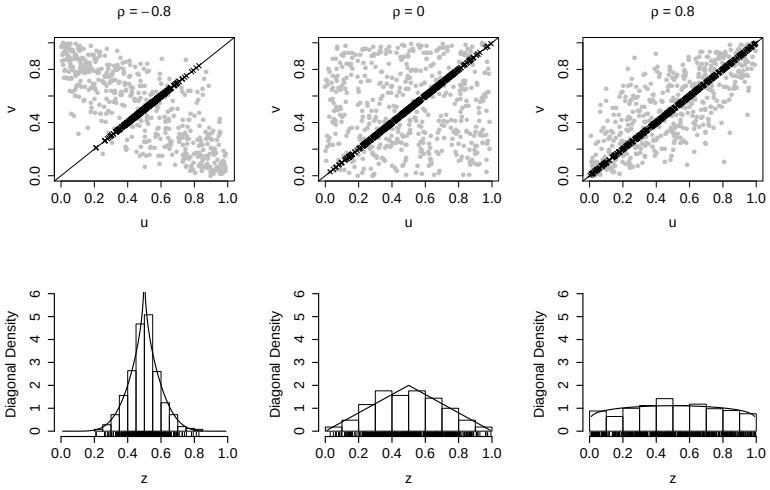


FIGURE 2. Top: Simulated data from a Gaussian copula. Bottom: True normal diagonal density along with a fitted mean-constrained histogram based on the simulated data.

Some comments on D-dimensional extensions and on inference are in order. In the latter setting it follows that $Z = D^{-1}\sum_{d=1}^{D} U_d$, and thus similarly to the bivariate setting it follows that $\mathrm{E}(Z) = 1/2$. And how can we learn about $F_Z(z) = P(Z \leq z)$ from data? By keeping in mind that the diagonal distribution needs to obey a moment constraint ($\mathrm{E}(Z) = 1/2$), we recommend using the mean-constrained density estimator on the unit interval from de Carvalho et al. (2013).

In the case of perfect dependence, the diagonal distribution is uniform, and in the case of independence it is a D-dimensional Bates distribution (Johnson et al., 1995, Section 26.9). Since the main diagonal density summarizes key features of the dependence within a random vector, as a byproduct, this paper also contributes to the literature on multivariate measures of association. Most published articles on measures of association focus on pairwise association of components of a random vector, including Spearman's rho, Kendall's tau, Blomqvist's beta, Gini's gamma, among other. Multivariate extensions can be found in Joe (1990) and Schmid and Schmidt (2007).

2 Illustration: FAANG against Crypto-Currencies

We now showcase our methods in practice. Two investment possibilities that have been receiving a substantial coverage in the financial media over the last few years are FAANG stocks and cypto-currencies. FAANG is an

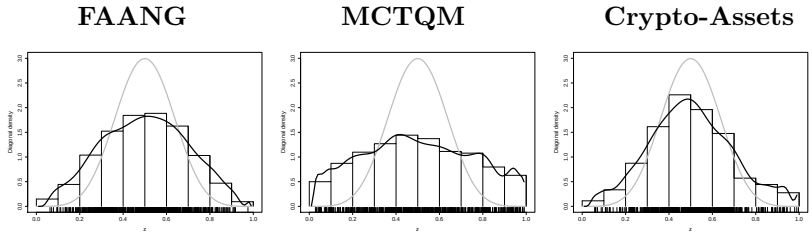


FIGURE 3. Smooth estimate for diagonal density (black), histogram estimate for diagonal density, and the diagonal density corresponding to independence (gray).

acronym for Facebook, Apple, Amazon, Netflix and Alphabet's Google. To have an idea on how FAANG compares with other portfolios of the same size and that also trade on the NASDAQ stock market, we will also be covering the following stocks: Marriott International (MAR), 21st Century Fox Class A (FOXA), Texas Instruments (TXN), Qualcomm (QCOM), and Microchip Technology (MCHP); we will refer to the latter stocks as MCTQM. The period under analysis consists of Aug 2015 to Oct 2017—thus leading to a total 544 observations; we focus on daily negative log-returns which can be regarded as a proxy for losses.

The fitted diagonal densities are presented in Figure 3; an important take-home-message from the fitted diagonal densities is that the FAANG portfolio can be less diversified than the selected portfolio of crypto-currencies (though the latter is subject to a much higher volatility).

Acknowledgments: The authors were financially supported by FCT—Fundação para a Ciência e a Tecnologia, Portugal—through the projects PTDC/MAT-STA/28649/2017 and UID/MAT/00006/2019.

References

de Carvalho, M., Oumow, B., Segers, J., and Warchol, M. (2013). A Euclidean likelihood estimator for bivariate tail dependence. *Communications in Statistics—Theory and Methods*, **42**, 1176–1192.

Joe, H. (1990). Multivariate concordance. *Journal of Multivariate Analysis*, **35**, 12–30.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*. New York: Wiley.

Schmid, F. and Schmidt, R. (2007). Multivariate extensions of Spearman's rho and related statistics. *Statistics & Probability Letters*, **77**, 407—416.

A copula-based multivariate hidden Markov model for modelling momentum in football

Marius Ötting[1], Roland Langrock[1], Antonello Maruotti[2]

[1] Bielefeld University, Germany
[2] Libera Universita Maria Ss. Assunta, Italy

E-mail for correspondence: marius.oetting@uni-bielefeld.de

Abstract: We investigate the potential occurrence of change points – commonly referred to as "momentum shifts" – in the dynamics of football matches. For that purpose, we model minute-by-minute in-game statistics of Bundesliga matches using hidden Markov models (HMMs). To further allow for within-state correlation of the variables considered, we formulate multivariate state-dependent distributions using a copula. The fitted HMMs comprise interpretable states which can be tied to different styles of play, and provide a potentially useful modelling framework allowing insights into causes of momentum shifts.

Keywords: Hidden Markov model; Copula; Football; Momentum; Sports analytics.

1 Introduction

Sports commentators and fans frequently use vocabulary such as "momentum", "momentum shift" or related terms to refer to change points in the dynamics of a match. Usage of such terms is typically associated with situations during a match where an event – such as a shot hitting the woodwork in a football match – changes the dynamics of the match, e.g. in a sense that a team which prior to the event had been pinned back in its own half suddenly seems to dominate the match.
Driven by the rapidly growing amount of freely available football data, several recent studies focused on modelling in-game statistics, e.g. to identify drivers of ball possession (see, e.g., Lago-Penas and Dellal, 2010), or to detect the main playing styles and tactics (Diquigiovanni and Scarpa, 2019). However, existing studies do not focus on the temporal dynamics of in-game statistics during a match, which is of special interest when analysing momentum.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

2 Data

In the present contribution, we analyse minute-by-minute in-game statistics of Bundesliga matches, taken from `www.whoscored.com`, to investigate whether momentum shifts actually do exist in a football match, and what kind of events lead to a shift. For that purpose, multivariate time series $\{\mathbf{y}_{mt}\}_{t=1,2,\ldots,T_m}$ are considered, where $\mathbf{y}_{mt} = (y_{mt1}, \ldots, y_{mtK})$ is the vector of variables observed at time t (in minutes) during match m, $m = 1, \ldots, 34$, with T_m denoting the total number of minutes played in match m. In our analysis, $K = 2$ variables are considered, namely the number of shots on goal and the number of ball touches. Figure 1 shows one example bivariate time series from the data set, which corresponds to the in-game statistics observed for Borussia Dortmund (in a match against SC Freiburg).

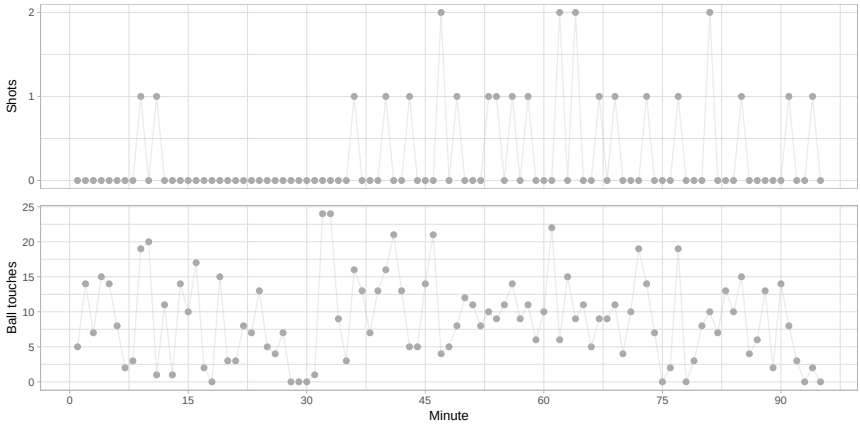


FIGURE 1. Bivariate time series of the number of shots on goal (top) and the ball touches (bottom) of Borussia Dortmund for one example match from the data set (Borussia Dortmund vs. SC Freiburg).

3 Model formulation

Figure 1 underlines that there are periods in the match where the team's number of ball touches and the number of shots on goal are fairly low (e.g. around minutes 20-30), as well as periods with relatively many ball touches and shots on goal (e.g. around minutes 30-45). Hidden Markov models (HMMs) hence constitute a natural modelling approach for the minute-by-minute bivariate time series data, as they accommodate the idea of a match progressing through different phases, with potentially changing momentum. HMMs involve two components: an unobserved Markov chain with N possible states, denoted by $\{s_{mt}\}_{t=1,2,\ldots,T_m}$, and an observed state-dependent process, whose observations are assumed to be generated by one

of N distributions as selected by the Markov chain. Furthermore, within these (multivariate) HMMs, we allow for within-state correlation of the observed variables $\mathbf{y}_{mt}$ by formulating a bivariate state-dependent distribution using a copula, i.e.:

$$F(\mathbf{y}_{mt} \mid s_{mt}) = C\big(F_1(y_{mt1} \mid s_{mt}), F_2(y_{mt2} \mid s_{mt})\big),$$

where F_1 and F_2 are marginal distributions and C is a copula. The corresponding joint p.m.f. is denoted by $f(\mathbf{y}_{mt} \mid s_{mt})$. Since the shots on goal and the ball touches are count variables, and to further account for possible over- and underdispersion, the Conway-Maxwell-Poisson distribution, with p.m.f.

$$\Pr(X = x) = \frac{1}{Z(\lambda, \nu)} \frac{\lambda^x}{(x!)^\nu},$$

is considered as marginal distribution for both variables considered, where $Z(\lambda, \nu) = \sum_{k=0}^{\infty} \lambda^k/(k!)^\nu$, $\lambda > 0$ and $\nu \geq 0$. Since we deal with discrete marginal distributions for the copula, differences are needed rather than derivatives when formulating the joint p.m.f of $\mathbf{y}_{mt}$ given state s_{mt} (see, e.g., Nikoloulopoulos 2013):

$$\begin{aligned} f(\mathbf{y}_{mt} \mid s_{mt}) = {} & C\big(F_1(y_{mt1} \mid s_{mt}), F_2(y_{mt2} \mid s_{mt})\big) \\ & - C\big(F_1(y_{mt1} - 1 \mid s_{mt}), F_2(y_{mt2} \mid s_{mt})\big) \\ & - C\big(F_1(y_{mt1} \mid s_{mt}), F_2(y_{mt2} - 1 \mid s_{mt})\big) \\ & + C\big(F_1(y_{mt1} - 1 \mid s_{mt}), F_2(y_{mt2} - 1 \mid s_{mt})\big), \end{aligned}$$

with F_1 and F_2 denoting the cumulative distribution functions of the two marginals. To also account for possible negative within-state correlation in $\mathbf{y}_{mt}$, the Frank copula with dependence parameter θ is chosen here, which is given by

$$C(u_1, u_2) = -\frac{1}{\theta} \log\Big(1 + \frac{(\exp(-\theta u_1) - 1)(\exp(-\theta u_2) - 1)}{\exp(-\theta) - 1}\Big).$$

Defining an $N \times N$ diagonal matrix $\mathbf{P}(\mathbf{y}_{mt})$ with the i-th diagonal element given by $f(\mathbf{y}_{mt}|s_{mt} = i)$, transition probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{ij})$ with $\gamma_{ij} = \Pr(s_{mt} = j|s_{m,t-1} = i)$ and $\boldsymbol{\delta} = \big(\Pr(s_{m1} = 1), \ldots, \Pr(s_{m1} = N)\big)$, the likelihood of our HMM for one match given by:

$$L = \boldsymbol{\delta}\mathbf{P}(\mathbf{y}_{m1})\mathbf{\Gamma}\mathbf{P}(\mathbf{y}_{m2}) \ldots \mathbf{\Gamma}\mathbf{P}(\mathbf{y}_{mT_m})\mathbf{1}$$

with column vector $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^N$ (see Zucchini et al., 2016). Calculation of this matrix product expression amounts to running the forward algorithm, which is a powerful recursive technique for efficiently calculating the likelihood of an HMM, at computational cost $\mathcal{O}(TN^2)$. We can thus fit

the copula-based bivariate HMM to data using numerical maximum likelihood estimation. To obtain the likelihood for the full data set, we assume independence between the individual matches such that the likelihood is given by the product of likelihoods for the individual matches.

4 Preliminary results

As a case study for assessing the potential of the copula-based bivariate HMM to investigate momentum in football, we analyse Bundesliga data from Borussia Dortmund (season 2017/18) with $N = 3$ states. Maximising the likelihood leads to the estimated state-dependent distributions shown in Figure 2. In addition, Table 1 displays the estimated parameters of the marginal distributions as well as the dependence parameter. According to the fitted model, in state 1 the mean number of shots on goal is ≈ 0, and the mean number of ball touches is 0.269. The corresponding means are 0.136 (shots) and 4.492 (ball touches) for state 2, and 0.185 (shots) and 8.640 (ball touches) for state 3. Thus, state 1 can be interpreted as a defensive only state, which can also be seen from Figure 2. State 2 refers to a state where the match is balanced, whereas state 3 refers to a state with a dominant style of play. In state 3, the estimated negative dependence between shots and ball touches may result from two different styles of dominant play: either Borussia Dortmund is controlling and passing the ball without much pressure on goal, or they go effectively straight for goal, without much controlled passing. The first possible style of play is a more defensive style of dominance, whereas the latter refers to a more offensive dominance. In addition, the estimated t.p.m. is given by

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.253 & 0.079 & 0.667 \\ 0.011 & 0.985 & 0.004 \\ 0.079 & 0.013 & 0.907 \end{pmatrix}.$$

Here, with $\hat{\gamma}_{22} = 0.985$ and $\hat{\gamma}_{33} = 0.907$, there is a high persistence of staying in state 3 (dominance state). Persistence is also high in state 2 (balanced state), whereas with $\hat{\gamma}_{11} = 0.253$, state 1 (defensive only state) is a transient state, where switching to state 3 (dominance state) is most likely.

5 Outlook

Current research focuses on including covariates in the state process, such that the probabilities γ_{ij} of switching between the underlying states depend on (e.g.) the intermediate score of the match and the strength of the opponent. Corresponding analyses may shed some light on what causes shifts in momentum, which would be of great interest to managers, bookmakers and sports fans.

Variable	State 1	State 2	State 3
Shots on goal	$\hat{\lambda} \approx 0, \hat{\nu} = 0.730$	$\hat{\lambda} = 0.120, \hat{\nu} \approx 0$	$\hat{\lambda} = 0.167, \hat{\nu} = 0.329$
Ball touches	$\hat{\lambda} = 0.212, \hat{\nu} \approx 0$	$\hat{\lambda} = 1.069, \hat{\nu} = 0.140$	$\hat{\lambda} = 1.636, \hat{\nu} = 0.253$
Dependence	$\hat{\theta} = 4.845$	$\hat{\theta} = 1.090$	$\hat{\theta} = -1.004$

TABLE 1. Parameter estimates for the state-dependent distributions.

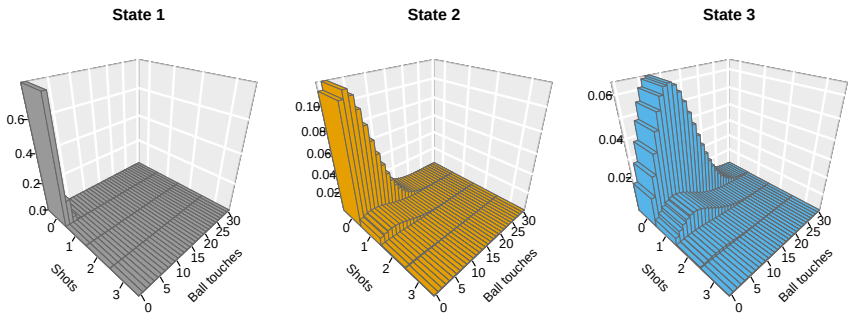


FIGURE 2. State-dependent distributions of the 3-state HMM.

References

Diquigiovanni, J., and Scarpa, B. (2019). Analysis of association football playing styles: An innovative method to cluster networks. *Statistical Modelling*, **19** (1), 28 – 54.

Lago-Penas, C., and Dellal, A. (2010). Ball possession strategies in elite soccer according to the evolution of the match-score: The influence of situational variables. *Journal of Human Kinetics*, **25** (1), 93 – 100.

Nikoloulopoulos, A. K. (2013). Copula-based models for multivariate discrete response data. In: *Copulae in Mathematical and Quantitative Finance*, Springer, Berlin, Heidelberg, 231 – 249.

Zucchini, W., MacDonald, I.L. and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction using R, 2nd Edition*. Boca Raton: Chapman & Hall/CRC.

Estimating mode effects from a sequential mixed-modes experiment

Paul S. Clarke[1] and Yanchun Bao[1]

[1] Institute for Social & Economic Research, University of Essex, U.K.

E-mail for correspondence: pclarke@essex.ac.uk

Abstract: The large-scale household panel study *Understanding Society* has, until recently, used interviewers to administer its questionnaires, but is now in the process of allowing individuals to participate using the web. Survey data are known to be affected by survey mode so a sequential-design experiment was carried out to evaluate the impact of web mode on the panel. We present a suite of approaches based on structural mean models for quantifying the impact of mode across a range of statistical analyses involving variables with different measurement scales. Adaptations of these methods to adjust for non-response bias and for data from complex sampling designs will also be presented.

Keywords: Instrumental variable; Selection effects; Structural mean model.

1 Introduction

The survey mode of the British Household Panel Survey and its successor, *Understanding Society*: The U.K. Household Longitudinal Study (UKHLS), has traditionally been (face-to-face) interview, but UKHLS is following other large-scale surveys by phasing in web mode as an option for the study participants. Unfortunately, survey mode is not neutral as far as the collection of survey data is concerned. The use of web mode could be positive in that respondents may answer sensitive questions more truthfully without an interviewer present, but could also be negative for complicated questions in which interviewer explanations and prompts could have helped to obtain more accurate answers (d'Ardenne et al. 2017). Whether a large mode effect represents a decline in data quality thus depends on the question, but the central issue here is that such changes could affect the resulting time series.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

To this end, a sequential mixed-modes experiment was carried out as part of UKHLS Wave 8. Households were initially randomised either to interview or web mode, but individual household members were then allowed to *non-comply* by choosing the other mode. This design, akin to the *encouragement designs* used for clinical trials, allows us to estimate mode effects by using randomisation as an instrumental variable to adjust for potentially non-random non-compliance. However, it is not straightforward to talk about *the* mode effect for a survey like UKHLS, where the questionnaire involves hundreds of questions and the data can be analysed in many different ways.

2 Mode Effects

Denote the observed values of the survey variables by $\mathbf{Y}_i = (Y_{1i}, \ldots, Y_{ki})^T$ for $i = 1, \ldots, n$. Using potential outcomes notation, let $\mathbf{Y}_i(0)$ denote the values recorded using interviewer mode and $\mathbf{Y}_i(1)$ those recorded using web mode. The **first** complicating issue is that we cannot observe both responses. If $M_i \in \{0, 1\}$ indicates the survey mode used by individual i, then $\mathbf{Y}_i(0)$ is counter-factual if $M_i = 1$ and $\mathbf{Y}_i(1)$ is counter-factual if $M_i = 0$ or, succinctly, $\mathbf{Y}_i = (1 - M_i)\mathbf{Y}_i(0) + M_i\mathbf{Y}_i(1)$. We use causal estimation techniques to addres this problem (see below).

The **second** complicating issue is to determine the relevant mode effect. If your analysis involves only the mean of one continuous/binary variable then $\mathbf{Y} = Y$ and the relevant mode effect for individual i is $Y_i(1) - Y_i(0)$, that is, the difference between the survey variable under the two modes *for the same individual*, and the average mode effect is $E\big\{Y_i(1) - Y_i(0)\big\}$. However, mode can also affect the spread of the survey variable distribution, which could be measured by, e.g., $\mathrm{var}\big\{Y_i(1)\big\}/\mathrm{var}\big\{Y_i(0)\big\}$.

There are many more ways to estimate mode effects for multivariate analyses. For example, if the aim were to estimate the coefficients of the multiple linear regression of Y_k on $Y_1, \ldots, Y_{k-1}$, the mode effects could be $\beta_j(1) - \beta_j(0)$ for $j = 1, \ldots, k - 1$, where $\beta_j(m)$ is the hypothetical estimand of the least-squares estimator of the coefficient of predictor j using $\{\mathbf{Y}_i(m) : i = 1, \ldots, n\}$ as data. More generally, the multivariate mode effect $\mathbf{Y}_i(1) - \mathbf{Y}_i(0)$ leads to an analysis-specific mode effect $\theta(1) - \theta(0)$, where θ is the parameter of analysis model $g(\mathbf{y}; \theta)$ for a family or class of models. We thus focus on mode effects for covariances because these are directly related to the parameters of the family of structural equation models (Vannieuwenhuyze 2015).

3 Estimating Mode Effects and Data

For an observational survey in which participants were free to choose, non-random mode selection would potentially confound the estimation of mode

effects if the factors driving selection were associated with the characteristics measured by the survey variables. A number of relatively simple methods for mode-effect adjustment would be available if it were possible to control for these factors using mode-invariant $\mathbf{Z}$ (Kolenikov and Kennedy 2014). The most powerful approach developed to date is based on multiple imputation of the counterfactual $\mathbf{Y}_i(0)$ among those who choose web (Park et al. 2016), but this is computationally intensive and requires many modelling assumptions about the survey variables and the mode effect.
One way to eliminate non-random selection is to conduct an experiment in which mode is randomly allocated. *Sequential designs* are more practicable: participants are randomly allocated to survey mode but, should they decline, are offered another mode in a pre-determined sequence until it is clear they do not wish to respond. The advantage of this design is that, while non-compliance can also be non-random, the initial randomisation can be used as an instrumental variable to adjust for selection effects even in the likely situation that no plausible $\mathbf{Z}$ is available. A sequential design was implemented during the first phase of fieldwork for UKHLS Wave 8. Our data are from a 60:40 randomisation of 5542 households (3298:2144) to give 5866 interview-first and 3917 web-first individual participants. The non-compliance among the interview-first and web-first individuals was, respectively, 6.6% and 33.9%.

4 Structural Mean Models

Vannieuwenhuyze (2015) reviews the use of instrumental variables (IVs) for the estimation of mode effects. Valid IVs for mode M_i affect $\mathbf{Y}_i$ but only through M_i and not any other pathway. We draw from the literature on encouragement designs which use randomisation, where Goetghebeur and Vansteelandt (2005) describe estimation of causal effects using structural mean models.
If R_i indicates the randomised allocation and M_i the mode chosen by individual i, with $R_i \neq M_i$ indicating non-compliance, a structural mean model (SMM) for the average mode effect is

$$E\{Y_i - Y_i(0) \mid M_i, R_i\} = M_i\mu,$$

where the estimator for $\mu = E\{Y_i(1) - Y_i(0) \mid M_i = 1\}$ is derived from the restriction $E\{Y_i(0) \mid R_i\} = E\{Y_i(0)\}$. The estimator $\hat{\mu}$ is identical to the standard two-stage least squares estimator used by other authors to estimate mode effects on the mean; however, unlike classical 'IV regression', the definition of the target parameter is unambiguous. While μ is the mode effect among those who choose web, not for the entire population, this is the relevant parameter if the aim is to gauge the impact of introducing web mode to what was a face-to-face survey. We also introduce the extension of this model to assess the impact on binary and nominal categorical variables.

The first novel estimator we introduce is for the effect of mode on the variance of a continuous variable. The log-linear structural variance model (SVM) is

$$\log\left[\frac{\text{var}\big(Y_i \mid M_i, R_i\big)}{\text{var}\big\{Y_i(0) \mid M_i, R_i\big\}}\right] = M_i\vartheta,$$

where the estimator for $\vartheta = \text{var}\big\{Y_i(1) \mid M_i = 1\big\}/\text{var}\big\{Y_i(0) \mid M_i = 1\big\}$ follows from $E\left\{Y_i(0) \mid R_i\right\} = E\left\{Y_i(0)\right\}$ *and* $E\left\{Y_i^2(0) \mid R_i\right\} = E\left\{Y_i^2(0)\right\}$. Moving on to the effect of mode on associations, we go on to consider linear, log-linear and logistic SMMs for the effect of mode on the association between two variables $Y_i = Y_{i1}$ and $X_i = Y_{i2}$, where X_i is either mode-invariant or only face-to-face. For cases where both Y_i and X_i are mixed-mode, we introduce the following linear structural covariance model (SCM) for the effect of mode on the covariance:

$$\text{cov}\left(X_i, Y_i \mid M_i, R_i\right) - \text{cov}\left\{X_i(0), Y_i(0) \mid M_i, R_i\right\} = M_i\varrho.$$

A linear SCM is appropriate if X and Y are not highly correlated (that is, do not take values close to ± 1). We derive an estimator for ϱ and its standard error based on $E\left\{Y_i(0) \mid R_i\right\} = E\left\{Y_i(0)\right\}$, $E\left\{X_i(0) \mid R_i\right\} = E\left\{X_i(0)\right\}$ and $E\left\{X_i(0)Y_i(0) \mid R_i\right\} = E\left\{X_i(0)Y_i(0)\right\}$. We note that Vannieuwenhuyze (2015) derived an alternative estimator for the effect on the covariance. However, we use semiparametric theory (Tsiatis 2006) to derive efficient estimating equations for all the models described above, and robust sandwich estimators for the standard errors of these models' parameters, which we extend to allow for stratified multi-stage cluster designs.

5 Results

We will summarise the results of a simulation study into the properties of the SMMs on UKHLS-like data; this study will also investigate the performance of confidence intervals based on asymptotic approximations and the bootstrap. However, the focus of our presentation will be on the 361 by 361 mode-effect array for 361 key UKHLS variables. The array's diagonal cells contain the mean and variance mode effects of the indexed variable, and the off-diagonal cells the covariance mode effects. We will show how this array can be used to investigate the impact of mode effects on analyses involving those variables most at risk.

Acknowledgments: This research was funded by U.K. Economic & Social Research Council (ESRC) grant ES/N00812X/1. *Understanding Society* is an initiative funded by the ESRC and various U.K. Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social

Research and Kantar Public. The research data are distributed by the UK Data Service.

References

d'Ardenne, J., Collins, D., Gray, M., Jessop, C., and Pilley, S. (2017) Assessing the risk of mode effects. *Understanding Society* Working Paper Series No. 2017-04. ISER, University of Essex: Colchester.

Goetghebeur, E., and Vansteelandt, S. (2005). SMMs for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statist. Methods Med. Res.*, **14**, 397 – 415.

Kolenikov, S., and Kennedy, C. (2014). Evaluating three approaches to statistically adjust for mode effects. *J. Surv. Statist. Methodol.*, **2**, 126 – 158

Park, S., Kim, J.-K., and Park, S. (2016). An imputation approach for handling mixed-mode surveys. *Ann. Appl. Statist.*, **10**, 1063 – 1085.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*, London: Springer

Vannieuwenhuyze, J. (2015). Mode effects on variances, covariances, standard deviations, and correlations. *J. Surv. Statist. Methodol.*, **3**, 1 – 21

Nonparametric inference in hidden Markov models for time series of counts

Timo Adam[1], Roland Langrock[1], Christian H. Weiß[2]

[1] Bielefeld University, Germany
[2] Helmut-Schmidt-University Hamburg, Germany

E-mail for correspondence: timo.adam@uni-bielefeld.de

Abstract: Hidden Markov models are popular tools for modeling time series where, at each point in time, a hidden state process selects among a finite set of possible distributions for the observations. Specifically for time series of counts, the Poisson family often provides a natural choice for the state-dependent distributions, though more flexible distributions such as the negative binomial or distributions with a bounded range can also be used. Choosing an adequate class of (parametric) distributions, however, is a complex task, and an inadequate choice can have severe negative consequences on the model's performance. To address this issue, we propose a nonparametric approach to fitting hidden Markov models to time series of counts, where the state-dependent distributions are estimated in a completely data-driven way without the need to specify a (parametric) family of distributions. To avoid overfitting, a roughness penalty is added to the likelihood. The suggested approach is illustrated in a real-data application, where the distribution of major earthquake counts is modeled over time.

Keywords: Count data; Nonparametric statistics; Penalized likelihood; State-space model; Time series modeling.

1 Introduction

Hidden Markov models (HMMs) constitute a versatile framework for modeling diverse types of time series data, including, *inter alia*, binary data, positive real-valued data, circular data, categorical data, compositional data, and count data. Depending on the application at hand, potential aims which can be addressed using HMMs include the prediction of future values of a time series, decoding of the hidden states underlying the observations, and inference on the drivers for example on the state-switching dynamics (Zucchini *et al.*, 2016).

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

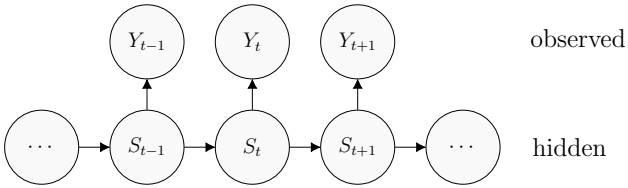


FIGURE 1. Dependence structure of a basic hidden Markov model.

A challenging task in HMMs is the specification of the state-dependent distributions, especially in cases where parametric distributions are not sufficiently flexible to capture complex distributional shapes. While a nonparametric solution based on penalized B-splines (Eilers and Marx, 1996) has been presented for continuous-valued time series (Langrock *et al.*, 2015, 2018), a solution for time series of counts is currently lacking. In this paper, we present such a nonparametric alternative for time series of counts, where the state-dependent distributions are estimated in a completely data-driven way without the need to specify a (parametric) class of distributions.

2 Methodology

2.1 Model formulation

Basic HMMs comprise two stochastic processes,

- a (hidden) state process, $\{S_t\}_{t=1,\ldots,T}$, which is usually modeled as a discrete-time, N-state Markov chain;
- the (observed) time series of interest, $\{Y_t\}_{t=1,\ldots,T}$, which in our specific case is a time series of counts.

Assuming the Markov chain to be time-homogeneous and of first order, the state process is specified by the transition probability matrix $\boldsymbol{\Gamma} = (\gamma_{ij})$,

$$\gamma_{ij} = \Pr\left(S_t = j | S_{t-1} = i\right),$$

$i, j = 1, \ldots, N$, and the initial distribution vector $\boldsymbol{\delta} = (\delta_i)$,

$$\delta_i = \Pr\left(S_1 = i\right),$$

$i = 1, \ldots, N$.
The states determine which of N possible distributions generates the observed count at any time point. The dependence structure of such a basic HMM is illustrated in Figure 1.
While it is common to consider some parametric distributional family such as the class of Poisson or negative binomial distributions, we drop this

assumption and instead assign a state-specific probability to each possible count on the bounded support $\{0, \ldots, K\}$,

$$\pi_{i,k} = \Pr(Y_t = k | S_t = i),$$

$i = 1, \ldots, N$ and $k = 0, \ldots, K$, where the upper bound, K, should at least cover all observed counts.

2.2 Model fitting

Defining the forward probabilities $\alpha_t(i) = \Pr(y_1, \ldots, y_t, S_t = i)$, which are summarized in the row vectors $\boldsymbol{\alpha}_t = (\alpha_t(1), \ldots, \alpha_t(N))$, the recursion

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\mathbf{P}(y_1); \quad \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(y_t),$$

where $\mathbf{P}(k) = \text{diag}(\pi_{1,k}, \ldots, \pi_{N,k})$ and $\mathbf{1} \in \mathbb{R}^N$ is a column vector of ones, can be applied to compute $\boldsymbol{\alpha}_T$, from which the likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = \Pr(y_1, \ldots, y_T | \boldsymbol{\theta}) = \sum_{i=1}^{N} \alpha_T(i) = \boldsymbol{\alpha}_T \mathbf{1} \tag{1}$$

is obtained by the law of total probability. Here $\boldsymbol{\theta}$ is the parameter vector comprising the initial state probabilities, the state transition probabilities, and the state-dependent probabilities of counts.
To avoid overfitting, a roughness penalty is added to the logarithm of the likelihood given in (1), which leads to the penalized log-likelihood

$$\log\left(\mathcal{L}_{\text{pen.}}(\boldsymbol{\theta})\right) = \log\left(\mathcal{L}(\boldsymbol{\theta})\right) - \sum_{i=1}^{N} \lambda_i \sum_{k=m}^{K} \left(\Delta^m \pi_{i,k}\right)^2,$$

where $\lambda_i, i = 1, \ldots, N$, is a smoothing parameter associated with the i-th state-dependent distribution, and where $\Delta^m \pi_{i,k} = \Delta^{m-1}(\Delta \pi_{i,k})$, $\Delta \pi_{i,k} = \pi_{i,k} - \pi_{i,k-1}$, denotes the m-th order differences between adjacent probabilities of counts (typically, $m = 3$). The smoothing parameters can be selected by cross validation over some grid of possible values, where the values corresponding to the highest average out-of-sample log-likelihood are chosen.

3 Illustrating example

To illustrate the suggested approach, we re-consider the running example from Zucchini *et al.* (2016) and model the count variable

$$y_t = \#\text{ of earthquakes worldwide with magnitude } \geq 7 \text{ in year } t$$

over time. The data cover the period from 1900 to 2006. For comparison, we fitted three different models,

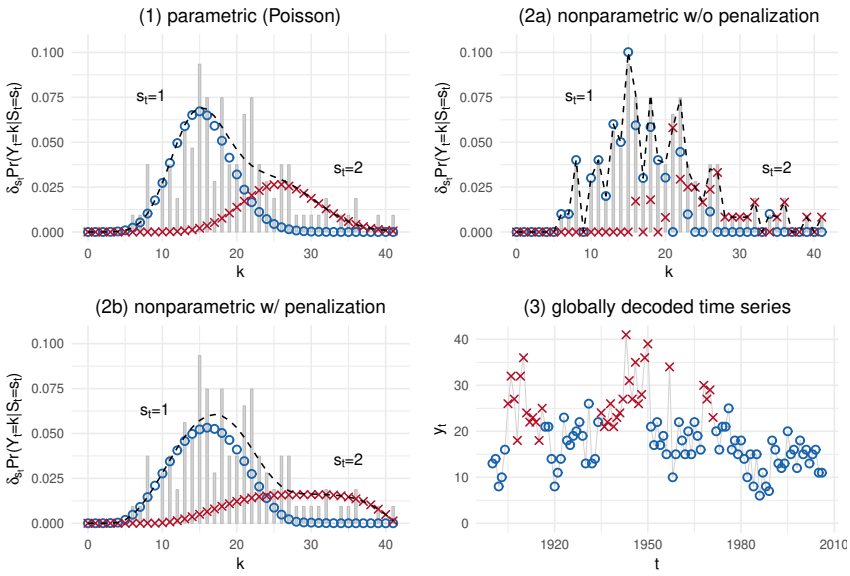


FIGURE 2. Fitted state-dependent probability mass functions (dots and crosses) under the different estimation approaches and decoded time series, with the decoding performed under the model fitted using the penalized nonparametric approach.

- a 2-state Poisson HMM (as a benchmark model);
- a nonparametric 2-state HMM without roughness penalization;
- a nonparametric 2-state HMM with roughness penalization.

Under the model fitted using the penalized nonparametric approach, the transition probability matrix was estimated as

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.934 & 0.066 \\ 0.128 & 0.872 \end{pmatrix}.$$

The associated stationary distribution, $\boldsymbol{\delta} = (0.660, 0.340)$, indicates that about 2/3 and 1/3 of the observations were generated in state 1 and 2, respectively.

The fitted state-dependent probability mass functions obtained by penalized maximum likelihood estimation and the decoded time series, with the decoding performed under the model fitted using the penalized nonparametric approach, are displayed in Figure 2. While the parametric model clearly lacks the flexibility to account for the overdispersion present in the data, the nonparametric model without penalization substantially overfits the data. The nonparametric model with penalization, in contrast, is able

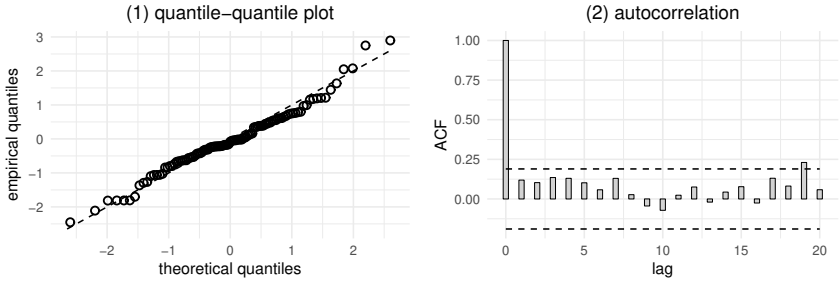


FIGURE 3. Quantile-quantile plot (left) and autocorrelation function (right) of the ordinary normal pseudo-residuals.

to capture the overdispersion specifically in state 2 and yields smooth functional shapes of the estimated state-dependent probability mass functions. Within each of the three different models considered, state 1 may be interpreted as a rather calm regime relating to periods of relatively low seismic activity, whereas state 2 corresponds to periods of relatively higher seismic activity.

Residual checks of the model fitted using the penalized nonparametric approach based on ordinary normal pseudo-residuals are displayed in Figure 3. Pseudo-residuals, which are commonly used for model checking in HMMs, indicate whether an observation is extreme relative to the conditional distribution under the fitted model, given all other observations, and approximately follow a standard normal distribution if the model fits the data well (Zucchini *et al.*, 2016). Overall, the plots do not reveal any substantial lack of fit.

4 Discussion

The proposed methodology constitutes a promising alternative to parametric HMMs for time series of counts (MacDonald and Zucchini, 1997). The increased flexibility to capture complex distributional shapes can improve the model's performance, but can also be regarded as an exploratory tool in applications where it is unclear which distributional family provides an appropriate choice.

On a final note, we wish to highlight that the presented approach is not restricted to modeling time series of counts; in fact, any discrete-valued time series where the observations are at least of ordinal scale (e.g. data on Likert-type scales) can, in principle, be modeled (in a similar spirit as presented in Simonoff, 1983).

References

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

Langrock, R., Adam, T., Leos-Barajas, V., Mews, S., Miller, D.L., and Papastamatiou, Y.P. (2018). *Spline-based nonparametric inference in general state-switching models*. Statistica Neerlandica, **72** (3), 179–200.

Langrock, R., Kneib, T., Sohn, A., and DeRuiter, S.L. (2015). *Nonparametric inference in hidden Markov models using P-splines*. Biometrics, **71** (2), 520–528.

MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov models and other models for discrete-valued time series*. Boca Raton: Chapman & Hall/CRC.

Simonoff, J.S. (1983). *A penalty function approach to smoothing large sparse contingency*. The Annals of Statistics, **11** (1), 208–218.

Zucchini, W., MacDonald, I.L., and Langrock, R. (2016). *Hidden Markov models for time series. An introduction using R*. Boca Raton: CRC.

Estimation of Latent Network Flow in Bike-Sharing Systems from Station Feeds

Marc Schneble[1], Göran Kauermann[1]

[1] Department of Statistics, Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: marc.schneble@stat.uni-muenchen.de

Abstract: Estimation of latent network flow is a common problem in statistical network analysis. Often, we know at least the margins, i.e. in- and outdegrees. In this paper, we develop a generalized mixed regression model to estimate integer temporal network flows if only the differences of in- and outdegrees are known. Estimation can be performed via an iterative penalized maximum likelihood approach. We apply our model to the Vienna Bike-Sharing network. The results show that station- and time-specific effects can be estimated well while it is harder to perform estimation for route-specific effets.

Keywords: Approximate EM-Algorithm; Bike-Sharing Networks; Generalized Additive Mixed Models; Network Flow; Skellam Distribution

1 Model and Notation

Consider a temporal network having N nodes (stations) and therefore N^2 possible edges (routes between stations), where we also allow for self-loops. For the discrete sequence of points in time $t = 0, 1, \ldots, T$ we observe a realization of the $\mathbb{N}_0$-valued random variable $C_i(t)$ (station feeds) on every node $i = 1, \ldots, N$. We denote with $Y_{ij}(t)$ the count of trips from station i to station j departing in the interval $[t-1, t)$ and choose each time interval to be one hour. Our aim is to estimate the network flows $Y_{ij}(t)$ based on the hourly station feeds $C_i(t)$. Since a bike trip departing in $[t-1, t)$ doesn't need to reach its destination within the same time interval, we account for these trips by installing a latent station "on the way", denoted by w. Hence, at every time point t, each bike in the network is either parked in one of the N physical stations or it is located at the latent station w. For the latter, we don't allow for self-loops such that a trip can span at most two time intervals in our model.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

We model the counts separately for each hour of the day and assume the log-linear Poisson model $Y_{ij}(t) \sim \text{Poi}(\lambda_{ij}(t))$ where

$$\lambda_{ij}(t) = \exp(\eta_{ij}(t)) = \exp\left(\eta(z_{ij}(t)) + u_i^{\text{out}} + u_j^{\text{in}}\right). \tag{1}$$

To account for unobserved station specific heterogeneity we specify random effects in (1) which are modeled as independently multivariate normally distributed, i.e. $\mathbf{u}_i = \left(u_i^{\text{out}}, u_i^{\text{in}}\right)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$.
The linear predictor $\eta(z_{ij}(t)) = z_{ij,\text{lin}}(t)\beta + \sum_{m=1}^{M} s_m(z_{ij,m}(t))$ is constructed from the row vector $z_{ij}(t) = (z_{ij,\text{lin}}(t), z_{ij,1}(t), \ldots, z_{ij,M}(t))$ which consists of an intercept as well as station-, route- and time-specific covariate values. The vector $z_{ij,\text{lin}}(t)$ contains linear effects, the scalars $z_{ij,m}(t)$ contain effects that are modeled semiparametrically. Furthermore, $s_m(\cdot)$ is some smooth function in $z_{ij,m}(t)$ represented by a B-spline basis and a vector of basis coefficients $\gamma^{(m)}$. A sum-to-zero constraint is enforced to ensure the identifiability of the M spline functions. According to Wood (2006), we specify a normal prior on the parameters $\gamma^{(m)}$ such that the estimation of Σ and the smoothing parameters ρ_m can be performed simultaneously. We penalize the second-order differences of $\gamma^{(m)}$ such that the variance of $\gamma^{(m)}$ is given by $\rho_m^{-1}\mathbf{K}_2^{(m)^-}$ where $\mathbf{K}_2^{(m)}$ is the second-order difference matrix.
We propose both a model to estimate the whole network flow and a model to estimate only the count of incoming and outgoing bikes at each station. First, the network flows are assumed to be independent given the covariates and random effects. Therefore, the counts of incoming bikes $Y_{\cdot i}(t)$ to station i in $[t-1, t)$ and the number of outgoing bikes $Y_{i\cdot}(t)$ from station i in $[t-1, t)$, respectively, are again Poission-distributed, so that

$$Y_{\cdot i}(t) = \sum_{j=1}^{N} Y_{ji}(t) + Y_{wi}(t) \sim \text{Poi}\left(\sum_{j=1}^{N} \lambda_{ji}(t) + \lambda_{wi}(t)\right) = \text{Poi}(\lambda_{\cdot i}(t)) \tag{2}$$

where $Y_{i\cdot}(t)$ is defined in a similar way. We obtain for the difference in the i'th station count

$$D_i(t) = C_i(t) - C_i(t-1) = Y_{\cdot i}(t) - Y_{i\cdot}(t)$$

a Skellam distribution with parameters $\lambda_{\cdot i}(t)$ and $\lambda_{i\cdot}(t)$, see Alzaid and Omair (2010). Furthermore, the differences of the physical station feeds imply the differences of the latent station's feeds by

$$D_w(t) = \sum_{j=1}^{N} Y_{jw}(t) - \sum_{j=1}^{N} Y_{wj}(t) = Y_{\cdot w}(t) - Y_{w\cdot}(t) = -\sum_{i=1}^{N} D_i(t).$$

Thus, our regression model results to $D_i(t) \sim \text{Skellam}(\lambda_{\cdot i}(t), \lambda_{i\cdot}(t))$ with

$$\mathbb{P}(D_i(t) = d_i(t)) = e^{-(\lambda_{\cdot i}(t) + \lambda_{i\cdot}(t))} \left(\frac{\lambda_{\cdot i}(t)}{\lambda_{i\cdot}(t)}\right)^{\frac{d_i(t)}{2}} I_{|d_i(t)|}\left(2\sqrt{\lambda_{\cdot i}(t)\lambda_{i\cdot}(t)}\right)$$

for $i \in \{1, \ldots, N, w\}, t \in \{1, \ldots, T\}$ and where $I_d(\cdot)$ denotes the modified Bessel function of the first kind. This modeling approach is further denoted as model 1.
In the second approach (model 2), we estimate the in- and outdegrees of the temporal network similar to model 1, i.e. $D_i(t) \sim \text{Skellam}(\lambda_{\cdot i}(t), \lambda_{i\cdot}(t))$. However, now the parameters are modeled as $\lambda_{\cdot i}(t) = \exp\left(\eta(z_i^{\text{in}}(t)) + u_i^{\text{in}}(t)\right)$ and $\lambda_{i\cdot}(t) = \exp\left(\eta(z_i^{\text{out}}(t)) + u_i^{\text{out}}(t)\right)$. The linear predictors $\eta(z_i^{\text{out}}(t))$ and $\eta(z_i^{\text{in}}(t))$ are built from $\eta(z_{ij}(t))$ where any route-specific effects are removed. Making use of the estimated in- and outgoing bikes at each station, one can also estimate the network flow. For example, Chen et al. (2017) estimate the network flow in bike-sharing systems using a type of lasso-and ridge-penalization technique if in- and outdegrees are actually observed.

2 Estimation

Following Fahrmeir and Tutz (2001), estimation of both models is performed by an approximate EM-algorithm. To do so, we first assign a flat prior to the parameters β and maximize the log-posterior $l(\delta) = f(\delta|d; \Sigma, \rho)$ with repect to $\delta = (\beta, \gamma^{(1)}, \ldots, \gamma^{(M)}, \mathbf{u}_1, \ldots, \mathbf{u}_N)$ given the observed differences d as well as current estimates of Σ and the smoothing parameters $\rho = (\rho_1, \ldots, \rho_M)$. It can be shown that

$$l(\delta) = \sum_{i=1}^{N} \sum_{t \in \mathcal{T}} l_D(d_i(t) \mid \delta) - \frac{1}{2} \sum_{m=1}^{M} \rho_m \gamma^{(m)\top} \mathbf{K}_2^{(m)} \gamma^{(m)} - \frac{1}{2} \sum_{i=1}^{N} \mathbf{u}_i^\top \Sigma^{-1} \mathbf{u}_i.$$

Here, $\mathcal{T}$ denotes the set of time points which belong to the evaluated hour of the day. The posterior which is not normal can be interpreted as a penalized log-likelihood where the penalties refer to the spline-parameters $\gamma^{(m)}$ and to the random effects $\mathbf{u}_i$, respectively. Second, we update estimates of Σ and ρ involving $\widehat{\delta}$ and the estimated covariance matrix of $\widehat{\delta}$. More precisely, in the p'th iteration we compute

$$\widehat{\Sigma}^{(p)} = \frac{1}{N+1} \sum_{i \in \{1, \ldots, N, w\}} \left(\widehat{\mathbf{V}}_{u_i u_i} + \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i \right)$$

where $\widehat{\mathbf{V}} = (\widehat{F_{\text{obs}}})^{-1}(\widehat{\delta})$ denotes the inverse of the observed Fisher matrix of δ and $\widehat{\mathbf{V}}_{u_i}$ denotes the diagonal elements of $\widehat{\mathbf{V}}$ related to $\mathbf{u}_i$. The update of ρ is calculated accordingly. This iterative procedure stops if a convergence criterion based on a matrix and a vector norm of Σ and ρ, respectively, is fullfilled. For model 1, we obtain an estimate of the latent network flow by inserting the estimate $\widehat{\delta}$ into (1). For model 2, we first estimate the margins by inserting $\widehat{\delta}$ into $\lambda_{\cdot i}(t)$ and $\lambda_{i\cdot}(t)$. Subsequently, one can estimate the flows $Y_{ij}(t)$ based on the estimated margins, e.g. by applying the method of Chen et al. (2017).

3 Application to the Vienna Bike-Sharing System

We apply our models to the Vienna Bike-Sharing network in the year 2014 consisting of $N = 120$ stations. The complete trip data is used for evaluation purposes. The network is very sparse: 99,1% of the observed $y_{ij}(t)$ are equal to zero which complicates the estimation. Furthermore, the provider redistributes bikes to keep the station feeds balanced. Therefore, we exemplarily evaluate our model for the hour from 5-6 pm on weekdays where the network is least sparse and service rides of the provider only account for 3% of the traffic. Figure 1 shows an extract of the considered bike network indicating the 30 most frequently traveled routes to that time. The larger the dots, the more the station is frequented. The linear predictor η also includes weather specific variables as temperature since they considerably affect the utlization of the system, see the right panel of Figure 1.

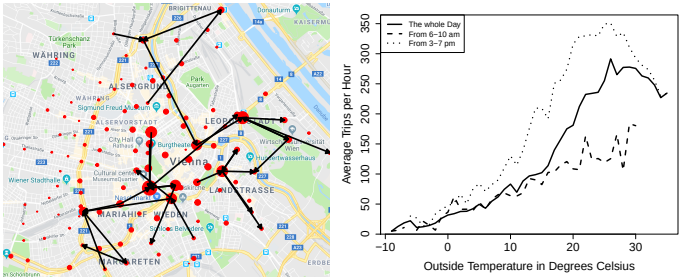


FIGURE 1. Left Panel: The Vienna Bike-Sharing System (5-6 pm on weekdays); Right Panel: Trips per Hour depending on the outside Temperature

4 Results

The total count of trips per hour can be estimated well with both methods whereas model 2 performs slightly better (top panel of Figure 2). From the middle panels of Figure 2 we can infer that estimating the in- and outdegrees directly (model 2) leads to better predictions of the cumulated margins as if we estimate the whole network flow. On average, Model 1 overestimates in- and outdegrees. Using model i, the Pearson correlation coefficients $r_{p(i)}$ of the observed and estimated time series in Figure 2 are $r_{p(1)} = 0.934$ and $r_{p(2)} = 0.938$, respectively. Calculating these coefficients for each station individually, the correlations are clearly lower with a mean of $r_{p(1)}$ ($r_{p(2)}$) equal to 0.50 (0.51) and a standard deviation equal to 0.13 (0.13). Anyhow, the more frequented a station is, the better is the prediction accuracy. The mean of $r_{p(1)}$ ($r_{p(2)}$) restricted to the 32 most frequented stations that determine half of the network flow is equal to 0.63 (0.62) with

a standard deviation of 0.08 (0.09). Although the correlations are similar, the average L1 error when using model 2 for estimating the in- and outdegrees is 9% lower.

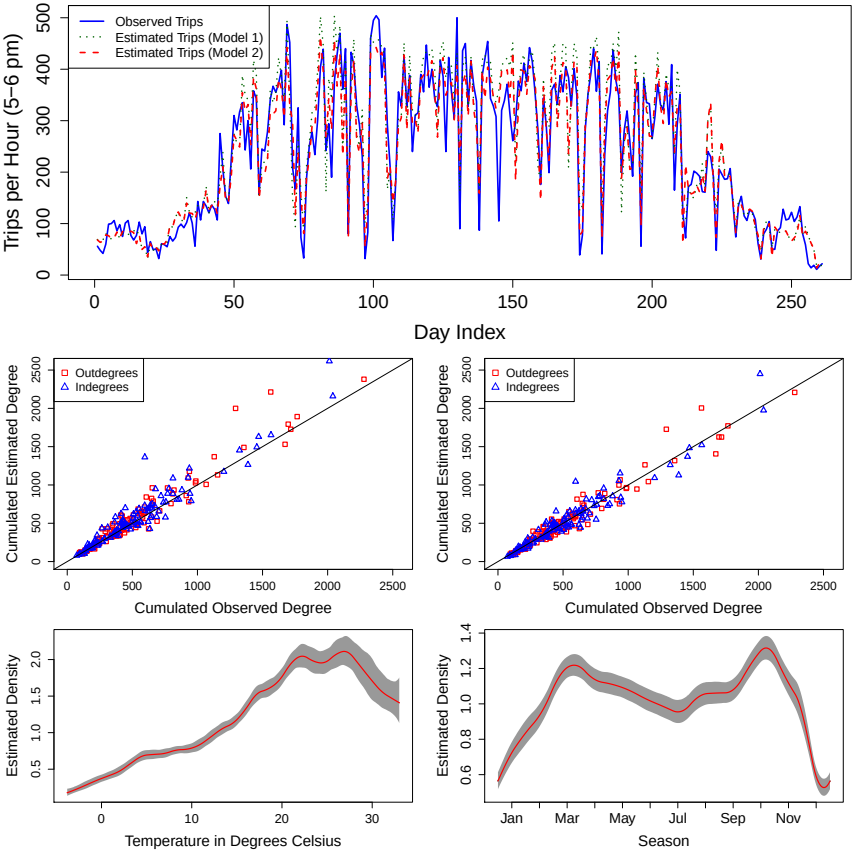


FIGURE 2. Top Panel: Time Series of Observed vs. Estimated Trips per Hour; Middle Panels: Observed vs. Estimated Cumulated In- and Outdegrees over the Estimation Period: Model 1 (left), Model 2 (right); Bottom Panels: Multiplicative Density Estimates of Smooth Effets with 95% Confidence Bands

In the bottom panels of Figure 2 we display the densities including 95% confidence bands of both smooth effects that we estimated with model 2. The smooth temperature effect fits to the corresponding empirical distribution (Figure 1) and the uncertainty grows with the temperature. Furthermore, the bottom right panel of Figure 2 shows that the system is mostly used in spring and fall disregarding all other effects included in the model.
Finally, we assess the performance of the model by the ability of detecting the most frequented routes in the network and estimating the distance effect. The intersection of the 100 most frequented routes that were es-

timated with Model 2 and the actually 100 most frequented routes is 33 routes. Table 1 shows that the count of trips less than 2 km and loops are underestimated. The weight of trips between 2 km and 5 km is slightly overestimated. Trips with a distance of more than 5 km (1/3 of all possible routes) are clearly overstimated.

TABLE 1. Percentage of Cumulated Trips by Distance in km

Distance	0	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$	>5
Observed	7.5%	10.0%	27.1%	23.0%	16.2%	8.5%	7.6%
Estimated	4.5%	5.03%	22.5%	25.8%	18.7%	10.7%	12.7%

5 Discussion

Model 2 produces more accurate estimates of the in- and outdegrees which can be explained by the more stringent independence assumption we postulate for model 1. Here, we assume that all counts of trips $Y_{ij}(t)$ are independent and not only the counts of in- and outgoing trips. Moreover, the estimation routine of Model 2 needs considerably less computation time than Model 1. For both methods, the EM-algorithm algorithm needs less than 10 iterations until convergence. Our results show that estimating the most frequent routes traveled in the bike-sharing system is hard making use of station feeds only. However, even though we do not input any information on routes traveled, the estimated coefficient for distance is significantly negative, i.e. routes covering longer distances get lower weights.

Acknowledgments: Special Thanks to ZAMG (Vienna) for providing the weather data and to Michael Sedlmair (University of Stuttgart) and Michael Oppermann (UBC Vancouver) for providing the station feed data.

References

Alzaid, A.A. and Omair, M.A. (2010). *On The Poisson Difference Distribution Inference and Applications.* Bulletin of the Malaysian Mathematical Sciences Society, **8**(33), 17-45.

Chen, L., et al. (2017) *Understanding bike trip patterns leveraging bike sharing system open data.* Frontiers of Computer Science, **11**(1), 38-48

Fahrmeir, L. and Tutz, G (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer Science & Business Media.

Wood, S.N. (2006). *Generalized Additive Models.* Chapman and Hall/CRC

Calibration of extreme values of simulated and real data

S. Pereira[1], P. Pereira[2], M. de Carvalho[3], P. de Zea Bermudez[1]

[1] CEAUL and Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] CEAUL and ESTSetúbal, Instituto Politécnico de Setúbal, Portugal
[3] CEAUL and School of Mathematics, University of Edinburgh, UK

E-mail for correspondence: sapereira@fc.ul.pt

Abstract: In a variety of situations of applied interest there is a need for combining real data with simulated data (say, obtained from a climate model); yet the marginal features of both data may differ—either in the bulk or in the tail. This article devises a covariate-adjusted equipercentile calibration method that gets both data on the same scale, and that can be used for learning about how the differences between the distributions—of simulated and real data—may change along with a covariate. Another byproduct of this article is a regression method that simultaneously models the bulk and the (right) tail of the response—of either the simulated data or the real data. The methods are illustrated on numerical experiments and on a case study with rainfall data.

Keywords: Bulk and Tails; Calibration; Equipercentile Transformation; Extended Generalized Pareto distribution; Extremes.

1 Equipercentile calibration

We start by laying the groundwork. Let Y be the observed data and Z be the simulated data. We start by noting that a simple way to calibrate the simulated data is via an equipercentile transformation (González et al., 2015) as follows

$$Z_i^* = F_Y^{-1}(F_Z(Z_i)), \quad i = 1, \ldots, n. \tag{1}$$

Below, we will refer to the Z_i^* obtained via this approach as the *calibrated simulated data*. Note that the Z_i^* have the same distribution as the observed data, Y, as indeed

$$P(Z_i^* \leq z) = P\{F_Y^{-1}(F_Z(Z_i)) \leq z\} = P\{U_i \leq F_Y(z)\} = F_Y(z), \quad z > 0,$$

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

by noting that $U_i = F_Z(Z_i)$ follows a standard uniform distribution, for $i = 1, \ldots, n$. For this setup, we could resort to Naveau et al. (2016) so to calibrate the data as in (1), which would have the nice feature of being able to 'fairly' calibrate at all quantiles—including the extremes.

2 Covariate-adjusted calibration

Suppose now we would like to calibrate the simulated data conditionally on covariates $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$. We proceed in a similar way as in (1), which yields the following *covariate-adjusted calibrated data*

$$Z_i^* = F_Y^{-1}(F_Z(Z_i \mid \mathbf{x}_i) \mid \mathbf{x}_i), \quad i = 1, \ldots, n. \tag{2}$$

To model the conditional distributions in (2) we would extend the approach in Naveau et al. (2016) to the conditional setting. To ease notation, we will only introduce the model for $F_Y(y \mid \mathbf{x})$, which is given by

$$F_Y(y \mid \mathbf{x}) = G_{\mathbf{x}}\left(H_\xi\left(\frac{y}{\sigma}\right)\right), \tag{3}$$

where $\{G_{\mathbf{x}}\}$ is a family of functions indexed by a covariate, obeying assumptions A, B, and C in Naveau et al. (2016), and

$$H_\xi(y) = \begin{cases} 1 - (1 + \xi y)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y), & \xi = 0. \end{cases}$$

One of the main advantages of Naveau et al. (2016) is that one bypasses the step of threshold selection. Since threshold selection is an even more challenging issue in the conditional setting—as it entails looking for $\{u_{\mathbf{x}}\}$—that advantage becomes even more notorious in our setting.

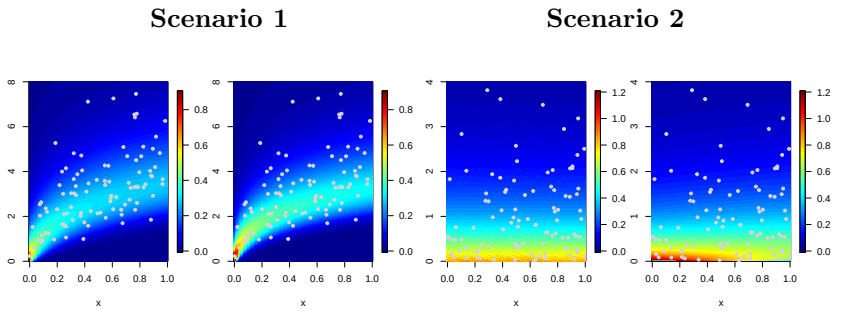


FIGURE 1. True conditional density (left) and estimated conditional density (right), for Scenarios 1 and 2; the gray points (•) represent the simulated data used to learn about the conditional density for each one-shot experiment.

An obvious consequence of Naveau et al. (2016, Eq. (7)), is that for $0 < p < 1$:

$$F_Y^{-1}(p \mid \mathbf{x}) = \begin{cases} \frac{\sigma}{\xi}[\{1 - G_{\mathbf{x}}^{-1}(p)\}^{-\xi} - 1], & \xi > 0, \\ -\sigma \log\{1 - G_{\mathbf{x}}^{-1}(p)\}, & \xi = 0. \end{cases} \tag{4}$$

From (4) we could then conduct the covariate-adjusted calibration as defined (2). In the implementations in Section 3 we focus the specification

$$G_x(u) = u^{\beta_0 + \beta_1 x}, \tag{5}$$

and resort to maximum likelihood estimation. In future implementations we aim to resort to a GAM (Generalized Additive Model) (Wood, 2006), based on the specification $G_{\mathbf{x}}(u) = u^{\kappa_{\mathbf{x}}} = u^{\beta_0 + \sum_{j=1}^{p} f_j(x_j)}$, with f_j denoting a smooth function corresponding the the jth covariate, for all j. An interesting aspect of (4) is that it is a simple model bridging *quantile regression* with *extremal quantile regression*. Quantile regression,

$$F_Y^{-1}(p \mid \mathbf{x}) = \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}(p), \quad 0 < p < 1,$$

is by now well understood (Koenker, 2005), but a limitation with its standard version is its inability to extrapolate into the tails of the conditional distribution. Extremal quantile regression are a class of models whose concern is precisely on modeling high quantiles, and which possess the ability to extrapolate into the tails of the conditional distribution (see e.g. Chernozhukov, 2005).

3 Experiments and case study

3.1 Numerical experiments

Key to our approach is the regression model in (3). Thus, to illustrate (3), under the specification in (5), we consider the following scenarios:

- **Scenario 1**: Simulation from a well-specificed setting, i.e., with data simulated according to (4). Specifically, we set $\sigma = 1$, $\xi = 0.1$, $\beta_0 = 1$ and $\beta_1 = 20$.
- **Scenario 2**: Simulation from a misspecificed setting, i.e., with data simulated according to (4) but allowing for $\sigma = \exp(\alpha_0 + \alpha_1 x)$. Specifically, we set $\beta_0 = 1$,$\beta_1 = 0$, $\xi = 0.1$, $\alpha_0 = 0.1$, and $\alpha_1 = 0.1$.

For comparing the estimated and the true conditional densities obtained in each scenario, 100 observations were simulated; the covariates were simulated from a standard uniform distribution. Figure 1 shows the true and the estimated conditional densities for Scenarios 1 and 2. As it can be seen from Figure 1 the method recovers satisfactorily well the true conditional density—especially keeping in mind that only 100 observations are simulated.

3.2 Illustration on real data

We now showcase how the method can be used in practice in a real data example. We consider real (y) and simulated (z) rainfall data from the county of Vizela (North of Portugal) from the year 2007; we focus only on $y > 0$. The data were gathered from Instituto Dom Luiz. The target is on assessing how real and simulated data compare over time (x).

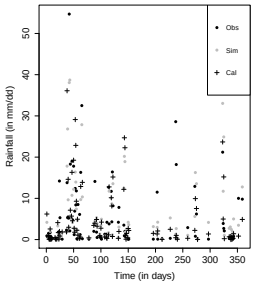


FIGURE 2. Observed (•), simulated (•), and calibrated (+) data.

Figure 2 depicts real and simulated data, along with calibrated data obtained according to (2).

Acknowledgments: The authors were financially supported by FCT—Fundação para a Ciência e a Tecnologia, Portugal—through the projects PTDC/MAT-STA/28649/2017 and UID/MAT/00006/2019.

References

Chernozhukov, V. (2005). Extremal quantile regression. *Annals of Statistics*, **33**, 806 – 839.

González, J., Barrientos, A. F., and Quintana, F. A. (2015). Bayesian nonparametric estimation of test equating functions with covariates. *Computational Statistics and Data Analysis*, **89**, 222 – 244.

Koenker, R. (2005). *Quantile regression*. Cambridge, MA: Cambridge University Press.

Naveau, P., Huser, R. , Ribereau, P., and Hannart, A. (2016). Modeling Jointly Low, Moderate, and Heavy Rainfall Intensities without a Threshold Selection. In: *Water Resources Research*, **52**, 2753 – 2769.

Wood, S. N. (2006). *Generalized additive models: An introduction with R.*. Boca Raton, FL: Chapman and Hall/CRC.

Mixtures of Generalized Nonlinear Models

Herwig Friedl[1], Sanela Omerovic[1]

[1] Graz University of Technology, Austria

E-mail for correspondence: hfriedl@tugraz.at

Abstract: The family of Mixtures of Generalized Nonlinear Models seems to be appropriate to provide predictions of the maximum gas consumption for extremely cold temperatures as they simultaneously face the problem of occurring heterogeneity arising from effects like sector-specific features (e.g. industrial or private consumer groups) or weekday-specific dependencies. The objective is to outline the statistical methods to enable the fitting of these models as well as to present a class of suitable applications.

Keywords: Finite Mixture Models; Generalized Nonlinear Models; EM Algorithm; Gas Consumption Data.

1 Introduction and Problem Specification

Finite mixture models (FMMs) represent a highly accommodative class of statistical models which gained strong interest in recent years. Due to their flexibility FMMs cover a large area of application. The particular group of mixtures of regression models has largely contributed to the gain in popularity of FMMs. This model class has been widely studied by *Bettina Grün* and *Friedrich Leisch* who developed the package `flexmix` in `R` for model-based clustering and mixtures of Generalized linear models (GLMs). As certain practical applications buttress the use of nonlinear regression functions the present work introduces the new model class of mixtures of Generalized nonlinear models (GNMs). It furthermore provides an efficient implementation of GNMs in `R` as an extension of the powerful package `flexmix`.

A suitable application of mixtures of GNMs is given by gas consumption data where gas suppliers agreed to model the load profile based on a sigmoid regression function. Figure 1 shows two examples of typical gas flow patterns (daily maximum gas flows displayed in dependence of the mean outside temperatures). The gas flow pattern exhibits in general a decreasing shape for increasing temperatures (sigmoidal structure) converging to a minimum consumption level. The present data samples outline specific

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

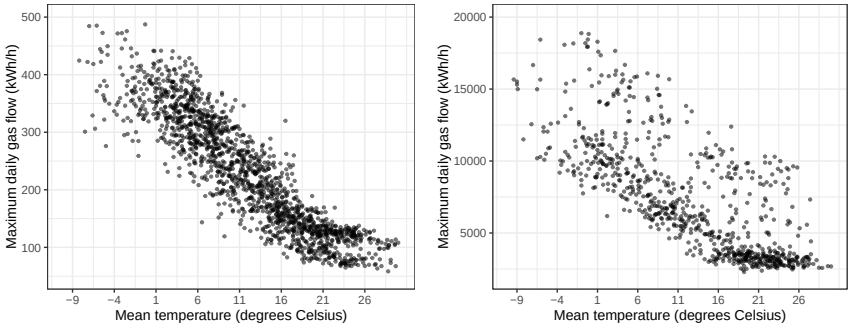


FIGURE 1. Two typical daily maximum gas consumption patterns depending on the average outside temperature

properties which can be addressed to heterogeneity due to latent classes. While the data set on the left shows two evident subgroups with different consumption levels, the second data set on the right side exhibits an increasing variability for low outside temperatures and shows two different minimum consumption levels. The aim of what follows is to present mixture models of GNMs as an appropriate statistical model to face such an occurring heterogeneity. The variability structure and the aim to model daily maxima will be taken into account by the use of gamma densities within mixtures of GNMs.

2 Mixtures of Generalized Nonlinear Models

Any K component mixture model can be marginally defined by the probability density function (pdf)

$$f(y_i|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f(y_i|\boldsymbol{\theta}_k), \qquad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ represents the observed values of the responses. Since we will model daily maxima we focus on mixtures of gamma pdf's. The component specific parameters are then $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ with $\theta_k = (\mu(\boldsymbol{\beta}_k), \phi_k)$ comprising the mean and dispersion in the k-th gamma component, with mixing weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. The present work takes up on the sigmoid function in Friedl et al. (2012) as nonlinear mean function in the k-th component, i.e.

$$E(y_i|k) = \mu_i(\boldsymbol{\beta}_k) = \beta_{k4} + \frac{\beta_{k1} - \beta_{k4}}{1 + \left(\frac{\beta_{k2}}{t_i - 40^o}\right)^{\beta_{k3}}}, \qquad i = 1, \ldots, n, \tag{2}$$

where t_i denotes the mean outside temperature. In order to model the variability structure, the conditional variance in the k-th component is assumed to be $Var(y_i|k) = \phi_k V(\mu_i(\boldsymbol{\beta}_k))$ with dispersion parameter ϕ_k and variance function $V(\mu_i(\boldsymbol{\beta}_k)) = \mu_i^2(\boldsymbol{\beta}_k)$.
Alternatively, define the indicator vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$ as

$$z_{ik} = \begin{cases} 1, & \text{if } y_i \text{ is from component } k, \\ 0, & \text{otherwise.} \end{cases}$$

Then the joint sample pdf can be written as

$$f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(y_i|\boldsymbol{\theta}_k)^{z_{ik}} \pi_k^{z_{ik}}.$$

In order to find the maximum likelihood estimates we apply the EM algorithm (Dempster et al., 1977) where the objective function to be iteratively maximised is given by

$$Q(\boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{\theta}^{(t)}, \boldsymbol{\pi}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(t)} \log(f(y_i|\boldsymbol{\theta}_k)\pi_k). \tag{3}$$

In a subsequent maximization step (M-step) the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are updated. In the t-th M-step the posterior probabilities

$$w_{ik}^{(t)} = \frac{f(y_i|\theta_k^{(t)})\pi_k^{(t)}}{\sum_{l=1}^{K} f(y_i|\theta_l^{(t)})\pi_l^{(t)}} \tag{4}$$

are considered to be fixed and maximization of (3) results in $\boldsymbol{\pi}^{(t+1)}$ and $\boldsymbol{\theta}^{(t+1)}$. This iterative process starts with some appropriate initial values $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\pi}^{(0)}$.

3 Marginal Confidence Intervals

An objective of this study is the accurate prediction of gas flow for low temperatures where the observations are typically sparse. The use of mixtures of GNMs enables the prediction of gas flow comprising individual differences in consumption levels within the identified components. The forecast of gas flow is therefore expressed by the general mean over the K-component mixture denoted as $\mu^M(\cdot)$ and evaluated at the component specific distribution parameters $\boldsymbol{\beta}_k$ and weighted by the prior probabilities π_k for $k = 1, \ldots, K$.
The predictions are subject to a specific level of uncertainty. In order to assess the variability of the mean predictions, the corresponding confidence

intervals are constructed by the use of the *Delta method*. Thus, the variance of the mean function $\mu^M(\hat{\boldsymbol{\beta}})$ can be approximated by its gradient and the variance-covariance matrix of the MLE $\hat{\boldsymbol{\beta}}$. The latter can be approximated by the following expression

$$Var(\mu^M(\hat{\boldsymbol{\beta}})) \approx \nabla(\mu^M(\hat{\boldsymbol{\beta}}))^\top Cov[\hat{\boldsymbol{\beta}}]\nabla(\mu^M(\hat{\boldsymbol{\beta}})),$$

where the gradient $\nabla(\mu^M(\hat{\boldsymbol{\beta}})) \in R^{KP}$ contains all the derivatives with respect to the parameters and is thus given by

$$\nabla(\mu^M(\boldsymbol{\beta})) = \left(\frac{\partial \mu^M(\boldsymbol{\beta})}{\partial \beta_{kp}}\right)_{k=1,\ldots,K;\ p=1,\ldots,P}.$$

The corresponding level $(1-\alpha)$ confidence interval for $\mu^M(\boldsymbol{\beta})$ can be derived as

$$(1-\alpha)\%\ CI(\mu^M(\boldsymbol{\beta})) \approx \left(\mu^M(\hat{\boldsymbol{\beta}}) \pm z_{1-\alpha/2} \cdot \sqrt{Var(\mu^M(\hat{\boldsymbol{\beta}}))}\right),$$

where $\mu^M(\hat{\boldsymbol{\beta}})$ corresponds to the predicted mean value of the maximum gas consumption given an average outside temperature x_i and z_α denotes the α quantile of the standard normal distribution.

4 Results

We now apply such a gamma mixture model to the real world gas flow data that have been already visualized in Figure 1. For this purpose the `R` package `flexmix` was extended for a new model class enabling the fitting of mixtures of GNMs. Figure 2 shows the original data together with the fitted nonlinear mean models for the application of two-component gamma mixture models.

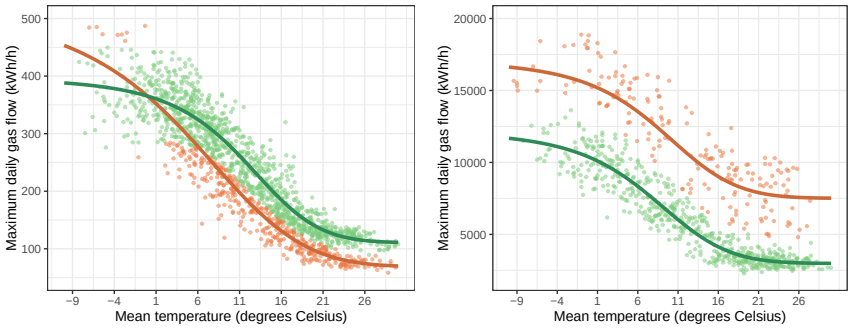


FIGURE 2. Fitted two-component gamma mixture models

The mean of each of both gamma components is modelled according to the sigmoidal function (2). A final component classification of all the responses is determined by their maximum posterior weights $\max_k w_{ik}$, $i = 1, \ldots, n$, as given in (4).
The first data set (left side) exhibits a dense structure whereas under the gamma mixture model two components with intersecting mean functions have been identified. The mixture model further succeeds to identify the two evident minimum consumption levels.
The second data set (right side) shows two well separated components with a band-like structure. Within the present data sample, as displayed in Figure 1, the gas flow attains temperatures up to a level of about -10 degrees Celsius (^{o}).
The fitted mean functions enable the prediction of the mean maximum gas flow for low temperatures, even below the observed temperatures. The predicted values for the temperatures -12^o, -14^o and -16^o are displayed in Table 1. In order to assess the variability of the predicted values, the respective 95% confidence intervals are also displayed as additional information.

TABLE 1. Predicted values and confidence intervals (in kWh/h)

	Temperatures in degrees Celsius (^{o})		
	-12^o	-14^o	-16^o
Data set 1	415 (396, 434)	419 (399, 439)	422 (400, 444)
Data set 2	13072 (12113, 14031)	13153 (12137, 14168)	13215 (12153, 14277)

5 Conclusions

Mixtures of GNMs prove as an adequate statistical model to incorporate heterogeneity due to latent classes whereas the application of GNMs enables the use of distributional shapes and models beyond the classic nonlinear regression model. The fitting can be also easily extended to $K > 2$ components whereas a direct comparison of different models is enabled through appropriate model selection criteria. Further extensions allow for modifications of the applied nonlinear mean function.

Acknowledgments: The authors are thankful to Bettina Grün for all the fruitful discussions and for sharing her knowledge regarding the implementation of the extension FlexMixNL within the `flexmix` package in `R`.

References

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

Friedl, H., Mirkov, R., and Steinkamp, A. (2012). Modelling and Forecasting Gas Flow on Exits of Gas Transmission Networks. *International Statistical Review*, **80**, 24 – 39.

Grün, B., and Leisch, F. (2007). Fitting Finite Mixtures of Generalized Linear Regressions in R. *Computational Statistics & Data Analysis*, **51**, 5247 – 5252.

Grün, B., and Leisch, F. (2008). FlexMix Version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, **28**, 1 – 35.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, **1**, 1 – 18.

Omerovic, S. (2019). Fitting Mixtures of Generalized Nonlinear Models. *Unpublished PhD Thesis*, Institute of Statistics, Graz University of Technology, Austria.

Neural Network Regression with an Application to Leukaemia Survival Data – An Unstructured Distributional Approach

Nadja Klein[1], Thorsten Simon[2] and Nikolaus Umlauf[2]

[1] Humboldt-Universität zu Berlin, Germany
[2] University of Innsbruck, Austria

E-mail for correspondence: nadja.klein@hu-berlin.de

Abstract: During the last decades there has been an increasing interest in distributional regression models that allow to model the entire data distribution conditional on covariates. In particular, the framework of structured additive distributional regression models enables to specify different types of effects such as linear, nonlinear or interaction effects on all the distribution parameters hence providing a very flexible and generic framework suited for many complex real data problems. However, when it comes to the question of variable selection, establishing a reasonable and 'good' distributional model is difficult in practice. In addition, the exact functional forms and possible interactions are often hard to fix in advance even with advanced expert knowledge. To overcome this drawback, we propose an extension of the structured additive regression predictors by a feed-forward neural network that allows to learn the functional forms and potential complex interactions of dependent variables from the data within the algorithm. We propose an efficient implementation that allows for sparsity through the elastic net. In an application on leukaemia survival data we show that the novel unstructured approach clearly outperforms a number of benchmark models.

1 Introduction

Semiparametric regression models offer considerable flexibility concerning the specification of additive regression predictors including effects as diverse as nonlinear effects of continuous covariates, spatial effects, random effects, or varying coefficients. Recently, such flexible model predictors have been combined with the possibility to go beyond pure mean-based analyses by specifying regression predictors on potentially all parameters of the response distribution in a distributional regression framework (Klein et al.,

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

2015). In these models, one assumes

$$y \sim \mathcal{D}\left(h_1(\theta_1) = \eta_1,\ h_2(\theta_2) = \eta_2, \ldots,\ h_K(\theta_K) = \eta_K\right),$$

where $\mathcal{D}$ denotes a parametric distribution for y with K parameters θ_k, $k = 1, \ldots, K$, that are linked to additive predictors using monotonic one-to-one transformations $h_k(\cdot)$. The k-th additive predictor is given by $\eta_k = \eta_k(\mathbf{x}; \boldsymbol{\beta}_k) = f_{1k}(\mathbf{x}; \boldsymbol{\beta}_{1k}) + \ldots + f_{J_k k}(\mathbf{x}; \boldsymbol{\beta}_{J_k k})$, Several approaches have been developed to allow for the inclusion of general tensor product interactions within the structured additive framework. However, one drawback of all these models is that the predictors need to be specified by the user which can be difficult since in most situations higher dimensional nonlinear interactions between covariates are hard to identify a priori and comparison of all potential model specifications is computationally infeasible.

2 Cox Model for Leukaemia Survival Data

For instance, for the analysis of leukaemia survival data of $n = 1,043$ patients in a study reported by the North West Leukaemia Register in the United Kingdom, the hazard of an event (status dead) at time t can be described with a relative additive risk model of the form:

$$\lambda(t) = \exp\left(\eta(t)\right) = \exp\left(\eta_\lambda(t) + \eta_\gamma\right),$$

i.e., a model for the instantaneous event risk conditional on being alive before time t. The probability to not survive after time t is

$$S(t) = Prob(T > t) = \exp\left(-\int_0^t \lambda(u)du\right).$$

Here, the hazard function is assumed to depend on a time-varying predictor $\eta_\lambda(t)$ and a time-constant predictor η_γ. In most survival models, the time-varying part $\eta_\lambda(t)$ represents the so-called baseline hazard and is a univariate function of time t. However, Henderson et al. (2002), who analysed this data before found that considerable between patient heterogeneity remains conditional on treatment and known prognostic factors despite effective therapies. Part of this heterogeneity may be linked to spatial effects but also complex and nonlinear interactions of demographic or clinical variables across time and space.

3 Unstructured Neural Network Predictors

To address these challenges we consider feedforward neural networks (FNN), which are extensively used in regression and classification applications in machine learning. The general idea is to use a FNN model term $f_{jk}(\mathbf{X}_{jk}; \boldsymbol{\beta}_{jk})$

additional to all other effects in a typical structured additive predictor η_k. A FNN model term has the simple structure $f_{jk}(\mathbf{X}_{jk}; \boldsymbol{\beta}_{jk}) = \boldsymbol{X}_{jk}\boldsymbol{\beta}_{jk}$, where the columns of $\mathbf{X}_{jk}$ are a composition of activation functions, e.g., using the sigmoid the l-th column (node) is

$$h_l(\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w}_l^\top \mathbf{x} + b_l))},$$

where $\mathbf{w}_l$ and b_l are inner weights and biases. However, estimation is usually difficult due to the high dimensionality of these parameters. To render inference in a distributional model feasible we follow Dudek (2017) and randomly sample the weights and biases such that the most nonlinear and steepest parts of the activation functions are inside the data region. In addition, to obtain sparsity and to avoid overfitting, we use elastic net regularization

$$\lambda_{jk1} \cdot J_{\mathrm{L}}(\boldsymbol{\beta}_{jk}) + \lambda_{jk2} \cdot J_{\mathrm{R}}(\boldsymbol{\beta}_{jk}),$$

%mbox with ridge penalties J_{R} and quadratic approximations of the lasso penalties J_{L}.

4 Model Specification and Results

For the leukaemia survival example, we use the following additive predictors

$$\begin{aligned}
\eta_\lambda &= f_1(\texttt{time}) + f_2(\texttt{time}, \texttt{sex}, \texttt{age}, \texttt{wbc}, \texttt{tpi}, \texttt{xcoord}, \texttt{ycoord}) \\
\eta_\gamma &= \beta_0 + \texttt{sex} + f_3(\texttt{age}) + f_4(\texttt{wbc}) + f_5(\texttt{tpi}) + \\
&\quad f_6(\texttt{xcoord}, \texttt{ycoord}) + f_7(\texttt{sex}, \texttt{age}, \texttt{wbc}, \texttt{tpi}, \texttt{xcoord}, \texttt{ycoord}).
\end{aligned}$$

Here, functions $f_2(\cdot)$ and $f_7(\cdot)$ represent a time dependent and a time constant neural network model term, respectively. A description of the variables can be found in Table 1.

TABLE 1. Variable description in the Leukaemia data set.

Variable	Description.
`time`	Survival time in days.
`cens`	Right censoring status 0=censored, 1=dead.
`xcoord`	Coordinates in x-axis of residence.
`ycoord`	Coordinates in y-axis of residence.
`age`	Age in years.
`sex`	male=1 female=0.
`wbc`	White blood cell count at diagnosis, truncated at 500.
`tpi`	The Townsend score for which higher values indicates less affluent areas.
`district`	Administrative district of residence.

We evaluate the performance of the neural network Cox model (GAM+Net) by randomly sampling 100 individuals that serve as a hold out sample and compare it with a pure network Cox model (Net), a random forest

and a Cox model without the network terms (GAM) using the Brier score (Figure 1). This is done 50 times and our proposed model GAM+Net clearly outperforms the three competitors. Finally, the predicted probabilities to

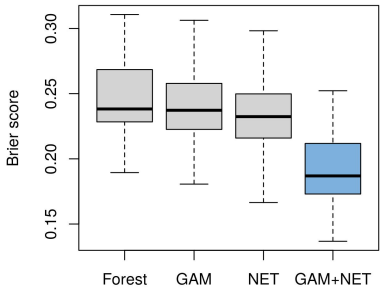


FIGURE 1. Out-of-sample Brier score.

not survive t for males (dashed) and females (solid) in two metropolitan areas Blackpool (blue) and Manchester (yellow) for the GAM model and the GAM+Net model (right) are shown in Figure 2 and are much more flexible than the ones of a GAM model.

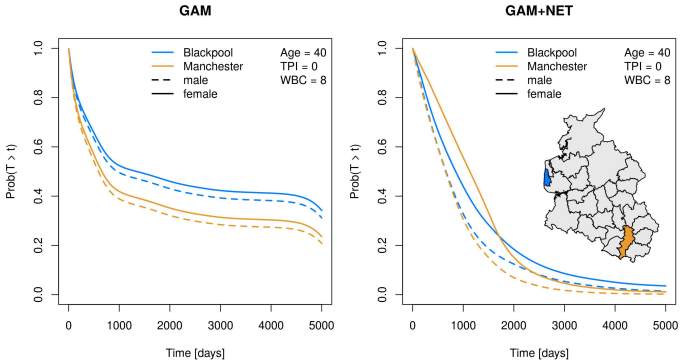


FIGURE 2. Probability to not survive after time t.

References

Dudek G. (2017). A method of generating random weights and biases in feedforward neural networks with random hidden nodes, arXiv:1710.04874.

Henderson R., Shimakura S. & and Gorst D. (2002). Modeling Spatial Variation in Leukemia Survival Data. *JASA* **70**(460).

Klein, N., Kneib, T., Lang, S. & Sohn, A. (2015). Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany. *AOAS*, **9**, 1024-1052.

Degradation Models in Reliability Analysis

Chien-Yu Peng[1]

[1] Institute of Statistical Science, Academia Sinica, Taiwan, Republic of China

E-mail for correspondence: chienyu@stat.sinica.edu.tw

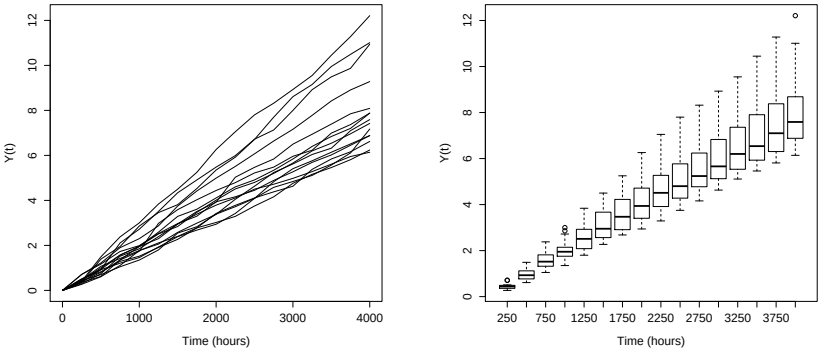
Abstract: Degradation data are widely analyzed using stochastic processes to assess the lifetime information of highly reliable products. In this article, we first review several stochastic degradation-based processes in the literature and then propose a general stochastic degradation-based process. This model is statistically plausible and demonstrates substantially improved fit when applied to real data. We give a consistent interpretation between physical/chemical mechanisms and statistical explanations. In addition, we provide a simple model-checking procedure to evaluate the appropriateness of the model assumptions. Several case studies are performed to demonstrate the flexibility and applicability of the proposed model with random effects and explanatory variables.

Keywords: First passage time; gamma process; inverse Gaussian process; Mean time to failure; Wiener process.

1 Background

High-quality products are frequently designed with high reliability and developed in a relatively short period of time. Manufacturers must achieve product reliability quickly and efficiently within a limited time for internal reliability tests. One problem with traditional life tests is the lack of sufficient time-to-failure data to effectively make inferences about a product's lifetime. Under this situation, if there are quality characteristics related to the degradation of physical characteristics over time, which are related to product reliability, an alternative option is to use sufficient degradation data to accurately estimate the product's lifetime distribution. General references for degradation models are included in Nelson (1990), Meeker and Escobar (1998), Bagdonavičius and Nikulin (2001) and the references therein. Other important applications of degradation models are in areas such as engineering, economics, environmental modeling, food and drugs.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.



a. Degradation paths b. Box plots

FIGURE 1. Laser data

2 A motivating example

We consider the laser data of an experiment described by Meeker and Escobar (1998, example 13.5) as a motivating example. The QC of a laser device is its operating current. To maintain nearly constant light output, the laser device contains a feedback mechanism for increasing the operating current when its light intensity degrades. Figure 1(a) and 1(b) display the degradation paths of the operating current over 4000 hours for 15 tested units and the box plot at each measurement time with mean line, respectively. Current values are recorded every 250 hours. When the operating current reaches a predefined threshold level $\omega = 10$, the device is considered to have failed. The primary objective of this experiment is to assess the lifetime information for lasers, such as the mean-time-to-failure ($MTTF$) or the qth quantile of the time-to-failure distribution. The accuracy and precision of the product's lifetime estimation mainly depends on modeling the degradation paths. For laser devices, the degradation (i.e., operating current) is considered the additive accumulation of damages caused by the feedback mechanism. Cumulative damage can be approximated as a stochastic process. See Singpurwalla (1995) and Bagdonavičius and Nikulin (2001) for more details. This approximation presents a physical interpretation of the stochastic processes and is applicable to address realistic problems.

3 Related literature

Generally speaking, non-monotonic and monotonic degradation paths are two well-known characteristics in the degradation data. The Wiener degradation-based process (or Gaussian process with specific

covariance structure) is used to describe a non-monotonic degradation path. For instance, Whitmore (1995) proposed a Wiener diffusion process, subject to measurement error, to model the declining gain of a transistor. Doksum and Normand (1995) presented two Wiener degradation-based processes to connect biomarker processes, event times and covariates of interest. Tseng and Peng (2004) described the light intensity of LED lamps of contact image scanners by using an integrated Wiener process. Peng and Tseng (2009) proposed a linear degradation model in which the unit-to-unit variation of all test units can be examined simultaneously with the time-dependent structure in degradation paths (see Cheng and Peng, 2012; Peng and Cheng, 2016).

In numerous applications, the gamma and inverse Gaussian (IG) processes are widely used when the degradation path is strictly increasing. Bagdonavičius and Nikulin (2000) constructed a degradation model by using a gamma process with time-dependent explanatory variables. Lawless and Crowder (2004) used a gamma degradation-based process that incorporates random effects on crack growth data. When neither the Wiener nor the gamma degradation-based processes adequately fit strictly monotonic degradation paths (see Wang and Xu, 2010), the IG process is an alternative degradation model that can be used to represent the strictly monotonic degradation paths. Peng (2015) proposed an IG degradation-based process with inverse normal-gamma random effects and derived the corresponding lifetime distribution and its properties.

4 Overview

In this work, we review several stochastic degradation-based processes and propose a general degradation-based process general as a new degradation model that is simple, flexible, and easily applied. The model parameters can vary from unit to unit by using random effects for degradation data. Some properties of the product's lifetime distribution are discussed based on the proposed degradation model. A Monte Carlo simulation study is conducted to demonstrate the performance of the estimation algorithm and the adequacy of the bootstrap procedure for constructing the confidence interval of a product's lifetime. Furthermore, we use a model selection criterion and provide a simple model-checking procedure to assess the validity of the proposed stochastic processes. Several case applications are used to illustrate the proposed degradation model with random effects and time-independent explanatory variables.

Acknowledgments: This work was supported by the Ministry of Science and Technology (MOST-106-2118-M-001-013-MY3) and Academia Sinica (AS-CDA-107-M09) of Taiwan, Republic of China.

References

Bagdonavičius, V. and Nikulin, M.S. (2000). Estimation in Degradation Models With Explanatory Variables. *Lifetime Data Analysis*, **7**, 85 – 103.

Bagdonavičius, V. and Nikulin, M.S. (2001). *Accelerated Life Models: Modeling and Statistical Analysis*. Boca Raton: Chapman & Hall/CRC.

Cheng, Y. S. and Peng, C. Y. (2012). Integrated Degradation Models in R Using iDEMO. *Journal of Statistical Software*, **49**, 1 – 22.

Doksum, K.A. and Normand, S.L.T. (1995). Gaussian Models for Degradation Processes—Part I: Methods for the Analysis of Biomarker Data. *Lifetime Data Analysis*, **1**, 131 – 144.

Lawless, J. and Crowder, M. (2004). Covariates and Random Effects in a Gamma Process Model With Application to Degradation and Failure. *Lifetime Data Analysis*, **10**, 213 – 227.

Meeker, W.Q. and Escobar, L.A. (1998). *Statistical Methods for Reliability Data*. New York: John Wiley & Sons.

Nelson, W. (1990). *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*. New York: John Wiley & Sons.

Peng, C.Y. (2015). Inverse Gaussian Processes With Random Effects and Explanatory Variables for Degradation Data. *Technometrics*, **57**, 100 – 111.

Peng, C. Y. and Cheng, Y. S. (2016). *Computational Network Analysis With R: Applications in Biology, Medicine and Chemistry*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, 83 – 124.

Peng, C.Y. and Tseng, S.T. (2009). Misspecification Analysis of Linear Degradation Models. *IEEE Transactions on Reliability*, **58**, 444 – 455.

Singpurwalla, N.D. (1995). Survival in Dynamic Environments. *Statistical Science*, **10**, 86 – 103.

Tseng, S.T. and Peng, C.Y. (2004). Optimal Burn-in Policy by Using an Integrated Wiener Process. *IIE Transactions*, **36**, 1161 – 1170.

Wang, X. and Xu, D. (2010). An Inverse Gaussian Process Model for Degradation Data. *Technometrics*, **52**, 188 – 197.

Whitmore, G.A. (1995). Estimating Degradation by a Wiener Diffusion Process Subject to Measurement Error. *Lifetime Data Analysis*, **1**, 307 – 319.

A longitudinal continuous time hidden Markov model on serum biomarkers for the early detection of hepatocellular carcinoma

Ruben Amoros[1], Ruth King[1], Hidenori Toyoda[2], Takashi Kumada[2], Philip J Johnson[3], Thomas G Bird[45]

[1] School of Mathematics, University of Edinburgh, United Kingdom
[2] Department of Gastroenterology, Ogaki Municipal Hospital, Japan
[3] Institute of Translational Medicine, University of Liverpool, United Kingdom
[4] Cancer Research UK Beatson Institute, United Kingdom
[5] MRC Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, United Kingdom

E-mail for correspondence: Ruben.Amoros@ed.ac.uk

Abstract: Early detection of hepatocellular carcinoma (HCC) is essential for successful treatment. The use of serum biomarkers or the combination of several biomarkers, age and sex in the so-called GALAD score, have been proposed to detect the presence of tumours. Previous static cut-off levels have been shown to be inefficient in detecting HCC due in part to the individual baseline heterogeneity, but an exploratory study suggests that analysing biomarker levels over time is a promising avenue for detecting the development of HCC. In this work we propose a Bayesian longitudinal hierarchical model for GALAD scores of patients under HCC screening to identify changes in the trend of this score indicating the development of HCC. The hidden states correspond to the absence or presence of HCC at the given time, with the later being an absorbent state. The model is additionally informed by the the diagnosis by standard clinical practice. We apply the proposed model to a Japanese cohort database of patients under HCC surveillance and show that the detection capability of this proposal is greater that using a fixed cut-off point on the GALAD score.

Keywords: Change-point model; Cancer detection; Hidden Markov model; Longitudinal.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

1 Introduction

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer in adults which kills more than 700,000 globally per year, and early detection is essential for successful treatment (Bray et al., 2018). The use of serum biomarkers, such as alpha-fetoprotein (AFP) or the combination of several biomarkers, age and sex in the so-called GALAD score (Johnson et al., 2014), have been proposed to detect the presence of tumours, with an increasing level indicative of potential cancer present.
Previous static cut-off levels have been shown to be inefficient in detecting HCC due in part to the individual baseline heterogeneity, but an exploratory study suggests that analysing biomarker levels over time is a promising avenue for detecting the development of HCC (Bird et al., 2016). In this work we propose a Bayesian hierarchical model for longitudinal GALAD scores of patients under HCC screening to identify changes in the trend of the GALAD score, that can be indicative of the development of HCC.

2 The model

For each patient $i = 1, \ldots, I$ and observation time $j = 1, \ldots, J_i$, the GALAD score B_{ij} is considered to follow a Gaussian distribution,

$$B_{ij}|\mu_{ij}, \sigma^2 \sim N(\mu_{ij}, \sigma^2)\,,$$

where the mean μ_{ij} is described as a general population mean ν plus a random personal baseline for each patient b_i. If a patient has developed cancer, this mean is linearly increased with a slope β according to the time elapsed since the development of the cancer τ_i.

$$\mu_{ij} = \nu + b_i + C_{ij} S_i \beta (t_{ij} - \tau_i)\,,$$

$$b_i|\sigma_b^2 \sim N(0, \sigma_b^2)\,,$$

with t_{ij} being the time of the observation j of patient i, S_i indicating whether patient i is susceptible of a trend change when developing HCC and C_{ij} a latent indicator variable that takes value 1 if the patient has developed HCC at time t_{ij} and 0 otherwise. This variable C_{ij} is modelled through an absorbent continuous time hidden Markov model, with transition matrix

$$\Gamma_{ij} = \begin{bmatrix} e^{-\lambda_i \Delta t_{ij+1}} & 1 - e^{-\lambda_i \Delta t_{ij+1}} \\ 0 & 1 \end{bmatrix},$$

where γ_{kl} is the probability of transitioning from state k in observation j to state l in the next observation $j + 1$. This part of the model takes the

form of a survival model with instant hazard function λ_i for the time until the development of HCC, with the peculiarity that this event is a hidden variable in our model. This hazard is a function of the patient baseline, so that patients with higher GALAD scores over time have higher probability of developing HCC,

$$\log(\lambda_i) = \zeta + \xi b_i \,,$$

with ζ and ξ to be estimated. The variables S_i are modelled through a Bernoulli distribution with unknown probability p_S. The prior distribution for the time of HCC development τ_i is defined to be uniform between the last observation without cancer $C_{ij-1} = 0$ and the first observation with cancer $C_{ij} = 1$ for the patient i.

Diagnose by standard clinical procedures is only possible when the tumor has grown to a certain size. Therefore, the probability of a patient being diagnosed by standard clinical procedures can be modelled in the fashion of a survival analysis by means of the instant hazard of being diagnosed δ and the time elapsed since the development of HCC $(t_{ij} - \tau_i)$. This informs the value of the variables C_{ij} taking into account possible false negatives of the diagnose by standard procedures. The variable od this diagnose, D_{ij}, can therefore be modelled with a Bernoulli distribution,

$$D_{ij}|q_{C_{ij},ij} \sim Bernoulli(q_{C_{ij},ij})$$

with probability of diagnose $q_{1,ij} = 1 - e^{-\delta(t_{ij}-\tau_i)}$ if the patient has HCC and 0 otherwise. Vague prior distributions are set for the parameters of the model.

3 Results and discussion

We applied our model on the dataset provided by the Ogaki Municipal Hospital, Japan, comprising of individual longitudinal data on the GALAD score and the absence or presence of a clinical diagnosis and collected at irregular times from patients with cirrhosis being screened for HCC. In total, data from 35001 observations of 2272 patients were available between the years 2009-2015. The model parameters were estimated using a training dataset comprising 75% of the patients of the original dataset. The posterior distributions of the parameters were used as priors for the prediction of the HCC states (values of the latent variables C_{ij}) of the rest of the patients, ignoring the information about clinical diagnosis.

The detection performance of our model was compared against the use of a threshold over the GALAD score (Johnson et al., 2014) using receiving operating characteristic (ROC) curves. Due to the longitudinal nature of the data, two different specificities (per patient and per observation) were considered. With comparable behavior for patient specificity, our model showed to outperform the use of a threshold with regards to the specificity

per observation, as seen in Figure 1. These results show the importance of modeling the longitudinal nature of the data, fitting the characteristic gradual increase of the GALAD score that occurs from the apparition of the tumour. Further studies considering the individual contribution of each of the biomarkers in a multivariate model will be done in the future to improve the detection power of the proposed method.

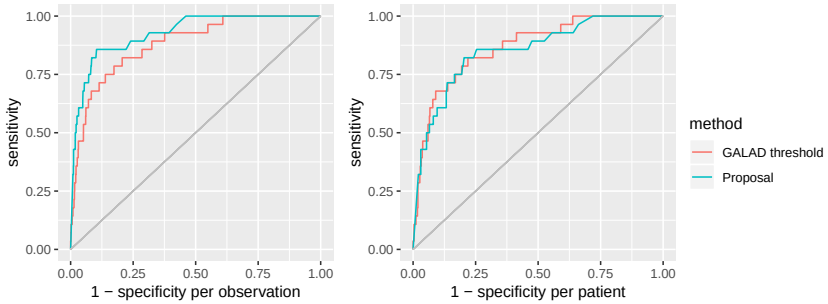


FIGURE 1. ROC curves using specificity per observation and per patient for the proposal (blue) and the threshold on the GALAD (red).

Acknowledgments: RA is supported by a Chief Scientist's Office Catalyst Award (CGA/17/19) and a Scottish Liver Transplant Unit Endowment Award. TGB is supported by the Wellcome Trust (107492/Z).

References

Bird, T.G., Dimitropoulou, P., Turner, R.M., Jenks, S.J., Cusack, P., Hey, S., Blunsum, A.,Kelly, S., Sturgeon, C., Hayes, P.C., Bird, S.M. (2016). *Alpha-Fetoprotein detection of hepatocellular carcinoma leads to a standardized analysis of dynamic AFP to improve screening based detection.* PLOS One **11**(6), e0156801 (2016).

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A. (2018). *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: A Cancer Journal for Clinicians **6**(68), 394 — 424.

Johnson, P.J., Pirrie, S.J., Cox, T.F., Berhane, S., Teng, M., Palmer, D., Morse, J., Hull,D., Patman, G., Kagebayashi, C., Hussain, S., Graham, J., Reeves, H., Satomura, S. (2014) *The detection of hepatocellular carcinoma using a prospectively developed and validated model based on serological biomarkers.* Cancer Epidemiology Biomarkers and Prevention **23**(1), 144 — 153.

A continuous-time capture-recapture model for annual movements of bottlenose dolphins

Sina Mews[1], Roland Langrock[1], Ruth King[2], Julia Schemm[1], Irina Janzen[1], Nicola Quick[3]

[1] Bielefeld University, Germany
[2] University of Edinburgh, UK
[3] Duke University Marine Lab, United States

E-mail for correspondence: sina.mews@uni-bielefeld.de

Abstract: Our modelling approach is motivated by individual sighting histories of bottlenose dolphins off the east coast of Scotland. The main objective here is to model the annual movement patterns of the dolphin population between different sites, as these migrations can be of conservation importance with regard to ongoing offshore development. Due to the irregularity of the capture occasions at hand, we formulate a capture-recapture model in continuous time and develop an approximate maximum likelihood approach for estimating the effect of time-varying covariates, which in the given example correspond to seasonal effects. While motivated by a particular data set and the associated conservation management problem, our modelling framework is much more generally applicable to irregularly sampled capture-recapture data subject to switches in underlying states.

Keywords: Capture-recapture; Continuous-time model; Maximum likelihood; Multi-state model; Time-varying covariates.

1 Introduction

Capture-recapture data consist of individual animals' sighting histories. When animals can be observed in different "states" (i.e. classes corresponding to location, behaviour, physiology, etc.), multi-state capture-recapture models allow for inference regarding the transitions between these states, but also to investigate potential differences in survival and detection probabilities across states. Typically, the Arnason-Schwarz model is fitted to such multi-state capture-recapture data, assuming a first-order Markov chain in

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

discrete time for the state process (Schwarz et al., 1993). However, in cases where the capture occasions are not regularly spaced in time, standard multi-state capture-recapture methods are not readily applicable.
Motivated by sighting histories of bottlenose dolphins off the east coast of Scotland, which do not follow a regular sampling protocol, we develop a continuous-time formulation of the Arnason-Schwarz model. Due to ongoing offshore development, conservation managers seek to assess the temporal movement patterns of the dolphin population between different sites, which constitute the states in our model (in addition to alive and dead). To investigate how these movement patterns depend on the time of year, we develop an approximate maximum likelihood approach for estimating the effect of time-varying covariates on the state transition rates.

2 Methodology

2.1 Basic model formulation

The capture-recapture setting can be regarded as a special case of a (partially) hidden Markov model (HMM), with the observed capture history of an individual as the state-dependent process and an underlying, partially observed state process, e.g. related to the movement of the individual between different sites as in our motivating example. Let n denote the total number of individuals observed, T the total number of survey occasions and $\mathcal{M} = \{1, \ldots, M\}$ the set of possible states while alive. Then for each individual $i = 1, \ldots, n$, at capture times $t = t_0, t_1, \ldots, t_T$, where $0 = t_0 < t_1 < \ldots < t_T$, the observed event is given by

$$x_{i,t} = \begin{cases} 0 & \text{if individual } i \text{ is not observed at time } t; \\ m & \text{if individual } i \text{ is observed in state } m \text{ at time } t, \end{cases}$$

and the true state by

$$s_{i,t} = \begin{cases} m & \text{if individual } i \text{ is alive and in state } m \text{ at time } t; \\ M+1 & \text{if individual } i \text{ is (presumed) dead at time } t. \end{cases}$$

For convenience we will drop the subscript i from now on, but will continue to refer to the individual.
The conditional probabilities of recapture during a survey, given that the area is searched (indicated by the dichotomous variable $a_m = 1$), are denoted by $p_m = \Pr(x_t = m | s_t = m, a_m = 1)$. The parameters p_m thus are the state-specific detection (or recapture) probabilities. We assume these to be constant over time, but this assumption can easily be relaxed. On survey occasions where one of the sites is not visited, recapture within that area is not possible (i.e. $p_m = 0$ if $a_m = 0$), and hence the probability of not observing an animal is one (i.e. $\Pr(x_t = 0 | a_m = 0) = 1$).

Due to the temporal irregularity of the survey occasions, we use a continuous-time Markov chain to model the state process $s_{t_0}, \ldots, s_{t_T}$, which in our example corresponds to a dolphin's location – one of M sites – at the time of the survey occasions. The transitioning between the different states is then governed by an underlying transition intensity matrix,

$$\mathbf{Q} = \begin{pmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,M} & q_{1,M+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ q_{M,1} & q_{M,2} & \cdots & q_{M,M} & q_{M,M+1} \\ 0 & 0 & \ldots & 0 & 0 \end{pmatrix},$$

where the state transition intensity $q_{j,k}$ describes the instantaneous probability to switch from state j to state $k \neq j$. Due to the constraints that $q_{j,k} \geq 0$ for $j \neq k$ and $\sum_{k=1}^{M+1} q_{j,k} = 0$, the diagonal entries are obtained as $q_{j,j} = -\sum_{k \neq j} q_{j,k}$. The last row in $\mathbf{Q}$ consists of zeros only because we assume the last state, corresponding to an individual's presumed death, to be an absorbing state.

Given a time-homogeneous intensity matrix $\mathbf{Q}$, the transition probability matrix (t.p.m.) $\mathbf{\Gamma}$ for a time interval between two consecutive capture occasions $[t_{u-1}, t_u]$, $u = 1, \ldots, T$, is obtained as a matrix exponential,

$$\mathbf{\Gamma}(t_{u-1}, t_u) = \exp\big(\mathbf{Q} \cdot (t_u - t_{u-1})\big) = \sum_{d=0}^{\infty} \mathbf{Q}^d (t_u - t_{u-1})^d / d!. \qquad (1)$$

The entries $\gamma_{j,k}(t_{u-1}, t_u)$ indicate the probability to move from state j at capture occasion t_{u-1} to state k at the next capture occasion t_u.

2.2 Likelihood evaluation

Since we formulate the capture-recapture model within a continuous-time HMM framework, we follow Jackson et al. (2003) in exploiting the convenient and efficient HMM-based forward algorithm for evaluating the likelihood. This yields the matrix product

$$\mathcal{L} = \boldsymbol{\pi}_{t_0} \Big(\prod_{u=1}^{T} \underbrace{\exp\big(\mathbf{Q} \cdot (t_u - t_{u-1})\big)}_{=\mathbf{\Gamma}(t_{u-1}, t_u)} \mathbf{P}(x_{t_u} | a_m) \Big) \mathbf{1}, \qquad (2)$$

where $\boldsymbol{\pi}_{t_0}$ is a row vector indicating the state at first capture, and $\mathbf{1} \in \mathbb{R}^{M+1}$ is a column vector of ones. Furthermore, $\mathbf{P}(x_t | a_m)$ is a diagonal matrix of dimension $\mathbb{R}^{(M+1)\times(M+1)}$ containing the elements $\Pr(x_t | s_t, a_m)$, $s_t = 1, \ldots, M+1$, with a_m indicating whether at time t area m was searched or not.

Assuming independence of the encounter histories, the likelihood over multiple capture histories is simply calculated as the product of the individual likelihoods $\mathcal{L}$ given in (2). The model parameters, namely the state transition intensities as well as the detection probabilities, are then estimated by numerically maximising the joint likelihood.

2.3 Incorporating time-varying covariates

In general, and in particular in our motivating example, the state transition intensities may depend on some time-varying covariate $z(t)$, e.g. such that $q_{j,k}(t) = \exp\big(\alpha_{jk0} + \alpha_{jk1} z(t)\big)$. However, incorporating such covariates into the continuous-time state process is rather challenging: Equation (1) then does not hold anymore and the likelihood function becomes intractable. An important exception is the case where the covariate of interest and hence also the intensities are piecewise constant over time (Langrock et al., 2013). We thus partition the time interval during which observations were made, $[0, t_T]$, into R intervals, $\tau_1, \ldots, \tau_R$, with $\tau_r = [b_{r-1}, b_r)$ and $b_0 = 0$, $b_R = t_T$, on which the (potentially continuously varying) transition intensities are approximated by a constant function. This approximation leads to a simple closed-form expression of the likelihood, without the need to evaluate integrals. Specifically, for piecewise constant transition intensities, we obtain the t.p.m. $\mathbf{\Gamma}(t_{u-1}, t_u)$ within (2) recursively as a product of t.p.m.s over which the intensities are constant (a consequence of the Chapman-Kolmogorov equation). For example, given $t_1 \in \tau_1$, $t_2 \in \tau_2$, it follows that $\mathbf{\Gamma}(t_1, t_2) = \exp\big(\mathbf{Q}_1 \cdot (b_1 - t_1)\big) \exp\big(\mathbf{Q}_2 \cdot (t_2 - b_1)\big)$. The approximation of the time-varying intensities by step functions thus allows us to estimate the parameters by numerically maximising a likelihood similar to (2), which is an approximation of the likelihood of the actual model of interest. Crucially, the approximation can be made arbitrarily accurate by decreasing the width of the intervals.

3 Annual movement of bottlenose dolphins

Regarding our motivating data, we are interested in the annual movements of bottlenose dolphins between two sites, namely the Moray Firth Special Area of Conservation (SAC) and Tayside and Fife (T&F). Our data set comprises $n = 322$ individual capture histories of dolphins that have been sighted at least six times between the years 1990 and 2015. Figure 1 illustrates such a capture history for one dolphin: The boat trips in both areas (indicated by waves) are irregularly spaced in time. At some of these capture occasions, the individual is observed (being identified based on natural marks), but at most occasions, the whereabouts of the dolphin remain unknown. Using our developed modelling approach, however, we can still make inference on the unobserved movement between the sites.

The dolphins' states, here with $M = 2$, correspond to their actual location, i.e. either SAC or T&F. Our covariate of interest is time of year. For the likelihood approximation using step functions, we partition the observation period into intervals with a length of 30 days each, here leading to $R = 326$, which provides a good balance between approximation accuracy and computational cost.

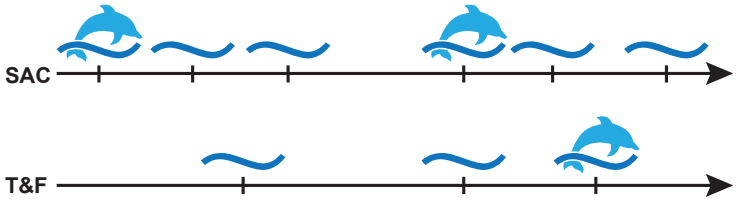


FIGURE 1. Illustration of a dolphin's capture history.

Denoting the midpoint (or centre) of each interval by $c_1, \ldots, c_R$, we then have

$$q_{j,k}(t) = \exp\left(\beta_{jk0} + \beta_{jk1}\sin\left(\frac{2\pi c_r}{365}\right) + \beta_{jk2}\cos\left(\frac{2\pi c_r}{365}\right)\right) \text{ for } t \in \tau_r.$$

The estimation results in Figure 2 reveal clear seasonal patterns with high intensities to move from T&F to SAC in summer, whereas intensities to move (back) to T&F are highest in autumn. The quantification of these migration patterns is of biological interest, and may also help to inform conservation management.

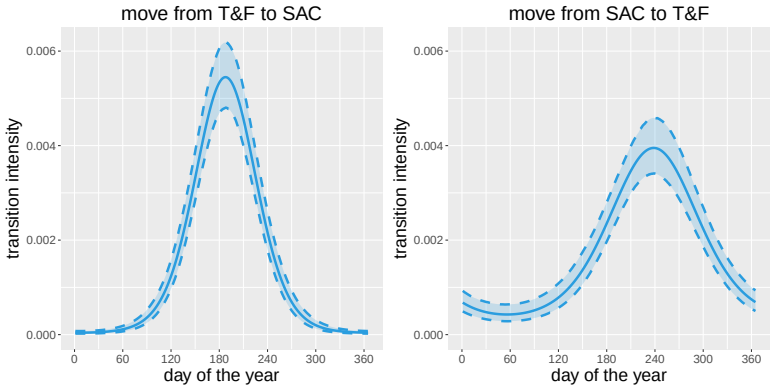


FIGURE 2. Seasonal pattern of the transition intensities between SAC and T&F.

4 Conclusion

When survey occasions within capture-recapture studies are irregularly spaced in time, it may be necessary to model the corresponding process in continuous time. In cases where individuals may additionally traverse through different states – e.g. corresponding to different sites, or the absence or presence of some infection – the corresponding continuous-time

multi-state model formulation becomes technically more involved than its discrete-time counterpart, the Arnason-Schwarz model. In this contribution, we developed a general modelling framework for multi-state capture-recapture modelling in continuous time, also allowing for time-varying covariates to affect the state transition intensities. A great advantage of embedding the capture-recapture setting within an HMM framework is that it not only facilitates the likelihood evaluation, but also renders other standard HMM tools applicable, such as the Viterbi algorithm for state decoding.

Acknowledgments: Special Thanks to Andrea Langrock, who turned a poor draft of the dolphin's capture history into a comprehensible illustration.

References

Langrock, R., Borchers, D.L., and Skaug, H.J. (2013). Markov-modulated nonhomogeneous Poisson processes for modeling detections in surveys of marine mammal abundance. *Journal of the American Statistical Association*, **108**(503), 840 – 851.

Jackson, C.H., Sharples, L.D., Thompson, S.G., Duffy, S.W., and Couto, E. (2003). Multistate Markov models for disease progression with classification error. *The Statistician*, **52**(2), 193 – 209.

Schwarz, C.J., Schweigert, J.F., and Arnason, A.N. (1993). Estimating migration rates using tag-recovery data. *Biometrics*, **49**(1), 177 – 193.

Continuous time hidden Markov models for astronomical gamma-ray light curves

Andrea Sottosanti[1], Mauro Bernardi[1], Luis Campos[2], Aneta Siemiginowska[3] and David van Dyk[4]

[1] University of Padova, Italy
[2] Harvard University, USA
[3] Harvard-Smithsonian Center for Astrophysics, USA
[4] Imperial College London, United Kingdom

E-mail for correspondence: sottosanti@stat.unipd.it

Abstract: We introduce a novel approach to study time varying γ-ray astronomical sources using continuous time hidden Markov models. The proposed method analyses the variation of signal from a source in time and successfully identifies different latent states that correspond to distinct physical mechanisms.

Keywords: Continuous time HMM; gamma-ray light curves; OU-process.

1 Introduction

The statistical analysis of time varying astronomical sources is an interdisciplinary field which combines both astronomical and statistical methods to investigate the physical mechanisms that characterise celestial objects. This type of analysis starts from collections of photons which fall on the detector surface of a telescope during the time. After that, the number of events is usually converted into flux in order to standardise the observations with respect to the size of the detector pixels and of the total observation period. The time series that describes the flux variation as a function of the time is called *light curve*.
Up to now, several authors [*Kelly et al. 2009*, *Meyer et al. 2014*, *Sobolewska et al. 2014*] considered X-ray light curves, that generally present moderate variations in the flux. However, lots of phenomena in the universe are high energetic, but a formal procedure to learn the physical processes of a source using γ-ray photons still does not exist, due to the complexity of these kind of light curves. In this paper, we propose a new statistical approach to the

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

analysis of γ-ray light curves of time varying sources using continuous time hidden Markov models.

2 Hidden Markov modelling approach

2.1 Continuous time observations

We consider a collection of n observations $\{y_{t_i}\}_{i=1}^n$ representing the flux of an astronomical source over a sequence of observation times $(t_1, \ldots, t_n) \in (0, T)$, where T is the entire observation period. The emission activity of a source is constantly monitored, but in practical data usually show big gaps in the observation times due to instrumental limits; thus, $\Delta_i = t_i - t_{i-1}$ is not constant.
This issue led *Kelly et al. (2009)* to consider a more appropriate model for continuous time observations. A generic process y is said to follow an *Ornstein-Uhlenbeck* (OU) process if its dynamic can be described by the stochastic differential equation $\partial y = \tau(\mu - y)\partial t + \sigma \partial Z_t$, where ∂Z_t is the increment of a Brownian motion with $Z_t \sim \mathcal{N}(0, \sigma^2)$, μ is a real parameter which represents the mean, σ is the volatility and τ is the speed of mean reversion. The solution to the differential equation above leads $y_{t_i+\Delta_t}|y_{t_i}$ to be Gaussian for every observation time t_i and for every arbitrary time interval Δ_t, which directly solves the time gaps problem inside the data.

2.2 Multiple states modelling

Although part of the literature focuses on single OU processes [*Kelly et al. 2009*, *Sobolewska et al. 2014*], more recently *Meyer et al. (2014)* proposed a multiple latent states model to represent different physical mechanisms of a source. It emerged that a two-states model well accomplishes for different states of variability in X-ray sources. In particular, the authors distinguish a prevalent state, whose flux activity is source dominated, from a baseline state, where the flux variation is due to measurement noise.
We propose to model the signal of time varying γ-ray sources through a two-states continuous time hidden Markov model [*Zucchini et al. 2016*, Chapter 11]. Our goal is to investigate a formal procedure to successfully fit γ-ray light curves and extrapolate the underlying information about the occurring physical phenomena. We adopt an OU-process for modelling the source flux y_t as *Meyer et al. (2014)*, but we consider also a continuous time model for the latent class variable $S_t \in \mathcal{S} = \{1, 2\}$. The latent Markov process we assume has initial probability vector $\boldsymbol{\delta}$ and generator matrix $\mathbf{Q} = \{q_{ij},\ i, j \in \mathcal{S}\}$, with $q_{ij} \geq 0$ when $i \neq j$, and $q_{ii} = -\sum_{j \neq i} q_{ij}$. Given the latent state S_t, for any time gap Δ_t the statistical model is $y_{t+\Delta_t}|S_{t+\Delta_t} = s, y_t \sim \mathcal{N}(\mu_{*,s}, \sigma^2_{*,s})$, where

$$\mu_{*,s} = y_{t_i} e^{-\tau_s \Delta_t} + \mu_s(1 - e^{-\tau_s \Delta_t}), \qquad \sigma^2_{*,s} = \frac{\sigma_s^2(1 - e^{-2\tau_s \Delta_t})}{2\tau_s}.$$

Let us denote with $\boldsymbol{f}(y_{t_i};\boldsymbol{\Theta})$ the 2×2 diagonal matrix whose s-th diagonal element is $\phi(y_{t_i};\mu_{*,s},\sigma^2_{*,s})$, ϕ is the density function of a Gaussian distribution, $\mathbf{1}$ is the unit vector of length 2 and $\boldsymbol{\Theta}=\cup_{s=1}^{2}(\mu_s,\sigma^2_s,\tau_s)$. According to *Zucchini et al (2016)*, the likelihood function for the model parameters $(\boldsymbol{\Theta},\mathbf{Q})$ is

$$\mathcal{L}(\boldsymbol{\Theta},\mathbf{Q})=\boldsymbol{\delta}^{\mathrm{T}}\exp\{\mathbf{Q}t_1\}\boldsymbol{f}(y_{t_1};\boldsymbol{\Theta})\prod_{i=2}^{n}\exp\{\mathbf{Q}(t_i-t_{i-1})\}\boldsymbol{f}(y_{t_i};\boldsymbol{\Theta})\mathbf{1}.$$

3 Parameters estimation via EM algorithm

We outline in this Section an efficient expectation-maximization (EM) algorithm for parameters estimation. Starting from $(\mathbf{Q},\boldsymbol{\Theta})^{(r-1)}$, the r-th iteration can be summarised into two steps as follows.

E-step: Given $\boldsymbol{\alpha}_0=\boldsymbol{\delta}^{\mathrm{T}}$ and $\boldsymbol{\beta}_{n+1}=\mathbf{1}$, update the *forward densities* $(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n)$ and the *backward densities* $(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_n)$ as

$$\begin{aligned}\boldsymbol{\alpha}_i&=\boldsymbol{\alpha}_{i-1}\exp\{\mathbf{Q}^{(r-1)}(t_i-t_{i-1})\}\boldsymbol{f}(y_{t_i};\boldsymbol{\Theta}^{(r-1)}),\\ \boldsymbol{\beta}_i&=\exp\{\mathbf{Q}^{(r-1)}(t_{i+1}-t_i)\}\boldsymbol{f}(y_{t_{i+1}};\boldsymbol{\Theta}^{(r-1)})\boldsymbol{\beta}_{i+1}.\end{aligned}\tag{1}$$

The quantities in Formula (1) are in practical numerically unstable, as they tend to zero or to infinity exponentially fast with the sample size n. According to *Roberts et al. (2006)*, we rescale $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ by a factor $c_i=\boldsymbol{\alpha}_i\mathbf{1}$, and we denote the new scaled quantities as $\tilde{\boldsymbol{\alpha}}_i$ and $\tilde{\boldsymbol{\beta}}_i$. Compute now $A^{(r)}_{s,s'}=\sum_{i=1}^{n}\tilde{\boldsymbol{\alpha}}_{i-1}\boldsymbol{\Lambda}^{(r-1)}\tilde{\boldsymbol{\beta}}_{i+1}$, where

$$\boldsymbol{\Lambda}^{(r-1)}=\int_{t_{i-1}}^{t_i}\exp\{\mathbf{Q}^{(r-1)}(t-t_{i-1})\}\mathbf{e}_s\mathbf{e}_{s'}^{\mathrm{T}}\exp\{\mathbf{Q}^{(r-1)}(t_i-t)\}\boldsymbol{f}(y_{ti};\boldsymbol{\Theta}^{(r-1)})dt.$$

$A^{(r)}_{s,s'}$ is the probability to transit from state s to s' during the total observation period $(0,T)$ and $\mathbf{e}_s$ is a vector of length 2 whose s-th element is 1 and the other is 0. To deal with the above matrix integral, we can factorise $\mathbf{Q}^{(r-1)}$ into $\mathbf{SDS}^{-1}$, where $\mathbf{S}$ is the matrix of eigenvectors and $\mathbf{D}$ is the diagonal matrix of eigenvalues. In this way, it is possible to show that the pq-th element of $\boldsymbol{\Lambda}^{(r-1)}$ becomes equal to

$$\Lambda^{(r-1)}_{p,q}=\sum_{u=1}^{2}\sum_{v=1}^{2}S_{pu}S^{-1}_{us}S_{s'v}S^{-1}_{vq}\phi(y_{t_i};\mu^{(r-1)}_{*,s'},{\sigma^2}^{(r-1)}_{*,s'})\mathcal{J}(d_u,d_v),$$

where d_u refers to the u-th eigenvalue of $\mathbf{D}$ and $\mathcal{J}(\cdot,\cdot)$ is given by *Rydén (1996)*.

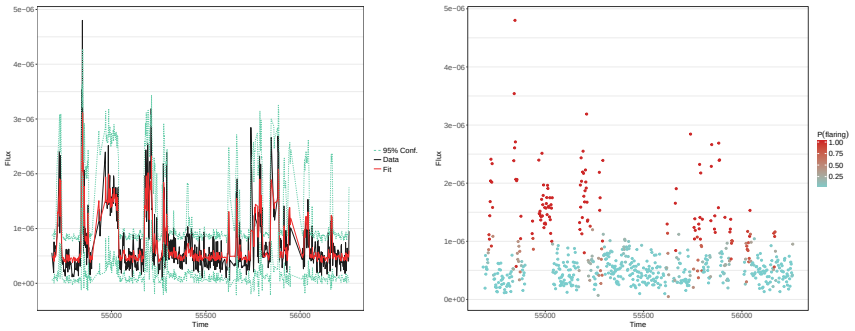


FIGURE 1. Left: γ-ray light curve of the blazar PKS 1510-05 (solid black line) against the model fitting (solid red line) and 95% prediction interval (dashed light blue line). Right: probability of each point to be a flaring observation.

M-step: Update the elements of $\mathbf{Q}$ out of the diagonal by doing $q^{(r)}_{s,s'} = q^{(r-1)}_{s,s'} A^{(r)}_{s,s'}/A^{(r)}_{s,s}$, for $s, s' = 1, 2$ and $s \neq s'$, while $q^{(r)}_{s,s} = -q^{(r)}_{s,s'}$. The solutions to the score equations for the mean μ_s and the volatility σ_s are available in closed form, while the speed of mean reversion τ_s requires a Newton-Raphson step to be updated. The estimated probability of being in the state s at the i-th observation time is $\mathbb{P}(S_{t_i} = s) = \tilde{\alpha}_{i,s}\tilde{\beta}_{i,s}$. Finally, we derive the estimate of δ_s by considering $\hat{\delta}_s = \mathbb{P}(S_{t_1} = s)$, which does not necessarily coincide with the stationary distribution of the latent Markov process.

4 Results

We present in this Section an application of our model to a real-case dataset, and in particular we consider a light curve from the blazar PKS 1510-05 detected by the Fermi LAT telescope. Blazars are very luminous and energetic sources characterised by an high variable signal, with heavy fluctuations in brightness on short time intervals. The available light curve of 630 observations is shown in the left plot of Figure 1, together with the model fitting and the 95% prediction interval obtained using a parametric bootstrap. The two-states model fits properly the prominent observations that come off from the resting flux concentrated around the mean value $4.779 \cdot 10^{-7}$. The only exception is made by the most evident flare recorded at time 54845.5, which does not fall into the 95% prediction interval and thus results as outlier with respect to the fitted model. We deduce that some external factors interacted with the light curve at that time, causing an extra amount of variability.
Table 1 displays the estimates of the model parameters. The second component, labelled as $s = 2$, has a larger and more variable flux activity than the

TABLE 1. Estimates of the model parameters in the two latent states. From left to right: mean, volatility, speed of mean reversion and probability to remain in the same state in a unitary time interval ($\Delta_t = 1$).

	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\tau}$	$\hat{p}_{s,s}$
$s = 1$	$4.779 \cdot 10^{-7}$	$5.198 \cdot 10^{-14}$	0.615	0.981
$s = 2$	$1.152 \cdot 10^{-6}$	$3.244 \cdot 10^{-13}$	0.359	0.950

first; moreover, it is faster mean reverting. This state mainly describes the flaring activity of the source, where a general augment of the average flux is anticipated by prominent spikes in the light curve. Solving $\exp\{\hat{\mathbf{Q}}\Delta_t\}$, we can access also to the estimated transition probability matrix of the Markov process for any arbitrary Δ_t. The last column of Table 1 displays the persistence probabilities assuming a unitary time gap, and confirms that both the identified states find large evidence. Contrary to *Meyer et al. (2014)*, none of the states we distinguish is noise dominated, but each one represents a specific phase of the emission activity of the source. Finally, the right plot in Figure 1 displays the probability of each point to be a flaring observation, and confirms the good separation between resting and flaring activity performed by our model.

References

Kelly, B. C., Bechtold, J. and Siemiginowska, A. (2009) Are the variations in quasar optical flux driven by thermal fluctuations? *The Astrophysical Journal*, **698(1)**.

Meyer, L., Witzel, G., Longstaff, F., and Ghez, A. (2014). A formal method for identifying distinct states of variability in time varying sources: Sgr A* as an example. *The Astrophysical Journal*, **791(1)**.

Roberts, W. J. J., Ephraim, Y., and Dieguez, E. (2006). On Rydén's EM algorithm for estimating MMPPs. *IEEE Signal Process. Lett.*, **13(6)**.

Rydén, T. (1996). An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis*, **21(4)**.

Sobolewska, M. A., Siemiginowska, A. , Kelly, B. C. and Nalewajko, K. (2014) Stochastic modeling of the Fermi/LAT γ-ray blazar variability. *The Astrophysical Journal*, **786(2)**.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2016). *Hidden Markov models for time series: an introduction using R*. Chapman and Hall / CRC.

Random effects dynamic panel models for unequally-spaced responses

Fiona Steele[1], Emily Grundy[2]

[1] London School of Economics & Political Science, UK
[2] University of Essex, UK

E-mail for correspondence: f.a.steele@lse.ac.uk

Abstract: A general random effects dynamic (autoregressive) model is proposed for handling unequally-spaced responses that are measured less frequently than time-varying covariates. The approach is suitable for continuous, binary or ordinal multivariate responses. The methodology is assessed in a simulation study, and applied to bivariate binary data on bidirectional exchanges of support between adult children and their non-coresident parents from the British Household Panel Survey and UK Household Longitudinal Study.

Keywords: latent autoregressive model; multivariate panel model; irregular panel; rotating module; intergenerational exchanges.

1 Introduction

Dynamic models, also known as autoregressive or lagged response models, are widely used in the analysis of longitudinal data in the health and social sciences. Standard discrete-time dynamic models assume that measurements of the response and time-varying covariates are equally spaced over time, but unequal spacing often arises by design or because of wave nonresponse. For example, it is common for household panel studies to use rotating modules to reduce survey costs and respondent burden, which leads to some variables being measured less frequently, and often at irregular intervals, than variables collected in the core questionnaire at each wave. This paper is concerned with dynamic models for the analysis of responses that are measured less frequently than time-varying covariates.

There has been little research on handling unequal spacing since early work by Rosner and Muñoz (1988). Recent work has proposed extensions to existing estimators for models for continuous equally-spaced responses

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

with fixed individual effects (e.g. Millimet and McDonough 2017; McKenzie 2001). While fixed effects models avoid the distributional assumptions of random effects models, they are less flexible in other respects such as the facility to handle categorical or multivariate outcomes, and are difficult to implement. We propose a random effects approach that can be generalised to ordinal and binary outcomes by specifying the model in terms of an underlying latent variable (Pudney 2008). The proposed model is applied in a study of the effects of changes in adult children's circumstances on help received from and given to their non-coresident parents.

2 Dynamic model for unequal spacing

Suppose that the underlying data generating process (DGP) for response y_{ti} at year t ($t = 1, 2, \ldots, T$) for individual i ($i = 1, \ldots, n$) takes the form of a linear first-order dynamic panel model

$$y_{ti} = \gamma y_{t-1,i} + \beta x_{ti} + u_i + e_{ti}, \quad t = 2, 3, \ldots, T \quad (1)$$

where γ is the autoregression parameter, x_{ti} is a time-varying covariate with coefficient β, $u_i \sim$ i.i.d. $N(0, \sigma_u^2)$ is an individual random effect capturing unmeasured time-invariant influences, and $e_{ti} \sim$ i.i.d. $N(0, \sigma_e^2)$ is a time-varying residual.

Let $m = 1, \ldots, M$ index the occasions at which y is measured, and denote by t_m the timing of measurement m and $\Delta t_m = t_m - t_{m-1}$ the gap in years between consecutive measurements. It can be shown that the DGP (1) implies the following model for the observed data

$$y_{mi} = \gamma^{\Delta t_m} y_{m-1,i} + \beta \sum_{k=0}^{\Delta t_m - 1} \gamma^k x_{t_m - k, i} + \left(\frac{1 - \gamma^{\Delta t_m}}{1 - \gamma} \right) u_i + \epsilon_{mi}, \quad (2)$$

where $\epsilon_{mi} = \sum_{k=0}^{\Delta t_m - 1} \gamma^k e_{t_m - k, i}$. Model (2) has two important features: (i) the coefficients and residual and random effect variances are nonlinear functions of the parameters of interest β, γ, σ_e^2 and σ_u^2; (ii) the predictors include values of x at t_m and each year since the time of the previous measurement of y. We consider a setting where x is available at each year (possibly with missing data).

In the application, we consider a generalisation where y is a bivariate binary response and (1) and (2) are types of random effects bivariate probit models for latent response y^* underlying y. We also model the initial condition y_{1i}. The time-varying covariate x is replaced by a set of indicators for transitions in individual characteristics for each year between $t_m - 1$ and t_m. For continuous y, (2) can be estimated via maximum likelihood, but Bayesian estimation is more flexible for binary y. All model estimation is carried out in JAGS (Plummer 2003) using the `rjags` R package.

3 Simulation study

A simulation study was carried out to assess the finite-sample performance of the proposed method for continuous and binary y. Data were generated for $T = 15$ from a model of form (1) for continuous y, or the corresponding latent autoregressive model (Pudney 2008) for binary y. Unequal spacing was generated by selecting observations of y at $t = 1, 6, 11, 13, 15$, corresponding to the spacing in the application. Combinations of three simulation conditions were considered for $n = 1000$ individuals: (i) balanced ($M = 5$) and unbalanced ($M_i \leq 5$) panels, (ii) strong and weak autocorrelation ($\gamma = 0.4, 0.8$), and (iii) moderate and low within-individual variation in x_{ti}. For all conditions, and for both continuous and binary data, estimates were found to be unbiased with good confidence interval coverage.

4 Application

We analyse exchanges of help between a respondent (child) and their non-coresident parent(s) using combined data from the British Household Panel Survey and its successor the UK Household Longitudinal Study for 2001-2015. Sample members were contacted at each year, but data on exchanges were collected in the family network module which was administered at less frequent and unequal intervals (in 2001, 2006, 2011, 2013 and 2015). The bivariate response consists of binary indicators of whether *any* help was given to or received from parents (based on a set of questions about different types of help). Covariates include respondent's gender and age and time-varying indicators of the presence of children, the age of the youngest child and annual changes in partnership and employment status. The time-varying covariates are based on data collected at each wave in the household and individual questionnaires.

TABLE 1. Parameter estimates from random effects bivariate probit model for any exchanges of help between children and their parents. Means and 95% credible intervals from 5 chains of 20k (burn-in=5k), thinning=5.

	To parents		From parents	
Parameter	Mean	95% CI	Mean	95% CI
Lag $y^*_{t_m-1}$	0.715	(0.658, 0.765)	0.681	(0.618, 0.733)
Partner at $t_m - 1$	-0.040	(-0.091, 0.011)	-0.095	(-0.146, -0.047)
Formation $(t_m - 1, t_m)$	-0.085	(-0.334, 0.163)	-0.363	(-0.595, -0.130)
Separation $(t_m - 1, t_m)$	-0.092	(-0.360, 0.170)	0.392	(0.132, 0.649)
Age (years) at t_m	0.013	(0.011, 0.016)	-0.029	(-0.034, -0.025)
Female	0.113	(0.071, 0.158)	0.121	(0.078, 0.166)
σ_u^2	0.187	(0.118, 0.273)	0.128	(0.073, 0.201)

Table 1 shows preliminary results from a model with gender, age and indicators of partnership transitions. We find that partnership transitions in the year before t_m are associated with the propensity to receive help from parents, but not the propensity to give help at t_m (conditional on helping behaviour in the previous year). The propensity to receive help decreases after partnership formation, but increases after a separation. Older respondents are more likely to give help and less likely to receive help, while women are more likely than men to engage in exchanges in either direction. There is strong evidence of reciprocity in exchanges with residual correlation estimates of approximately 0.45 at both the time and individual levels.

References

McKenzie, D.J. (2001). Estimation of AR(1) models with unequally spaced pseudo-panels. *Econometrics Journal*, **4**, 89 – 108.

Millimet, D.L. and McDonough, I.K. (2017). Dynamic panel data models with irregular spacing: with an application to early childhood development. *Journal of Applied Econometrics*, **32**, 725 – 743.

Rosner, B. and Muñoz, A. (1988). Autoregressive modelling for the analysis of longitudinal data with unequally spaced examinations. *Statistics in Medicine*, **7**, 59 – 71.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Pudney, S. (2008). The dynamics of perception: modelling subjective well-being in a short panel. *Journal of the Royal Statistical Society, Series A*, **171**, 21 – 40.

Density regression via penalised splines dependent Dirichlet process mixture of normals models

Vanda Inácio de Carvalho[1], María Xosé Rodríguez–Álvarez [2], Nadja Klein [3]

[1] School of Mathematics, University of Edinburgh, UK
[2] Basque Center for Applied Mathematics & IKERBASQUE, Spain
[3] School of Business and Economics, Humboldt Universität zu Berlin, Germany

E-mail for correspondence: vanda.inacio@ed.ac.uk

Abstract: We propose a novel Bayesian nonparametric method for density regression combining dependent Dirichlet process mixtures of normals and penalised splines. A practically important feature of our method is that, since the full conditional distributions for all model parameters are available in closed form, it allows for ready posterior simulation through Gibbs sampling. An application to a study concerning the association of a toxic metabolite on preterm birth is provided.

Keywords: Density regression; Dependent Dirichlet process mixtures; Gibbs sampling; Penalised splines.

1 Introduction

In many real-life applications, it is of interest to study how the distribution of a continuous (real-valued) response variable changes with covariates. Dependent Dirichlet process mixtures of normals models, a Bayesian nonparametric method, successfully address such goal. Roughly speaking, and in its full generality, these models can be thought of infinite mixtures of normal regression models where both the weights associated to the mixture components as well as the components' parameters are covariate dependent. The approach of considering covariate independent weights, also known as the single-weights dependent Dirichlet process mixture of normals model, it is very popular due to its computational convenience, but can have limited flexibility in practice (MacEachern, 2000). In order to obtain accurate

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

predictions, flexible forms for the components' parameters, mainly for the mean functions, are needed. In turn, formulations also allowing the weights to be dependent on covariates provide virtually all the flexibility needed for most data applications. However, such flexibility comes at a computational cost, with limited availability of simple algorithms for tractable posterior inference. In this work, to overcome the lack of flexibility, but retaining the computational tractability, we develop a single weights dependent Dirichlet process mixture of normals model where the components' means are modelled using Bayesian penalised splines (P-splines), so that the smoothness associated with each covariate can be learned automatically. A practically important feature of our P-splines dependent Dirichlet process mixture of normals model is that all parameters have conjugate full conditional distributions thus leading to straightforward Gibbs sampling.

2 Penalised splines dependent Dirichlet process mixture of normals model

Let y be a continuous response variable and $\mathbf{x} = (x_1, \ldots, x_p)'$ be a p-dimensional vector of covariates. For the sake of simplicity, we assume that all covariates are continuous. However, our modelling procedure can easily deal with categorical covariates, as well as, the interaction between continuous and categorical covariates.

In a single-weights dependent Dirichlet process mixture of normals model (De Iorio et al., 2004), the conditional density function is modelled as

$$f(y \mid \mathbf{x}) = \int \phi(y \mid \mu(\mathbf{x}, \boldsymbol{\beta}), \sigma^2) \mathrm{d}G(\boldsymbol{\beta}, \sigma^2),$$

where $\phi(\cdot \mid \mu, \sigma^2)$ is the density function of the normal distribution with mean μ and variance σ^2, and G follows a Dirichlet process prior with centring distribution $G_0(\boldsymbol{\beta}, \sigma^2)$ and precision parameter $\alpha > 0$. For ease of modelling we express G in the truncated stick-breaking form and therefore

$$f(y \mid \mathbf{x}) = \sum_{l=1}^{L} \omega_l \phi(y \mid \mu(\mathbf{x}, \boldsymbol{\beta_l}), \sigma_l^2),$$

where $(\boldsymbol{\beta}_l, \sigma_l^2) \overset{\text{iid}}{\sim} G_0$ and the weights are such that $\omega_1 = v_1, \omega_l = v_l \prod_{h<l}(1 - v_h)$, for $l = 2, \ldots, L$; the inputs of the weights are distributed according to a Beta distribution, i.e., $v_1, \ldots, v_{L-1} \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, and $v_L = 1$. Note that L is not the exact number of components we expect to observe but rather an upper bound on such number.

Regarding the specification of $\mu(\mathbf{x}, \boldsymbol{\beta_l})$ the usual, but somewhat rigid, choice is to assume a linear combination of the covariates in each component, i.e.,

$$\mu(\mathbf{x}, \boldsymbol{\beta_l}) = \mu_l(\mathbf{x}) = \beta_{l0} + \beta_{l1} x_1 + \ldots + \beta_{lp} x_p, \quad l = 1, \ldots, L.$$

However, this formulation implies that that the expected value of the response changes linearly with the covariates. To allow for nonlinear effects of the covariates we propose to model the mean of each component as an additive combination of smooth functions, i.e.,

$$\mu_l(\mathbf{x}) = f_{l1}(x_1) + \cdots + f_{lp}(x_p), \quad l = 1, \ldots, L,$$

where each $f_{lj}(\cdot)$, $j = 1, \ldots, p$, is approximated by a linear combination of cubic B-splines basis functions. More specifically, given a sequence of knots $x_{j,\min} = \xi_{j,0} < \xi_{j,1} < \ldots < \xi_{j,r_j} < \xi_{j,r_j+1} = x_{j,\max}$, we may write

$$f_{lj}(x_j) = \sum_{k=1}^{m_j} \beta_{ljk} B_{jk}(x_j), \quad m_j = r_j + 4, \quad l = 1, \ldots, L, \quad j = 1, \ldots, p,$$

where $B_k(x)$ denotes the kth cubic B-spline basis evaluated at x. It is well-known that estimates depend heavily on the number and location of the knots. Although a prior can be placed on the number of knots and their position, this could be challenging to implement efficiently in practice (e.g., involving reversible jump Markov chain Monte Carlo). As an alternative, P-splines rely on using a large number of equidistant knots in combination with a penalty on the regression coefficients to avoid overfitting. This is the approach we follow in this work. Bayesian P-splines (Lang and Brezger, 2004) are the Bayesian analogue to B-splines penalised by q-order differences (Eilers and Marx, 1996) and are constructed around q-order Gaussian random walks. In particular, we consider a second-order random walk prior to the spline coefficients, that is

$$\beta_{ljk} = 2\beta_{lj,k-1} - \beta_{lj,k-2} + u_{ljk}, \quad k = 3, \ldots, m_j,$$

where $u_{ljk} \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_{lj}^2)$. The random walk variance τ_{lj}^2 acts as an inverse smoothing parameter, with small values leading to heavy smoothing and large values allowing for considerable variation in the estimated function. To ensure identifiability of the additive structure, all functions $f_{lj}(\cdot)$, $j = 1, \ldots, p$, are centred around zero. To complete our model specification, we let

$$\alpha \sim \Gamma(a_\alpha, b_\alpha), \quad \sigma_l^{-2} \sim \Gamma(a_{\sigma^2}, b_{\sigma^2}), \quad \tau_{lj}^{-2} \sim \Gamma(a_{\tau^2}, b_{\tau^2}),$$

where $\Gamma(a, b)$ denotes the Gamma distribution with shape parameter a and rate parameter b. We use the blocked Gibbs sampler for posterior sampling and, as already mentioned, all full conditional distributions have simple conjugate forms.

3 Epidemiology application

Our method is applied to a dataset (comprised of 2312 observations) aimed at relating DDE (dichlorodiphenyldichloroethylene) concentration in maternal serum to the risk of premature delivery (Longnecker et al., 2001).

The DDE is a persistent metabolite of the pesticide DDT, which is used against malaria transmitting mosquitoes in endemic malaria areas, in spite of evidence suggesting adverse effects on premature delivery. The response variable is the gestational age at delivery (GAD) and births occurring before the 37th week (corresponding to approximately 260 days) are considered as preterm. Our interest is in modelling how the GAD distribution changes with DDE levels, with a particular focus placed on the left tail in order to assess the effect on preterm deliveries. Figure 1 provides inference for the conditional distribution of the GAD given the DDE, evaluated at the 10th, 60th, and 99th percentiles of DDE. It can be observed that the estimated conditional densities nicely follow the histograms and there is evidence that the left tail of the GAD distribution becomes fatter as DDE increases.

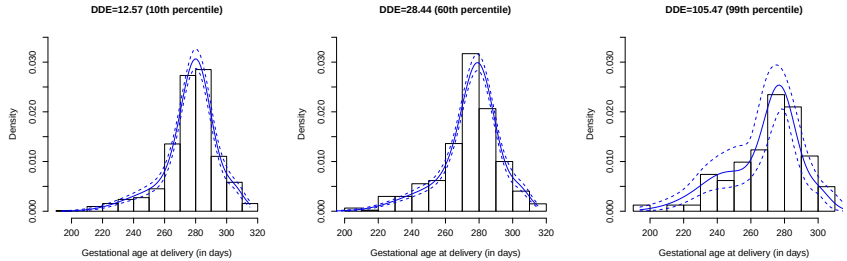


FIGURE 1. Histograms of the GAD for selected DDE intervals, along with the posterior mean (continuous blue line) and 95% posterior credible bands (dashed lines) of the conditional density of GAD given DDE, computed for the 10th, 60th, and 99th percentile of DDE.

Acknowledgments: The work of V Inácio de Carvalho was partially supported by FCT (Fundação para a Ciência e a Tecnologia, Portugal), through the projects PTDC/MAT-STA/28649/2017 and UID/MAT/00006/2019. MX Rodríguez was funded by project MTM2017-82379-R (AEI/FEDER, UE), by the Basque Government through the BERC 2018-2021 program and by the Spanish Ministry of Science, Innovation, and Universities (BCAM Severo Ochoa accreditation SEV-2017-0718).

References

De Iorio, M., Johnson, W. O., Muller, P. and Rosner, G. L. (2009). Bayesian nonproportional hazards survival modeling. *Biometrics*, **65**, 762 – 771.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines

and penalties. *Statistical Science*, **11**, 89 – 121.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **358**, 110 – 114.

Longnecker, M. P. , Klebenoff, M. A., Zhou, H. and Brock, J. W. (2001). Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *The Lancet*, **13**, 183 – 212.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Department of Statistics, Ohio State University.

Non-parametric Frailty Models for Cardiac Allograft Vasculopathy Data

Wenyu Wang[1], Ardo van den Hout[1]

[1] University College London, United Kingdom

E-mail for correspondence: w.wang.16@ucl.ac.uk

Abstract: The frailty model is a good approach to measure unobserved heterogeneity in survival analysis. Non-parametric frailty models define the latent frailty classes. We propose an extension of the model for class membership. In the application, we illustrate these frailty models for a disease process.

Keywords: Multi-state model; Longitudinal data; Survival analysis.

1 Introduction

Multi-state models are widely used in survival analysis to describe individuals change of status over time. There are two types of effects when describing the hazards for change of status: fixed effects and random effects. For the fixed-effects multi-state model, the characteristics of individuals are usually considered as covariates, such as age, gender and education level. However, there are still some differences of the hazards between different individuals, in addition to what we have measured with fixed effects. These differences can vary depending on the characteristics of individuals which are not in the data, or some variables which can not be measured and collected, or some information which researchers did not realize that may affect the results, or just the measured errors of explanatory variables. This unobserved heterogeneity can be taken into account as random effects. Models with both fixed effects and random effects in survival analysis are called frailty models. See Putter and Van Houwelingen (2011) for details of the role of frailty in multi-state model and whether we need them.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

2 Hazard function

For the frailty model, the hazard function for individual i in transition $r \to s$ can be defined by

$$h_{rs.i}(t|\boldsymbol{x}) = h_{rs.0}(t)\exp(\boldsymbol{\beta}_{rs}^{\top}\boldsymbol{x})B_{rs.i}\,, \tag{1}$$

where $\boldsymbol{x}$ is the vector with covariates values, $\boldsymbol{\beta}_{rs}$ is a parameter vector, $h_{rs.0}(t)$ is the baseline hazard. $B_{rs.i}$ is the frailty variable, where $B_{rs.i} > 0$, since (1) is a hazard function, which must be positive. Note that $B_{rs.i}$ can be changed to $B_{rs.g}$ for a group-shared random effect.
Frailties are often assumed to be parametrically distributed. For instance, frailty $B_{rs.i}$ following a lognormal distribution: $V_{rs.i} \sim N(0, \sigma_{rs}^2)$ where $\exp(V_{rs.i}) = B_{rs.i}$. Or one parameter gamma distribution $B_{rs.i} \sim \text{Gamma}(\theta)$. Here for this study, we mainly focus on a non-parametric frailty model. For the non-parametric model, define that there are a number of frailties B_k according to the classes C_k, where $k = 1, 2, \ldots, K$. For each individual $(i \in C_k)$ or group $(g \in C_k)$, the probability distribution π_k of frailties B_k is unknown. A major benefit of the non-parametric frailty model is that it is less restricted about the form of the distribution. The non-parametric model fits data well whether the frailty has a normal trend or not. However, there is no proper way to determine the optimal K when we fit the model. A good approach to solve this problem is fitting several models with different K and use model selection criteria.

3 Likelihood function

Kalbfleisch and Lawless (1985) presented the Markov assumption for analysis of panel data. For the non-parametric frailty model with K classes C_k, where $k = 1, 2, \ldots, K$, the likelihood contribution for individual i under the Markov assumption is given by

$$\begin{aligned} L_i(\boldsymbol{\theta}|i, \boldsymbol{y}, \boldsymbol{x}) &= P(Y_J = y_J, \ldots, Y_2 = y_2|Y_1 = y_1, i, \boldsymbol{\theta}, \boldsymbol{x}) \\ &= \sum_{k=1}^{K} P(Y_J = y_J, \ldots, Y_2 = y_2|Y_1 = y_1, i \in C_k, \boldsymbol{\theta}, \boldsymbol{x})\pi_k \\ &= \sum_{k=1}^{K}\prod_{j=2}^{J} P(Y_j = y_j|Y_{j-1} = y_{j-1}, i \in C_k, \boldsymbol{\theta}, \boldsymbol{x})\pi_k, \end{aligned} \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ is a vector which combines the fixed-effects parameters and class-specific parameters. $\boldsymbol{\theta}_0$ denotes the fixed-effect parameters, $\boldsymbol{\theta}_1 = \{B_1, \pi_1\}, \ldots, \boldsymbol{\theta}_K = \{B_K, \pi_K\}$. $\pi_k = P(i/g \in C_k)$ for each individual i or group g.

Therefore, the likelihood for N individuals is given by

$$L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x}) = \prod_{i=1}^{N} L_i(\boldsymbol{\theta}|i,\boldsymbol{y},\boldsymbol{x}).$$

In this study, we use the General purpose optimisation `optim` in the R software to maximise the likelihood function.

4 Models for CAV Data

We fit a non-parametric frailty model to the cardiac allograft vasculopathy (CAV) data. CAV is a disease that limits survival for cardiac transplant recipients. Sharples et al. (2003) defined it by three living states, which are the levels of CAV at each time. Figure 1 shows the multi-state process. State 1 to 3 are defined by no CAV, moderate CAV, severe CAV, respectively. State 4 represents dead. In this study, we define a progressive process with no backward transitions. Therefore, there are 5 transitions: $1 \to 2$, $1 \to 4$, $2 \to 3$, $2 \to 4$, $3 \to 4$.

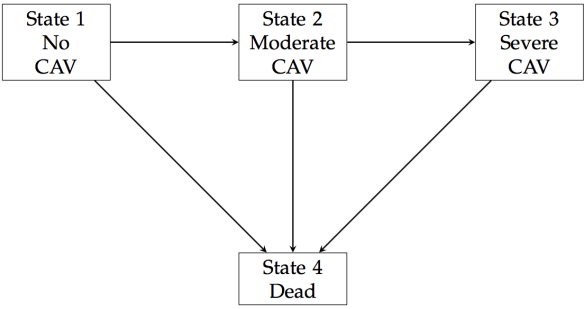


FIGURE 1. Transitions in the four-state model for cardiac allograft vasculopathy (CAV) data.

Computationally, it is easier to optimise over an unrestricted parameter space when maximising the loglikelihood function. Given the probabilities $\sum_{k=1}^{K} \pi_k = 1$, we use the logit link for probabilities π_k for class k. For example, for a model with $K = 2$, the two probabilities π_k can be represented by a parameter η with no restriction.

$$\pi_1 = \frac{1}{1+\exp(\eta)} \qquad , \qquad \pi_2 = 1 - \pi_1$$

The probabilities defined in above equations do not distinguish the differences for different people. The model will be more useful if the probabilities

of each frailty are related to people's characteristics. Bartolucci and Farcomeni (2015) proposed an approach to consider parameterizing η to a linear predictor,

$$\eta = \boldsymbol{\delta}^T \boldsymbol{x},$$

where $\boldsymbol{\delta}$ is a vector of parameters, $\boldsymbol{x}$ is a vector of covariates including an intercept.

In application, we fit three models with the same fixed-effect covariates: patients' age, patients' baseline age and donor's age. (*i*) a fixed-effect multistate model. (*ii*) a 2-class non-parametric frailty model without parameterizing η. The frailty is defined in transition $1 \to 2$. (*iii*) a 2-class gender-specific non-parametric frailty model with the gender-specific frailty defined in transition $1 \to 2$. Table 1 shows the results. The AIC values denote that both frailty models are better than the fixed-effects model. Regarding Model (*ii*), the frailty parameters illustrate that the probability of a random patient being a mover is 61.8% ($b_1 > 1$) as well as 38.2% ($b_2 < 1$) chance to be a stayer during transition $1 \to 2$. Model (*iii*) has the lowest AIC value. In this model, female patients are more likely to be stayers than movers during transition $1 \to 2$, since $\pi_1 < \pi_2$. In contrast, males have a higher probability to be a mover rather than stayer ($\pi_1 > \pi_2$).

TABLE 1. The -2loglik and AIC value and estimates (standard errors) of parameters for models for cardiac allograft vasculopathy (CAV) data. (i) is the fixed-effect model (ii) is the 2-class frailty model (iii) is the gender-specific 2-class frailty model.

Model	-2loglik	AIC	Frailty	
(*i*)	3446.7	3472.7		
(*ii*)	3438.5	3468.5	$b_1 = 2.237(0.539)$	$b_2 = 0.447(0.539)$
			$\pi_1 = 0.618(0.108)$	$\pi_2 = 0.382(0.175)$
(*iii*)	3432.8	3464.8	$b_1 = 3.411(0.798)$	$b_2 = 0.293(0.069)$
			for female:	
			$\pi_1 = 0.359(0.023)$	$\pi_2 = 0.641(0.023)$
			for male:	
			$\pi_1 = 0.613(0.084)$	$\pi_2 = 0.387(0.084)$

More comparison for movers ($b > 1$) and stayers ($b < 1$) can be illustrated by transition probabilities. They are the probabilities for each transition during a certain time interval. In the application, transition probabilities can be presented in a 4×4 matrix, where rows represent current states and columns represent the next states. Conditional on the mean of baseline age 47.1 and donor's age 30.6, transition probabilities for movers ($b_1 = 3.411$)

and stayers ($b_2 = 0.293$) in Model (*iii*) in 2 years after transplant are

$$P(t|b_1) = \begin{pmatrix} 0.688 & 0.188 & 0.043 & 0.081 \\ 0 & 0.586 & 0.272 & 0.142 \\ 0 & 0 & 0.578 & 0.422 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$P(t|b_2) = \begin{pmatrix} 0.902 & 0.019 & 0.004 & 0.075 \\ 0 & 0.586 & 0.272 & 0.142 \\ 0 & 0 & 0.578 & 0.422 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where $t = 2$, hazards are fixed midway the interval. It is easy to see the difference of transition probabilities from the matrixes above. For example, for individuals who transplant at 47.1 with the donor at 30.6, the probabilities of staying in state 1 are 68.8% (movers) versus 90.2% (stayers).
For the future work, we plan to explore and extend non-parametric frailty models. For instance, 3-class non-parametric models, the linear predictor η with more covariates, and multivariate frailty models.

References

Bartolucci, F., and Farcomeni, A. (2015). A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. *Biometrics*, **71(1)**, 80 – 89.

Kalbfleisch, J. D., and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, **80(392)**, 863 – 871.

Putter, H., and van Houwelingen, H. C. (2015). Frailties in multi-state models: Are they identifiable? Do we need them? *Statistical methods in medical research*, **24(6)**, 675 – 692.

Sharples, L. D., Jackson, C. H., Parameshwar, J., , Wallwork, J., and Large, S. R. (2003). Diagnostic accuracy of coronary angiography and risk factors for post-heart-transplant cardiac allograft vasculopathy. *Transplantation*, **76**, 679 – 682.

Robust Penalized Signal Regression

Brian D. Marx[1], Bin Li[1]

[1] Department of Experimental Statistics, Louisiana State University, USA

E-mail for correspondence: bmarx@lsu.edu

Abstract: A multivariate calibration problem from a soil characterization study motivated the proposed tractable and robust variants of penalized signal regression (PSR) using a class of nonconvex Huber-like loss function criteria. Standard methods may not be reliable, especially with heavy-tailed errors. We present a computationally efficient algorithm to solve this nonconvex problem. Simulation and empirical examples are extremely promising and show the proposed algorithm substantially improves the PSR performance under heavy-tailed errors.

Keywords: Huber loss; Multivariate calibration; P-splines; Robust regression.

1 Introduction

We revisit the rich regression problem where the p ordered regressors ensemble a signal contained in X with a scalar response, also known as multivariate calibration. As far as we know little or no work has been done in implementing robust smoothing into penalized signal regression. An assortment of loss functions — in addition to squared error loss — have been applied to the penalized spline and regression splines, e.g. Huber (1979), Härdle and Gasser (1984), Silverman (1985) and Hall and Jones (1990).

2 The Motivating Example

The dataset contains a total of 675 soil samples collected from Seward County (Nebraska), Kern County (California), and Lubbock County (Texas) in 2014. Ten physicochemical properties were measured for all 675 soil samples. They are: soil cation exchange capacity (CEC), electrical conductivity (EC), sand, among others. The reflectance spectra were measured from 360 to 2490 nm at 10 nm intervals. We aim to predict the nine soil physicochemical properties from spectra.
For illustration, PSR was applied to the spectra to predict each of the nine responses. Figure 1 shows the normal q-q plots of the PSR residuals

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

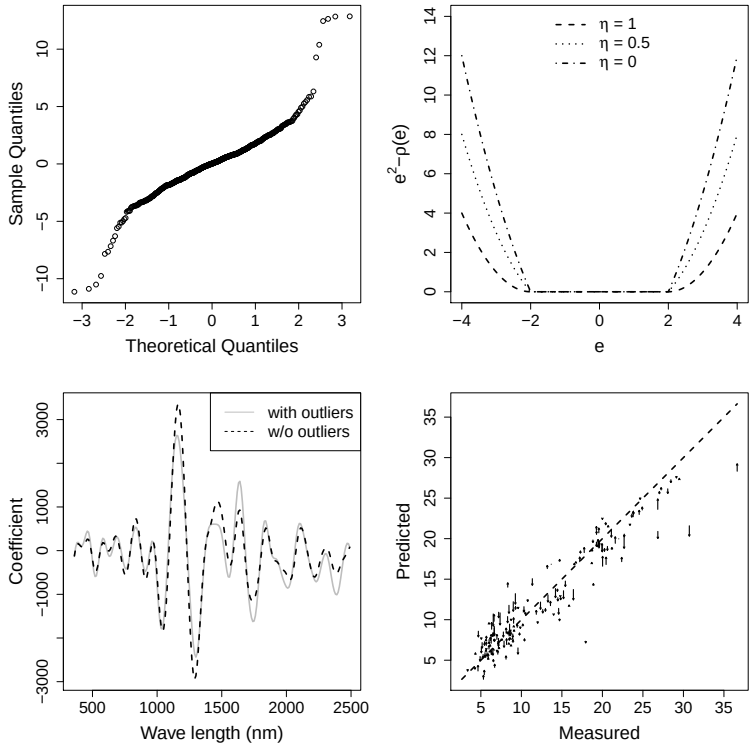


FIGURE 1. *Normal q-q plot of the CEC residuals with PSR fit (top, left); generalized Huber loss (trailing from squared error loss) with three different values of* η *at* $K = 2$ *(left); signal coefficient plots with (solid) and without (dash) the outliers for PSR (bottom, left) and rPSR (bottom left) for CEC; the prediction plots on test samples with and without the outliers for PSR (bottom, right). The arrows start from the predicted CEC values including all the training samples (with outlying samples) to the predicted CEC values excluding the samples with outlying residuals.*

for CEC, with some outlying residuals. Figure 1 also compares the coefficient plots and prediction on test samples with and without the outliers in the CEC case for PSR and rPSR. Outliers are those whose standardized residuals exceed 3 in absolute value. Penalized signal regression (Marx and Eilers, 1999) uses an objective function $S(\alpha) = ||y - XB\alpha||^2 + \lambda||D\alpha||^2$, reexpressing $\beta = B\alpha$, using a rich (n) B-spline basis, along the signal index. The solution is $\hat{\alpha} = (B'X'XB + \lambda D'D)^{-1}B'X'y$, with effective regressors $U = XB$. The optimal choice of λ can be made using cross-validation.

3 Generalized Huber Loss

Robust regression employs a criterion that is more resistant (to unusual observations) than those found using least squares. Li and Yu (2009) generalized the Huber loss criterion to a class of M-estimators,

$$\rho_\eta(e) = \begin{cases} e^2 & |e| < K \\ K^2 + 2\eta K(|e| - K) & |e| \geq K, \end{cases} \tag{1}$$

where $0 \leq \eta \leq 1$. The upper, right panel of Figure 1 illustrates the family of generalized Huber loss with three different values of η at $K = 2$, as the difference between squared error loss and $\rho_\eta(e)$. Although the generalized Huber criterion is not convex (in $e \in \mathbb{R}$) for $0 \leq \eta < 1$, it can be expressed as a difference of two convex functions (of e) as follows:

$$\rho_\eta(e) = e^2 - \mathbf{I}(|e| > K)\left[e^2 + 2\eta K(K - |e|) - K^2\right], \tag{2}$$

where $\mathbf{I}(\cdot)$ is an indicator. The leading convex function is the square loss function, where the second term of (2) is convex and *K-insensitive*.

4 Difference Convex Programming

The difference convex (d.c.) programming, developed by An and Tao (1997), addresses the problem of minimizing an objective function, which can be expressed as a difference of two convex functions, on the whole space. Consider minimizing an objective function $h(\mathbf{a})$ which is a difference of two convex functions, i.e. $h(\mathbf{a}) = h_1(\mathbf{a}) - h_2(\mathbf{a})$ where both $h_1(\mathbf{a})$ and $h_2(\mathbf{a})$ are convex in $\mathbf{a}$. The key idea of d.c. programming is to construct a sequence of subproblems, which are obtained by replacing the trailing convex function, e.g. $h_2(\mathbf{a})$, by its first order approximation function $h_2(\mathbf{a}^{(o)}) + \langle \mathbf{a} - \mathbf{a}^{(o)}, \bigtriangledown h_2(\mathbf{a}^{(o)})\rangle$ and solve them iteratively:

$$\mathbf{a}^{new} = arg\min_{\mathbf{a}} h_1(\mathbf{a}) - [h_2(\mathbf{a}^{cur}) + \langle \mathbf{a} - \mathbf{a}^{cur}, \bigtriangledown h_2(\mathbf{a}^{cur})\rangle]. \tag{3}$$

Note that after removing the constant terms in (3), minimization is equivalent to $\mathbf{a}^{new} = arg\min_{\mathbf{a}} h_1(\mathbf{a}) - \langle \mathbf{a}, \bigtriangledown h_2(\mathbf{a}^{cur})\rangle$.

5 Algorithm of Robust P-Splines

We replace the least square criterion by the generalized Huber criterion described in (1) within the PSR framework. Hence, the robust penalized smoothing splines (rPSR) minimizes

$$Q(\alpha) = \left\{\sum\nolimits_{i=1}^{m} \rho_\eta(y_i - U_i'\alpha)\right\} + \lambda\alpha' D_d' D_d \alpha, \tag{4}$$

which can be represented as a difference of two convex functions as follows:

$$Q(\alpha) = h_1(\alpha) - h_2(\alpha), \quad \text{where} \tag{5}$$

$$h_1(\alpha) = \sum_{i=1}^{m} e_i^2 + \lambda\alpha' D' D\alpha, \tag{6}$$

$$h_2(\alpha) = \sum_{i=1}^{m} \mathbf{I}(|e_i| > K)\left[e_i^2 + 2\eta K(K - |e_i|) - K^2\right], \tag{7}$$

and $e_i = y_i - U_i'\alpha$. The subgradient of $h_2(\alpha)$ with respect to α is

$$\bigtriangledown h_2(\alpha) = \frac{\partial h_2}{\partial e} \cdot \frac{\partial e}{\partial \alpha} = 2\sum_{i=1}^{m} \mathbf{I}(|e_i| > K)[e_i - \eta K \text{Sign}(e_i)]U_i. \tag{8}$$

The vector U_i' is the ith row of $U = XB$, with n elements $\{u_{ij}\}_{j=1}^{n}$. Let V be a column vector of length m with elements $\{\mathbf{I}(|e_i| > K)[e_i - \eta K \text{Sign}(e_i)]\}_{i=1}^{m}$. It follows that the right side of (8) above can be expressed as $2U'V$. The inner product of α and subgradient $\bigtriangledown h_2(\alpha)$ is then

$$\langle \alpha, \bigtriangledown h_2(\alpha) \rangle = -2\alpha' U' V. \tag{9}$$

Through d.c. programming, the minimization of the objective function (4) translates to the minimizing of a sequence of subproblems

$$\hat{\alpha} = arg\min_{\alpha}(Y - U\alpha)'(Y - U\alpha) + \lambda\alpha' D' D\alpha - 2\alpha' U' V. \tag{10}$$

Setting the first order derivative of the right side of (10) above to zero, we have the closed form of the solution

$$\hat{\alpha} = (U'U + \lambda D'D)^{-1} U'(Y - V) = (U'U + \lambda D'D)^{-1} U' Y^A. \tag{11}$$

The right side of (11) further shows that the subproblem solution is itself a modified PSR solution, one with the *adjusted* responses Y^A defined as

$$Y^A = \begin{bmatrix} y_1 - \mathbf{I}(|e_1| > K)[e_1 - \eta K \text{Sign}(e_1)] \\ \vdots \\ y_m - \mathbf{I}(|e_m| > K)[e_m - \eta K \text{Sign}(e_m)] \end{bmatrix}_{m \times 1}. \tag{12}$$

Note that only the observations with the residuals greater than K (in absolute value) will be "adjusted." Further if K is greater than all the residuals $\{e_i\}$, then rPSR and PSR solutions are the same.

Robust PSR Algorithm

1. Initializations:
   - Choose the tuning parameter value λ and η.
   - Construct B using a rich set of n B-spline basis functions of degree q on equally-spaced knots and penalty order d. Default $q = d = 3$.
   - Calculate $U = XB$
   - Calculate $\hat{\alpha} = \text{PSR}(U, Y, \lambda, d, n, q)$.
2. Cycle until convergence of $\hat{\alpha}$:
   - Calculate residuals $\{e_i\}_{i=1}^m$.
   - Find the K based on residuals.
   - Update the adjusted response vector Y^A according to η and K.
   - Update $\hat{\alpha} = \text{PSR}(U, Y^A, \lambda, d, n, q)$.
3. Prediction: $\hat{y}^{new} = x^{new'} B\hat{\alpha}$

End algorithm.

The algorithm terminates when $\max\{|(\hat{\alpha}_j^{cur} - \hat{\alpha}_j^{pre})/\hat{\alpha}_j^{pre}|\}_{j=1}^n < \epsilon$, where ϵ or set tolerance.The cutoff value K is chosen based on the proportion, γ-quantile, of the outliers among the residuals. In our algorithm, the $1.5\times IQR$ *rule* (interquartile range) is used to identify the outlying residuals. Lee and Oh (2007) and Tharmaratnam et al. (2010) provide advice to choose γ. The optimal value for η can be tuned through a grid search based on CV. Our algorithm usually converges within a few iterations. Updating $\hat{\alpha}$ is only a matrix-vector multiplication $(U'U + \lambda D'D)^{-1} U'$, and hence is computationally efficient.

6 Simulation Studies

Simulation studies showed that both the proposed rPSR is competitive with PSR for the normal errors, and alos when large errors exist (e.g. the error term has a heavy-tailed distribution), rPSR achieved better performance than PSR in terms of both prediction accuracy as well as model stability. We used the VisNIR spectra of the soil data and the PSR model with $\lambda = 10^{-5}$ (based on 10-fold CV) was fitted using the CEC response. The predicted values $\{\hat{f}_i\}_{i=1}^{675}$ from the PSR model are used as the "true" values for this simulation study. The data were then randomly split into a training set (506 observations or approximately 75% of the total sample size) and into a test set (169 observations). For the training samples, we created some *artificial responses* y_i^* by adding random errors e_i to the "true" values $\hat{f}_i$ (i.e. $y_i^* = \hat{f}_i + e_i$). Three types of error distributions are considered in this study: a normal distribution (i.e. $e_i \sim N(0, 2.39^2)$, a mixed normal distribution, and a slash distribution. Note that 2.39 is the standard deviation of the residuals from the PSR model. The mixed normal errors are generated from $0.95N(0, 2.39^2) + 0.05N(0, 23.9^2)$, that is the error constituted with 95% from $N(0, 2.39^2)$ and 5% from $N(0, 23.9^2)$. The

slash distribution is defined as a standard normal variate divided by an independent standard uniform variate (i.e. $N(0,1)/U(0,1)$), well-known for its heavy tail and extreme outliers.

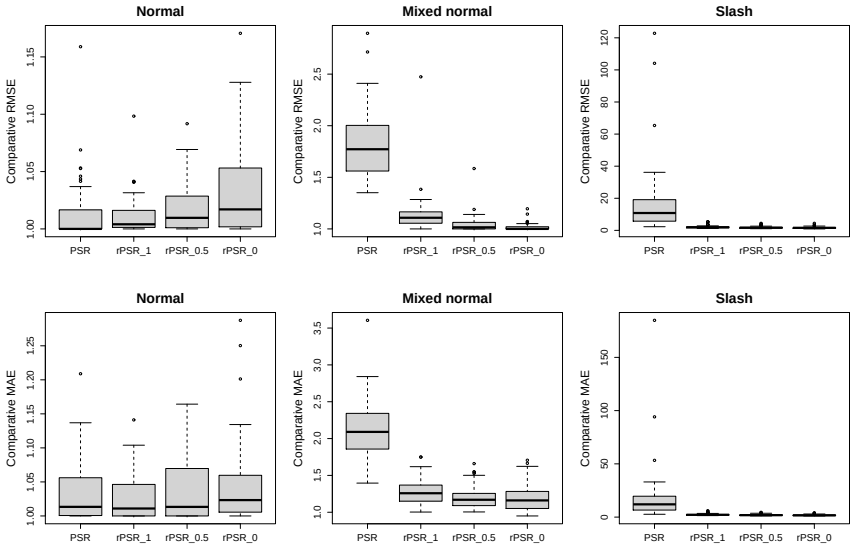


FIGURE 2. *Boxplots of comparative RMSE for rPSR and PSR. by distributions.*

References

An, L. and Tao, P. (1997).
Solving a class of linearly constrained indefinite quadratic problems by d.c. algorithms, *Journal of Global Optimization*, **11**, 253–285.

Hall, P. and Jones, M. (1990). Adaptive M-estimation in nonparametric regression, *The Annals of Statistics*, **18**(4), 1712–1728.

Härdle, W. and Gasser, T. (1984). Robust non-parametric function fitting. *JRSS, B*, **46**, 42–51.

Huber, P. (1979). In: *Robust Smoothing Robustness in Statistics*, Launer, R. and Wilkinson, G. (Eds.), Academic Press, 33–47.

Lee, T. and Oh, H. (2007). Robust penalized regression spline fitting with application to additive mixed modeling, *Computational Stat*, **22**(1), 159–171.

Li, B. and Yu, Q. (2009). Robust and sparse bridge regression, *Statistics and Its Interface*, **2**(4), 481–491.

Marx, B.D. and Eilers, P.H.C. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach, *Technometrics*, **41**(1), 1–13.

Silverman, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *JRSS, B*, **47**(1), 1–52.

Tharmaratnam, K., Claeskens, G., Croux, C. and Salibián-Barrera, M. (2010). S-estimation for penalized regression splines. *JCGS*, **19**(3), 609–625.

Nonparametric cure rate estimation when cure is partially known

Wende Clarence Safari[1], Ignacio López-de-Ullibarri[2], María Amalia Jácome[3]

[1] Universidade da Coruña, MODES group, Department of Mathematics, Faculty of Computer Science, A Coruña, Spain.
[2] Universidade da Coruña, MODES group, Department of Mathematics, Escuela Universitaria Politécnica, Ferrol, Spain.
[3] Universidade da Coruña, CITIC, MODES group, Department of Mathematics, Faculty of Science, A Coruña, Spain.

E-mail for correspondence: wende.safari@udc.es

Abstract: A nonparametric estimator for the cure rate in the presence of a covariate is introduced for the cure rate model, with cure partially known. Some properties are given, and the method is applied to a sarcoma dataset from the Cancer Epigenomics from Translational Medical Oncology (OMT) group, Health Research Institute of Santiago (IDIS) and the University Hospital of Santiago (CHUS) in Spain.

Keywords: Local maximum likelihood; Bandwidth; Censored data.

1 Introduction

A common assumption in standard survival modeling is that all individuals can experience the event if observed for enough time. Cure models have been developed because there might be situations where the standard survival model is not true. For example, in cancer studies, due to advances in cancer treatment there might be a proportion of patients who will get cured.
A common aspect in traditional cure models is that cured and uncured subjects cannot be distinguished within the censored observations. Hence, the cure indicator is usually modeled as a latent variable. However, sometimes this assumption is not entirely valid, when some extra information allows to conclude that some individuals with censored lifetimes are cured

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

or long-term survivors. One typical example is the case if individuals are assumed to be cured when their survival time is larger than a given threshold (e.g., 5 years when considering recurrence in some types of cancer). In this paper, a nonparametric estimator of the cure rate in the presence of a known cure fraction and conditional on a covariate is introduced.

2 Nonparametric cure rate estimator

Suppose Y is a random variable representing time to event of interest, $S(t) = P(Y > t)$ is the survival function, and C is the censoring time. Y and C are independent given a covariate X. It is assumed that the studied population is a mixture of individuals: those who will and those who will not experience the event of interest. According to this assumption the survival function can be written as

$$S(t|x) = 1 - p(x) + p(x)S_0(t|x)$$

where $S_0(t|x)$ is the survival function of the uncured or latency conditional on $X = x$, and $1-p(x)$ is the probability of being cured. The estimation of the model is usually performed with parametric or semiparametric models. Xu and Peng (2014) and López-Cheda et al. (2017) proposed a nonparametric mixture cure model which ignored the existence of known cures. In the presence of known cures the observations are

$$\{(X_i, T_i, \delta_i, \xi_i, \xi_i\nu_i) : i = 1, \ldots, n\}$$

where X is a covariate, $T = \min(Y, C)$ is the observed time, $\delta = \mathbf{1}(Y \leq C)$ is an uncensoring indicator, ξ is a binary variable which indicates the cure status is known ($\xi = 1$) or not ($\xi = 0$), and ν is the cure indicator. Therefore, $\xi\nu = 1$ indicates that the individual is known to be cured. Given the observations, the proposed estimator of $1 - p(x) = P(Y = \infty|X = x)$ is,

$$1-\widehat{p}_h(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} B_{h[i]}(x)}{\delta_{[i]} B_{h[i]}(x) + \sum_{j=i+1}^{n} B_{h[j]}(x) \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 0\right) + B_h^c(x)} \right)$$

where $B_h^c(x) = \sum_{j=1}^{n} B_{h[j]}(x) \mathbf{1}\left(\xi_{[j]}\nu_{[j]} = 1\right)$ is the sum of the weights of all the individuals known to be cured,

$$B_{h[i]}(x) = \frac{K_h\left(x - X_{[i]}\right)}{\sum_{j=1}^{n} K_h\left(x - X_{[j]}\right)}$$

are the Nadaraya-Watson weights with $K_h(.) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$ a rescaled kernel with bandwidth h. Finally, $\delta_{[i]}, X_{[i]}, \xi_{[i]}$ and $\nu_{[i]}$ are the concomitants of the ordered observed times $T_{(1)} < T_{(2)} < \ldots < T_{(n)}$.
It can be proved that $1 - \widehat{p}_h(x)$ is the nonparametric local maximum likelihood estimator of the cure rate. The proposed estimator of the cure rate has the following properties:

- If there are no known cures, it reduces to the nonparametric cure rate estimator proposed by Xu and Peng (2014) and López-Cheda et al. (2017).
- If there is no censoring it is equal to the sum of the weights of the known cures.
- In an unconditional setting and if the lifetime is greater than a known fixed time, it reduces to the generalized maximum likelihood estimator of the cure probability proposed by Laska and Meisner (1992).
- If there exists a common specific known cure threshold, it reduces to nonparametric cure rate estimator by Xu and Peng (2014) and López-Cheda et al. (2017).

3 Application to Sarcoma data

While this estimator can be applied into different research areas, the motivation in this paper was from a data set related to patients with sarcomas. There were 233 patients with sarcoma in the data set, with an outcome of interest, lifetime since diagnosis to death from sarcoma. A total of 59 (25.2%) patients died from sarcoma, and 174 (74.8%) patients were censored. Of the censored patients, a total of 18 patients were known to be long-term survivors, as they were tumor free for more than five years.
The covariate age (20 to 90 years) was used to estimate the probability of being cured. The proposed estimator was compared with the semiparametric estimator proposed by Bernhardt (2016), who assumed a logistic regression model for fitting cure probability.
Figure 1 shows the cure probability estimates obtained with the proposed nonparametric estimator for different choices of the smoothing parameter. These estimates are compared with the estimate given by the semiparametric estimator of Bernhardt (2016). Although the later estimate suggests a uniformly decreasing effect of the age on the cure rate, the curves from the proposed estimator are more consistent with a pattern characterized by a sharp decrease of the cure rate at younger ages until reaching a plateau at older ages.

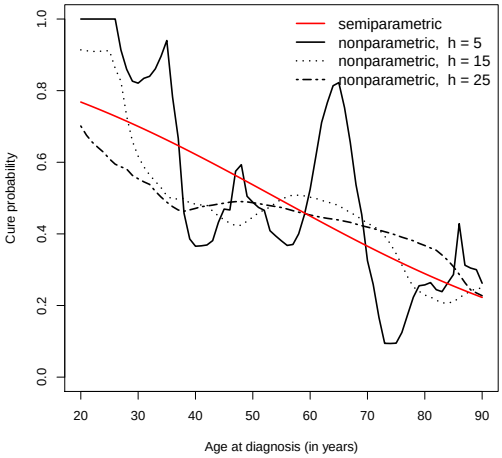


FIGURE 1. Estimation of the cure probability with the nonparametric estimator for different bandwidths (in black) and with the semiparametric estimator of Bernhardt (2016) (in red).

Acknowledgments: We would like to thank Ángel Díaz-Lagares and Yolanda Vidal-Insua from the Cancer Epigenomics of Translational Medical Oncology (OMT) group, Health Research Institute of Santiago (IDIS) and the University Hospital of Santiago (CHUS), Spain, for providing the sarcoma dataset. This research has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

Bernhardt, P. (2016). A flexible cure rate model with dependent censoring and a known cure threshold. *Statistics in Medicine*, **35(25)**, 4607–4623.

Laska, E.M. and Meisner, M.J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, **48(4)**, 1223–1234.

López-Cheda, A., Cao, R., Jacóme, M.A. and Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics and Data Analysis*, **105**, 144–165.

Xu, J., and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics*, **42(1)**, 1–17.

A skewness and kurtosis comparison for continuous distributions

Fernanda De Bastiani[1], Robert A. Rigby[2] Dimitrios M. Stasinopoulos[2] and Gillian Z. Heller[3]

[1] Statistics Department, Universidade Federal de Pernambuco, Recife/PE, Brazil,
[2] STORM, London Metropolitan University, London, UK
[3] Department of Mathematics and Statistics, Macquarie University, Australia

E-mail for correspondence: `debastiani@de.ufpe.br`

Abstract: The main goal of this paper is to compare the skewness and kurtosis of continuous distributions. It compares their moment skewness and kurtosis and compares their centile skewness and kurtosis. It shows the flexibility in skewness and kurtosis of different continuous distributions (within the **gamlss** `R` package), which helps the selection of an appropriate distribution.

Keywords: GAMLSS, leptokurtic, platykurtic, skew

1 Introduction

This paper presents a way to compare the skewness and kurtosis of continuous distributions, and an application to select an appropriate distribution for a response variable. Section 2 discusses the moment and centile definitions for skewness and kurtosis. In Section 3, a comparison of different theoretical distributions is presented, showing their flexibility in modelling skewness and kurtosis, where six distributions from the **gamlss.dist** `R` package are compared. An application to a real data is presented in section 4. Section 5 presents conclusions.
For a thorough investigation of the concept of skewness see MacGillivray (1986), where a wide variety of skewness measures and orderings are given. For a thorough investigation of the concept of kurtosis see MacGillivray and Ballanda (1988) and Balanda and MacGillivray (1990).

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

2 Skewness and kurtosis

2.1 Skewness

A distribution of a random variable Y is defined to be right skewed (i.e. 'positively skewed') if Y is more skew to the right than $-Y$, according to a particular skewness ordering. The problem is that 'more skew' is not well defined as there are many different skewness orderings which are often not equivalent. The following are two criteria for comparing the skewness of two distributions.

Moment skewness: Moment skewness of a random variable Y is defined by

$$\gamma_1 = \mu_3/(\mu_2)^{1.5}, \tag{1}$$

where μ_k is the kth central moment of Y. It is also known as Pearson's moment coefficient of skewness. The distribution of Y_2 is 'more moment skew to the right' than the distribution of Y_1 if $\gamma_1(Y_2) > \gamma_1(Y_1)$, where $\gamma_1(Y_i)$ is the moment skewness of Y_i. [Using this measure, 'moment positive skewness' is defined by $\gamma_1 > 0$, provided γ_1 is finite.]

Centile skewness The centile skewness function of a random variable Y is defined as

$$s_p = \frac{(y_p + y_{1-p})/2 - y_{0.5}}{(y_{1-p} - y_p)/2}, \tag{2}$$

for $0 < p < 0.5$, where $y_p = F_Y^{-1}(p)$ and $F_Y^{-1}(\cdot)$ is the inverse cdf of Y.

The distribution of Y_2 is 'more centile skew to the right' than the distribution of Y_1 if $s_p(Y_2) \geq S_p(Y_1)$ for all $0 < p < 0.5$. [Using this measure, 'centile positive skewness' is defined by $s_p \geq 0$ for all $0 < p < 0.5$, with $s_p > 0$ for some p.]

One important case is $p = 0.25$ in (2) giving Galton's measure of skewness $\gamma = s_{0.25} = \dfrac{(Q_1 + Q_3)/2 - m}{(Q_3 - Q_1)/2}$, we call this central centile skewness.

The two criteria above for comparing the skewness of two distributions are not equivalent. For example, when $Y \sim \texttt{TF}(\mu, \sigma, \nu)$, the moment skewness is finite only if $\nu > 3$.

2.2 Kurtosis

A distribution is defined to be leptokurtic (platykurtic) if it is more (less) kurtotic than the normal distribution, according to a particular kurtosis ordering. The problem is that 'more kurtotic' is not well defined as there are many different kurtosis orderings which are often not equivalent. The following are two criteria for comparing the kurtosis of two distributions.

Moment kurtosis: Moment excess kurtosis of random variable Y is defined by $\gamma_2 = \mu_4/(\mu_2)^2 - 3$, where μ_k is the kth central moment.

The distribution of Y_2 is 'more moment kurtotic' than the distribution of Y_1 if $\gamma_2(Y_2) > \gamma_2(Y_1)$, where $\gamma_2(Y_i)$ is the moment excess kurtosis of Y_i.

When $\gamma_2 < 0$, this indicates moment platykurtic, and $\gamma_2 > 0$ indicates moment leptokurtic.

Centile kurtosis: The centile kurtosis function (MacGillivray, 1986) of a random variable Y is defined by

$$k_p(Y) = \frac{y_{1-p} - y_p}{y_{0.75} - y_{0.25}}, \tag{3}$$

for $0 < p < 0.5$, where $y_p = F_Y^{-1}(p)$.

The distribution of Y_2 is 'more centile kurtotic' than the distribution of Y_1 if

$$k_p(Y_2) \geq k_p(Y_1), \text{ for all } 0 < p < 0.25, \tag{4}$$

and

$$k_p(Y_2) \leq k_p(Y_1), \text{ for all } 0.25 < p < 0.5, \tag{5}$$

with $k_p(Y_2) \neq k_p(Y_1)$ for some p.

Note condition (4) is one definition of Y_2 having heavier tails than Y_1, while condition (5) is one definition of Y_2 being more peaked than Y_1 around their medians. One important case is $p = 0.01$ in (3) giving $\delta = k_{0.01}$. The normal distribution has centile kurtosis $k_{0.01} = 3.449$. Hence the centile excess kurtosis $ek_{0.01}$ is given by $ek_{0.01} = k_{0.01} - 3.449$. Figure 1 presents the regions of combinations of moment excess kurtosis, and (positive) moment skewness for five distributions, and centile excess kurtosis and central centile skewness for six distributions.

3 Skewness and kurtosis comparison

Here six continuous distributions with range $(-\infty, \infty)$ implemented in the **gamlss** package in `R` are considered. For more details about **gamlss** and its distributions, see Stasinopoulos et al (2017) and Rigby et al (2019). All six distributions are location-scale family distributions each with parameters: μ the location shift parameter, σ the scaling parameter (μ and σ do not affect moment (or centile) skewness and kurtosis) and ν and τ parameters which control the skewness and kurtosis.

Figure 1 (a) shows the regions of combinations of excess moment kurtosis, and (positive) moment skewness for five distributions: (i) the exponential generalised beta type 2, `EGB2`, (ii) the Johnson SU, `JSU`, (iii) the Skew t type 3, `ST3`, (iv) the skew power exponential type 3 , `SEP3`, and (v) the sinh-arcsinh, `SHASHo`. Figure 1 (b) shows the central centile skewness and excess centile kurtosis of six distributions, the five distributions presented in Figure 1 (a) plus the stable distribution, `SB`, because its moment based kurtosis-skewness plot is not possible. [The corresponding regions for negative skewness are given by reflections of Figure 1 (a) and 1 (b) about a vertical axis at zero skewness.]

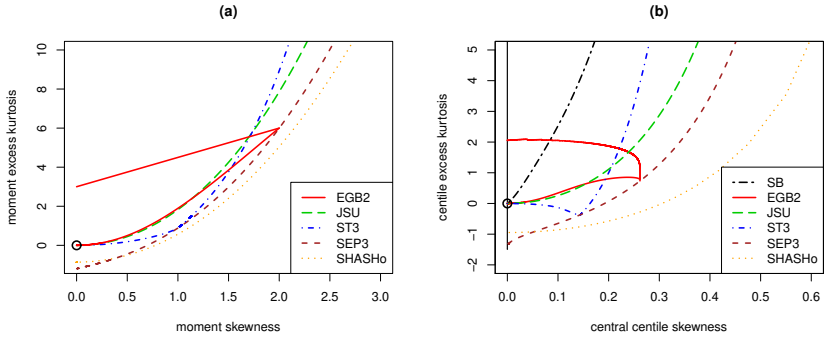


FIGURE 1. The regions of combinations of (a) moment excess kurtosis, and (positive) moment skewness for five distributions: `EGB2, JSU, ST3, SEP3` and `SHASHo`. (b) Centile excess kurtosis and central centile skewness for six distributions `SB, EGB2, JSU, ST3, SEP3` and `SHASHo`.

A modified version of the plots shown in Figure 1 can be useful to decide (at an exploration stage) about the adequacy of a fitted model, in terms of skewness and kurtosis. Within the GAMLSS package, **gamlss.dist**, there are two function that help: `checkMomentSK()` and `checkCentileSK()`. The functions take as argument either a response variable (with no explanatory variables) or a fitted GAMLSS model. In the latter case the quantile residuals of the model are extracted and analysed. The sample transformed skewness and excess kurtosis of the variable (or residuals) are plotted with the allowable regions of moment or centile skewness and kurtosis of the theoretical distributions. An assessment can then be made on whether the skewness and kurtosis of the variable (or residuals) are adequately fitted or not. However, these plots only tell us about the skewness and/or kurtosis of the variable or residuals. They are not designed for checking the location and scale for the variable or residuals. For information related to location and scale, worm plots and Q-statistics are appropriate. These plots provide information about the behaviour of the location and scale parameter, as well as skewness and kurtosis, see Chapter 16 of Stasinopoulos *et al.* (2017).

4 Application

An illustration of the FTSE returns data from 1991 to 1998. The original FTSE index is one of the four financial indices given in the `R` data set `EuStockMarkets`. Returns are calculated by the first difference of the natural logarithm of the series, $R_t = \log(Y_t/Y_{t-1})$. The goal is to find an appropriate distribution to the FTSE returns.
A histogram of the original data is shown in Figure 2. The moment skewness of the reruns is 0.1095, while the moment excess kurtosis is 2.63976.

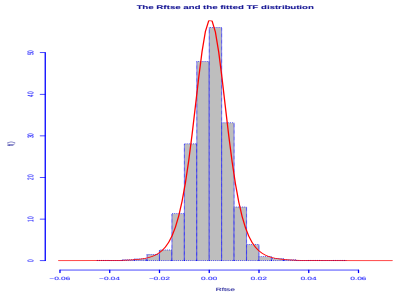


FIGURE 2. The FSTE returns with a fitted t family distribution.

The Jarque-Bera test for testing whether simultaneously there is skewness and kurtosis in the data has a value of 543.4, which compared to a $\chi^2(2)$ value of 5.99, it is significant. An automatic search for an appropriate distribution to the returns resulted in a t family distribution, denoted in GAMLSS as `TF`(μ, σ, ν), where μ and σ are location and scale parameters, respectively, while ν is the degrees of freedom. The fitted distribution is shown in Figure 2. Figure 3(a) and (b) shows two different plots created by the function `checkMomentSK()`. Figure 3(a) shows the skewness and kurtosis plot for the original values of the returns, `Rftse`, while Figure 3(b) shows the residuals of the fitted t family distribution, respectively. The background of the function is a standardised version of figure 1(a) reflected about the y-axis, so both negative and positive skewness can be shown. The vertical axis of Figure 3(a) and (b) is the transformed moment kurtosis γ_{2t} and the horizontal axis is the transformed moment skewness γ_{1t}. $[\gamma_{jt} = \gamma_j/(1 - |\gamma_j|)$, for $j = 1, 2]$. In the middle of the figure there is an elliptic region around the zero values of γ_{2t} and γ_{1t}. This region represents a 95% region for γ_{2t} and γ_{1t} based on the Jarque-Bera test, assuming a normal distribution for the variable (or residuals) with $(\gamma_{2t}, \gamma_{1t}) = (0, 0)$. If any $(\hat{\gamma}_{2t}, \hat{\gamma}_{1t})$ falls in this region then we accept the the null hypothesis of the normal distribution, i.e. there is no skewness and excess kurtosis in the variable/residuals. In Figure 3(a) $(\hat{\gamma}_{2t}, \hat{\gamma}_{1t})$ of the returns fall in the upper middle quarter of the plot, indicating that no skewness is present but hight kurtosis (leptokurtosis). The cloud of points around $(\hat{\gamma}_{2t}, \hat{\gamma}_{1t})$ for the `Rftse` variable are 99 values obtained from bootstrapping `Rftse` values. The cloud gives an indication of the variability of the skewness and kurtosis measures. In this case, since the bootstrap points crosses the vertical y-axis, the cloud indicates that skewness is not a problem for the variable `Rftse`. However, there are evidences for leptokurtosis. Figure 3(b) shows that the fitted model t family distribution is within the Jarque-Bera test region, and therefore skewness and excess kurtosis have been eliminated. Note that similar conclusions could be reached using centile measure of skewness and kurtosis with the `checkCentileSK()`.

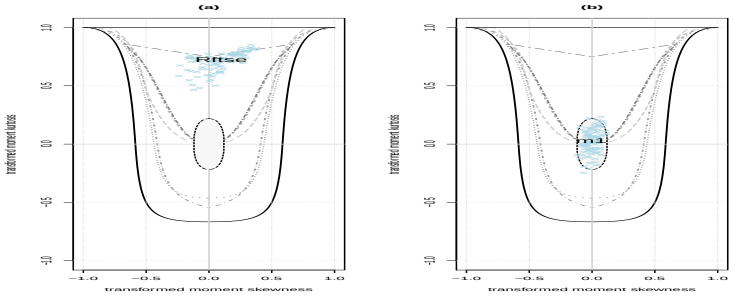


FIGURE 3. The FSTE returns with a fitted t family distribution.

5 Conclusion

The excess moment kurtosis against the moment skewness, and the excess centile kurtosis against the centile skewness are given for four important parameter for continuous distributions with range $(-\infty, \infty)$, implemented in **gamlss** package. The `SHASHo` and `SEP3` are flexible enough to model a response variable which can exhibit a wide range of skewness and kurtosis, while the `SB` and `ST3` are more appropriate to model a response variable with high kurtosis and low skewness. A visual method for detecting skewness and kurtosis in practical situations is also proposed. More details and discussion concerning skewness and kurtosis within a distributional regression model like GAMLSS can be found in Rigby et al (2019).

Acknowledgments: The partial financial support from Propesq/UFPE

References

Balanda, K. P. and MacGillivray (1990). Kurtosis and Spread, *Canadian Journal of Statistics*, **18**, 17 – 30

MacGillivray, H. L. (1986). Skewness and Asymmetry: measures and orderings, *Annals of Statistics*, **14**, 994 – 1011.

MacGillivray, H. L. and Balanda, K. P. (1988). The relationships between skewness and kurtosis, *Australian Journal of Statistics*, **30**, 319 – 337.

Rigby, R.A., Statsinopoulos, D.M., Heller, G.Z. and De Bastiani, F. (2019). *Distributions for Modelling Location, Scale, and Shape: Using GAMLSS in R*, Chapman and Hall.

Stasinopoulos, D.M., Rigby, R.A., Heller, G.Z., Voudouris, V. and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in* `R`, Chapman and Hall.

The competing risks model with interval sampling

Jacobo de Uña-Álvarez[1]

[1] Universidade de Vigo, Spain

E-mail for correspondence: jacobo@uvigo.es

Abstract: Interval sampling often occurs with registry data, induces double truncation on event times, and may result in a gross estimation bias. In this work we consider suitable corrections for such a potential bias in the scope of the competing risks model. Estimation of cumulative incidence functions and related curves is considered. Regression approaches for cause-specific and subdistribution hazards are discussed too. The properties of the proposed estimators are investigated both theoretically and through simulations. Applications to cancer registry data are included.

Keywords: Double truncation; Nonparametric estimation; Regression models.

1 Competing risks

Competing risks naturally arise in survival analysis when there exist several types of endpoints or events. One of the most well-known examples is found in oncology; when analysing the progression-free survival, one computes the time to progression or death without prior progression, whatever occurs first. Then, the separate investigation of each of the two transitions (disease progression, death) motivates the competing risks model, which has been the focus of a hugh literature. See for example Tsiatis (1975), Gray (1988), Pepe and Mori (1993), Lunn and McNeil (1995), Fine and Gray (1999), Putter et al. (2007) or Beyersmann et al. (2012).

Let T and η denote the absortion time and the even type, respectively, and assume that there exist K different types of events, so $\eta \in \{1, \cdots, K\}$. Then, the competing risks process is characterized by the joint distribution of (T, η), that is, by the K subdistributions of T restricted by η. These subdistributions are the so-called cumulative incidence functions $F_j(t) =$

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

$P(T \leq t, \eta = j)$, $1 \leq j \leq K$, which can be nonparametrically estimated under censoring by the method of Aalen and Johansen (1978).

2 Interval sampling

In this work we consider, rather than the standard situation of right-censoring, the issue of interval sampling. With interval sampling the data are limited to individuals with events within a certain observational window, determined by two particular calendar dates d_0 and d_1. This means that the available sample (T_i, η_i), $1 \leq i \leq n$, is formed by iid observations following the conditional distribution of (T, η) given $V - \tau \leq T \leq V$, where V is the time from onset to d_1 (assumed to be independent of T) and $\tau = d_1 - d_0$ is the width of the sampling interval. Besides, it is assumed that the dates of onset are observable, so the sample is completed with V_i, $1 \leq i \leq n$, iid observations following the conditional distribution of V given $V - \tau \leq T \leq V$. Interval sampling frequently occurs with registry data; indeed, epidemiological data often restrict to events (disease diagnosis, for example) within a (typically short) time interval. This may result in a sampling bias, in the sense that very small or large T-values will be hardly observed. This is in general the situation with doubly truncated data; see Zhu and Wang (2014) and references therein.

3 Cumulative incidences

In order to consistently estimate the cumulative incidence functions with interval sampling the aforementioned potential sampling bias must be taken into account. In particular, it happens that the naive application of the Aalen-Johansen estimator to interval sampling data is consistent for a weighted version of $F_j(t)$, specifically for

$$F_j^*(t) = \int_0^t w_j(t) F_j(dt),$$

where the weight function $w_j(t)$ may depend on the event type j (de Uña-Álvarez, 2018). Therefore, a consistent estimator for $F_j(t)$ can be constructed by downweighting in the Aalen-Johansen estimator the T_i's with relatively larger $w_j(T_i)$'s. This is not immediate however, since the weight functions $w_j(t)$, $1 \leq j \leq K$, are unknown and must be estimated from the (T_i, η_i, V_i)'s. See Efron and Petrosian (1999) for more on the difficulties behind nonparametric estimation from doubly truncated data.

In this work we address the construction of a consistent nonparametric estimator for $F_j(t)$ and we investigate its finite-sample and asymptotic properties both theoretically and through simulations. The estimation of the cumulative cause-specific hazards attached to (T, η) is discussed too.

4 Modelling for regression

Another question of much interest is how to perform regression analysis for the competing risks under interval sampling. Here, several modelling approaches can be considered, ranging from proportional cause-specific (resp. subdistribution) hazards models (see e.g. Fine and Gray, 1999) to time-varying coefficients models (Scheike et al., 2008). In this work we introduce proper corrections of such approaches so they can provide consistent estimates with interval sampling data. Again, the general idea is to use the estimated weight functions, $\hat{w}_j(t)$ say, in the score equations to recover the true (unbiased) regression coefficients. Mandel et al. (2018) exploited this idea for the standard Cox regression setting in which there exists an unique endpoint.

5 Application

As mentioned, the analysis of competing risks with registry data may be seriously affected by interval sampling. In this work we apply the proposed models and estimation techniques to cancer registry data, where several cancer groups are treated as competing risks. The data correspond to all the children diagnosed from cancer in the region of North Portugal between January 1, 1999, and December 31, 2003, and were gathered by RORENO (the cancer registry for that area). Cases were grouped according to the International Classification of Childhood Cancer (ICCC). Specifically, the cancer groups are leukaemias (group I), lymphomas (II), central nervous system (III), neuroblastoma (IV), and other less frequently observed cancers (ICCC groups V-XII). In the application it becomes clear that (i) the correction for the sampling bias is critical, and that (ii) the underlying assumptions which determine whether or not the weight functions $w_j(t)$ are free of the event indicator j may greatly influence the final estimates. Practical recommendations are given.

Acknowledgments: The author acknowledges financial support by the Grant MTM2017-89422-P (MINECO/AEI/FEDER, UE)

References

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

Beyersmann, J., Allignol, A., and Schumacher, M. (2012). *Competing Risks and Multistate Models with R*. Springer.

de Uña-Álvarez, J. (2018). Nonparametric estimation of the cumulative incidences of competing risks under double truncation. *Submitted*.

Efron, B. and Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association*, **94**, 824 – 834.

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.

Gray, R.J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics*, **16**, 1141 – 1154.

Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks. *Journal of the American Statistical Association*, **94**, 824 – 834.

Mandel, M., de Uña-Álvarez, J., Simon, D.K., and Betensky, R.A. (2018). Inverse probability weighted Cox regression for doubly truncated data. *Biometrics*, **74**, 481 – 487.

Pepe, M.S. and Mori, M. (1993). Kaplan-Meier, marginal or conditional-probability curves in summaryzing competing risks failure time data. *Statistics in Medicine*, **12**, 737 – 751.

Putter, H., Fiocco, M., and Geskus, R.B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389 – 2430.

Scheike, T.H., Zhang, M.-J., and Gerds, T.A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, **95**, 205 – 220.

Tsiatis, A. (1975). A nonindentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the U.S.A.*, **72**, 20 – 22.

Zhu, H. and Wang, M.-C. (2014). Nonparametric inference on bivariate survival data with interval sampling: association estimation and testing. *Biometrika*, **101**, 519 – 533.

Modular Regression – A Lego System for Building Structured Additive Distributional Regression Models with Tensor Product Interactions

Thomas Kneib[1], Nadja Klein[2], Nikolaus Umlauf[3] and Stefan Lang[3]

[1] Chair of Statistics, Georg-August-Universität Göttingen, Germany
[2] Chair of Applied Statistics, Humboldt-Universität zu Berlin, Germany
[3] Department of Statistics, Universität Innsbruck, Austria

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

Abstract: Semiparametric regression models offer considerable flexibility concerning the specification of additive regression predictors including effects as diverse as nonlinear effects of continuous covariates, spatial effects, random effects, or varying coefficients. In this paper, we discuss a generic concept for defining interaction effects in such semiparametric distributional regression models based on tensor products of main effects. These interactions can be anisotropic, i.e. different amounts of smoothness will be associated with the interacting covariates. We study identifiability and the decomposition of interactions into main effects and pure interaction effects (similar as in a smoothing spline analysis of variance) to facilitate a modular model building process. The decomposition is based on orthogonality in function spaces which allows for considerable flexibility in setting up the effect decomposition. Inference is based on Markov chain Monte Carlo simulations with iteratively weighted least squares proposals under constraints to ensure identifiability and effect decomposition. The performance of modular regression is demonstrated along the construction of spatio-temporal interactions for the analysis of precipitation sums and extreme precipitation events.

Keywords: Markov chain Monte Carlo simulations; Penalized splines; Smoothing spline analysis of variance; Space-time regression; Tensor product interactions.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

1 Introduction

In regression analyses, a model includes an interaction of two covariates ν_1 and ν_2, say, if the effect of ν_1 depends on the value observed for ν_2 (and vice versa) which typically leads to a specification such as

$$\eta_i = \ldots + \nu_{i1}\gamma_1 + \nu_{i2}\gamma_2 + \nu_{i1}\nu_{i2}\gamma_3 + \ldots$$

where η_i is some regression predictor, γ_1 and γ_2 are the main effects parameters of the covariates ν_1 and ν_2, respectively, and γ_3 represents the interaction effect. Importantly, the inclusion of an interaction only requires the definition of a new covariate given by the product of the two original covariates.

In semiparametric regression models comprising for example nonlinear or spatial effects, things turn out to be more difficult and in particular there exist a variety of different types of interaction effects that can be considered. For example, in a varying coefficient model

$$\eta_i = \ldots + \nu_{i1} f(\nu_{i2}) + \ldots$$

the effect of covariate ν_1 (the interaction variable) is varying with respect to the second covariate ν_2 (the effect modifier). This extends the standard product form of interactions by allowing for nonlinear or spatial changes of the effect of ν_1 depending on the value of ν_2 (which may also reflect spatial location of observations). Importantly, this type of interaction is now asymmetric in the sense that ν_1 is still assumed to have a linear effect albeit the dependence on the specific value observed for ν_2. This is overcome in tensor product interaction surfaces

$$\eta_i = \ldots + f(\nu_{i1}, \nu_{i2}) + \ldots$$

where the joint effect of ν_1 and ν_2 is represented by a nonlinear surface and therefore each combination of covariate values ν_1 and ν_2 may give rise to a completely different value of the joint effect $f(\nu_{i1}, \nu_{i2})$.

Some more advanced types of interaction effects include spatio-temporal effects $f(t, s_i)$ where t represents time and s_i geographical information on the observation of interest. In functional random effects models, we include effects $f_c(\nu_{ic})$ to allow for distinct nonlinear effects of covariate ν relative to some clustering variable c.

Such types of interactions allow for considerable flexibility in setting up a regression model, but they are also associated with a number of specific challenges. For the sake of interpretation, it would be beneficial to decompose the interaction effect into main effects and (maybe several) interaction effects. Furthermore, models involving multiple interactions are often not identifiable without imposing appropriate constraints.

In this paper, we will develop a generic and general framework for working with these and other types of interactions based on tensor products of

main effects. This will allow us to incorporate the interaction effects in the framework of structured additive distributional regression (Klein et al., 2015) and to benefit from efficient modes of Bayesian inference based on Markov chain Monte Carlo simulations. In addition, we will take particular care of the separation of interactions into main effects and pure interaction effects to facilitate the interpretation of estimated models and we will study the identifiability of specific models.

2 Tensor Product Interactions

2.1 Structured Additive Regression

In structured additive regression, we consider regression models where the regression predictor η_i obeys an additive structure such that

$$\eta_i = f_1(\nu_1) + \ldots + f_p(\nu_p)$$

where $f_1(\nu_1), \ldots, f_p(\nu_p)$ are different types of (potentially nonlinear, spatial or random) regression effects defined on covariates $\nu_1, \ldots, \nu_p$ representing different types of regression effects. Each of these effects is then approximated by an expansion in J basis functions, i.e. (ignoring the function/covariate index for simplicity)

$$f(\nu) = \sum_{j=1}^{J} \gamma_j B_j(\nu).$$

To enforce specific properties of the estimate such as moothness or shrinkage, we assign an informative, multivariate prior

$$p(\boldsymbol{\gamma}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{\frac{\text{rank}(\boldsymbol{K})}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right)$$

with precision matrix $\boldsymbol{K}$ and smoothing variance τ^2 to the vectors of basis coefficients $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)'$. The basis functions and the precision matrix are then chosen to represent a variety of effect types, see Fahrmeir et al. (2013, Ch. 9) for an overview.

2.2 Tensor Product Interactions

Turning to the construction of generic interaction effects of the two main effects

$$f_1(\nu_1) = \sum_{j_1=1}^{J_1} \gamma_{j_1} B_{j_1}(\nu_1), \qquad f_2(\nu_2) = \sum_{j_2=1}^{J_2} \gamma_{j_2} B_{j_2}(\nu_2)$$

with priors

$$p(\boldsymbol{\gamma}_d|\tau_d^2) \propto \left(\frac{1}{\tau_d^2}\right)^{\frac{\text{rank}(\boldsymbol{K}_d)}{2}} \exp\left(-\frac{1}{2\tau_d^2}\boldsymbol{\gamma}_d'\boldsymbol{K}_d\boldsymbol{\gamma}_d\right), \quad d = 1, 2,$$

we define the tensor product interaction of these two effects as

$$f(\nu_1, \nu_2) = \sum_{j_1=1}^{J_1}\sum_{j_2=1}^{J_2} \gamma_{j_1 j_2} B_{j_1 j_2}(\nu_1, \nu_2)$$

with tensor product basis functions

$$B_{j_1 j_2}(\nu_1, \nu_2) = B_{j_1}(\nu_1) B_{j_2}(\nu_2).$$

This in fact resembles the structure of common interaction since we use all pairwise products of main effect basis functions as interaction basis functions but, as mentioned above, has the drawback that the interaction effect usually includes the main effects as special cases which makes both interpretation and identification challenging. We therefore develop appropriate linear constraints to remove basically arbitrary portions of the interaction effect while still allowing for efficient Bayesian inference via Markov chain Monte Carlo simulation techniques.

3 Spatio-Temporal Analysis of Precipitation

We will illustrate the application of the developed methodology for tensor product interactions along the spatio-temporal analysis of precipitation in Germany. In a first analysis, we study daily precipitation sums in the period 1986 to 2015 and include all stations from *Deutscher Wetterdienst* above 900m sea level, while for stations below 900m we selected a representative subset with good coverage all over Germany. This results in a total of 164 meteorological stations and over 1.6 million spatio-temporally aligned observations.

For the response variable (total amount of precipitation), we applied a square-root transformation to improve the fit of a censored normal model where

$$y_{st} = \max(0, y_{st}^\star)$$

and

$$y_{st}^\star \sim \mathcal{N}(\mu_{st}, \sigma_{st}^2)$$

with $s = 1, \ldots, 164$ indexing the spatial locations of the meteorological stations and $t = 1, \ldots, T$ indexing the daily measurement time points. The predictor structure for both the "location" and the "scale" parameter is then given by

$$\eta = \beta_0 + f_1(\texttt{alt}) + f_2(\texttt{day}) + f_3(\texttt{lon}, \texttt{lat}) + f_4(\texttt{day}, \texttt{lon}, \texttt{lat}),$$

where `alt` represents altitutde, `day` is the day of the year and `lon`, `lat` represent longitude and latitude. As an exemplary result, Figure 1 shows the predicted precipitation climatology for January and July 10th.

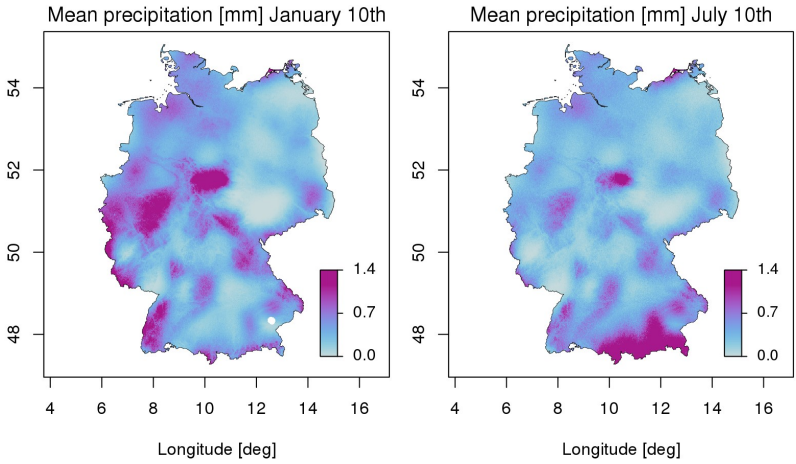


FIGURE 1. Predicted mean precipitation climatology for January and July 10th

In a second analysis, we focus on the spatio-temporal variation in 100 year return levels of precipitation in Germany based on roughly 1.1 millon observations of 569 meteorological stations. Here we assume a generalized Pareto model

$$\mathtt{P}_i \sim \mathrm{GP}(\xi(\mathbf{x}_\mathrm{i}), \sigma(\mathbf{x}_\mathrm{i}))$$

with the following predictor structure for both parameters:

$$\eta = \beta_0 + f_1(\mathtt{alt}) + f_2(\mathtt{year}) + f_3(\mathtt{day}) + f_4(\mathtt{lon}, \mathtt{lat}) + f_5(\mathtt{day}, \mathtt{lon}, \mathtt{lat}).$$

Figure 2 shows the temporal main effects (solid black lines) together with the spatio-temporal interaction variation around the main effect. To highlight the north-south gradient in the spatio-temporal interaction, the station-specific effects are coloured according to their north-south orientation.

References

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013). *Regression - Models, Methods and Applications*, Springer, Heidelberg.

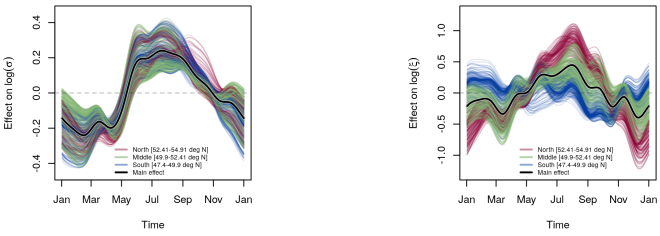


FIGURE 2. Estimated temporal main effect (solid black lines) together with spatio-temporal interaction effects (coloured lines) for the different measurement stations.

Klein, N., Kneib, T., Lang, S. and Sohn, A. (2015). Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany. *Annals of Applied Statistics*, **9**, 1024–1052.

Kneib, T., Klein, N., Umlauf, N. and Lang, S. (2019). Modular Regression – A Lego System for Building Structured Additive Distributional Regression Models with Tensor Product Interactions. *TEST*, to appear.

Regression-based Network Reconstruction with Covariates and Random Effects

Michael Lebacher[1], Göran Kauermann[1]

[1] Department of Statistics, LMU Munich, Ludwigstrasse 33, Munich, Germany

E-mail for correspondence: michael.lebacher@stat.uni-muenchen.de

Abstract: Network reconstruction is a general problem which occurs if matrix entries need to be predicted given the margins of the matrix. We show that the predictions obtained from the Maximum Entropy approach or equivalently using Iterative Proportional Fitting (IPF) can be obtained by restricted maximum likelihood estimation. Based on that we extend the framework towards regression and allow for covariates and random heterogeneity effects. The performance of the estimator is evaluated with a simulation study. Additionally, we apply the approach to interbank lending data and show that the inclusion of exogenous information leads to superior predictions in comparison to the IPF solution.

Keywords: ECM; IPF; Network analysis; Maximum entropy; Random effects

1 Model Derivation

We are interested in predicting $N = n(n-1)$ unobserved dyadic variables $x_{ij} \geq 0$ for $i, j = 1, ..., n$ and $i \neq j$. Let $\mathbf{x} = (x_{12}, \ldots, x_{n(n-1)})^{\mathrm{T}}$ be the corresponding column vector and stack the observed row- and column sums in the column vector $\mathbf{y} = (y_1, ..., y_{2n})^{\mathrm{T}}$. Furthermore, we define the binary $(2n \times N)$ routing matrix $\mathbf{A}$ with rows $\mathbf{A}_r$, allowing to denote the marginal restrictions by $\mathbf{A}_r\mathbf{x} = y_r$, for $r = 1, ..., 2n$. In order to predict the unknown $\mathbf{x}$ based on the observed marginals $\mathbf{y}$, we build on Golan and Judge (1996) and search for the density f that maximizes the Shannon entropy functional

$$H[f] = -\int_{\mathcal{X}} f(\mathbf{x}) \log(f(\mathbf{x}))\mathrm{d}\mathbf{x},$$

with support $\mathcal{X} \in \mathbb{R}_+^N$. We require that the density f integrates to unity

$$\int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}\mathbf{x} = 1, \tag{1}$$

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

and constrain the expectation $\boldsymbol{\mu} = (\mu_{12}, ..., \mu_{n(n-1)})^{\mathrm{T}}$ according to the marginal restrictions

$$\int_{\mathcal{X}} \mathbf{A}_r \mathbf{x} f(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbf{A}_r \boldsymbol{\mu} = y_r \text{ for } r = 1, ..., 2n. \tag{2}$$

Combining constraints (1) and (2) results in the Lagrangian functional

$$\mathcal{L}[f] = -\int_{\mathcal{X}} f(\mathbf{x}) \log(f(\mathbf{x})) \mathrm{d}\mathbf{x} - \lambda_0 \Big(\int_{\mathcal{X}} f(\mathbf{x}) \mathrm{d}\mathbf{x} - 1 \Big) - \sum_{r=1}^{2n} \lambda_r \Big(\int_{\mathcal{X}} \mathbf{A}_r \mathbf{x} f(\mathbf{x}) \mathrm{d}\mathbf{x} - y_r \Big),$$

with Lagrange multipliers $\lambda_r > 0$ for $r = 0, ..., 2n$. Maximization with respect to f using the Euler-Lagrange equation (Dym and Shames, 2013) provides the Maximum Entropy distribution

$$\hat{f}(\mathbf{x}) = \frac{1}{c(\boldsymbol{\lambda})} \exp \Big\{ -\sum_{r=1}^{2n} \lambda_r \mathbf{A}_r \mathbf{x} \Big\}, \text{ for } \mathbf{x} \in \mathcal{X},$$

where $c(\boldsymbol{\lambda}) = \exp(1+\lambda_0) = c(\lambda_1, ..., \lambda_{2n})$ is the normalization constant that ensures restriction (1). The parameters $\boldsymbol{\lambda}$ can be found by IPF (Koller et al., 2009). We, however, re-sort the sufficient statistics in order to obtain

$$\sum_{r=1}^{2n} \lambda_r \mathbf{A}_r \mathbf{x} = \sum_{i \neq j} (\lambda_i + \lambda_{n+j}) x_{ij} = \sum_{i \neq j} \frac{x_{ij}}{\mu_{ij}}$$

with $\mu_{ij} = (\lambda_i + \lambda_{n+j})^{-1}$ for $i \neq j$. Hence, the distribution of the network can be represented by a product of exponentially distributed random variables X_{ij}:

$$\hat{f}(\mathbf{x}) = \exp \Big\{ -\sum_{i \neq j} \frac{x_{ij}}{\mu_{ij}} - \sum_{i \neq j} \log(\mu_{ij}) \Big\} = \prod_{i \neq j} \frac{1}{\mu_{ij}} \exp \Big\{ -\frac{x_{ij}}{\mu_{ij}} \Big\}, \tag{3}$$

with observed margins $\mathbf{A}\mathbf{x} = \mathbf{A}\boldsymbol{\mu} = \mathbf{y}$ and $x_{ij} \geq 0 \; \forall \; i \neq j$.

2 Estimation and Inference

Given exogenous covariates $\mathbf{z}_{ij}$, we can model the expectation of model (3) with the parameter vector $\boldsymbol{\theta} = (\delta_1, ..., \delta_n, \gamma_1, ..., \gamma_n, \boldsymbol{\beta})^{\mathrm{T}}$ through

$$\mathbb{E}[X_{ij}] = \mu_{ij}(\boldsymbol{\theta}) = \exp(\delta_i + \gamma_j + \mathbf{z}_{ij}^{\mathrm{T}} \boldsymbol{\beta}), \text{ for } i, j = 1, ..., n \text{ and } i \neq j, \tag{4}$$

with δ_i and γ_j being subject-specific sender- and receiver-effects and $\boldsymbol{\beta}$ is the parameter vector for exogenous covariates $\mathbf{z}_{ij}$. For estimation, we propose to use an iterative procedure similar to the Expectation Conditional Maximization (ECM, Meng and Rubin, 1993) algorithm. Starting with an

initial estimate $\boldsymbol{\theta}_0$ that satisfies $\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}_0) = \mathbf{y}$, we form the expectation of the log-likelihood derived from (3)

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = \sum_{i \neq j} \left(- (\delta_i + \gamma_j + \mathbf{z}_{ij}^{\mathrm{T}} \boldsymbol{\beta}) - \frac{\mathbb{E}_{\boldsymbol{\theta}_0}[X_{ij}]}{\exp(\delta_i + \gamma_j + \mathbf{z}_{ij}^{\mathrm{T}} \boldsymbol{\beta})} \right).$$

Then, the maximization problem is given by

$$\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}_0) \text{ subject to } \mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{y}. \tag{5}$$

A suitable optimizer that allows for maximization under non-linear constraints is available by the augmented Lagrangian (Hestenes, 1969)

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\xi}_k, \zeta, \boldsymbol{\theta}_k) = -Q(\boldsymbol{\theta}; \boldsymbol{\theta}_k) - \boldsymbol{\xi}_k^T (\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \mathbf{y}) + \frac{\zeta}{2} ||\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \mathbf{y}||_2^2,$$

with $\boldsymbol{\xi}_k$ and ζ being auxiliary parameters. The augmented Lagrangian method decomposes the constrained problem (5) into iteratively solving unconstrained problems. In each iteration the algorithm starts with an initial parameter $\boldsymbol{\xi}_k$ in order to find the preliminary solution $\boldsymbol{\theta}_{k+1}$. Then, the algorithm updates $\boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_k + \zeta(\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}_{k+1}) - \mathbf{y})$ in order to increase the accuracy of the estimate. An implementation in `R` is given by the package `nloptr` by Johnson (2014).

The combination of ECM and augmented Lagrangian can easily be extended to allow for random effects, for example by assuming that the sender- and receiver-effects are jointly normally distributed

$$\begin{pmatrix} \delta_i \\ \gamma_j \end{pmatrix} \sim \mathcal{N}_2 \left(\mathbf{0}, \begin{pmatrix} \sigma_\delta^2 & \sigma_{\delta,\gamma}^2 \\ \sigma_{\delta,\gamma}^2 & \sigma_\gamma^2 \end{pmatrix} \right), \text{ for } i, j = 1, ..., n \text{ and } i \neq j. \tag{6}$$

Furthermore, prediction intervals for the unknown matrix entries x_{ij} can be obtained via parametric bootstrap. In order to do so, we define the prediction error as $e_{ij} = x_{ij} - \hat{\mu}_{ij}$ and construct prediction intervals for the unknown x_{ij} based on the quantiles of the empirical distribution of

$$\hat{\mu}_{ij} + e^*_{(b),ij} = \hat{\mu}_{ij} + x^*_{(b),ij} - \hat{\mu}^*_{(b),ij}, \text{ for } b = 1, ..., B,$$

where B represents the number of bootstrap samples $x^*_{(b),ij}$ and $\hat{\mu}^*_{(b),ij}$ represents the corresponding bootstrap estimates.

3 Performance of the estimator

We hope to see improvements in the predictions, relative to IPF, if the variation in $\mathbf{z}_{ij}$ is able to explain variation in the unknown x_{ij}. However, including $\mathbf{z}_{ij}$ with a low association to x_{ij} might lead to inferior predictions. For the simulation study, we use the following data generating process

$$\delta_i \sim N(0,1),\ \gamma_j \sim N(0,1),\ \mathbf{z}_{ij} \sim N(0,1), \text{ for } i, j = 1, ..., 10 \text{ and } i \neq j$$
$$\mu_{ij}(\beta) = \exp(\delta_i + \gamma_j + \mathbf{z}_{ij}\beta),\ x_{ij} \sim Exp(\mu_{ij}(\beta)).$$

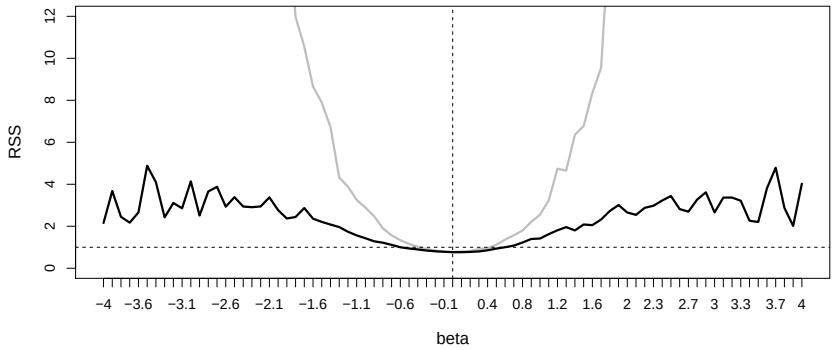


FIGURE 1. Median (solid black) and mean (solid grey) of the relative squared error $RSS_s(\beta)$ (vertical axis) for different values of β (horizontal axis).

Since the association between $\mathbf{z}_{ij}$ and the unknown x_{ij} is crucial, we vary the parameter β from -4 to 4 and denote with $\mu_{ij}(\beta)$ the mean based on β. For each parameter β we re-run the simulation $S = 1\,000$ times and calculate the IPF solution $\check{\mu}_{s,ij}(\beta)$ and the restricted maximum likelihood solution $\hat{\mu}_{s,ij}(\beta)$. Based on that, we calculate the ratio of the squared errors

$$RSS_s(\beta) = \frac{\sum_{i \neq j}(x_{s,ij} - \check{\mu}_{s,ij}(\beta))^2}{\sum_{i \neq j}(x_{s,ij} - \hat{\mu}_{s,ij}(\beta))^2}, \text{ for } s = 1, ..., 1\,000.$$

This ratio is smaller than one if the IPF estimates yield a lower mean squared error than the restricted maximum likelihood estimates and higher than one if the exogenous information improves the predictive quality.
In Figure 1, we show the median (solid black) and the mean (solid grey) of $RSS_s(\beta)$ for different values of β as well as a horizontal line indicating the value one (dashed black) and a vertical line for $\beta = 0$ (dashed black). It can be seen, that the mean and the median of $RSS_s(\beta)$ are below one for values of β that are roughly between -0.5 and 0.5 but increase strongly with higher absolute values of β. Apparently, the distribution of $RSS_s(\beta)$ is skewed with a long tail since the mean is much higher than the median. With very high or low values of β, the median of the relative mean squared error becomes more volatile and partly decreases.

4 Model Application

We use a multivariate time series of 52 networks consisting of the 21 most important countries from the *locational banking statistics* (LBS) provided by the Bank for International Settlements (www.bis.org). Within each

country, the LBS accounts for outstanding claims (x_{ij}) and liabilities (x_{ji}) of internationally active banks located in reporting countries. Additionally, we use logarithmic *gross domestic product* (gdp_i) (International Monetary Fund, www.imf.org) and logarithmic *dyadic trade flows* ($trade_{ij}$) between states (Correlates of War Project, www.correlatesofwar.org) as covariates. The models to reconstruct the LBS networks are fitted for all 52 time

TABLE 1. Comparison of models with the LBS Dataset (values scaled by 10^{-5}).

	Covariates	Rand. eff.	average L_1	SE	average L_2	SE
(I)	-	-	80.850	12.564	10.445	1.168
(II)	-	(6)	80.850	12.564	10.445	1.168
(III)	gdp_i, gdp_j, $trade_{ij}$	-	**63.466**	9.246	**7.802**	1.085
(IV)	gdp_i, gdp_j, $trade_{ij}$	(6)	63.763	**9.222**	7.850	**1.052**

points separately. By knowing the real matrix entries in this example we can compare the models in terms of their average L_1 and L_2 errors and the corresponding standard errors (SE). In Table 1 we show the IPF model (I), the regression model with random effects (II) and the model including exogenous covariates without (III) and with random effects (IV). It can be seen that the model (III) performs best, i.e. the inclusion of exogenous information increases the predictive quality as compared to the IPF model. In Figure 2 we visualize some results from model (III). In the top row on the left the predicted values are plotted against the real ones for the most recent network, together with gray prediction intervals. Matrix entries not covered by the prediction intervals are highlighted. It is also shown that the coverage of the prediction intervals is quite good for all time points under study (top, right). The estimated coefficients (bottom) provide the intuitive result that the claims from country i to country j increase with gdp_i and gdp_j and the trade volume between them ($trade_{ij}$).

5 Discussion

We propose a regression model for predicting matrix entries based on the marginals and exogenous information. Using a simulation study and real data we demonstrate that the approach has the potential to increase the predictive power relative to IPF. Furthermore, the approach allows for uncertainty quantification via bootstrap and comes with parameter estimates. Those are, however, to be interpreted with care because they are obtained in a setting with much less information compared to "common" regression settings. As a further caveat it is important to note that the usage of non-informative exogenous information might decrease the predictive power.

Acknowledgments: We thank Samantha Cook and Kimmo Soramäki (www.fna.fi) for providing data and discussing the problem with us.

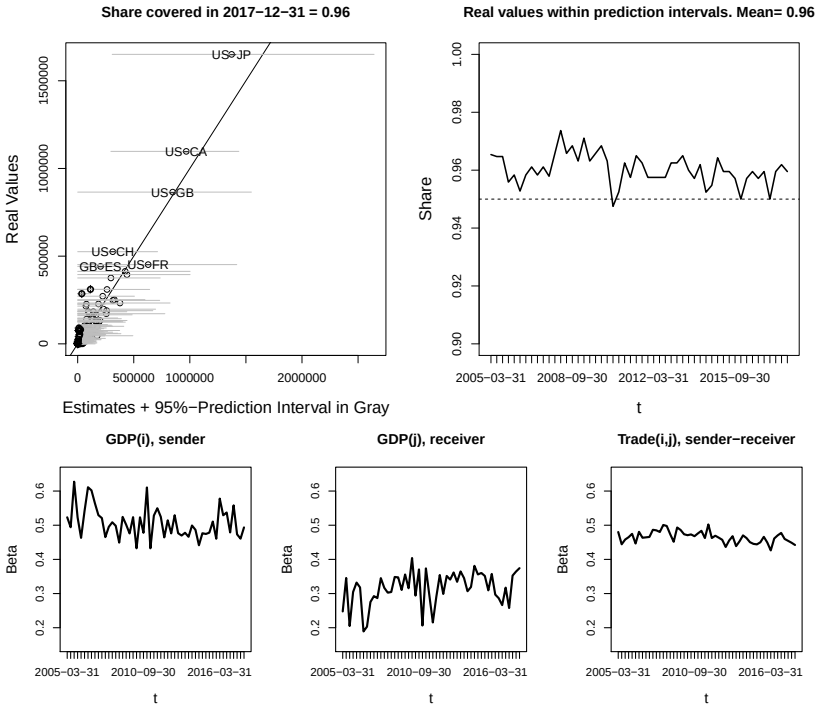


FIGURE 2. Actual vs. predicted values for the most recent observation (top, left). 95%-predictions intervals in gray, matrix entries not covered marked by $\oplus$. Share of values covered by the prediction interval against time (top right). Estimated coefficients (gdp_i, gdp_j, $trade_{ij}$) against time in the second row.

References

Dym, C. L., and Shames, I. H. (2013). *Introduction to the calculus of variations*. In: *Solid Mechanics* (Chap. 2). New York: Springer.

Golan, A., and Judge, G. (1996). Recovering information in the case of underdetermined problems and incomplete economic data. *Journal of Statistical Planning and Inference*, **49**, 127 – 136.

Hestenes, M. R. (1969). Multiplier and gradient methods. *Journal of optimization theory and applications*, **4**, 303 – 320.

Johnson, S. G. (2014). *The NLopt nonlinear-optimization package*. Version 1.2.1. URL: cran.r-project.org/web/packages/nloptr

Koller, D., Friedman, N., and Bach, F. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge: MIT Press.

Meng, X. L., and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267 – 278.

Distributional Trees for Circular Data

Lisa Schlosser[1], Moritz N. Lang[1,2], Torsten Hothorn[3], Georg J. Mayr[2], Reto Stauffer[1], Achim Zeileis[1]

[1] Department of Statistics, Universität Innsbruck, Innsbruck, Austria
[2] Department of Atmospheric and Cryospheric Science, Universität Innsbruck, Innsbruck, Austria
[3] Epidemiology, Biostatistics and Prevention Institute, Universität Zürich, Zürich, Switzerland

E-mail for correspondence: Lisa.Schlosser@uibk.ac.at

Abstract: For probabilistic modeling of circular data the von Mises distribution is widely used. To capture how its parameters change with covariates, a regression tree model is proposed as an alternative to more commonly-used additive models. The resulting distributional trees are easy to interpret, can detect non-additive effects, and select covariates and their interactions automatically. For illustration, hourly wind direction forecasts are obtained at Innsbruck Airport based on a set of meteorological measurements.

Keywords: Distributional Trees; Circular Response; Von Mises Distribution.

1 Motivation

Circular data can be found in a variety of applications and subject areas, e.g., hourly crime rate in the social-economics, animal movement direction or gene-structure in biology, and wind direction as one of the most important weather variables in meteorology. Circular regression models were first introduced by Fisher and Lee (1992) and further extended by Jammalamadaka and Sengupta (2001) and Mulder and Klugkist (2017) among others. While most of the already existing approaches are built on additive regression models, we propose an adaption of regression trees to circular data by employing distributional trees.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

2 Methodology

Distributional trees (Schlosser et. al, 2019) fuse distributional regression modeling with regression trees based on the unbiased recursive partitioning algorithms MOB (Zeileis et. al, 2008) or CTree (Hothorn et. al, 2006). The basic idea is to partition the covariate space recursively into subgroups such that an (approximately) homogeneous distributional model can be fitted to the response in each resulting subgroup. To capture dependence on covariates, the association between the model's scores and each available covariate is assessed using either a parameter instability test (MOB) or a permutation test (CTree). In each partitioning step, the covariate with the highest significant association (i.e., lowest significant p-value, if any) is selected for splitting the data. The corresponding split point is chosen either by optimizing the log-likelihood (MOB) or a two-sample test statistic (CTree) over all possible partitions.
In this study distributional trees are adapted to circular responses by employing the von Mises distribution, also known as "the circular normal distribution". Based on a location parameter $\mu \in [0, 2\pi]$ and a concentration parameter $\kappa > 0$ the density for $y \in [0, 2\pi]$ is given by:

$$f_{\text{vM}}(y; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \, e^{\kappa \cos(y-\mu)} \tag{1}$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and order 0 (see, e.g., Jammalamadaka and Sengupta 2001, for a more detailed overview).
In each subgroup maximum likelihood estimators $\hat{\mu}$ and $\hat{\kappa}$ are obtained by maximizing the corresponding log-likelihood $\ell(\mu, \kappa; y) = \log(f_{\text{vM}}(y; \mu, \kappa))$. The model scores are given by $s(y; \mu, \kappa) = (\partial_\mu \ell(\mu, \kappa; y), \partial_\kappa \ell(\mu, \kappa; y))$. In a subgroup of size n, evaluating the scores at the individual observations and parameter estimates $s(y_i; \hat{\mu}, \hat{\kappa})$ yields an $n \times 2$ matrix that can be employed as a kind of residual, capturing how well a given observation conforms with the estimated location $\hat{\mu}$ and precision $\hat{\kappa}$, respectively. Hence MOB or CTree can assess whether the scores change along with the available covariates. If so, by maximizing a partitioned likelihood the parameter instabilities are incorporated into the model. This procedure is repeated recursively until there are no significant parameter instabilities or until another stopping criterion is met (e.g., subgroup size or tree depth).

3 Application

Wind is a classical circular quantity and accurate forecasts of wind direction are of great importance for decision-making processes and risk management, e.g., in air traffic management or renewable energy production. This study employs circular regression trees to obtain hourly wind direction forecasts at Innsbruck Airport. Innsbruck lies at the bottom of a deep

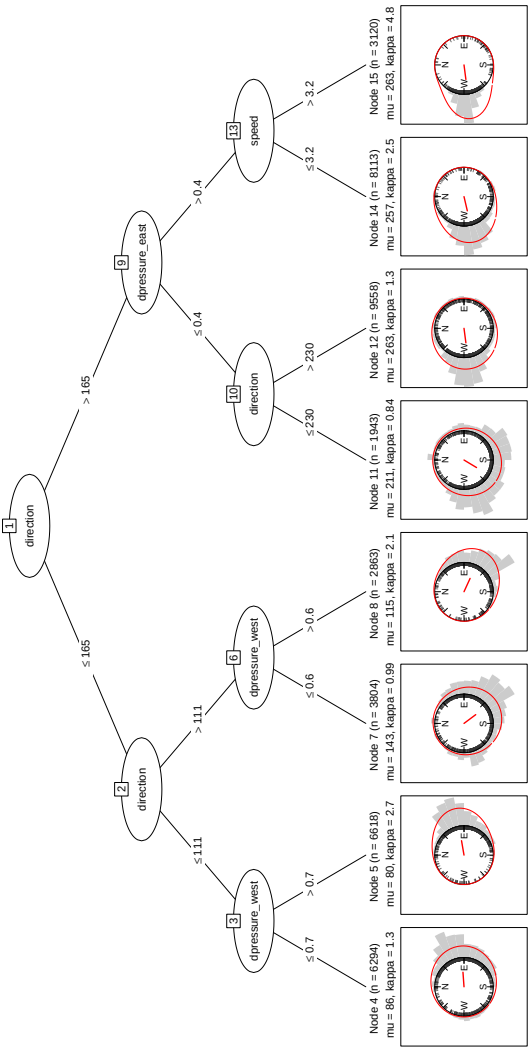


FIGURE 1. Fitted tree based on the von Mises distribution for wind direction forecasting. In each terminal node the empirical histogram (gray) and fitted density (red line) are depicted along with the estimated location parameter (red hand). The covariates employed are wind direction (degree), wind speed (ms^{-1}), and pressure gradients (dpressure; hPa) west and east of the airport, all lagged by one hour.

valley in the Alps. Topography channels wind along the west-east valley axis or along a tributary valley intersecting from the south. Hence, pressure gradients to which valley wind regimes react both west and east of the airport are considered as covariates along with other meteorological measurements at the airport (lagged by one hour), such as wind direction and wind speed at Innsbruck Airport. Note that in the meteorological context wind direction is defined on the scale $[0, 360]$ degree and increases clockwise from North (0 degree).

Figure 1 depicts the resulting distributional tree, including both the empirical (gray) and fitted von Mises (red) distribution of wind direction in each terminal node. Based on the fitted location parameters $\hat{\mu}$, the subgroups can be distinguished into the following wind regimes: (1) Up-valley winds blowing from the valley mouth towards the upper valley (from east to west, nodes 4 and 5). (2) Downslope winds blowing across the Alpine crest along the intersecting valley towards Innsbruck (from south-east to north-west, nodes 7 and 8). (3) Down-valley winds blowing in the direction of the valley mouth (from west to east, nodes 12, 14, and 15). Node 11 captures observations with rather low wind speeds that cannot be distinguished clearly into wind regimes and consequently are associated with a very low estimated concentration $\hat{\kappa}$. In terms of covariates, the lagged wind direction ("persistence") is mostly responsible for distinguishing the broad wind regimes listed above while the pressure gradients and wind speed separate between subgroups with high vs. low precision.

4 Discussion and outlook

Distributional trees for circular responses are established by coupling model-based recursive partitioning with the von Mises distribution. The resulting trees can capture nonlinear changes, shifts, and potential interactions in covariates without prespecification of such effects. This is particularly useful for modeling wind direction in mountainous terrain where wind shifts can occur due to turns of the pressure gradients along a valley.

4.1 Ensembles and random forests

A natural extension are ensembles or forests of such circular trees that can improve forecasts by regularizing and stabilizing the model. Random forests introduced by Breiman (2001) average the predictions of an ensemble of trees, each built on a subsample or bootstrap of the original data. A generalization of this strategy is to obtain weighted predictions by adaptive local likelihood estimation of the distributional parameters (Schlosser et. al, 2019). More specifically, for each possibly new observation x a set of "nearest neighbor" weights $w_i(x)$ is obtained that is based on how often x is assigned to the same terminal node as each learning observation $y_i, i \in \{1, \ldots, n\}$.

The parameters μ and κ are then estimated for each (new) observation x by weighted maximum likelihood based on the adaptive nearest neighbor weights:

$$\underset{\mu,\kappa}{\operatorname{argmax}} \sum_{i=1}^{n} w_i(x) \cdot \ell(\mu, \kappa; y_i). \tag{2}$$

Therefore, the resulting parameter estimates can smoothly adapt to the given covariates x whereas $w_i(x) = 1$ would correspond to the unweighted full-sample estimates and $w_i(x) \in \{0, 1\}$ corresponds to the abrupt splits from the tree.

4.2 Splits in circular covariates

In order to obtain more parsimonious and more stable trees another possible extension for *circular covariates* (with or without a *circular response*) is to consider their circular nature when searching the best split into two segments. In general, searching the best separation of a covariate into a "left" and "right" daughter node tries to maximize the segmented log-likelihood:

$$\max \left(\sum_{y \in left} \ell(\hat{\mu_1}, \hat{\kappa_1}; y) + \sum_{y \in right} \ell(\hat{\mu_2}, \hat{\kappa_2}; y) \right) \tag{3}$$

where $\hat{\mu_1}$, $\hat{\kappa_1}$, $\hat{\mu_2}$, $\hat{\kappa_2}$ are the estimated parameters of the von Mises distribution in the two daughter nodes. Searching a single split point ν in a circular covariate $\in [0, 2\pi)$ only considers linear splits into the intervals $left = [0, \nu]$ and $right = (\nu, 2\pi)$, thus enforcing a potentially unnatural separation at zero. This can be avoided by searching for two split points ν and τ considering a split into one interval $left = [\nu, \tau]$ and its complement $right = [0, \nu) \cup (\tau, 2\pi)$, encompassing zero. The latter strategy is invariant to the (often arbitrary) definition of the direction at zero.

When one split point ν is sufficiently close to zero and the other τ sufficiently far away, a simple linear split typically suffices to capture such a split (as seen for the lagged wind direction in Figure 1). If both ν and τ differ clearly from zero, two linear splits should also lead to a reasonable (but less parsimonious) fit. However, if both ν and τ are rather close to zero, a linear split strategy might miss such a pattern.

The required test statistic to maximally select two split points simultaneously is straightforward to accommodate in the CTree framework by providing all binary indicators corresponding to the splits into *left*/*right* intervals. However, this will become increasingly slow for larger sample sizes but it might be possible to speed up computations by exploiting the particular covariance structure similar to Hothorn and Zeileis (2008). In the MOB framework the extension is not quite as straightforward but one strategy could be to adapt double maximum tests à la Bai and Perron (2003).

Hence, the splitting idea can be naturally extended to a two-point search, however, for an unbiased and inference-based selection the corresponding testing strategies might need further adaption.

Computational details: R packages implementing the proposed methods are currently under development at `https://R-Forge.R-project.org/projects/partykit/`.

Acknowledgments: This project was partially funded by the Austrian Research Promotion Agency (FFG) grant no. 858537.

References

Bai, J., and Perron, P. (2003). Computation and Analysis of Multiple Structural Change Models. *Journal of Applied Econometrics*, **18**, 1 – 22.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 1, 5 – 32.

Fisher, N. I., and Lee, A. J. (1992). Regression Models for an Angular Response. *Biometrics*, **48**, 3, 665 – 677.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, **15**, 3, 651 – 674.

Hothorn, T., and Zeileis, A. (2008). Generalized Maximally Selected Statistics. *Biometrics*, **64**, 4, 1263 – 1269.

Jammalamadaka, S. R., and Sengupta, A. (2001). *Topics in Circular Statistics*. World Scientific.

Mulder, K., and Klugkist, I. (2017). Bayesian Estimation and Hypothesis Tests for a Circular Generalized Linear Model. *Journal of Mathematical Psychology*, **80**, 4 – 14.

Schlosser, L., Hothorn, T., and Zeileis, A. (2019). Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain. arXiv:1804.02921, *arXiv.org E-Print Archive*.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, **17**, 2, 492 – 514.

Bayesian functional PCA clustering with applications in neuroscience

Nicolò Margaritella[1], Vanda Inácio De Carvalho[1], Ruth King[1]

[1] University of Edinburgh, School of mathematics, UK

E-mail for correspondence: N.Margaritella@sms.ed.ac.uk

Abstract: The modelling of brain activity in recent years has started benefiting from extraordinary technological advances. With a remarkable amount of spatio-temporal data recordable from several parts of the brain, researchers are challenged to find models that can capture meaningful patterns behind such complexity. Thus motivated, we aim at modelling neuroscientific data employing functional data analysis within a Bayesian perspective; in particular, we exploit the flexibility of a Dirichlet process mixture model for clustering functional principal component scores to account for spatial dependence among curves. Our approach offers a general clustering procedure and a higher level of understanding of brain activity data. We present results from a simulation study and a resting-state fMRI dataset recorded from a healthy subject.

Keywords: Functional PCA; Dirichlet process; Clustering; Hierarchical model; Spatio-temporal data, Neuroscientific data.

1 Background

Most of the recoding tools in neuroscience produce a remarkable amount of spatio-temporal data that are obtained simultaneously from several parts of the brain. Large datasets require new advanced statistical methods to efficiently extract useful information.
Functional data analysis (FDA) deals with observed, noise-corrupted signals $Y_i(t)$ for curve $i = 1, ..., n$ at the time interval $t = 1, ..., T$. These observations can be expressed by an additive error model:

$$Y_i(t) = \mu(t) + X_i(t) + \epsilon_i(t) \tag{1}$$

where $\epsilon_i(t)$ is a white noise process, $\mu(t)$ is the underlying mean and $X_i(t)$ is a realisation of a mean-zero smooth stochastic process. If functional Principal Components analysis (fPCA) is employed then $X_i(t)$ has the following

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

expansion:

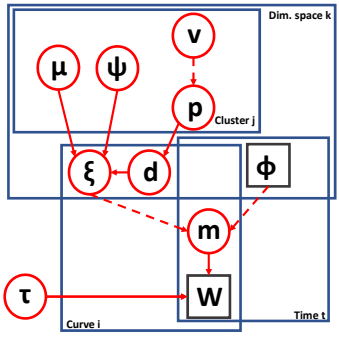
$$X_i(t) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t), \tag{2}$$

where $\{\phi_k(t)\}_{k=1}^{\infty}$ are orthogonal eigenfunctions called functional principal components (fPCs) and ξ_{ik} are fPC scores with variance given by the eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$.
Functional data models typically assume $X_i(t)$ to be independent and identically distributed which implies $\mathrm{Cov}(\xi_{ik}, \xi_{jk}) = 0, \forall\ i, j, k$. Recently, Liu et al. (2017) modelled spatial dependence among curves through a suitable covariance function for the fPC scores and estimating the relative parameters.
Our proposed model extends the standard Bayesian functional data model with independent fPC scores (Crainiceanu and Goldsmith, 2010) to a simple, computationally feasible hierarchical model which allows for dependence among fPC scores through a Dirichlet process mixture prior specification.

2 Methods

Let $W_i(t) = Y_i(t) - \widehat{\mu}(t)$, then a Bayesian hierarchical model that allows for clustering of the fPC scores can be specified as shown in the panel below. We highlight here some important aspects of this approach.

$$\begin{cases} W_{it}|m_{it}, \tau \sim N(m_{it}, \tau^{-1}), \\ m_{it} = \sum_{k=1}^{K} \xi_{ik}\phi_{kt}, \\ \xi_{ik}|d_{ik} \sim N(\mu_{jk}, \psi_{jk}^{-1}), \\ d_{ik} \sim \mathrm{Cat}(\underline{p}_k) \\ \mu_{jk} \sim N(r, w), \\ \psi_{jk} \sim \mathrm{Gamma}(\gamma, \beta) \\ p_{jk} = v_{jk}\prod_{l<j}(1 - v_{lk}) \\ v_{jk} \sim \mathrm{Beta}(1, \alpha) \\ \tau \sim \mathrm{Gamma}(\gamma, \beta) \end{cases}$$



Our model employs a truncated approximation of the well known infinite Gaussian mixture (Rasmussen, 2000) but we shift the mixture from data to the fPC scores. For every eigendimension k, uncertainty in the latent component weights p_j is accounted for by v_j, the inputs into the construction of the stick-breaking weights. This approach has at least two main advantages: first, clustering over the fPC scores allows a flexible definition of their dependence without imposing any model or constraint on its form (e.g. positive-semidefiniteness of the covariance structure). It follows that

in the presence of any underlying structural dependence among curves, improvements in curve reconstruction are expected compared to the standard model as also shown by Liu et al. (2017) in a frequentist framework. We report results of curve reconstruction from a simulation study in Section 3. Second, our model can be seen as a generalisation of the Bayesian infinite mixture model based clustering as the fPC scores are clustered for every mode of variation (eigenfunction) independently, resulting in a potentially much finer classification. In fact, in the case where two curves have the relative fPC scores allocated to the same clusters *for all* K eigendimensions considered, the classification procedure reduces to the standard curve classification. Results of clustering fPC scores in a resting-state fMRI dataset are presented in the next section.

3 Simulation study and fMRI data analysis

We performed a Monte Carlo simulation study to assess the performance of the proposed model and compare it to the standard Bayesian fPCA model in terms of curve reconstruction and classification. We generated three groups (Group 1.1, 1.2 and 2) of $n = 100$ time series of length $T = 150$. We applied a random Gaussian noise and tested the models with high and low signal-to-noise ratios (STN).
Results of curve reconstruction in the high noise scenario (STN=1) show improvements in the Integrated Mean Square Error (IMSE) of all curves in the proposed model compared to the standard model; results of low noise scenario also support the use of the proposed model with 80% of curves with IMSE improved (Figure 1). Similar results were also obtained from correlation reconstruction.

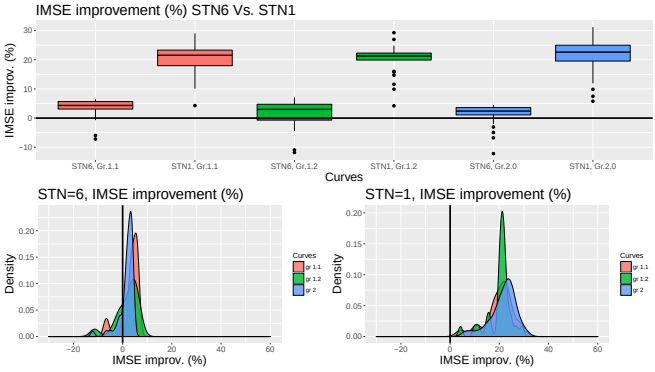


FIGURE 1. Simulation study: curve reconstruction. IMSE distribution stratified for noise level (STN=6, STN=1) and curve group (1.1, 1.2, 2).

A thirty year old healthy woman volunteered for the fMRI study. She underwent a resting-state recording at the Department of Radiology, Scientific

Institute Santa Maria Nascente, Don Gnocchi Foundation (Milan, Italy). After preprocessing, one minute length time series were extracted according to the Automated Anatomical Labeling (AAL90) coordinates. The resulting dataset was input to fPCA for curve smoothing and dimension reduction first and subsequently analysed with the proposed model. We identified 21% of curves in the first eigendimension belonging to a separate cluster (Figure 2, panel B). These curves pertain to brain areas from the occipital, parietal and temporal lobe which are known to be highly involved in resting-state brain networks (Figure 2, panel C-D).
Overall, results from simulation and fMRI analyses support the usefulness of fPCA clustering in curve reconstruction and exploration of complex spatio-temporal patterns in neuroscientific data.

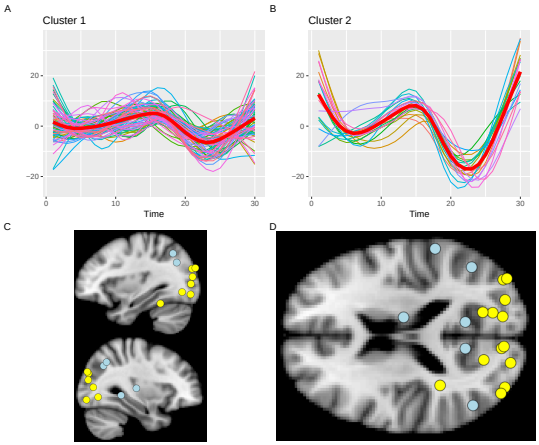


FIGURE 2. fMRI data analysis: cluster identification in the 1st eigendimension according to MAP and pairwise probabilities. Panel A-B: curves in cluster 1 and 2 according to the fPC scores partition. Panel C-D: 3D localisation of cluster 2 over sagittal and axial slices of human brain. Blue dots: areas identified by MAP only.

References

Crainiceanu, C.M., and Goldsmith, A.J. (2010). Bayesian functional data analysis using WinBUGS *Journal of Statistical Software*, **32**(11).

Liu, C., Surajit, R., and Hooker, G. (2017). Functional principal component analysis of spatially correlated data. *Statistics and Computing*, **27**(6), 1639 – 1654.

Rasmussen, C.E. (2000). The infinite Gaussian mixture model. In: *Advances in neural information processing systems*, 554 – 560.

Non-parametric learning algorithm for evaluating the influence of environmental factors on sudden medical emergencies

Mátyás Constans[1], Attila Lovas[2], Péter Sótonyi[3], Brigitta Szilágyi[4]

[1] Department of Analysis, Budapest University of Technology and Economics, Budapest, Hungary
[2] Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest, Hungary
[3] Department of Vascular Surgery, Semmelweis University, Budapest, Hungary
[4] Department of Geometry, Budapest University of Technology and Economics, Budapest, Hungary

E-mail for correspondence: lovas@math.bme.hu

Abstract: We develop a non-parametric Cox process model for sparse events in time. By assuming that the incidence of certain medical emergencies is influenced by a stochastic process which we interpret as being the environment, our model can be applied to a population in which each entity reacts to a variety of different environmental factors in a similar way. Furthermore, the incidence of events follows an unknown global trend which can be tracked back to changes in the population such as migration, aging or changing habits. Moreover, these changes are supposedly much slower than fluctuations caused by the environmental parameters. We propose a generalized EM algorithm to infer the global trend and the influence of the environment. Finally, we demonstrate the capabilities of our methodology on real medical data.

Keywords: Statistical learning; epidemiology; non-parametric approach; Cox process model; EM algorithm.

1 Introduction

From birth to death, human body is continuously exposed to environmental factors such as weather and air pollution. Over the span of the last decades, statistical modeling has become a considerable part of research

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

in medical circles, its primary goal being the exact quantification of these effects. Many attempts have been made to understand the exact causal relationship between our surrounding environment and the effect it has on our bodies. Hence, a good quantitative model avers to be crucial since it can help us separate significant and negligible factors from each other and let's us draw adequate and more precise conclusions about the cause of those effects.

The most used analytical tool in epidemiology nowadays is the Generative Additive Model (GAM), which is used heavily in different studies relating to public health. It was originally developed by Trevor Hastie and Robert Tibshirani to mix the benefits of generalized linear models and additive models (Hastie and Tibshirani, 1986). These methods however suffer from serious limitations such as the presence of confounding variables and concurvity. Furthermore, by construction, complex non-linear interactions between explicative variables are ruled out (Jalila, 2011).

In order to alleviate these issues, we propose a robust non-parametric alternative based on a Cox process model, where the non-parametric intensity is the product of a multidimensional link function and a slowly varying hidden trend. The non-parametric nature of the model enables it to possibly learn complex trends that parametric models could not determine.

We demonstrated the efficiency of our algorithm using data sets concerning pulmonary embolism and weather reports in the concerned area. Our approach seems suitable for forecasting medical emergencies, provided that predictions for environmental parameters are available. Potential uses of this method include the personalization of asthmatic people's treatment via an application, considering their medical history and different external conditions. Finally, our model could predict the effect of different diseases on a specific population during climate change.

2 Description of the model

Let N_t be the registered number of events on the t^{th} day and $X_t \in \mathbb{R}^p$ the actual value of the environmental parameters, $t = 1, \ldots, T$, where p denotes the number of influencing factors taken into account and T is the length of the observation period. The conditional distribution of N_t given X_t is assumed to be Poisson with parameter $\lambda(t, X_t)$. We also assume that the intensity parameter can be written as a product in which the dependence of λ on t and x is separated, that is: $\lambda(t, x) = f(t)^2 g(x)$. Here $g : \mathbb{R}^p \to [0, \infty)$ describes the effect of X_t on N_t and $f : \mathbb{N} \to [0, \infty)$ captures the slowly varying hidden trend which we cannot observe directly. We propose the following EM type algorithm to infer the functional form of f and g:

E-step: We define $g^{(j+1)}(x) = \mathbb{E}(N_t \mid X_t = x) / \left(f_t^{(j)}\right)^2$, where the right hand side does not depend on t.

M-step: We also define $f^{(j+1)} \in \mathbb{R}^T$, which minimizes the utility function

$$U_{N|X;g^{(j+1)}}(f) = \sum_{t=1}^{T} f_t^2 g^{(j+1)}(X_t) - N_t \log(f_t^2 g^{(j+1)}(X_t)) + \beta \sum_{t=1}^{T} (f_t - f_{t-1})^2,$$

where $f_1 - f_0 := 0$ and $0 \log(0)$ is defined to be zero.

The utility function is essentially the negative log-likelihood function with a regularization term, that measures the complexity of f and the parameter $\beta > 0$ is responsible for avoiding overfitting.

Now, we prove that the expected utility function given $X_1, \ldots, X_T$ converges to a minimum. The M step minimizes the utility pointwise, hence

$$\mathbb{E}_{N|X}\left(U_{N|X;g^{(j+1)}}(f^{(j)})\right) \geq \mathbb{E}_{N|X}\left(U_{N|X;g^{(j+1)}}(f^{(j+1)})\right)$$

trivially holds. On the other hand, the E step does not increase the expected utility. We consider the estimate

$$\begin{aligned}
\Delta U &:= U_{N|X;g^{(j+1)}}(f^{(j)}) - U_{N|X;g^{(j)}}(f^{(j)}) \\
&= \sum_{t=1}^{T} \left(f_t^{(j)}\right)^2 \left(g^{(j+1)}(X_t) - g^{(j)}(X_t)\right) - N_t \log \frac{g^{(j+1)}(X_t)}{g^{(j)}(X_t)} \\
&\leq \sum_{t=1}^{T} \left[\left(f_t^{(j)}\right)^2 - \frac{N_t}{g^{(j+1)}(X_t)}\right] \left(g^{(j+1)}(X_t) - g^{(j)}(X_t)\right),
\end{aligned}$$

where we used $-\log x \leq 1/x - 1$, $x > 0$. By the definition of $g^{(j+1)}$, we have

$$\mathbb{E}\left[\left(f_t^{(j)}\right)^2 - \frac{N_t}{g^{(j+1)}(X_t)} \middle| X_t\right] = 0, \quad t = 1, \ldots, T$$

hence $\mathbb{E}_{N|X}(\Delta U) \leq 0$. Taking into account that the expected utility function is bounded from below:

$$\mathbb{E}_{N|X}\left(U_{N|X;g}(f)\right) \geq \sum_{t=1}^{T} \mathbb{E}(N_t \mid X_t) - \mathbb{E}(N_t \mid X_t) \log(\mathbb{E}(N_t \mid X_t)),$$

where the lower bound does not depend on f and g, we can conclude that $\mathbb{E}_{N|X}\left(U_{N|X;g^{(j)}}(f^{(j)})\right)$ converges almost surely as $j \to \infty$.

3 Implementation

Suppose that we are given a data set N_t, X_t, $t = 1, \ldots, T$. We set $f^{(0)} = (1, \ldots, 1)^T$. In the E step, we estimate g as follows:

$$\hat{g}^{(j+1)}(x) = \sum_{t=1}^{T} \frac{e^{-\alpha d(X_t, x)}}{\sum_{s=1}^{T} e^{-\alpha d(X_s, x)}} N_t / \left(\hat{f}_t^{(j)}\right)^2,$$

where $d : \mathbb{R}^p \times \mathbb{R}^p \to [0, \infty)$ is the Mahalonobis distance and $\alpha > 0$ is a model parameter. Such kind of estimators are popular in non-parametric statistics and under mild circumstances they are weakly consistent (Stone, 1977).

In the M step, $\hat{f}^{(j+1)} = \underset{f}{argmin}\, U_{N|X;\hat{g}^{(j+1)}}(f)$ is obtained by a nonlinear conjugate gradient method using the Polak–Ribière scheme (Polak and Ribière, 1969). Iteration is stopped when the Euclidian norm of the gradient became smaller than 10^{-3} or the number of iterations reached 100.

Let $\bar{X}$ be the time average of X_t, $t = 1, \ldots, T$. We define the hidden trend as $\hat{q}(t) = \hat{g}(\bar{X})\hat{f}(t)^2$ and introduce

$$\hat{r}(x) = \left(\frac{\hat{g}(x)}{\hat{g}(\bar{X})} - 1 \right) \times 100\%$$

that measures the relative percentage growth of $\hat{\lambda}(t, x)$.

4 Applications

In our previous work we analyzed fatal pulmonary embolism (PE) data, but with a different methodology. The target groups included cases of PE in the capital Budapest. Based on the database of the Department of Forensic and Insurance Medicine, Semmelweis University there were 23.892 cases autopsied between 1st January 2001 and 31st December 2010. Among these cases there were 467 PE defined as cause of death in this time period. Meteorological data were obtained from the gridded E-OBS datasets. Daily atmospheric air pressure and atmospheric air pressure change between days with PE death and the previous days were analyzed from the region of capital Budapest. We found that cumulative number of registered PE cases follows a power law in time, moreover, there is a definite link between the cold temperature and the increasing incidence of fatal pulmonary embolism. For a more detailed description of the survey, we refer the reader to Törő et. al (2016).

We performed a simulation on the same dataset with the parameters $\alpha, \beta = 1$. The tolerance bound for $\|\hat{f}_{\text{next}} - \hat{f}\|_\infty$ was set to 10^{-3}. Iteration converged in 3056 steps. We can see in Figure 1. that the regulation term becomes constant while the utility decreases rapidly.

Figure 2 shows the estimated percentage change in daily PE cases depending on the daily average temperature and atmospheric pressure change between the actual and previous day. Estimated intensities and cumulative trends are presented in Figure 3. So far, the present results align well with our previous findings. However, they provide a better picture of the relationship between weather and the incidence of fatal PE due to the non-parametric nature of the model.

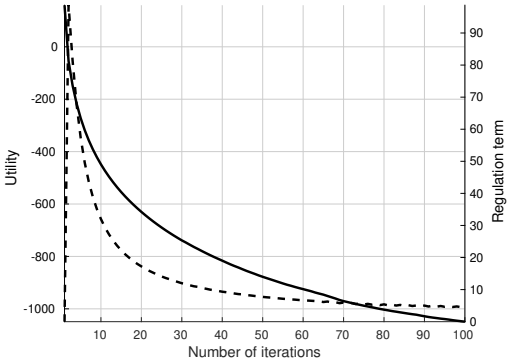


FIGURE 1. Utility function (left axis) and the regulation term (right axis).

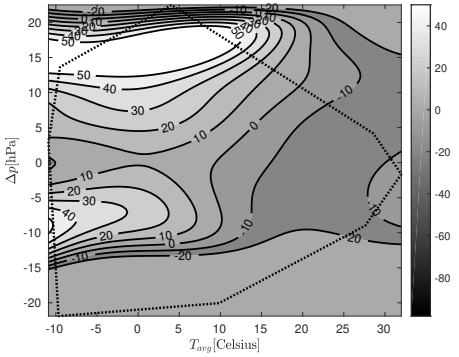


FIGURE 2. Estimated r-values. The interior of the convex hull (dotted line) of weather data can be considered as the domain of validity of the model.

5 Further Applications

We considered other uses for our results such as an application for asthmatic people that could prescribe a treatment based on the forthcoming weather conditions. A precise dosage of medication on days with abnormal or extreme weather conditions could have an overall huge impact on the number of medical emergencies on a larger population. Another use of our research could be found in the prediction of larger global phenomenons concerning the effect of diseases on a certain population considering local factors such as the weather and different specific regional traits.

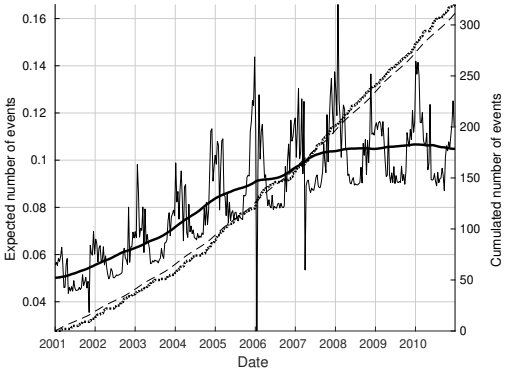


FIGURE 3. Estimated intensities (left axis): daily intensity modulated by weather – solid line, hidden trend – thick line. Cumulative number of events (right axis): estimated – dashed line, actual – dotted line.

Acknowledgments: This research was supported by the Higher Education Excellence Program of the Ministry of Human Capacities in the frame of Biotechnology research area of Budapest University of Technology and Economics (BME FIKP-BIO). The first author enjoyed the support of the Lendület grant LP 2015-6 of the Hungarian Academy of Sciences.

References

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, **1**(3), 297–310.

Jalila, J. and Adlouni, S.E. (2011). Generalized Additive Models in Environmental Health: A Literature Review. *IntechOpen*, DOI: 10.5772/38811.

Polak, E. and Ribière (1969). Note sur la convergence de directions conjuguèe. *Rev. Francaise Informat Recherche Operationelle*, **3**(16), 35–43.

Stone, C. J. (1977). Consistent Nonparametric Regression. *The Annals of Statistics*, **5**(4), 595–620.

Törő, K., Pongrácz, R., Bartholy, J., Váradi-T, A., Marcsa, B., Szilágyi, B., Lovas, A., Dunay, G. and Sótonyi, P. (2016). Evaluation of meteorological and epidemiological characteristics of fatal pulmonary embolism. *International Journal of Biometeorology*, **60**, 351 – 9.

Boosting health-related quality of life via distributional beta regression

Andreas Mayr[1], Leonie Weinhold[1], Stephanie Titze[2], Matthias Schmid[1]

[1] Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, Germany
[2] Department of Nephrology and Hypertension, FAU Erlangen-Nürnberg, Germany

E-mail for correspondence: amayr@uni-bonn.de

Abstract: We present a distributional beta regression approach to analyse bounded outcome variables and combine it with an adapted boosting algorithm. The approach allows to model both, the expected value and the precision parameter based on covariates. The boosting algorithm leads to data-driven variable selection and works for high-dimensional data, while the resulting model is in the same way interpretable as if it was fitted via classical inference schemes. We analyse the health-related quality of live of patients with chronic kidney disease from an newly developed German registry, focusing on variable selection and interpretation of effects.

Keywords: beta regression; bounded outcome; variable selection.

1 Introduction

The modelling of interval-bounded outcome variables like proportions or scores is a common issue in practical data analysis. In this work we will focus on health-related quality of life scales that are usually bounded between 0 (lowest possible quality of live) and 100 (highest possible value). One possibility is to transform the response in order to use classical Gaussian regression approaches, however, a more suitable technique is to directly apply beta regression (Ferrari and Cribari-Neto, 2004). The most common parametrization of the beta distribution in this context is

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1 \ ,$$

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

with mean parameter $0 < \mu < 1$ and the precision $\phi > 0$, leading to $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \frac{\mu(1-\mu)}{1+\phi}$.

2 Distributional beta regression

In classical beta regression, the mean parameter $\mu_i(x_i) = \mathrm{E}(y|x = x_i)$ is modelled – treating the dispersion parameter ϕ as fixed and nuisance (Ferrari and Cribari-Neto, 2004). Distributional beta regression, in contrast, follows the framework of Generalized Additive Models for Location, Scale and Shape (GAMLSS, Rigby and Stasinopoulos, 2005) relating also the precision parameter to an additive predictor:

$$\begin{aligned} \log\left(\frac{\mu_i}{1-\mu_i}\right) &= \eta_\mu(x_i) &&= \beta_{0\mu} + \sum_{j=1}^{p_\mu} f_{j\mu}(x_{ij}) \\ \log\left(\phi_i\right) &= \eta_\phi(x_i) &&= \beta_{0\phi} + \sum_{j=1}^{p_\phi} f_{j\phi}(x_{ij}) \end{aligned}$$

As a result, one estimates two different additive predictors, η_μ and η_ϕ, representing the two distribution parameters. This approach is favourable when explanatory variables do not only affect the location of the outcome distribution but also its shape. Additionally, it can lead to case-specific prediction intervals, as not only the center of the interval but also its size depends on covariates (Mayr et al., 2012).

3 Boosting algorithm

We applied a modified component-wise gradient boosting algorithm with linear as well as spline-based non-linear base-learners for distributional regression (Mayr et al., 2012). Each base-learner refers to one candidate variable and is fitted one-by-one to the gradient of the Likelihood. Due to this design, the algorithm works also for high-dimensional data with more candidate variable p than observations n.

The iterative algorithm basically circles through the different parameter dimensions and computes any possible update – carrying out only the best-performing one with respect to the increase of the likelihood (Thomas et al., 2018). By stopping the algorithm before it converges, variables (and base-learners) that have never been selected to be updated are effectively excluded from the final model. This way, we incorporate automatic data-driven variable selection that works simultaneously for η_μ and η_ϕ.

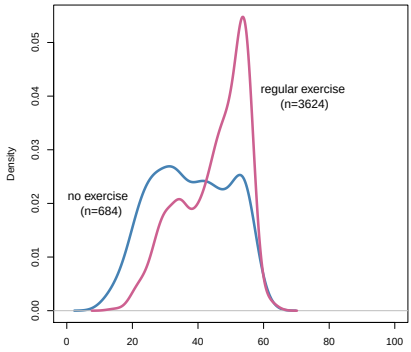


FIGURE 1. Estimated density of the health related quality of life measured via the Physical Composite Score (PCS) in the German chronic kidney disease study: Separate densities were plotted for patients regularly doing physical exercise ($n = 3264$) and patient without regular exercise ($n = 684$).

4 German chronic kidney disease study

We analysed the health-related quality of life in chronic kidney disease patients from an ongoing cohort study ($n = 3,947$) . The outcome variable is the Physical Composite Score (PCS), theoretically ranging from 0 to 100. The distribution of the PCS values at baseline is left-skewed and shows clear deviations from normality which also depends on explanatory variables (see Figure 1).

The set of potential explanatory variables consists of socio-demographic variables, clinical variables and laboratory measurements obtained from blood and urine samples. Altogether, there are 54 explanatory variables, leading to $2^{54} > 10^{13}$ potential predictor combinations for η_μ and η_ϕ.

To evaluate variable selection properties of our approach, we used the bootstrap to generate 1000 random samples. On each of the bootstrap samples we fitted both a classical beta-regression model and an distributional one for the patient's PCS score. All 54 explanatory variables were considered as potential predictors (continuous variables as spline effects, categorical variables as categorical effects). The distributional regression model yielded higher pseudo R^2 values (on test data) than classical beta regression with fixed ϕ.

Regarding variable selection, Figure 2 displays the selection rates for η_μ and η_ϕ. One can clearly observe, that more potential predictors are included in the mean model than in the precision part. The average size of mean model was 26 explanatory variables (median, range:12-46) while the average size of the precision model was 13 explanatory variables (median,range: 4-36).

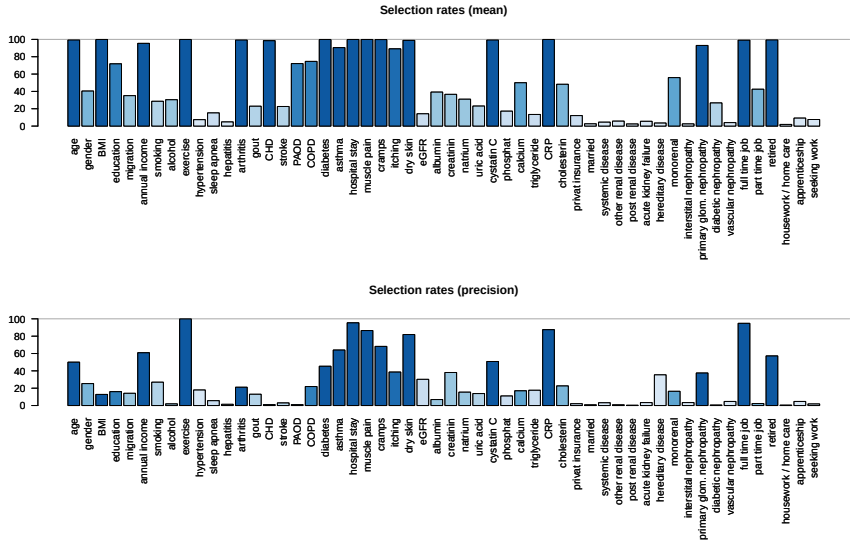


FIGURE 2. Selection rates of the candidate variables both for η_μ (upper plot) and η_ϕ (lower plot) estimated from 1000 bootstrap samples.

Furthermore, there is a clear tendency to include mostly variables in η_ϕ, that have also an effect on the mean.
The variable selection itself seems to be rather stable, many predictors are selected in almost all bootstrap samples. One of these variable is the physical exercise, Figure 3 displays the partial effects on both model-domains after re-fitting the model on the complete data set. The effect estimate shows a positive effect of exercise on the quality of life as well as a higher variance when the patients do not regularly exercise (cf., the empirical distribution in Figure 1).

5 Implementation

Distributional beta regression could be estimated via various inference schemes. The standard `betareg` package provides a simple and fast implementation for linear models (Cribari-Neto and Zeileis, 2010) but also the classical framework for distributional regression (`gamlss`, Rigby and Stasinopoulos, 2005) as well as the Bayesian counterpart (`bamlss`, Umlauf et al., 2018) contain beta regression as a special case.
Our approach is implemented in the R add-on package `betaboost` (Mayr et al., 2018), building up on the general boosting implementations `mboost` and `gamboostLSS`. This new package aims at facilitating the usage of boosting for practitioners, trying further to bridge the gap between methodology and

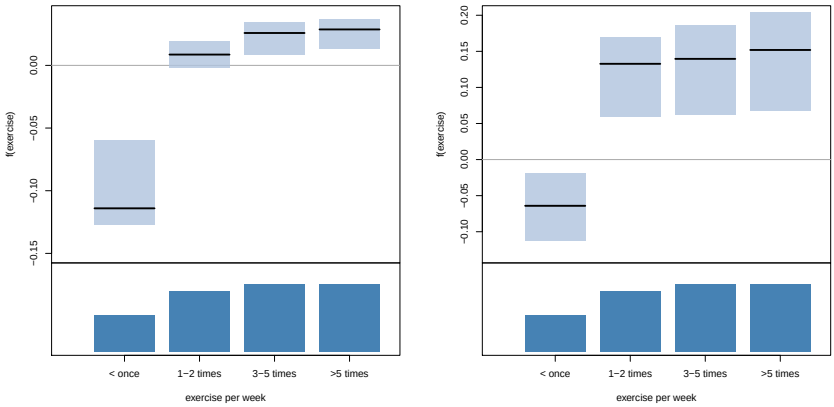


FIGURE 3. Resulting partial effect of physical exercise in η_μ (left plot, logit link) and η_ϕ (right plot, log link), the lower bar-plots reflect the variable's empirical distribution. The grey confidence regions were estimated via 1000 bootstrap samples.

applications (in the spirit of Groll et al., 2018), sparing the user unnecessary technical details. For example, the user can include the model formula in a very similar way to the classical packages, without specifying different base-learners as it is typically done in boosting implementations.

6 Discussion

We have presented a boosting approach for distributional beta regression which can be used to model health related quality of life scales. The advantage of our algorithm is, that it can carry out automated variable selection and works for high-dimensional data. Although relying a machine learning algorithm, it leads to interpretable statistical models. A limitation of our approach is the lack of standard errors for effect estimates, making it necessary to use work-around methods like resampling procedures or permutation tests to construct confidence intervals or p-values (Hepp et al., 2019), leading to longer run-times.

Further research is warranted on enhancing the variable selection properties of the algorithm. The resulting models were relatively big: The algorithm selected on average half of the candidate variables for the mean model. One way do deal with this could be to incorporate the stability selection approach on top (cf. Thomas et al., 2018). However, a more elegant solutions might tackle the problem in the algorithm itself, e.g., by adding an additional penalty for updates on variables that up-to-this iteration have never been selected.

Acknowledgments: The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, SCHM2966/1-2) and the Interdisciplinary Center for Clinical Research of the FAU Erlangen-Nürnberg (IZKF, project J49). The GCKD study was funded by grants from the German Ministry of Education and Research (BMBF, 01ER0804) and the KfH Foundation for Preventive Medicine.

References

Cribari-Neto, F. and Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, **34**(2), 1-24.

Ferrari, S.L.P and Cribari-Neto, F. (2004). Beta-regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799-815.

Groll, A., Kneib, T. and Mayr, A. (2018). Editorial 'Bridging the gap between methodology and applications: Tutorials on semiparametric regression'. *Statistical Modelling*, **18**(3-4), 199-202.

Hepp, T., Schmid, M. and Mayr, A. (2019). Significance Tests for Boosted Location and Scale Models with Linear Base-Learners. *The International Journal of Biostatistics*, doi: 10.1515/ijb-2018-0110.

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.

Mayr, A., Weinhold, L., Hofner, B., Titze, S., Gefeller, O. and Schmid, M. (2018). The betaboost package – a software tool for modelling bounded outcome variables in potentially high-dimensional epidemiological data. *International Journal of Epidemiology*, **47**(5), 1383-1388.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**, 507-554.

Thomas, J., Mayr, A., Bischl, B. , Schmid, M., Smith, A. and Hofner, B. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, **28**, 673-687.

Umlauf, N., Klein, N. and Zeileis, A. (2018), BAMLSS: Bayesian Additive Models for Location, Scale and Shape (and Beyond). *Journal of Computational and Graphical Statistics*, **27** (3), 612-627.

Stable Randomized Generalized Autoregressive Conditional Heteroskedastic Models

Pedro A. Morettin, Jhames M. Sampaio

[1] University of São Paulo, Brazil
[2] Federal University of Bahia, Brazil

E-mail for correspondence: pam@ime.usp.br

Abstract: In this paper, we propose the class of RS-GARCH models, an extension of the R-GARCH models, where both returns and volatility have stable distribution. We present the indirect inference method to estimate the RS-GARCH models, some simulations and an empirical application.

Keywords: Stable distribution; RS–GARCH models; Indirect inference.

1 Introduction

The main motivation for introducing the α-stable distribution family is that it allows asymmetry, tails much heavier than other popular distributions (as Student's t) and it is closed under linear combinations. Another important characteristic of stable distributions is its ability to accommodate the leptokurtic feature present in financial data.
There is a simple way to obtain the characteristic function of a stable distribution,

$$\ln \Phi_X(t) = \begin{cases} it\mu - \sigma^\alpha |t|^\alpha \left[1 - i\beta \mathrm{sgn}(t) \tan\left(\frac{\pi\alpha}{2}\right)\right], & \text{if } \alpha \neq 1, \\ it\mu - \sigma|t| \left[1 + i\beta \mathrm{sgn}(t) \ln(t)\right], & \text{if } \alpha = 1, \end{cases} \quad (1)$$

and depends on four parameters: $\alpha \in (0, 2]$, measuring the tail thickness (thicker tails for smaller values of the parameter), $\beta \in [-1, 1]$ determining the degree and sign of asymmetry, $\sigma > 0$ (scale) and $\mu \in \mathbb{R}$ (location). The distribution will be denoted as $S_\alpha(\sigma, \beta, \mu)$.
Weron and Weron(1995) proposed an algorithm that makes quite straightforward to simulate stably distributed pseudo-random numbers.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

2 SR–GARCH models

This work proposes the hypothesis of stability for the errors of the model, so we will be able to choose the shape of our distribution, it being more or less heavy-tailed and more or less leptokurtic. We will call this class RS–GARCH and we will propose the indirect inference to estimate the model parameters in addition to generalize some results of R-GARCH models.
The RS-GARCH(r,p,q), (r,p,q $\in \mathbb{N}$) model is defined by the equations

$$r_t = h_t^{1/\lambda}\epsilon_t,\ 1 < \lambda \leq 2,\ t = 0, \pm 1, \pm 2, \cdots, \tag{2}$$

$$h_t = \sum_{i=1}^{r} \theta_i \eta_{t-i} + \sum_{j=1}^{p} \phi_j h_{t-j} + \sum_{k=1}^{q} \psi_k |r_{t-k}|^{\lambda}, \tag{3}$$

where $r \geq 1, p \geq 0, q \geq 0, \theta_r > 0, \theta_i \geq 0, i = 1, \cdots, r-1, \phi_p > 0, \phi_j \geq 0, j = 1, \cdots, p-1, \psi_q > 0, \psi_k \geq 0, k = 1, \cdots, q-1$, the innovations ϵ_t are i.i.d. $S_\lambda(\sigma, 0, 0)$ random variables, the innovations η_t are positive i.i.d. random variables and $\{\epsilon_t\}$ and $\{\eta_t\}$ are independent.
It is easy to see that the distribution of r_t conditioned on $\mathcal{F}_{t-1} = \sigma\{\epsilon_t, \eta_t : s \leq t-1, s \in \mathbb{Z}\}$, is α-stable $r_t|\mathcal{F}_{t-1} \sim S_\lambda\left(\sigma h_t^{1/\lambda}, 0, 0\right)$. Thus, the conditional expectation $E(r_t|\mathcal{F}_{t-1})$ is constant and equal to zero and the conditional variance h_t of the models RS-GARCH depends on the past as a linear function of past innovations $\eta_{t-1}, \ldots, \eta_{t-r}$, the conditional variances $h_{t-1}, \ldots, h_{t-p}$ and also of past observations of returns $|r_{t-1}|^\lambda, \ldots, |r_{t-q}|^\lambda$.

3 The RS-GARCH model with stable innovations

Let us assume the RS-GARCH(r,p,0) model, with strictly stable random variables η_t totally skewed to the right, distributed as

$$\eta_t \sim S_{\alpha/2}\left(2\left(\cos\frac{\pi\alpha}{4}\right)^{2/\alpha}, 1, 0\right), \tag{4}$$

where $0 < \alpha < 2$. This means that the index of stability of η_t is smaller than one and the first moment does not exist.

Theorem 2.1. The unconditional distribution of r_t in the defined process RS-GARCH(r, p, 0) with stable innovations given by (4) is symmetric stable

$$r_t \sim S_{\frac{\lambda\alpha}{2}}\left(2^{\frac{1}{\lambda}}\sigma\left(\sum_{j=1}^{\infty}\delta_j^{\alpha/2}\right)^{\frac{2}{\lambda\alpha}}, 0, 0\right). \tag{5}$$

If we think the SR-GARCH model as representing daily returns, one could be interested about the unconditional distributions of weekly, monthly, etc.

returns. Defining m as the number of trading days within the given interval, the theorem below give us the desired distribution.

Theorem 2.2. The uncoditional distribution of the sum $\sum_{k=0}^{m-1} r_{t-k}$, $m \geq 1$, in the RS-GARCH(r,p,0) process with innovations η_t given by formula (4) is symmetric α-stable

$$\sum_{k=0}^{m-1} r_{t-k} \sim S_{\frac{\lambda\alpha}{2}} \left(2^{\frac{1}{\lambda}} \sigma \left[\sum_{j=1}^{m-1} \left(\sum_{i=1}^{j} \delta_i \right)^{\alpha/2} + \sum_{j=m}^{\infty} \left(\sum_{i=j-m+1}^{j} \delta_i \right)^{\alpha/2} \right]^{\frac{2}{\lambda\alpha}}, 0, 0 \right). \tag{6}$$

After these properties it is desirable that the process is stationary. Indeed, this is confirmed in the corollary below.

Corollary 2.1. The RS-GARCH(p,q,0) process with innovations given by formula (4) is symmetric α-stable and stationary, $0 < \alpha < 2$ and $1 < \lambda \leq 2$.

4 Indirect inference for RS-GARCH process

The main purpose of this section is to introduce brieflythe indirect inference approach to estimate the parameters of a SR-GARCH model. As we observed some interesting asymptotic properties for the model SR-GARCH(1,1,0) we will illustrate the idea for this model, with the innovations ϵ_t i.i.d. $S_\lambda(\sigma, 0, 0)$, the innovations η_t i.i.d. stable random variables distributed as (4) and $\{\epsilon_t\}$ and $\{\eta_t\}$ are independent.
The indirect inference has the potential to be an intensive computationally method to overcome difficulties associated with stable distributions. Gourieroux et al. (1993), Lombardi and Calzolari (2009) and Sampaio and Morettin (2015) are some recent works that confirm this claim and we take them as basic references for indirect inference.
Here, we have the likelihood of the model of interest (IM),, which is not available or difficult to handle. We consider an auxiliary model (AM), which has a likelihood that is easy to handle. The idea is to use simulation going to the AM and back to the IM model until convergence occurs. If θ is the vector of parameters of the IM and ζ is the vector of parameters of the AM, the procedure involves updating the parametrs in order to minimize

$$\left[\hat{\zeta} - \hat{\zeta}_S\right]' \Omega \left[\hat{\zeta} - \hat{\zeta}_S\right],$$

where Ω is a symmetric non negative definite matrix defining the metrics. For a given estimate $\hat{\theta}^{(p)}$, the procedure yields $\hat{\theta}^{(p+1)}$; this is then repeated until the series of estimates $\hat{\theta}^{(p)}$ converges.
For proper implementation of the method the first question to be answered is the identification of an appropriate AM. The parameter vector of the

AM must be greater than or equal to the dimension of θ in order for the solution to be unique.
Since we have a simulation-based approach, we decided to use an auxiliary model which is close to a SR-GARCH(1,1,0) model, namely the GARCH(1,1) model with Student's t innovations.
Denoting the parameters of the AM by $\zeta = (\nu, \alpha_0, \beta_1)$ and the parameters of the original model by $\theta = (\alpha, \theta_1, \phi_1)$ we see that $\dim \zeta = \dim \theta$ and therefore the indirect inference estimator is independent of the symmetric nonnegative matrix Ω. Thus we choose Ω as the identity matrix.

5 Simulation

In this section we conduct a simulation in order to investigate the properties of indirect inference estimators of the parameters α, θ_1 and ϕ_1 for the SR-GARCH(1,1,0) model. As we wish a leptokurtic distribution for the residuals and we desire a more efficient estimation, we will choose the values $\lambda = 1.8$ and $\sigma = 0.5$ for the parameters.
The estimates were based on 100 independent replications, 10,000 observations, $\nu = 4$ for the AM and taking $S = 10$ vectors. We can see the chosen values and their estimates in Table 1. We decided to use a large number of observations, since this will be the case for high frequency data. In the same way, we have chosen θ_1 small, since this will be the case in practical situations. We see, from the table, that the bias and standard errors are small, showing the good performance of the method. ,

	$\boldsymbol{\alpha = 1.5}$	$\boldsymbol{\theta_1 = 0.0000001}$	$\boldsymbol{\phi_1 = 0.7}$
Mean	1.5003	1.3718×10^{-7}	0.7586
Standard error	0.0214	3.7875×10^{-8}	0.0233
	$\boldsymbol{\alpha = 1.6}$	$\boldsymbol{\theta_1 = 0.0000001}$	$\boldsymbol{\phi_1 = 0.6}$
Mean	1.5985	1.6802×10^{-7}	0.6245
Standard error	0.0291	2.6712×10^{-8}	0.0134
	$\boldsymbol{\alpha = 1.7}$	$\boldsymbol{\theta_1 = 0.0000001}$	$\boldsymbol{\phi_1 = 0.8}$
Mean	1.7091	1.4225×10^{-7}	0.8472
Standard error	0.0116	4.86×10^{-8}	0.0251

TABLE 1. Monte Carlo mean and Standard error for different parameter values.

6 An application

In this section we are interested in evaluating the performance of the SR-GARCH(1,1,0) model relative to the GARCH models, also relative to the R-GARCH and R-GARCH-t models which were discussed in Sampaio and

Morettin (2015). To do this we will adjust the model SR-GARCH(1,1,0) to intraday Brazilian stock index Ibovespa logreturn series, sampled at each 15 minutes based on 100 independent replications and $S = 4$ vectors assuming the stable distribution $S_{1.8}(0.5, 0, 0)$ for the error. The total of 37,960 observations are taken from April 1998 to June 1998. We can see the logreturns in Figure 1.

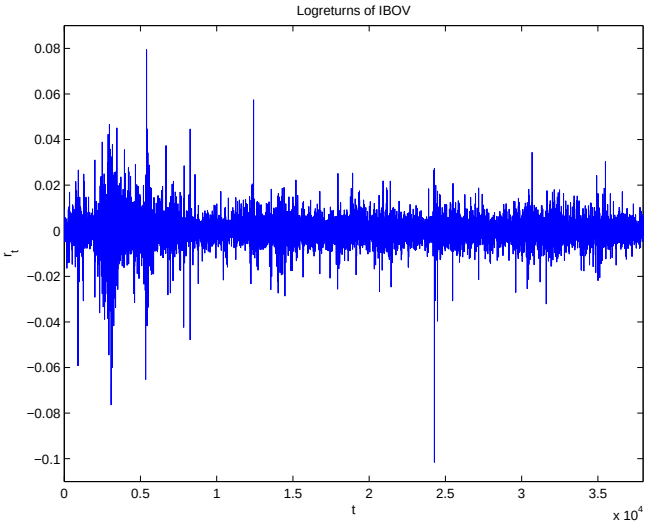


FIGURE 1. Logreturns of Ibovespa.

We present, in Table 2, the Monte Carlo mean values and standard errors of the estimated parameters $\hat{\alpha}$, $\hat{\theta}_1$ and $\hat{\phi}_1$ where $\nu = 3$ in the auxiliary model.

	$\boldsymbol{\alpha}$	$\boldsymbol{\theta_1}$	$\boldsymbol{\phi_1}$
Mean	1.5031	1.5140×10^{-6}	0.7578
Standard error	0.0040	2.3607×10^{-7}	0.0036

TABLE 2. Monte Carlo mean and standard errors parameters values for the fitted SR-GARCH(1,1,0) model.

Next, in Table 3, follows the comparison of the mean squared error for the respective models GARCH(0,1), R-GARCH(1,1,0), R-GARCH-t(1,1,0) and SR-GARCH(1,1,0).
We can observe that the mean square error of the SR-GARCH(1,1,0) model is somewhat lower than others. We present the estimated volatility in Figure 2.

MSE for GARCH(0,1)	2.4778×10^{-5}
MSE for R-GARCH(1,1,0)	3.5512×10^{-6}
MSE for R-GARCH-t(1,1,0)	6.7555×10^{-7}
MSE for SR-GARCH(1,1,0)	6.2983×10^{-7}

TABLE 3. MSE of residuals from the fitting of GARCH(0,1), R-GARCH(1,1,0), R-GARCH-t(1,1,0) and SR-GARCH(1,1,0) to the Ibovespa logreturns.

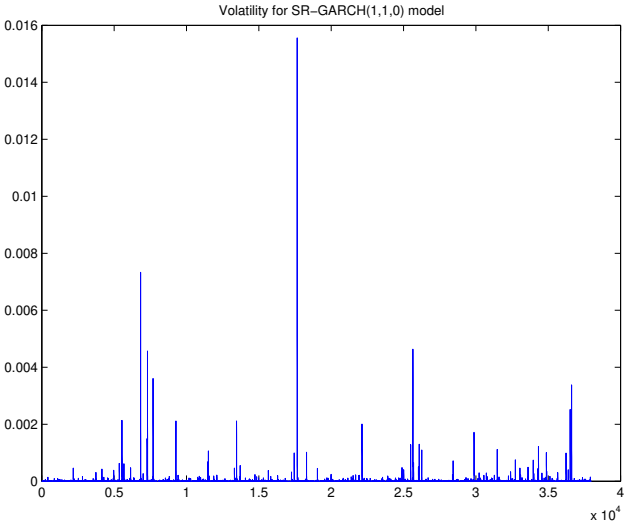


FIGURE 2. Estimated volatility for SR-GARCH(1,1,0) model.

Acknowledgments: Special thanks to Fapesp grant 2018/04654-9

References

Gourieroux, C., Monfort, A. and Renault, E. (1993), Indirect inference. Journal of Applied Econometrics 8: 85-118.

Lombardi, M. and Calzolari, G. (2009), Indirect estimation of alpha-stable stochastic volatility models. Computational Statistics and Data Analysis.Vol 53; Issue 6: 2298–2308.

Sampaio, J.M. and Morettin, P.A. (2015), Indirect estimation of randomized generalized autoregressive conditional heteroskedastic models. Journal of Statistical Computation and Simulation 85: 2702–2717.

Weron, A. and Weron, R. (1995), Lectures of Levy alpha-stable Variables and Processes. Lectures Notes in Physics 457 Springer-Verlag, pp. 379–392.

Intervention analysis based on INAR models with applications in public health

David Moriña[1], Juan M. Leyva-Moral[2], Maria Feijoo-Cid[2], Pedro Puig[1]

[1] Barcelona Graduate School of Mathematics (BGSMath), Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB), Spain
[2] Nursing Department, Faculty of Medicine, Universitat Autònoma de Barcelona (UAB), Spain

E-mail for correspondence: david.morina@uab.cat

Abstract: It is common in many fields to be interested in the evaluation of the impact of an intervention over a particular phenomenon. In the context of classical time series analysis a possible choice might be intervention analysis. In this work, we propose a modified INAR model that allows to quantify the effect of an intervention.

Keywords: count data; INAR models; intervention analysis; public health.

1 Introduction

This work is focused on the evaluation of the impact of an intervention over the number of occurrences of a particular phenomenon by using discrete time series techniques. Therefore, unlike in many other applications of time series, the main interest is not in forecasting but in the estimation of the effect of the intervention and its further inference. Many models of discrete time series have been considered in the literature (see McKenzie (2003)), although we focus on *Integer Autoregressive* ($INAR$) models, which are a natural extension to the well known AR models, and are often easily interpretable in practical contexts. It is usual in many contexts such as public health or sociology to design and conduct an intervention to change some phenomenon behaviour. When dealing with continuous valued time series or series with large counts, intervention analysis may be used with this purpose. When the time point where the potential change occurs is unknown, some authors have proposed the change-point analysis (see Csörgö

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

and Horváth (1997) and Horváth and Rice (1997)). However, for count data there is not a clear analogous methodology, although there have been some recent contributions (Liboschik et al. (2016), Vasileios (2015)) and an application of the change-point techniques to $INAR$ models (Hudecová et al. (2015)). Nonetheless, most of these models are focused on real-time monitoring for structural changes in the series while we are interested on transient or definitive changes in the new observed cases after an intervention through a retrospective analysis.

2 Model definition and goodness of fit

$INAR(k)$ models are defined by

$$X_t = p_1 \circ X_{t-1} + \ldots + p_k \circ X_{t-k} + W_t, \tag{1}$$

where $p_1, \ldots, p_k$ are fixed parameters, $0 < p_1, \ldots, p_k < 1$ and W_t is assumed to follow a Poisson distribution with a fixed mean λ. A recent review of INAR models can be found in Scotto et al. (2015). For each intervention we are interested in, we define a dummy variable I_t, which takes the value 1 if t is within an intervention period or 0 otherwise. The proposed model is a variation of $INAR(k)$ model (1):

$$X_t = p_1 \circ X_{t-1} + \ldots + p_k \circ X_{t-k} + W_t(\lambda^{'}), \tag{2}$$

where $\lambda^{'} = \lambda + \sigma \cdot I_t$. The parameters of the model (2), $\theta = (\lambda, \sigma, p_1, \ldots, p_k)$, can be estimated by using the method of conditioned maximum likelihood and the main interest in our context will be to test the null hypothesis $H_0 : \sigma = 0$ using the standard error corresponding to $\hat{\sigma}$, obtained from the inverse of the Hessian matrix. The proposed model is focused on detecting changes in the innovations (W_t), while a methodology for detecting changes in the parameters p_i, $i = 1, \ldots, k$ in (1) is proposed in Hudecová et al. (2015), introducing a method for monitoring structural changes in INAR(1) processes. The goodness of fit of the selected model can be assessed through a discretised version of the Cox-Snell residuals (Cox and Snell (1968)), computed from the estimated conditional distribution. The normal pseudo-residual segment $[z_n^-, z_n^+]$ can be obtained as

$$z_n^- = \Phi^{-1}(\hat{P}(Y_n < y_n \mid (Y_1, \ldots, Y_{n-1}, Y_{n+1}, \ldots, Y_T))) = \Phi^{-1}(u_n^-) \tag{3}$$

$$z_n^+ = \Phi^{-1}(\hat{P}(Y_n \leq y_n \mid (Y_1, \ldots, Y_{n-1}, Y_{n+1}, \ldots, Y_T))) = \Phi^{-1}(u_n^+), \tag{4}$$

where Φ is the standard normal distribution function. The mid-pseudo-residuals, defined by

$$z_n^m = \Phi^{-1}\left(\frac{u_n^- + u_n^+}{2}\right) \tag{5}$$

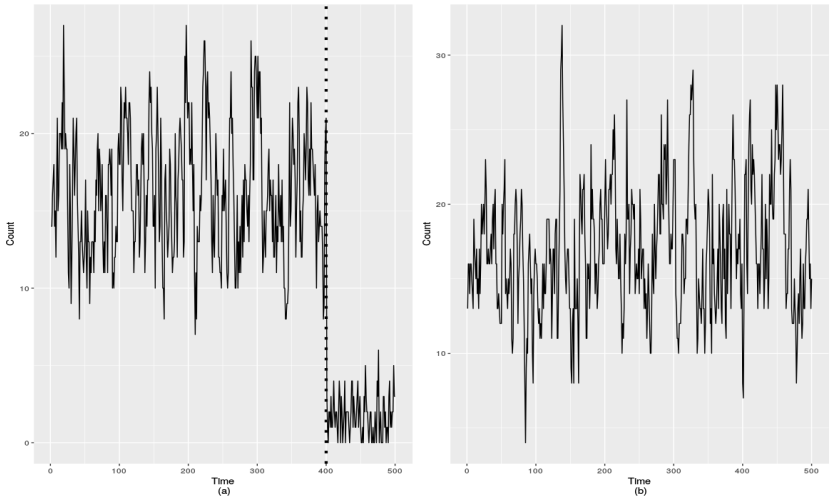
are commonly used in practice to check the validity of a fitted model, as a white noise behavior is expected.

3 Examples

3.1 Simulated data

Two INAR(1) processes were simulated, with parameters $p_1 = 0.7$ and $\lambda = 5$ for $t = 1, \ldots, 400$ and $p_1 = 0.7$ and $\lambda = 1$ for $t = 401, \ldots, 500$. The simulated series is shown in Figure 1 (a).

FIGURE 1. Simulated INAR(1) process with an intervention at time $t = 400$ (a) and no intervention (b).



Therefore, a change is expected to be detected for $t > 400$, and the fitted model is

$$X_t = p_1 \circ X_{t-1} + W_t(\lambda^{'}),$$

where $\lambda^{'} = \lambda$ for $t \leq 400$ and $\lambda^{'} = \lambda + \sigma$ for $t > 400$. The estimate for σ is -5.89 with 95% confidence interval (-6.82, -4.96), meaning that the intervention at time $t = 400$ had a significant impact on the series.

Similarly, an INAR(1) process consisting of 500 observations was simulated with parameters $p_1 = 0.7$ and $\lambda = 5$. The simulated process can be seen in Figure 1 (b). In that case, as expected, an estimate of $\hat{\sigma} = 0.39$ $(-0.25, 1.03)$ is obtained, which can be interpreted as no effect of the hypothetical intervention at time $t = 400$.

Another approach would be to test for structural changes in the series, following for example the methodology described in Hudecová et al. (2015). In this case, as the change in series (a) is structural and the series does not return to the pre-intervention values after the intervention, this methodology is able to detect a change at $t = 396$. No change is observed for series (b), as expected. It is important to notice that our approach is retrospective and the potential intervention time is known, while the real-time monitoring alternatives are able to estimate when the change occurred.

3.2 Venereal lymphogranuloma and massive events in Barcelona

Venereal lymphogranuloma (LGV) is a STI caused by the bacteria *chlamydia trachomatis*. Due to the popularity that the so called *circuit* parties have reached recently, especially among gay and bisexual men, the impact of these massive events over the number of cases of this and other STI is a public health concern. The analysed data correspond to the number of LGV cases registered in the Barcelona area from January 2007 to December 2014. The time evolution of these data is shown in Figure 2, and no trend or seasonal behavior is observed.

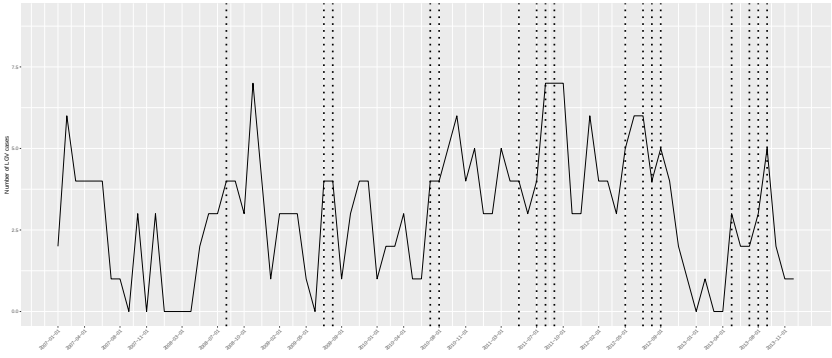


FIGURE 2. Observed number of LGV cases in Barcelona (2007-2014).

According to the ACF and PACF of the process a model of order 1 seems to be appropriate, so the model

$$X_t = p_1 \circ X_{t-1} + W_t(\lambda'), \tag{6}$$

where $\lambda' = \lambda + \sigma \cdot I(t)$ is proposed. In this case, the indicator variable takes the value 1 for all periods of time after the celebration of any circuit festival in Barcelona within the LGV incubation time (between 3 and 30 days). We have $\hat{\sigma} = 1.51$ $(0.55; 2.47)$ so a significant effect of the celebration of circuit parties over the number of new LGV cases is detected. The AIC of this model is 706.01, while the AIC corresponding to the standard INAR(1) model is 716.74. To check the goodness of fit of the model the mid pseudo-residuals approach was used, and the results are shown in Figure 3, supporting its suitability.

Following the approach described in Hudecová et al. (2015) no significant change in the parameters can be detected at a 95% confidence level. This fact highlights the difference between the two approaches, as the focus here is in detecting changes in the innovations due to an intervention or several interventions but then possibly returning to the original pre-intervention stage.

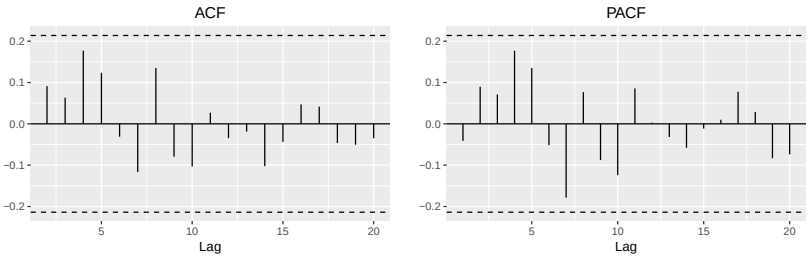


FIGURE 3. ACF and PACF of the mid-pseudo-residuals of the model for the number of LGV cases in Barcelona 2007-2013.

Acknowledgments: David Moriña acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the María de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445) and from Fundación Santander Universidades.

References

Cox, D. R. and Snell, E. J. (1968). A General Definition of Residuals. *Journal of the Royal Statistical Society, Series B*, **30(2)**, 248 – 275.

Csörgö, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. New York: Wiley & Sons.

Horváth, L. and Rice, G. (2014). Extensions of some classical methods in change point analysis. *Test*, **23(2)**, 219 – 255.

Hudecová, S., Huskova, M., and Meintanis, S. (2015). Detection of changes in inar models. In: *Stochastic Models, Statistics and Their Applications*, **122**, 11 – 18.

Liboschik, T., Fried, R., Fokianos, K., and Probst, P. (2016). tscount: Analysis of Count Time Series. *R package version 1.3.0.*

McKenzie, E. (2003). Discrete variate time series. In: *Handbook of statistics*, **21**, 573 – 606.

Scotto, M. G., Weiss, C. H., and Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling*, **15(6)**, 590 – 618.

Vasileios, S. (2015). acp: Autoregressive Conditional Poisson. *R package version 2.1.*

Selecting the Number of Factors in Bayesian Factor Analysis

Emmanuel Lesaffre[1], Luis Adrian Quintero[1], Geert Verbeke[1]

[1] KU Leuven, I-BioStat, Leuven 3000, Belgium

E-mail for correspondence: emmanuel.lesaffre@kuleuven.be

Abstract: We propose a Bayesian method to infer model dimensionality in factor analysis that is computationally not demanding, and where the ordering of the outcomes does not influence the solution. Implementation of our approach is simple via an efficient Gibbs algorithm. The performance of our approach is assessed via simulations and using a medical study related to burnout of nurses.

Keywords: Gibbs sampling; Model dimensionality; Spike-slab prior.

1 Introduction

Factor analysis (FA) aims to describe the dependence among a set of outcomes in terms of a lower number of latent factors. Selecting the number of factors underlying the data is, however, quite challenging in both frequentist and Bayesian approaches. Frequentist procedures for model dimensionality selection are based on the likelihood ratio test (LRT), which assumes that the factor loadings matrix is of full rank. If violated, the method retains too many factors. Similar problems are seen with AIC and BIC, where it is also not clear how to compute the penalty term.
Many Bayesian approaches have been suggested in the literature to determine number of factors. Lee and Song (2002) proposed to estimate the Bayes factor via path sampling for comparing factor models. Lopes and West (2004) proposed a reversible jump MCMC algorithm to move between models with different number of factors. In genomics, Carvalho et al. (2008) proposed an algorithm for factor analysis in high-dimensional settings. But, all of these approaches assume a lower triangular structure for the factor loadings matrix in order to ensure identifiability of the models. With this assumption, the order of the variables introduces unintended prior information. Recently, also other approaches have been suggested.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

However, all above approaches either still depend on the lower triangular representation, or suffer from identification issues or are computationally not straightforward. Recently, Quintero et al. (2018) suggested a method, which does not impose the lower triangular condition for the factor loadings matrix. Simulations indicate that our method performs substantially better than other Bayesian techniques and is efficient without much computational burden.

2 Factor Model Specification

The factor analytic model assumes that the $p-$dimensional observation $\mathbf{y}_i$ can be explained by a $m-$dimensional vector $\boldsymbol{\eta}_i \sim N_m(\mathbf{0}, \mathbf{I}_m)$ of latent factors as follows

$$\mathbf{y}_i = \boldsymbol{\Delta}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\boldsymbol{\Delta}$ is a $p \times m$ matrix of factor loadings, $\boldsymbol{\varepsilon}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ is a residual vector with covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\boldsymbol{\eta}_i$ is independent from $\boldsymbol{\varepsilon}_i$ for $i = 1, \ldots, n$. It is assumed here that the outcomes are standardized (mean=0, SD=1) avoiding the intercept in (1).
The marginal distribution of $\mathbf{y}_i$ integrating out the latent factors is $N_p(\mathbf{0}, \boldsymbol{\Omega})$ with $\boldsymbol{\Omega} = \boldsymbol{\Delta}\boldsymbol{\Delta}' + \boldsymbol{\Sigma}$. Hence, the dependence in the outcomes is exclusively explained by the common latent factors. In practice, the number of factors is smaller than the number of outcomes ($m < p$).
When the factor loadings matrix is not of full rank, i.e. $\text{rank}(\boldsymbol{\Delta}) = r < m$, the parameters in $\boldsymbol{\Sigma}$ are underidentified. Indeed, let $\mathbf{R}$ be a $m \times (m-r)$ matrix such that $\boldsymbol{\Delta}\mathbf{R} = \mathbf{0}_{p\times(m-r)}$ and $\mathbf{R}'\mathbf{R} = \mathbf{I}_{m-r}$. Then

$$\boldsymbol{\Omega} = \boldsymbol{\Delta}\boldsymbol{\Delta}' + \boldsymbol{\Sigma} = (\boldsymbol{\Delta} + \mathbf{M}\mathbf{R}')(\boldsymbol{\Delta} + \mathbf{M}\mathbf{R}')' + (\boldsymbol{\Sigma} - \mathbf{M}\mathbf{M}'), \tag{2}$$

for any $p\times(m-r)$ matrix $\mathbf{M}$ with mutually orthogonal rows. This represents a serious practical problem because it is not possible to determine a priori the maximum value of m. This may affect the validity of Bayesian results in a similar way as for LRT in the frequentist setting.

3 Sparse Model Representation

Instead of considering (2) as a difficulty in FA, it can be used as a tool to determine model dimensionality. For any $m \times m$ orthogonal matrix $\mathbf{P}$, model (1) can be re-expressed by its rotated solution $\mathbf{y}_i = \boldsymbol{\Delta}\mathbf{P}\boldsymbol{\eta}_i^* + \boldsymbol{\varepsilon}_i$ with $\boldsymbol{\eta}_i^* = \mathbf{P}'\boldsymbol{\eta}_i$. Now assume $r < m$, then there exists a $m \times (m-r)$ matrix $\mathbf{R}$ such that $\boldsymbol{\Delta}\mathbf{R} = \mathbf{0}_{p\times(m-r)}$ and $\mathbf{R}'\mathbf{R} = \mathbf{I}_{m-r}$. In addition, let $\mathbf{Q}$ be any $m \times r$ matrix for which $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_r$ and $\mathbf{Q}'\mathbf{R} = \mathbf{0}_{r\times(m-r)}$. Then, $\mathbf{P} = (\mathbf{Q}_{m\times r}\ \mathbf{R}_{m\times(m-r)})$ is orthogonal and model (1) can be re-expressed as

$$\mathbf{y}_i = \boldsymbol{\Delta}(\mathbf{Q}_{m\times r}\ \mathbf{R}_{m\times(m-r)})\boldsymbol{\eta}_i^* + \boldsymbol{\varepsilon}_i = (\boldsymbol{\Delta}\mathbf{Q}_{m\times r}\ \mathbf{0}_{p\times(m-r)})\boldsymbol{\eta}_i^* + \boldsymbol{\varepsilon}_i. \tag{3}$$

Thus, for overfitting models there must exist a rotated solution in which $(m-r)$ columns of the factor loadings matrix are null. We introduce auxiliary parameters $\nu_1, \ldots, \nu_m$ that control the null columns in the factor loadings matrix. More specifically, we fit the factor model

$$\mathbf{y}_i = \begin{pmatrix} \nu_1 \boldsymbol{\delta}_1 & \ldots & \nu_m \boldsymbol{\delta}_m \end{pmatrix} \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i = \boldsymbol{\Delta}\mathbf{N}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \tag{4}$$

with factor loadings matrix $\boldsymbol{\Delta}\mathbf{N}$ where $\boldsymbol{\delta}_k$ corresponds to the kth column of $\boldsymbol{\Delta}$ $(k = 1, \ldots, m)$ and $\mathbf{N} = \text{diag}(\nu_1, \ldots, \nu_m)$. When $\nu_k = 0$, the kth column $\nu_k \boldsymbol{\delta}_k$ of the factor loadings matrix is rendered null in model (4). If the number of factors is $r \leq m$, there exists a solution of model (4) with $\sum_k I\{\nu_k \neq 0\} = r$. For all other rotated solutions, $\sum_k I\{\nu_k \neq 0\} \geq r$. Our approach for inferring dimensionality in factor analysis (IDIFA) employs the inferential model (4), inducing prior sparsity for the ν_k components in order to obtain the representation (3) of the model. For this, we use a variable selection approach for the ν_k components.

4 Prior Specification

To find the "effective rank" of the factor loadings matrix via MCMC methods, we use the concept of "practically null" as defined by George and McCulloch (1988). Each component of $\mathbf{N}$ in (4) is assigned a normal mixture prior as

$$\nu_k | \gamma_k \sim (1-\gamma_k) N(0, \tau_k^2) + \gamma_k N(0, c_k^2 \tau_k^2) \text{ for } k = 1, \ldots, m, \tag{5}$$

with $p(\gamma_k = 1) = 1 - p(\gamma_k = 0) = \pi_k$. Setting τ_k^2 small and c_k^2 to be large implies that if $\gamma_k = 0$ then ν_k is "practically null" and when $\gamma_k = 1$ probably $\nu_k \neq 0$. The elements δ_{jk} $(j = 1, \ldots, p;\ k = 1, \ldots, m)$ are mutually independent having a standard normal prior distribution and are independent from ν_k. Hence, if $\gamma_k = 0$, all components of $\nu_k \boldsymbol{\delta}_k$ in (4) are "practically null" and the corresponding factor is effectively switched off. But, if $\gamma_k = 1$, the kth factor appears as important in the model. The "effective rank" of $\boldsymbol{\Delta}\mathbf{N}$ is then $r_\gamma = \sum_k \gamma_k$. We assume a constant inclusion probability for all columns, i.e. $\pi_1 = \cdots = \pi_m = \pi$ and assigned a prior $\pi \sim \text{Beta}(a/m, b)$ where m is the number of potential factors. The values a and b are selected such that a small number of relevant columns r_γ are preferred a priori, supporting a sparse rotated solution (3). An inverse gamma prior is chosen for the idiosyncratic variances, namely $\sigma_1^2, \ldots, \sigma_p^2 \sim IG(d, e)$. The density in the slab component corresponds to the distribution of the product of two normal variables, i.e. $\nu_k | \gamma_k = 1 \sim N(0, c_k^2 \tau_k^2)$ and $\delta_{jk} \sim N(0,1)$. Note that the slab prior is also peaked around zero. This is crucial because, for important factors with $\gamma_k = 1$, some of the factor loadings $\nu_k \delta_{jk}$ can still be sampled close to zero. Note that each underlying factor explains some of the outcomes in practice and not all factor loadings need to be considerably large.

In the IDIFA approach, we estimate only model (4) in contrast to other approaches in which it is necessary to estimate all models with 0, 1, ..., m factors. All prior distributions are conditionally conjugate in IDIFA, so the conditional posteriors for Gibbs sampling correspond to closed form distributions. In each Gibbs iteration, we compute the number of important columns in the factor loadings matrix as $r_\gamma = \sum_k \gamma_k$.

5 Performance of the IFIDA approach

A simulation study set up to evaluate the performance of our approach in comparison to other approaches, showed that IDIFA finds more often the correct number of factors.
We also evaluated our approach on the data of the RN4CAST project, which is a European nurse workforce study. Burnout was measured according to the Maslach Burnout Inventory (MBI) scale based on 22 items. All items are answered in terms of frequency on a seven-point scale. MBI assumes the presence of three latent components underlying the data: emotional exhaustion, depersonalization and reduced personal accomplishment. It is of interest to confirm if the hypothesized three factor model is valid in Belgian university hospitals. In this analysis we select a subset of 11 items and added the university hospital as a covariate. Implementing the IDIFA approach with $m = 4$ suggests a three factor model with posterior probability equal to 95%. This was also suggested using the Bayes factor via path sampling. However, when changing the ordering of items, the Bayes factor favored a two factor model in contrast to the IDIFA approach.

References

Quintero, A., Verbeke, G., and Lesaffre, E. (2018). Selecting the number of factors in Bayesian factor analysis. *Submitted*

Carvalho, C.M., Chang, J., Lucas, J.E. , Nevins, J.R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, **103**, 1438 – 1456.

George, E.I. and McCulloch, R.E. (1988). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881 – 889.

Lee, S.Y., and Song, X.Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika*, **29**, 23 – 39.

Lopes, H.F., and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**, 41 – 67.

Multivariate surveillance using a multivariate integer-valued time series model

Xanthi Pedeli[1,2], Dimitris Karlis[1]

[1] Athens University of Economics and Business, Greece
[2] Ca' Foscari University of Venice, Italy

E-mail for correspondence: xpedeli@aueb.gr

Abstract: In this paper, we suggest a multivariate integer-valued autoregressive model that allows for both serial and cross correlation between the series and can easily accommodate overdispersion and covariate information. Moreover, its structure implies a natural decomposition into an endemic and an epidemic component, a common distinction in dynamic models for infectious disease counts. Detection of disease outbreaks is achieved through the comparison of surveillance data with one-step-ahead predictions obtained after fitting the suggested model to a set of clean historical data.

Keywords: Count data; Integer-valued time series; Multivariate surveillance.

1 Surveillance using a new multivariate integer-valued autoregressive model specification

Traditionally, statistical models for health surveillance data aim to effectively capture the endemic and epidemic dynamics of disease risk. In principle, the endemic component explains a baseline rate of cases with stable temporal pattern. The epidemic component on the other hand aims to introduce infectiousness, that is explicit dependence between events. Therefore the epidemic component is driven by the observed past and is identified with the autoregressive part of the model (Meyer et al., 2017).
This additive decomposition of disease risk is well embodied in the multivariate integer-valued autoregressive model (Pedeli and Karlis, 2013) $\boldsymbol{X}_t = \boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \quad t \in \mathbb{Z}$, where $\boldsymbol{A}$ is assumed to be a $n \times n$ diagonal matrix with independent elements and $\{\boldsymbol{\epsilon}_t\}_{t\in\mathbb{Z}}$ is a sequence of non-negative integer-valued random vectors, independent of $\boldsymbol{A}\circ\boldsymbol{X}_{t-1}$, that follow jointly a discrete multivariate distribution. However, the assumption of a diagonal

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

matrix $\mathbf{A}$ weakens the ability of the model to capture the epidemic dynamics of disease risk since it ignores the relationship with time lag between series that is typical in disease transmission. Moreover, inference for this model is based on a pairwise likelihood approach which is not appropriate for prediction purposes.

To balance between effectiveness and attractiveness of the model, we consider here another specification. In particular, we assume that the correlation matrix $\mathbf{A}$ is non-diagonal and we relax the degree of complexity of the model by assuming that the innovation series $\boldsymbol{\epsilon}_t$, i.e. the endemic components, are uncorrelated. The resulting model admits a realistic epidemiological interpretation and is extremely advantageous in terms of practical implementation since the distribution of the innovations becomes a product of univariate mass functions.

The newly defined multivariate INAR(1) process can be used for modeling clean historical data and make one-step-ahead forecasts that can be used for prediction-based monitoring. The suggested outbreak detection statistical process comprises of two steps: In the first step, the available series of data in the set-up phase (clean historical data) is modeled through a multivariate INAR(1) process and a parameter vector of maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ is obtained. In the second step, the actually observed realization $\mathbf{x}_{t+1}$ is assessed against a multivariate prediction threshold derived from the model fitted in the first step in order to define whether an alarm should be triggered. More specifically, for each multivariate observation $\mathbf{x}_{t+1}$ in the operational phase, we estimate the one-step-ahead predictive distribution $\hat{P}(\mathbf{X}_{t+1} = \mathbf{x}_{t+1}|\mathbf{x}_t, \hat{\boldsymbol{\theta}})$, $\mathbf{x} \in \mathbb{N}_0^n$ and obtain the marginal predictive probabilities $\hat{P}(X_{i,t+1} = x_{i,t+1}|\mathbf{x}_t, \hat{\boldsymbol{\theta}})$, $i = 1, \ldots, n$. For each observation $X_{i,t+1}$, we construct an $(1-\alpha)\%$ prediction interval with upper bound $x_{i,t+1}^{UB}$ equal to the $(1-\alpha)$-quantile of the corresponding marginal predictive distribution, where α is a prespecified significance level. The lower bound of the prediction interval is set equal to 0 since we are only interested in detecting positive deviations from the in-control model. Each series flags an alarm at time $t+1$ if the corresponding observation lies outside the prediction interval, i.e. if

$$x_{i,t+1} > x_{i,t+1}^{UB}.$$

Finally, for the overall alarm, a majority rule can be defined, i.e. flagging an alarm if a certain percentage of the series signals an alarm at the same point in time (Vial et al., 2016).

2 Application using syndromic data

To illustrate the suggested methodology we consider data that is part of the syndromic surveillance data collected during Athens 2004 Olympic Games. For the purpose of the current analysis we consider three distinct syndromes recorded in a specific hospital that are significantly correlated to each other

(cross-correlations ranging from 0.31 to 0.48). In particular, we consider respiratory infection with fever, febrile illness with rash and other syndrome with potential interest for public health. The latter is a general category including all symptoms that could not be classified in any of the other prespecified categories.

Our monitoring period starts on March 2, 2004 and ends on September 28, 2004 while the period between August 1, 2002 and August 29, 2003 is considered as the set-up phase. During both periods syndromes were recorded every three days so that the historical and surveillance data consist of $t_0 = 127$ and $t_1 = 71$ observations respectively.

In the following we fit a trivariate INAR(1) model with indepedent Poisson innovations for modeling and prediction using the historical syndromic surveillance data. To account for regressors usually related to infectious disease data we express the expectation of the innovation series as function of the available covariate information, i.e. $E(\epsilon_{it}) = \exp(\mathbf{z}_t'\boldsymbol{\beta})$, $i = 1, 2, 3$, where $\mathbf{z}_t$ as a vector of covariates with associated regression parameters $\boldsymbol{\beta}$. As candidate covariates we consider terms for seasonality and a binary indicator for the day of the week on which the recording of syndromes was implemented (weekdays vs. weekends). We don't consider time trends since time series plots do not suggest the presence of any trend in our data. Therefore, each marginal series is modeled as $X_{it} = \sum_{j=1}^{3} \alpha_{ij} \circ X_{j,t-1} + \epsilon_{it}, i = 1, 2, 3$, where ϵ_{it} are independent Poisson random variables with mean

$$E(\epsilon_{it}) = \exp\left\{\beta_{i0} + \beta_{i1}\text{Weekday} + \beta_{i2}\cos\left(\frac{2\pi t}{122}\right) + \beta_{i3}\sin\left(\frac{2\pi t}{122}\right)\right\} \quad (1)$$

for $t = 1, \ldots, t_0$. For comparison purposes we also employ a univariate surveillance approach based on fitting three indepedent INAR(1) regression models with Poisson innovations. Covariate information is incorporated in the univariate models in the same way, i.e. through (1). With both approaches, the marginal one-step-ahead predictive distributions are used for the construction of $(1 - \alpha)\%$ prediction intervals, the upper bounds of which serve as thresholds for outbreak detection. We assume a type I error of $\alpha = 0.01$ and for the overall alarm we set a rule of 2/3 that is an alarm is triggered if at least two out of the three series flag an alarm at the same point in time.

The parameter estimates and corresponding standard errors obtained with the two modeling approaches indicate significant first-order autocorrelations under both fittings. The cross-correlation parameters estimated by the trivariate INAR(1) model are also significant indicating the appropriateness of the multivariate approach.

The surveillance plots obtained under the two models are shown in Figure 1. Red dashed lines represent the upper bounds of the corresponding 99% prediction intervals while blue crosses indicate the time points at which an overall alarm is raised. The two alarms signalled with the trivariate INAR(1) fitting are also trigerred when three independent INAR(1)

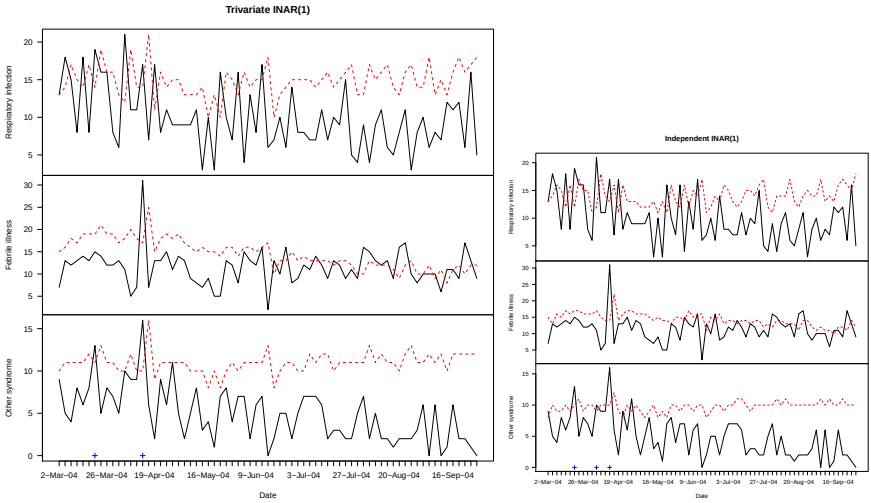


FIGURE 1. Surveillance plots as obtained after fitting a trivariate INAR(1) regression model with independent Poisson innovations (left panel) or three independent Poisson INAR(1) regression models (right panel) to the historical data.

models are fitted to the historical data. However, with the later approach an additional alarm is also raised. This additional alarm might be due to the ignorance of cross-correlation between the series that results in narrowing down the corresponding prediction intervals and thus increasing the number of false alarms.

Acknowledgments: This project has received funding from the Athens University of Economics, Action II Funding and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 699980.

References

Meyer, S. and Held, L. and Höhle, M. (2017). Spatio-Temporal Analysis of Epidemic Phenomena Using the `R` Package `surveillance`. *Journal of Statistical Software*, **77**, 1 – 55.

Pedeli, X. and Karlis, D. (2013). On composite likelihood estimation of a multivariate INAR(1) model. *Journal of Time Series Analysis*, **34**, 206 – 220.

Vial, F. and Wei, W. and Held, L. (2016). Methodological challenges to multivariate syndromic surveillance: a case study using Swiss animal health data *BMC Veterinary Research*, **12**, 288.

Fast sequential Bayesian analysis of football scores illustrating the evolution of the styles and strengths of each of the British premier league football sides over two decades

Ridall, P. G.[1], George, M.[1], Pettitt A. N. [2]

[1] Lancaster University, UK
[2] Queensland University of Technology, Australia

E-mail for correspondence: g.ridall@lancs.ac.uk

Abstract: West, Harrison and Wigan (1985) introduce a class of dynamic generalised linear models where dynamic updates of the sufficient statistics can be made through the exploitation of conjugacy. We extend this methodology to models where the dynamic parameters do not have sufficient statistics but where the full conditional posteriors of each parameter are known distributions. We illustrate our methods using premier league football data collected over the last two decades. We formulate updates for proxys for the sufficient statistics of each of the dynamic parameters. We validate our model, test its predictivity and examine critically our assumptions by examining the out of sample Pearson residuals. The outcome of our analysis are a set of informative trellis plots, showing the evolution of strength and style of each of the premier league sides over the last two decades.

1 Introduction

Our principal objective is to develop a sequential conjugate Bayesian dynamic model. Quasi-sufficient statistics are constructed for the dynamic parameters and are allowed evolve upon realisations of the scores of each game. Our model has similarities to Dixon and Coles (1997), Karlis and Ntzoufras (2003), Crowder et. al (2002) and Koopman and (2013). However unlike them we use a univariate Bayesian dynamic state space model and make dynamic updates avoiding sampling based methods. Our diagnostic analysis indicate that the univariate model gives an adequate description of the data. In section 2. we develop a Bayesian static model and apply

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

it just to one season's data. In section 3 we formulate our dynamic model and in section (4) we present our results and make conclusions.

1.1 Data description

We use the last 23 years of UK Premier League data from the web site, `http://www.football-data.co.uk/englandm.phhp`.

2 Theory and background

State space models have two components: a state equation that describes the evolution of the parameter of interest : $\theta_t \sim \pi(. \mid \theta_{t-1})$ and an observation equation $Y_t \sim f(. \mid \theta_t)$ which conditions on the underlying state. In order to make updates tractable, West et al., (1998) considered observation equations from the exponential family.

$$f(y_t|\theta_t, \tau_t) = h(y_t, \tau_t) \exp\left\{\tau_t \left[T(y_t)^T \eta(\theta_t) - b(\theta_t)\right]\right\}.$$

where θ_t is the dynamic parameter, $\eta(\theta_t)$ is the natural parameter, $b(\theta_t) \in \Re$ is the log partition function and τ_t is a dispersion parameter which can be assumed known. The conjugate prior is given by

$$\pi(\theta_t \mid \tilde{p}_t, \tilde{q}_t) = \frac{1}{c(\tilde{p}_t, \tilde{q}_t)} \exp\{\tilde{p}_t \eta(\theta_t) - \tilde{q}_t b(\theta_t)\}$$

where $\tilde{p}_t$ and $\tilde{q}_t$ are the prior sufficient statistics derived from the posterior from the last observation. The parameter $\tilde{q}_t$ is known as the prior precision parameter. The location parameter of the prior us $\mu_t = \frac{\tilde{p}_t}{\tilde{q}_t}$ Then we can show the posterior has the form

$$\pi(\theta_t \mid p_t, q_t) = \frac{1}{c(p_t, q_t)} \exp\{p_t \eta(\theta_t) - q_t b(\theta_t)\}$$

which ensures that the updates for the sufficient statistics for the state becomes

$$p_t \leftarrow \tilde{p}_t + \tau_t T(y) \qquad\qquad q_t \leftarrow \tilde{q}_t + \tau_t \tag{1}$$

Note that here that the updates of the sufficient statistics, p_t and q_t from prior to posterior distribution for densities from the exponential family involve quantities that are known. In this paper we look at extending this idea to models where no such sufficient statistics exist.

2.1 The evolution or state process

We use discount or a forgetting parameters to describe changes of form over time. The amount of forgetting, can be allowed to increase with increasing separation in the time between observations: $\omega_t = \exp(-k\Delta t)$, where k is a fixed parameter. We formulate the extension of the posterior of the last observation to the prior of the current observation by

$$\tilde{p}_t \leftarrow \omega_t p_{t-1} \qquad \tilde{q}_t \leftarrow \omega_t q_{t-1}$$

Note that although the parameter loses precision, the mean is left unchanged. The updating of the parameters for the prior to the parameters of the posterior as

$$p_t \leftarrow \tilde{p}_t + \tau_t T(y) \qquad q_t \leftarrow \tilde{q}_t + \tau_t.$$

3 Example: Premier League football

Our aim is to develop a good predictive model of football scores encompassing the style and strength of play of each team over the history of the league. Our data is sourced from http://www.football-data.co.uk/englandm.php.

A stationary model We start this section by looking at models in the stationary setting where the parameters are assumed fixed over the season. Let $i \in \{1, 2, \ldots, 20\}$ denote the home team and $j \in \{1, 2, \ldots, 20\}$ denote the away team and let the games of the season, in chronological order, be labelled as $t = 1, \ldots, 380$. Let α_i be the attacking strength of team i, β_j be the defensive strength of team j and γ be the common home ground advantage. Then the home goals $X_{i,j}$ and away goals $Y_{i,j}$ at time t are described by two univariate Poisson distributions given by

$$X_{i,j} \sim \text{Poisson}\,(\alpha_i \beta_j \gamma), \quad Y_{i,j} \sim \text{Poisson}\,(\alpha_j \beta_i).$$

with $\beta_1 \to 1$ to maintain identifiability. We set the priors of the attacking and defensive strengths of all teams and the HGA to be

$$\begin{aligned} \alpha_i &\sim \text{Gamma}\,(\delta, \delta),\ i = 1, 2, \ldots, 20, \\ \beta_i &\sim \text{Gamma}\,(\delta, \delta),\ i = 2, 3, \ldots, 20, \\ \gamma &\sim \text{Gamma}\,(\delta, \delta). \end{aligned} \tag{2}$$

Then the posterior at end of season is

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma \mid \mathbf{x}, \mathbf{y}) \propto \underbrace{\exp\Big(-\sum_{\{i,j\}\in\mathcal{S}} [\alpha_i \beta_j \gamma + \alpha_j \beta_i]\Big) \times \prod_{\{i,j\}\in\mathcal{S}} \{[\alpha_i \beta_j \gamma]^{x_{i,j}} \times [\alpha_j \beta_i]^{y_{i,j}}\}}_{\mathrm{L}(\theta)\text{ Likelihood}}$$
$$\times \underbrace{\left[\prod_{i=1}^{20} \alpha_i^{\delta-1} \exp(-\alpha_i \delta)\right]}_{\text{Prior Attacking Strengths}} \times \underbrace{\left[\prod_{j=2}^{20} \beta_j^{\delta-1} \exp(-\beta_j \delta)\right]}_{\text{Prior Defensive Strengths}} \times \underbrace{\gamma^{\delta-1} \exp(-\gamma\delta)}_{\text{HGA}}$$

We have found that $\delta \to 1$ gives almost identical estimates of posterior means to their MLEs, calculated numerically. A comparison is displayed in Figure (1).

The dynamic model Now we assume that the $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$, γ_t evolve over time and are each described by two quasi-sufficient statistics. We describe the evolution of these parameters in terms of their quasi-sufficient statistics. For game t the model is given by.

$$\begin{aligned} X_{i,j} &\sim \text{Poisson}\left(\alpha_{i,t}\beta_{j,t}\gamma_t\right) && \text{(Home Goals)} \\ Y_{i,j} &\sim \text{Poisson}\left(\alpha_{j,t}\beta_{i,t}\right) && \text{(Away Goals)} \end{aligned}$$

The priors of the dynamic parameters at each time point are set at:

$$\begin{aligned} \alpha_{i,t} &\sim \text{Gamma}\left(\tilde{p}^{\alpha}_{i,t}, \tilde{q}^{\alpha}_{i,t}\right) && \text{(AS Home)} \\ \alpha_{j,t} &\sim \text{Gamma}\left(p^{\alpha}_{j,t}, q^{\alpha}_{j,t}\right) && \text{(AS Away)} \\ \beta_{j,t} &\sim \text{Gamma}\left(p^{\beta}_{j,t}, q^{\beta}_{j,t}\right) && \text{(DS Away)} \\ \beta_{i,t} &\sim \text{Gamma}\left(\tilde{p}^{\beta}_{i,t}, \tilde{q}^{\beta}_{i,t}\right) && \text{(DS Home)} \\ \gamma_t &\sim \text{Gamma}\left(\tilde{p}^{\gamma}_{t}, \tilde{q}^{\gamma}_{t}\right) && \text{(HGA)} \end{aligned}$$

The updates of all five dynamic parameters following the observation of the game are made using their quasi-sufficient statistics

$$\begin{aligned} p^{\alpha}_{i,t} &\leftarrow \tilde{p}^{\alpha}_{i,t} + x_{i,j} & q^{\alpha}_{i,t} &\leftarrow \tilde{q}^{\alpha}_{i,t} + \hat{\gamma}_t\hat{\beta}_{j,t}, && \text{(AS Home)} \\ p^{\alpha}_{j,t} &\leftarrow \tilde{p}^{\alpha}_{j,t} + y_{i,j} & q^{\alpha}_{j,t} &\leftarrow \tilde{q}^{\alpha}_{j,t} + \hat{\beta}_{i,t}, && \text{(DS Away)} \\ p^{\beta}_{i,t} &\leftarrow \tilde{p}^{\beta}_{i,t} + y_{i,j} & q^{\beta}_{i,t} &\leftarrow \tilde{q}^{\beta}_{i,t} + \hat{\alpha}_{j,t}, && \text{(DS Home)} \\ p^{\beta}_{j,t} &\leftarrow \tilde{p}^{\beta}_{j,t} + x_{i,j} & q^{\beta}_{j,t} &\leftarrow \tilde{q}^{\beta}_{j,t} + \hat{\gamma}_t\hat{\alpha}_{i,t} && \text{(DS Away)} \\ p^{\gamma}_{t} &\leftarrow \tilde{p}^{\gamma}_{t-1} + x_{i,j} & q^{\gamma}_{t} &\leftarrow \tilde{q}^{\gamma}_{t-1} + \hat{\alpha}_{i,t}\hat{\beta}_{j,t} && \text{(HGA)} \end{aligned} \tag{3}$$

where for instance $\hat{\beta}_{j,t}$ refers to the expectation of $\beta_{j,t}$ at the previous observation which is $\hat{\beta}_{j,t} = \frac{p^{\beta}_{j,t-1}}{q^{\beta}_{j,t-1}}$.

3.1 Between and within season variability

We denote all the dynamic parameters by $\boldsymbol{\theta}_t = \{\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \gamma_t, \}$. After each game five of these dynamic parameters are updated through their sufficient statistics $\mathbf{p}_t$ and $\mathbf{q}_t$. Note that at any time the expectation of the dynamic parameters can be estimated from their sufficient statistics $\hat{\boldsymbol{\theta}}_t = \frac{\mathbf{p}_t}{\mathbf{q}_t}$. Updates are carried out by repeatedly applying the two steps below.

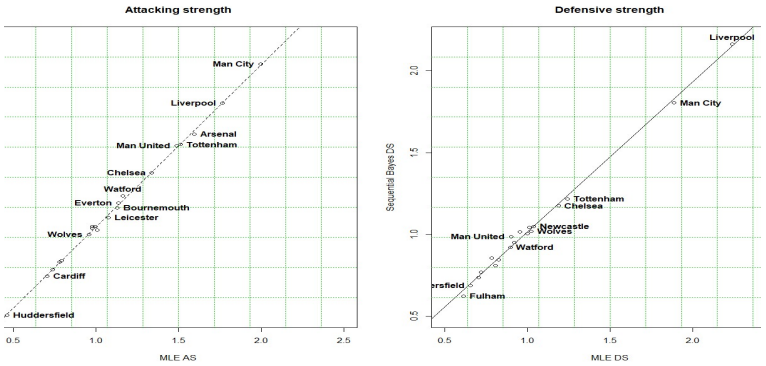


FIGURE 1. A comparison of maximum likelihood estimates and sequential Bayesian estimates of attacking and defensive strengths using only the data from the current season with no parameter discounting and the parameter of the prior from Equations (2) set to $\delta \to 2$.

1 Extend the prior In this step the means of the posteriors dynamic parameters from the previous game are extended in such a way that the means of $\boldsymbol{\theta}_{t-1}$ are preserved whilst increasing their variances. The updating from the posterior of last observation to the prior of the new observation are driven by the within season and between season fixed forgetting or volatility parameters: $0 \leq \omega_w, \omega_b \leq 1$, which can be allowed to vary by team and season. $\forall t \in \{2, \ldots, 380\}$ and $\forall s \in \{1, 2, \ldots, 23\}$ .We set $\omega_w \to .98$ and $\omega_b \to .67$.

$$\begin{aligned}
&\tilde{\mathbf{p}}_{s,t} \leftarrow \omega_w \mathbf{p}_{s,t-1}, && \tilde{\mathbf{q}}_{s,t} \leftarrow \omega_w \mathbf{q}_{s,t-1} && \\
&\tilde{\mathbf{p}}_{s,1} \leftarrow \omega_b \mathbf{p}_{s-1,380}, && \tilde{\mathbf{q}}_{s,1} \leftarrow \omega_b \mathbf{q}_{s-1,380} && \text{(For surviving teams)} \\
&\tilde{\mathbf{p}}_{s,1} \leftarrow \delta, && \tilde{\mathbf{q}}_{s,1} \leftarrow \delta && \text{(For promoted teams)}
\end{aligned}$$

2. Update and predict Updates are made on the sufficient statistics for the dynamic parameters using Equations (3) in Section (3).

4 Results and conclusion

We have carried out thorough diagnostics of the out of sample Pearson residuals and find little evidence of either over-dispersion or correlation, suggesting that our univariate Poisson dynamic model is adequate for our purposes. We have compared our Bayesian estimates of just the current season to their MLE estimates of the same season and have displayed them in Figure (1). We have analyzed the evolution of style and strengths of form of all the participating teams in the premier league, but have just presented the results of Manchester United in Figure (2).We believe that

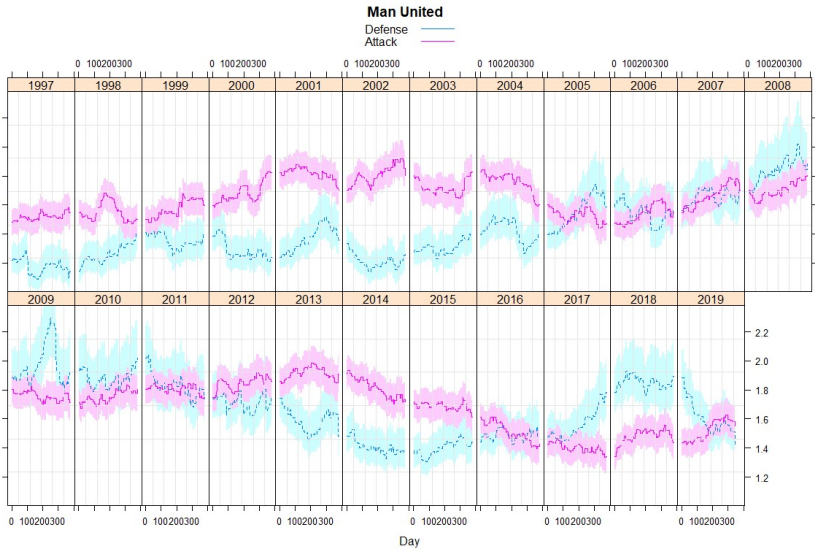


FIGURE 2. 50 % credible intervals for $\alpha_{i,t}$ and $1/\beta_{i,t}$ illustrating the evolution of the attacking and defensive strengths of Manchester United over the last two decades or so.

such an analysis will be a useful tool for the evaluation of the coaching regimes of each team over the history of this league and any other football competitions where the data is available.

References

Crowder, M , Dixon, M. and Ledford, A. and Robinson, M. (2002). Dynamic modelling and prediction of English football *Journal of the Royal Statistical Society. Series D (The Statistician)* **51**, 157–168.

Dixon, M. ,G., and Coles, S. G. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society, Series C*, **46**, 265–280.

Karlis, M. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models *Journal of the Royal Statistical Society. Series D (The Statistician)*, **52**, 381–393.

Koopman, S. J. and Lit, R. (2013). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League *Journal of the Royal Statistical Society. Series A*, **178** , 167–186.

West, M, Harrison, P. J. & Migon H.S. (1985). Dynamic Generalized linear models for Bayesian forecasting. *Journal of the American Statistical Society,* **80**, *67–86.*

Joint latent class trees: A tree-based approach to joint modeling of time-to-event and longitudinal data

Jeffrey S. Simonoff[1], Ningshan Zhang[1]

[1] New York University, USA

E-mail for correspondence: jsimonof@stern.nyu.edu

Abstract: Joint modeling of longitudinal and time-to-event data provides insights into the association between the two quantities. The joint latent class modeling approach assumes that conditioning on latent class membership, the trajectories of longitudinal data such as biomarkers are independent of survival risks. The resulting latent classes provide a data-dependent clustering of the population, which is also of interest in clinical studies of, for example, precision (personalized) medicine. Existing joint latent modeling approaches are parametric and suffer from high computational cost. We propose a nonparametric joint latent class modeling approach based on trees (JLCT). JLCT is fast to fit, and can use time-varying covariates in all of its modeling components. Based on simulations JLCT has similar performance to current approaches when using only time-invariant covariates, but can take advantage of the prognostic value of using time-varying covariates. We apply JLCT to a real application and see evidence of JLCT's strong predictive performance, while being orders of magnitude faster than the standard latent class model approach.

Keywords: Biomarker; Conditional independence; Recursive partitioning; Survival data; Time-varying covariates.

1 Introduction

Clinical studies often collect three types of data on each subject: the time to the event of interest (possibly censored), longitudinal measurements on a continuous response (for example, some sort of biomarker viewed as clinically important), and an additional set of covariates (possibly time-varying) about the subject. Analysis then focuses on the relationship between the time-to-event and the longitudinal responses, using the additional covariates.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

The most common approach for the joint modeling problem is the shared random effects model (SREM); for discussion see Tsiatis and Davidian (2004). A different line of research focuses on the joint latent class model (JLCM); for discussion see Proust-Lima et al. (2014). JLCM assumes that the population of subjects consists of multiple latent classes. A subject's time-to-event and longitudinal responses are independent conditioning on his or her latent class membership. In addition, JLCM assumes the latent classes are homogeneous, so subjects within a latent class follow the same survival and longitudinal model. The latent class membership is modeled by a multinomial logistic regression model. Blanche et al. (2015) showed that JLCM and SREM are each special cases of a general parametric joint modeling of the longitudinal and time-to-event outcomes, with the variable that ties the two together either continuous (SREM) or discrete (JLCM). What makes JLCM interesting conceptually, however, is the idea of latent class membership, which can be used to identify clinically important groups useful in, for example, precision (personalized) medicine. JLCM is restricted to time-invariant covariates for both latent class membership and time-to-event in current software implementation, and is quite computationally intensive.
In this work, we propose the joint latent class tree (JLCT) method. JLCT, like JLCM, is based on the key assumption that conditioning on latent class membership, time-to-event and longitudinal responses are independent. JLCT therefore looks for a tree-based partitioning such that within each estimated latent class defined by a terminal node, the time-to-event and longitudinal responses display a lack of association. Once the tree is constructed, we assign each observation to a latent class (i.e. terminal node), and independently fit survival and linear mixed-effects models, using the class membership information.

2 Joint Latent Class Trees

The joint latent class modeling problem makes the key assumption that a subject's time-to-event and longitudinal outcomes are independent conditioning on his or her latent class membership $(g_{it}) \in \{1, \ldots, G\}$. Without controlling the latent class membership, time-to-event and longitudinal outcomes may appear to be correlated because each is related to the latent class, but given it the two are independent of each other. The modeling of the time-to-event and the longitudinal outcome are therefore separated conditioning on group membership, greatly simplifying things. We assume the longitudinal outcomes come from a linear mixed-effects model. Conditioning on latent class membership, we assume the time-to-event depends on a subset of covariates observed at all time points through the extended Cox model for time-varying covariates (Cox, 1972), although once a tree is constructed the user can decide which type of survival models and which covariates to use within each terminal node, providing additional flexibility to the analyst.
Under these assumptions, JLCT looks for a tree-based partitioning such that within each estimated class defined by a terminal node the time-to-event and longitudinal outcomes display a lack of association. Tree-based methods are fast to construct, able to uncover nonlinear relationships between covariates, and are intuitive and easy to explain, making them ideal for this purpose (see chapter 9 of Hastie et al., 2009, for background on trees).

We consider binary trees, where each node is recursively split into two children nodes based on a splitting criterion. The splitting criterion ensures that the two children nodes are more "homogeneous" than their parent node. The measure of "homogeneity" in JLCT is based on the conditional independence between the time-to-event and the longitudinal outcomes: the more apparently independent the two variables are conditioning on the node, the more "homogeneous" the node is. The splitting criterion repeatedly uses the likelihood ratio test statistic for the hypothesis test

$$H_0 : b_y = 0, \quad \text{vs.} \quad H_1 : b_y \neq 0,$$

under the extended Cox model,

$$h(t, \mathbf{X}_i, \mathbf{Y}_i) = h_0(t) e^{Y_i(t) b_y + X_i(t) b_x},$$

where $h(t)$ is the hazard (risk) at time t, h_0 is a baseline hazard function, and $X_i(t)$ and $Y_i(t)$ indicate values at t . The coefficient b_y is the slope associated with the longitudinal outcomes, and thus $b_y = 0$ corresponds to the longitudinal outcome having no relationship with the time-to-event in the node given the other covariates. Note that this time-to-event formulation is **only** being used as a splitting criterion, **not** as a representation of the true relationship between $\mathbf{Y}$ and T.

We will denote the test statistic of the hypothesis test as `TS`. The smaller the value of `TS` is, the less related longitudinal outcomes are to the time-to-event data given the covariates and current node. JLCT seeks to partition observations such that `TS` is small within each leaf node, but stops partitioning when `TS` is less than a specified stopping parameter. The stopping criterion is based on the property that under the null model the distribution of `TS` is approximately a χ^2_1 distribution. This criterion can be tuned by changing the nominal significance level of the test, resulting in more or less aggressive splitting of nodes; here $\alpha = 0.05$ is taken as a default.

Simulations demonstrate that the performance of JLCT is comparable to that of JLCM when there are no time-varying covariates for the time-to-event or latent class, but can greatly outperform it when there are such covariates. In addition, when the underlying latent classes follow a tree structure, JLCT is very successful in recovering that structure, which makes inferences for the longitudinal and time-to-event variables within these estimated classes reasonably effective. Further, the JLCT algorithm is orders of magnitude faster than is the JLCM algorithm when the latter can be fit.

3 Application

In this section, we apply JLCT to a real dataset, the PAQUID dataset from the `R` package `lcmm`, which was also examined in Proust-Lima et al. (2017), The dataset collects five time-varying cognitive test score values along with age at visit, and three time-invariant covariates. The time-to-event is the age at dementia diagnosis or last visit. The goal is to jointly model the trajectories of the Mini-Mental State Examination score (`normMMSE`) as a biomarker (longitudinal outcomes) and the risk of dementia (time-to-event), using the remaining covariates.

We consider two JLCM, an SREM, and three JLCT models: the time-invariant model in Proust-Lima et al. (2017) ($JLCM_1$) and its corresponding shared random effects ($SREM_1$) and tree version ($JLCT_1$), the extension of $JLCM_1$ ($JLCM_2$) that uses the first occurrence of the time-varying covariates as time-invariant covariates and its corresponding tree version ($JLCT_2$), and the full version of JLCT ($JLCT_3$) that uses all of the values of time-varying covariates. More complex SREM models could not be fit with available software. For the two JLCM models, the number of latent classes is chosen from 2 to 6 according to the BIC selection criterion. For the four JLCT models, we set the stopping threshold to 3.84 and prune the trees to have no more than 6 terminal nodes.
Table 1 summarizes the results. When using only time-invariant covariates, $JLCT_1$ performs similarly to its counterpart $JLCM_1$ in prediction accuracy, while SREM outperforms both in terms of survival prediction. When using the same predictors $JLCT_1$ is orders of magnitude faster than the two parametric methods. By adding four "time-invariant" covariates (which are converted from time-varying ones) to the class membership and survival models, the performance of $JLCT_2$ remains similar, but the performance of $JLCM_2$ becomes much worse, mainly because JLCM failed to converge when optimizing the log-likelihood function (in fact, its performance is worse than a simple prediction of $\hat{S} = 0.5$ for every observation, which gives IBS = 0.25. When using the original time-varying covariates in the class membership and survival models, however, $JLCT_3$ improves its time-to-event prediction accuracy and outperforms all other methods by a significant margin on that measure. Further, JLCT is much faster than JLCM: fitting using JLCT took less than 2 minutes even for the most complex model, while fitting using JLCM took from 40 to 60 minutes. The experiments are performed on a desktop with 2.26GHz CPU and 32GB of memory.

TABLE 1. Performance of JLCM, SREM, and JLCT methods on the PAQUID dataset based on 10-fold CV; IBS refers to Integrated Brier Score of time-to-event prediction and RMSE refers to root-mean squared error for biomarker prediction.

	$JLCM_1$	$JLCM_2$	$SREM_1$	$JLCT_1$	$JLCT_2$	$JLCT_3$
IBS	0.1731	0.4467	0.1262	0.1611	0.169	0.0966
RMSE	14.759	18.354	14.669	14.550	14.291	14.501
Time (secs)	2448.7	4107.6	97.1	1.7	40.9	87.9

Figure 1 gives the tree associated with the construction of the latent classes. The tree splits into three nodes based only on `age`, splitting at ages 82 and 90, suggesting that people transition into different dementia statuses as they get older, which are reflected in differences in the distributions of both cognitive test score (`normMMSE`) and time until a dementia diagnosis.

4 Conclusion

In this paper we have proposed a tree-based approach to jointly model longitudinal outcomes and time-to-event with latent classes. JLCT performs comparably

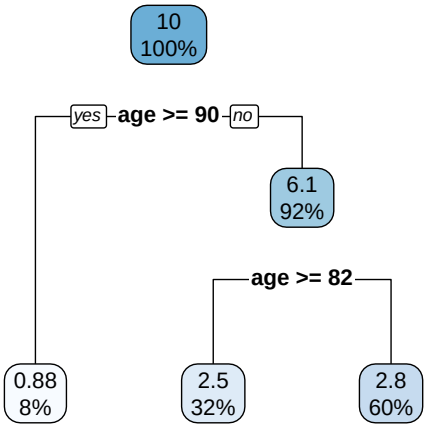


FIGURE 1. JLCT using time-varying covariates.

to its parametric counterpart JLCM, but makes full use of time-varying information when it is available and can significantly outperform JLCM as a result. JLCT is orders of magnitude faster than JLCM, and is highly flexible, allowing the data analyst to fit any longitudinal and time-to-event models they wish at each terminal node of the tree. Interesting generalizations of this tree-based approach to joint modeling include situations where there are competing hazards risks (the PAQUID data is actually an example of this type, since there is a risk of death before dementia is observed), where time-to-event is only known to within an interval of time (interval-censoring), and where several potentially prognostic longitudinal variables (biomarkers) are available for a subject (this is also the case for the PAQUID data, as there are other cognitive test variables in addition to `normMMSE` that could be used as biomarkers).
An `R` package, `jlctree`, that implements JLCT is available at CRAN.

References

Blanche, P., Proust-Lima, C. , Loubère, L., Berr, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2015). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*. **71**, 102 – 113.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187 – 220.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, 2nd edition*. New York: Springer New York Inc.

Proust-Lima, C., Séne, M., Taylor, J.M. and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, **23**, 74 – 90.

Proust-Lima, C., Philipps, V. , and Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software*, **78(2)**,1 – 56.

Tsiatis, A.A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, **14**, 809 – 834.

Two-step method for Joint Models of Longitudinal and Time-to-event Data

Srimanti Dutta[1], Arindom Chakraborty[1]

[1] Department of Statistics, Visva-Bharati University, Santiniketan, India

E-mail for correspondence: srimantid@gmail.com

Abstract: Motivated by an empirical analysis of Duchenne muscular dystrophy (DMD) data collected in a study, we propose a joint modeling technique for estimating the association between two responses: a continuous longitudinal one and a time-to-event indicator subject to censoring. We propose a two-stage approach to handle this type of data sets using all available information. At the first stage, we summarize the longitudinal information with the linear mixed-effects model, and at the second stage, we include the Empirical Bayes estimates of the subject-specific parameters as predictors in the accelerated failure time (AFT) model. We conclude that either joint modeling or the simpler two-stage multilevel approach can be used to estimate conditional associations between growth and later outcomes, but that only joint modeling is unbiased with nominal coverage for unconditional associations.

Keywords: Duchenne muscular dystrophy; Two-stage approach; AFT model.

1 Section 1

In clinical and epidemiological studies we often come across types of data where we perceive repeated evaluations of outcomes of a particular characteristic of a subject in time, along with an event of interest. Joint modelling deals with these two processes i.e., longitudinal and time-to-event processes simultaneously, as separate analysis and estimation of them could lead to biased and misleading estimates. In our present work we have extended the bivariate approach of joint modelling to a data set with $n + 1$ components where the first n components define the longitudinal process, and the last component is the time-to-event.

1.1 Section 1.1

Here, Y_{ij} denotes the longitudinal trajectory for subject $i = 1, .., m$ at time j for $j = 1, ..., n_i$. Let T_i denotes the time-to-event outcome for the i^{th} individual,

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

where δ_i is denoted as the censoring parameter. Then the joint distribution of longitudinal and time-to-event data, given subject-specific random effects $\boldsymbol{b}_i$ can be modeled by using the multivariate normal specification

$$(\boldsymbol{Y}_i', \log T_i)' \sim N(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, \Sigma_i),$$

$$\boldsymbol{V}_i = \begin{pmatrix} \boldsymbol{Y}_i \\ \log T_i \end{pmatrix} = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \tag{1}$$

Here, the design matrix for the fixed effects and random effects are denoted by $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ respectively, and $\boldsymbol{\beta}$ is the vector of regression parameters. Further,

$$\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \Sigma_i) \text{ with } \Sigma_i = \begin{pmatrix} \boldsymbol{\Sigma}_{Y_i} & \boldsymbol{\sigma}_{1i} \\ \boldsymbol{\sigma}_{1i}' & \sigma_{T_i}^2 \end{pmatrix} \tag{2}$$

The vector $\boldsymbol{\sigma}_{1i}$ signifies the association structure between $\boldsymbol{Y}_i$ and $\log T_i$. The association between these two simultaneous processes has till now been captured through the subject specific random effects $\boldsymbol{b}_i$ in literature. In our proposed framework, mainly the longitudinal correlation is captured by $\boldsymbol{b}_i$. So it indicates that dependency between the longitudinal and survival part can still be captured using the conditional distribution even if these two processes do not share common $\boldsymbol{b}_i$. Owing to difficulties due to positive definiteness constraints and high-dimensional complexities it is cumbersome to model the entire covariance matrix for each subject. This issue can be addressed by factorization of the joint distribution of $(\boldsymbol{Y}_i, \log T_i)$. In our proposed modeling framework, we factor the joint distribution of $\boldsymbol{Y}_i$ and $\log T_i$ into two components: a marginal model for $\boldsymbol{Y}_i$ and a correlated regression model for $\log T_i$ given $\boldsymbol{Y}_i$. In the presence of subject specific random effects $\boldsymbol{b}_i$, let

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{X}_{i1} & 0 \\ 0 & \boldsymbol{X}_{i2} \end{pmatrix}, \boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{Z}_{i1} & 0 \\ 0 & \boldsymbol{Z}_{i2} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

Then by implementing the Bartlett decomposition of a covariance matrix, the new models can be expressed as:

$$\begin{aligned} \boldsymbol{Y}_i|\boldsymbol{b}_i &= \boldsymbol{X}_{i1}\boldsymbol{\beta}_1 + \boldsymbol{Z}_{i1}\boldsymbol{b}_i + \boldsymbol{\epsilon}_{i1} \\ \log(T_i|\boldsymbol{Y}_i, \boldsymbol{b}_i) &= \boldsymbol{X}_{i2}\boldsymbol{\beta}_2 + \boldsymbol{Z}_{i2}\boldsymbol{b}_i + \boldsymbol{B}_i(\boldsymbol{Y}_i - \boldsymbol{X}_{i1}\boldsymbol{\beta}_1 - \boldsymbol{Z}_{i1}\boldsymbol{b}_i) + \epsilon_{i2} \end{aligned} \tag{3}$$

where $\boldsymbol{B}_i = \sigma_{1i}'\Sigma_{\boldsymbol{Y}_i}$ is the vector reflecting structural association between these two processes. Here we also capture the local dependency through non-zero $\boldsymbol{Z}_{i1}$ and $\boldsymbol{Z}_{i2}$. Further let us assume, $\boldsymbol{\epsilon}_{i1} \sim N(\boldsymbol{0}, \Sigma_{\boldsymbol{Y}_i})$ and $\epsilon_{i2} \sim N(0, \sigma_T^2)$. In the present work, to capture longitudinal correlation, we have assumed $\boldsymbol{b}_i = (b_{i1}, b_{i2})'$ are from $N_2(0, \Sigma_b)$. The covariance matrix Σ_b is denoted as

$$\begin{pmatrix} 1 & \rho\sigma_b \\ \rho\sigma_b & \sigma_b^2 \end{pmatrix} \tag{4}$$

the variance of b_{i1} being set to 1 for identifiability issue, σ_b^2 is the variance of b_{i2}, and ρ is the correlation coefficient between the two random effect components. Subsequently, we assume $\boldsymbol{Z}_{i1} = (1, 1)$ and $\boldsymbol{Z}_{i2} = (\nu_1, \nu_2)$. Let $f_0(.)$, $S_0(.)$ and $h_0(.)$ denote the density, survival, and hazard functions of random error ϵ_{i2} in equation (3), respectively. Let $f(.)$, $S(.)$ and $h(.)$ denote the density, survival,

and hazard functions of T, respectively. The contribution to the likelihood can be expressed as:

$$f(logT_i; y_{ij}, b_i) = h(t_i|\boldsymbol{b_i})^{\delta_i} S(t_i|\boldsymbol{b_i})$$

where $h(t_i|\boldsymbol{b_i})$ is the conditional hazard and $S(t_i|\boldsymbol{b_i})$ is the conditional survival function. Under the log-normal assumption, we have

$$S(t_i|\boldsymbol{Y}_i, \boldsymbol{b_i}) = 1 - \Phi(\frac{\log t_i - \boldsymbol{X}_{i2}\boldsymbol{\beta}_2 - \boldsymbol{Z}_{i2}\boldsymbol{b}_i + \boldsymbol{B}_i(\boldsymbol{Y}_i - \boldsymbol{X}_{i1}\boldsymbol{\beta}_1 - \boldsymbol{Z}_{i1}\boldsymbol{b}_i)}{\sigma_T})$$

The AFT (accelerated failure time) structure in joint modeling is troublesome to deal with compared to the Cox model since $f(logT_i; y_{ij}, b_i)$ is more complicated and unlike the Cox model, the baseline function involves unknown quantities. As a result, it is not possible to use the point mass function with masses assigned to all uncensored survival times t_i for the baseline hazard function., The Complete data Likelihood for the i^{th} individual can be expressed as:

$$L_i = \Big(\prod_{j=1}^{n_i} f(y_{ij}|b_i)\Big) f(logT_i|y_{ij}, b_i) f(b_i; \sigma^2, \rho)$$

Assuming independence among subjects, we can take $\Sigma_{Y_i} = \sigma_y^2 I$, where I is a $n_i \times n_i$ matrix., For notational simplicity, we let $\boldsymbol{y} = \{y_{ij}\} \cup \{T_i\} \cup \{\delta_i\}$ be the observed data and $\boldsymbol{\Psi} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \sigma_y^2, \sigma_b^2, \rho, \sigma_T^2, \nu_1, \nu_2)'$ be the parameter vector.

1.2 Section 1.2

The frequentist method of estimation poses as a major hurdle leading to issues of non-convergence or slow convergence of the model parameters. Moreover, when the dimension of the random-effects is high or the parameter space is large it poses as a serious difficulty for the classical estimation procedure to yield satisfactory estimates (due to the intractability of the Hessian matrix). Here in the first stage Linear mixed-effects regression for the longitudinal data is fitted using lme4 package. In the second stage, the estimates thus obtained are used as plug-in estimates in the survival part of our model. Survival parameters are estimated by adopting Gradient-Descent algorithm (with adaptive learning rate)

2 Section 2

We had conducted an extensive simulation study in order to check the efficacy of our proposed model and had also performed the robustness check under model misspecification (as we had used parametric AFT model). We had performed two data studies, i.e., the popular AIDS data and Duchenne muscular dystrophy (DMD) data. The DMD data consists of composite scores based on six different muscles (neck, deltoids, bicep, iliopsoas, quadriceps, and hamstrings) observed at different time points and are actually responsible for all movements. We had used three settings for the analysis purpose, i.e the usual joint model (proposed), the joint model with local independence (where the two processes are linked by conditional dependence only) and the fully independent model. The table on

the DMD data study below displays comparable estimates on the joint models and displays tight 95% confidence intervals thus ensuring the precision of the estimates. The parameter estimates of the survival part in the fully independent model fail to converge which is predictable. The simulation study and the robustness check under model misspecification also yielded satisfactory results with negligible bias, low mean squared error and standard error for the first two settings (thus asserting our claim).

TABLE 1. Duchenne Muscular Dystrophy data study

Parameter	Joint Model			JM with local independence		
	Estimate	SE	95% CI	Estimate	SE	95% CI
longitudinal						
α	4.74537	0.34276	[4.07,5.42]	4.74537	0.34276	[4.07,5.42]
β	0.10298	0.06024	[-0.02,0.22]	0.10298	0.06024	[-0.02,0.22]
σ_b^2	0.03339	0.0065	[0.02,0.05]	0.03339	0.0065	[0.02,0.05]
ρ	-0.29	0.318	[-0.7328,0.34]	-0.29	0.318	[-0.7328,0.34]
time-to-event						
β_0	1.1	0.001	[1.097,1.102]	1.1	0.00109	[1.097,1.102]
$\sigma_{(1)}^2$	0.03	0.0012	[0.027,0.032]	0.029	0.0010957	[-0.0036,0.061]
σ_T^2	1.702	0.01586471	[1.67,1.733]	1.7	0.016649	[1.667,1.732]
ν_1	-0.199	0.00109	[-0.2011,-0.1968]	-	-	-
ν_2	-0.799	0.00109	[-0.8011,-0.7968]	-	-	-

Parameter	Independent Model		
	Estimate	SE	95% CI
longitudinal			
α	4.74537	0.34276	[4.07,5.42]
β	0.10298	0.06024	[-0.02,0.22]
σ_b^2	0.03339	0.0065	[0.02,0.05]
ρ	-0.29	0.318	[-0.7328,0.34]
time-to-event			
β_0	not conv.	-	-
$\sigma_{(1)}^2$	-	-	-
σ_T^2	not conv.	-	-
ν_1	-	-	-
ν_2	-	-	-

References

Huong, P.T.T., Nur, D., Pham, H., & Branford, A. (2018). A modified two-stage approach for joint modelling of longitudinal and time-to-event data. *Journal of Statistical Computation and Simulation*, **88**, 3379 – 3398.

Murawska, M., Rizopoulos, D., & Lesaffre, E. (2012). A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies. *Journal of Probability and Statistics*, **2012**

Bayesian Probit Classification Trees

Mauro Bernardi[1,2], Daniele Durante[3], Paola Stolfi[2]

[1] Department of Statistical Sciences, University of Padova, Padova, Italy
[2] Istituto per le Applicazioni del Calcolo "Mauro Picone" - CNR, Roma, Italy
[3] Department of Decision Sciences, Bocconi University, Milano, Italy

E-mail for correspondence: p.stolfi@iac.cnr.it

Abstract: Ensemble of decision trees are popular techniques for regression and classification either because of their forecasting performances and their ability to account for complex nonlinear dependence structures among predictors. Leveraging on the Bayesian Additive Regression Trees (BART) approach, we propose new methods to deal with binary classification for CART and BART. Specifically, we introduce a new representation for the Probit classification model that avoid the data augmentation scheme. The proposed approach is illustrated and validated through comparison with alternative methods on real datasets.

Keywords: Bayesian additive trees, classification, unified skew–normal distribution, probit regression.

1 Introduction

Decision trees and their ensemble counterparts, [**?**] and [**?**], have been originally proposed for binary classification and regression and extended in several directions, for modelling conditional quantiles or to include high–order approximating polynomials for the conditional mean function on each terminal node. On the likelihood–based side, the Bayesian estimation of decision trees have been initially proposed for both classification and regression and extended to additive trees (BART), see Chipman et al. (1998), Denison et al. (1998), Chipman et al. (2010). The main novelty of this latter approach relies on exploiting the likelihood of parametric models where regressors splitting rules play the role of hard thresholding operators that partition the overall model into local models. For the binary classification problem previous algorithms directly apply to the augmented representation of the Probit link function of Albert and Chib (1993). Therefore, unlike the regression trees, classification trees suffer the major drawback that the marginal likelihood for sampling the tree structure is only available up to the latent factors.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

In this paper, we propose new methods for dealing with binary classification within the context of Bayesian additive regression trees (BART) of Chipman et al. (2010). Specifically, leveraging the results of Durante (2018) on Probit regression, we introduce a new representation for the Probit classification that avoid the data augmentation scheme of Albert and Chib (1993), thereby leading to a sampling scheme for the tree structure which relies on the proper marginal likelihood, i.e., the normalising constant of the posterior distribution of the model parameters.

2 Binary regression tree

We consider the following formulation of the BART for classification with Probit link function. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be the vector of observations on the response variable Y and let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ be the associated matrix of covariates of dimension $(n \times q)$, then the likelihood function of the i–th observation y_i can be factorised as follows:

$$Y \mid \mathbf{X} = \mathbf{x} \sim \text{Ber}\left(1, \psi\left(\mathbf{x}\right)\right) \tag{1}$$

$$\psi\left(\mathbf{x}\right) = \mathsf{P}\left(Y = 1 \mid \mathbf{x}\right) = \Phi\left[\eta\left(\mathbf{x}\right)\right] \tag{2}$$

$$\eta\left(\mathbf{x}\right) \approx \sum_{j=1}^{m} g\left(\mathbf{x}, \mathcal{T}_j, \mathcal{M}_j\right), \tag{3}$$

where $\Phi(\cdot)$ denotes the Probit–link function and m denotes the number of trees. For $m = 1$ we get the CART algorithm. In equation (3) we assume that the probit transformation of the response variable is a function of the regression trees, which is composed by a tree structure, denoted by $\mathcal{T}$, and the parameters of the terminal nodes, denoted by $\mathcal{M}$. The tree structure $\mathcal{T}$ contains information on how any observation y_i, in a set of n independent and identically distributed observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, recurses down the tree specifying a splitting rule for each non–terminal node. We denote by $\mathcal{M} = \{\mu_1, \mu_2, \ldots, \mu_b\}$ the set of parameters associated to the b terminal nodes of the tree, where μ_l, for $l = 1, 2, \ldots, b$ denotes the parameter associates to the l–th terminal node.
The classification tree specified in equations (1)–(3) provides a natural framework for likelihood–based inference on the set of regression parameters, i.e., the location parameters associated to the terminal nodes of the tree. Due to the complexity of the logistic link function in equation (2), the resulting posterior density for the regression parameters does not admit a closed form representation for the full conditional distributions, and needs to be sampled by using MCMC–based algorithms. To develop their BART probit for classification Chipman et al. (2010) provide a Gibbs sampling algorithm that relies on the data augmentation of scheme of Albert and Chib (1993). Our algorithm for simulating the posterior distribution of the classification tree instead exploits the representation of the likelihood function of the probit regression model as a Unified Skew–Normal recently provided by Durante (2018). In particular, all the key results we will consider in developing improved computational methods for BART and CART rely on the following Theorem 2, which proves that the full conditional distribution for the terminal nodes parameters in μ, given the trees $T = (T_1, \ldots, T_m)$,

belongs to the well known class of unified skew-normal (SUN) random variables Arellano et al (2006), Gupta et al. (2013).

theorem Let $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ denote independent response variables from a BART model (1)–(3), and assume $(\mathbf{mu} \mid T) \sim \mathrm{N}_p(\mathbf{0}, \sigma_\mu^2 \mathbf{I}_p)$, with $p = b_1 + \cdots + b_m$. Then

$$(\mu \mid T, \mathbf{y}, \mathbf{x}) \sim \mathrm{SUN}_{p,n}(\mathbf{f}0, \sigma_\mu^2 \mathbf{I}_p, \sigma_\mu \mathbf{S}^\intercal, \mathbf{0}, \sigma_\mu^2 \mathbf{S}\mathbf{S}^\intercal + \mathbf{I}_n), \tag{4}$$

where $\mathbf{S} = \mathrm{diag}(2y_1 - 1, \ldots, 2y_n - 1)\mathbf{D}$ and $\mathbf{D}$ is the $n \times p$ design matrix with rows $\mathbf{D}_i^\intercal$, $i = 1, \ldots, n$ representing the terminal nodes assignments for each unit $i = 1, \ldots, n$ based on the known trees structures T.

3 Numerical comparison

TABLE 1. Assessment on computational efficiency of BART with $m = 20$. For each sampling scheme under analysis, total running time in seconds and statistics summarizing the effective sample sizes (ESS) computed from the produced chains for the coefficients. The number of draws is $10,000$.

	Running time	Mixing via ESS					
	Time in secs	Min	1st quartile	Median	Aver.	Aver./Time	M.–H. Acc. rate
Pima–Indians data							
GS DA	297	51	57	57	58	0.1939	0.0144
GS SUN	567	960	1130	2110	2350	4.138	0.0183
BMKA data							
GS DA	137	147	154	157	156	1.1417	0.13147
GS SUN	259	980	2340	4610	4620	17.864	0.0535

We compare the performance of the proposed CART and BART algorithms for classification with the original versions of Chipman et al. (1998) and Chipman et al. (2010). We consider four illustrative datasets: the Pima Indians, see Kapelner and Bleich (2016), and the gene expression (BMKA) dataset, see Martinez et al. (2005). The datasets have been chosen because of their characteristics in terms of sample size and dimension. Specifically, the Pima Indians data consists of $n = 768$ subjects, of which 268 were diagnosed with diabetes, the binary response, and of $p = 8$ predictors, thus it is an example of small n–small q dataset. The BMKA dataset consists of $n = 74$ gene expressions of normal and cancerous biological tissues at $p = 516$ different tags, thus it represents an example where $p \gg n$. Results are reported in Table 1.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, 88(422):669–679.

Arellano-Valle, R. B. and Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33:561–574.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J. H. , Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika*, 85(2):363–377.

Durante, D. (2018). Conjugate bayes for probit regression via unified skew-normals. *arXiv:1802.09565*.

Gupta, A. K., Aziz, M. A., and Ning, W. (2013). On some properties of the unified skew-normal distribution. *Journal of Statistical Theory and Practice*, 7:480–495.

Kapelner, A. and Bleich, J. (2016). bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software, Articles*, 70(4):1–40.

Martinez, R., Christen, R., Pasquier, C., and Pasquier, N. (2005). Exploratory analysis of cancer SAGE data. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05), Discovery Challenge*, Porto, Portugal.

Robust inference for ROC regression

de Carvalho, V. I.[1], Lourenço, V. M.[2], de Carvalho, M.[1]

[1] University of Edinburgh, United Kingdom
[2] NOVA University of Lisbon (UNL) and Center for Mathematics and its Applications(CMA), Portugal

E-mail for correspondence: vmml@fct.unl.pt

Abstract: The receiver operating characteristic (ROC) curve is the most popular tool for evaluating the diagnostic accuracy of continuous biomarkers. Often, covariate information that affects the biomarker performance is also available and several regression methods have been proposed to incorporate covariates in the ROC framework. In this work, we propose robust inference methods for ROC regression, which can be used to safeguard against the presence of outlying biomarker values. Simulation results suggest that the methods perform well in recovering the true conditional ROC curve and corresponding area under the curve, on a variety of data contamination scenarios. Methods are illustrated using data on age-specific accuracy of glucose as a biomarker of diabetes.

Keywords: Receiver operating characteristic curve; M-regression; B-splines.

1 Methods

1.1 Biomarker accuracy assessment

Let $Y_D \sim F_D$ and $Y_{\bar{D}} \sim F_{\bar{D}}$, be the biomarker values of diseased and non-diseased subjects, that is $F_D(c) = P(Y_D \leq c)$ and $F_{\bar{D}}(c) = P(Y_{\bar{D}} \leq c)$. Without loss of generality, we proceed with the assumption that, at any cutoff value c, a subject is classified as diseased when his/her test outcome is equal or greater than c and as non-diseased when it is below c. Then, the sensitivity and the false positive fraction associated with this decision criterion are

$$\mathrm{Se}(c) = \Pr(Y_D \geqslant c) = 1 - F_D(c), \quad \mathrm{FPF}(c) = \Pr(Y_{\bar{D}} \geqslant c) = 1 - F_{\bar{D}}(c).$$

Formally, the ROC curve consists of the set of points $\{(1 - F_{\bar{D}}(c), 1 - F_D(c)) : c \in \mathbb{R}\}$. Setting $p = \mathrm{FPF}(c) = 1 - F_{\bar{D}}(c)$, it rewrites as

$$\mathrm{ROC}(p) = 1 - F_D(c) = 1 - F_D\{F_{\bar{D}}^{-1}(1-p)\}, \quad 0 \leqslant p \leqslant 1. \tag{1}$$

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

In order to assess biomarker accuracy, the area under the ROC curve (AUC), which is given by

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(p)\ \mathrm{d}p = P(Y_D > Y_{\bar{D}}), \tag{2}$$

is usually computed. The greater the AUC the better the discriminanting ability of the biomarker. When covariate information is available, ROC regression is usually employed as a way to evaluate biomarker accuracy as a function of such covariate (e.g., age, gender). The induced regression approach of Pepe (1998) is a popular method that addresses the incorporation of covariate information in the ROC curve. Since the presence of outliers may put at risk the reliability of the inferences and, in addition, the covariate effects may not necessarily be linear, a flexible approach to the method of Pepe that incorporates robust regression techniques and B-splines modelling is proposed and described in the next subsection.

1.2 Flexible induced ROC regression

We define the covariate-adjusted sensitivity and covariate-adjusted FPF as

$$\begin{cases} \mathrm{Se}(c \mid \mathbf{x}) & = \Pr(Y_D > c \mid \mathbf{x}) = 1 - F_D(c \mid \mathbf{x}), \\ \mathrm{FPF}(c \mid \mathbf{x}) & = \Pr(Y_{\bar{D}} > c \mid \mathbf{x}) = 1 - F_{\bar{D}}(c \mid \mathbf{x}), \end{cases}$$

where $F_D(c \mid \mathbf{x})$ and $F_{\bar{D}}(c \mid \mathbf{x})$ are the conditional distributions of the marker, given the predictor $\mathbf{x}$, in the diseased and non-diseased populations, respectively. The key object of interest here is given by the covariate-adjusted ROC surface, formally defined as the plot

$$\{(p, \mathbf{x}, \mathrm{ROC}(p \mid \mathbf{x})) : p \in [0,1], \mathbf{x} \in \mathbb{R}^q\},$$

where

$$\mathrm{ROC}(p \mid \mathbf{x}) = 1 - F_D\{F_{\bar{D}}^{-1}(1 - p \mid \mathbf{x}) \mid \mathbf{x}\}, \tag{3}$$

with $F_{\bar{D}}^{-1}(1 - p \mid \mathbf{x}) = \inf\{y : F_{\bar{D}}(y \mid \mathbf{x}) \geqslant 1 - p\}$. Similarly, the covariate-adjusted AUC is defined as $\mathrm{AUC}(\mathbf{x}) = \int_0^1 \mathrm{ROC}(p \mid \mathbf{x})\, \mathrm{d}p$. If there are different covariates for diseased and non-diseased subjects, say $\mathbf{x}_D$ and $\mathbf{x}_{\bar{D}}$, the definition in (3) requires only small adjustments.

Suppose we observe data of the type $\{(Y_{D,i}, \mathbf{x}_{D,i}^T)\}_{i=1}^{n_D}$ and $\{(Y_{\bar{D},j}, \mathbf{x}_{\bar{D},j}^T)\}_{j=1}^{n_{\bar{D}}}$ where $Y_{D,i}$ and $Y_{\bar{D},j}$ are biomarker values for diseased and non-diseased subjects and $\mathbf{x}_{D,i}^T$ and $\mathbf{x}_{\bar{D},j}^T$ are covariates for the corresponding populations of interest. The general ROC regression approach assumes a location-scale model of the form

$$\begin{cases} Y_{D,i} = \mu_D(\mathbf{x}_{D,i}) + \sigma_D(\mathbf{x}_{D,i})\varepsilon_{D,i}, & \varepsilon_{D,i} \overset{iid}{\sim} G_D, \\ Y_{\bar{D},j} = \mu_{\bar{D}}(\mathbf{x}_{\bar{D},j}) + \sigma_{\bar{D}}(\mathbf{x}_{D,j})\varepsilon_{\bar{D},j}, & \varepsilon_{\bar{D},j} \overset{iid}{\sim} G_{\bar{D}}, \end{cases}$$

for $i = 1, \ldots, n_D$ and $j = 1, \ldots, n_{\bar{D}}$, with μ_D, $\mu_{\bar{D}}, \sigma_D$ and $\sigma_{\bar{D}}$ functions of the $q+1$ covariates (intercept included), G_D and $G_{\bar{D}}$ left unspecified and errors assumed to verify $E(\varepsilon_{D,i}) = E(\varepsilon_{\bar{D},j}) = 0$, $\mathrm{var}(\varepsilon_{D,i}) = \mathrm{var}(\varepsilon_{\bar{D},j}) = 1$ and $\varepsilon_D \perp \varepsilon_{\bar{D}}$. Under this model, the ROC curve and the AUC rewrite to

$$\mathrm{ROC}(p \mid \mathbf{x}) = 1 - F_D\{F_{\bar{D}}^{-1}(1 - p \mid \mathbf{x}) \mid \mathbf{x}\} = 1 - G_D(a(p, \mathbf{x})), \tag{4}$$

and

$$\mathrm{AUC}(\mathbf{x}) = \int_0^1 \mathrm{ROC}(p \mid \mathbf{x})\ \mathrm{d}p = 1 - \int_0^1 G_D(a(p, \mathbf{x}))\,\mathrm{d}p, \tag{5}$$

where

$$a(p, \mathbf{x}) = \frac{\mu_{\bar{D}}(\mathbf{x}) - \mu_D(\mathbf{x})}{\sigma_D(\mathbf{x})} + \frac{\sigma_{\bar{D}}(\mathbf{x})}{\sigma_D(\mathbf{x})} G_{\bar{D}}^{-1}(1-p). \tag{6}$$

The method of Pepe (1998) assumes the simple location-scale model, i.e., $\mu_D(\mathbf{x}) = \mathbf{x}_D^T \boldsymbol{\beta}_D$, $\mu_{\bar{D}}(\mathbf{x}) = \mathbf{x}_{\bar{D}}^T \boldsymbol{\beta}_{\bar{D}}$, $\sigma_D(\mathbf{x}) = \sigma_D$ and $\sigma_{\bar{D}}(\mathbf{x}) = \sigma_{\bar{D}}$. Here, estimation of the unknown parameters is done via least squares. Once the estimates of $\boldsymbol{\beta}_D$, $\boldsymbol{\beta}_{\bar{D}}$, σ_D and $\sigma_{\bar{D}}$ are obtained, the distribution of the errors are estimated on the basis of the empirical distribution of the standardized residuals, i.e., as

$$\widehat{G}_D(e) = \frac{1}{n_D} \sum_{i=1}^{n_D} I\left(\frac{y_i - \mathbf{x}_{D,i}^T \widehat{\beta}_D}{\widehat{\sigma}_D} \leqslant e \right), \quad \widehat{G}_{\bar{D}}(e) = \frac{1}{n_{\bar{D}}} \sum_{j=1}^{n_{\bar{D}}} I\left(\frac{y_j - \mathbf{x}_{\bar{D},j}^T \widehat{\beta}_{\bar{D}}}{\widehat{\sigma}_{\bar{D}}} \leqslant e \right)$$

where sample quantiles $\widehat{G}_{\bar{D}}^{-1}(1-p)$ are evaluated as in Hyndman (1996). Inference can then be done through bootstrap techniques. Our proposed approach, firstly approximates functions $\mu_D(\mathbf{x})$ and $\mu_{\bar{D}}(\mathbf{x})$ by a linear combination of cubic B-spline basis functions over a sequence of knots (De Boor et al., 1978) and subsequently estimates the unknown parameters of the location-scale model via M-regression Maronna, 2006).

2 Simulation and Results

The simulation study is conducted considering several linear and non-linear scenarios over 1000 replications. For illustration we set $n_D = n_{\bar{D}} = 100$ for the diseased and non-diseased populations contemplating several % of location and scatter outliers. See Figure 1 for a summary of our main results. There is good evidence of the ability of the proposed method in recovering the true functional form of the covariate adjusted AUC as well as providing confidence intervals with smaller amplitudes containing the true conditional AUC.

3 Diabetes data

To illustrate our method we resort to data from a population-based survey in Cairo, Egypt (Smith & Thompson, 1996) where postprandial glucose measurements (biomarker values) were obtained from a fingerstick on 286 subjects. According to the World Health Organization diagnostic criteria for diabetes, 88 subjects were classified as diabetic and 198 subjects as non-diabetic. This information is used as gold-standard in our evaluation of the accuracy of glucose as a biomarker for diabetes. In particular, we are interested in assessing the change in the accuracy of the biomarker with age.

Figure 2 suggests that the accuracy of the biomarker decreases with age. Our robust method is not only able to capture a slight non-linearity of the covariate effect but also hints that the decrease of biomarker accuracy is more evident from around 67 years onwards.

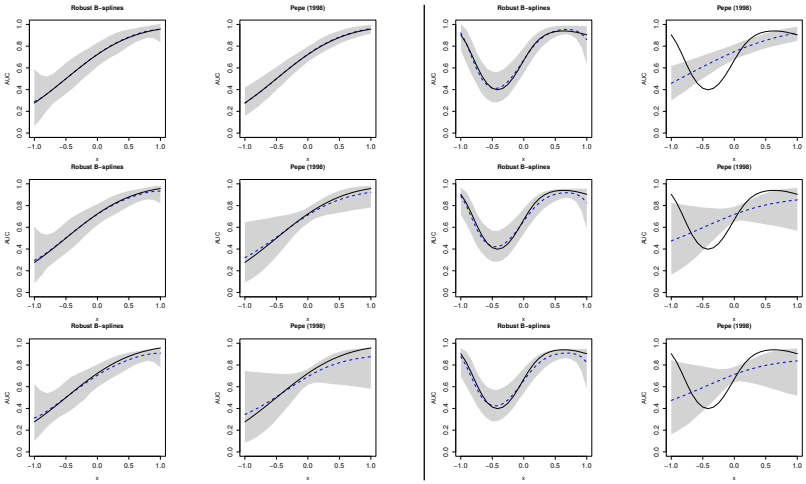


FIGURE 1. True conditional AUCs (solid) and estimated conditional AUCs (blue) using the proposed robust B-spline (with no interior knots) and Pepe's approaches. Left and right panels refer to the linear and non-linear scenarios, respectively; grey areas refer to the 95% confidence bands computed from the distribution simulated quantiles; each row refers to 0, 2 and 5% levels of location outliers. **Results referring to scatter outliers are similiar.**

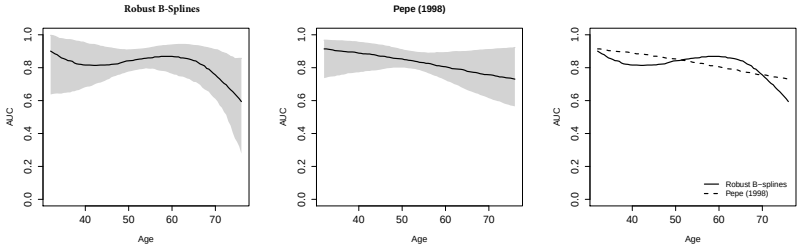


FIGURE 2. Estimated robust and classical conditional AUCs (left and center, respectively) with 95% bootstrap confidence bands; and estimated AUC curve comparison (right).

Acknowledgments: This work received partial financial support from: (i) Fundação para a Ciência e a Tecnologia through project UID/MAT/002 97/2013 (Centro de Matemática e Aplicações) and sabbatical grant SFRH/BSAB/142919/2018; and (ii) Erasmus Contracts 29191/0 02/2017/STT and 29191/036/2018/STT.

References

De Boor, C. et al. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.

Hyndman, R.J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician* **50(4)**, 361 – 365.

Maronna R.A., Martin, D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. Chichester: Wiley.

Pepe, M.S. (1998). Three approaches to regression analysis of ROC curves for continuous test results. *Biometrics* **54(1)**, 124 – 135.

Smith, P. and Thompson, T. (1996). Correcting for confounding in analyzing ROC curves. *Biometrical Journal* **38(7)**, 857 – 863.

A cohesive Bayesian approach to competing risks models

Janet van Niekerk[1], Haakon Bakka[1] and Håvard Rue[1]

[1] CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia

E-mail for correspondence: Janet.vanNiekerk@kaust.edu.sa

Abstract: Complexities in survival data is present in most clinical environments. One of these complexities, is the presence of multiple competing events. This forms various dependence structures since the realization of an event could influence the hazards of the other competing events. Additionally, features like clustering of individuals or spatial clustering present further complications. In this study, we present a cohesive framework in which we define competing risks models as latent Gaussian models. This definition enables the efficient implementation of these models, with or without complicated features, in the R-INLA framework.

Keywords: Competing risk; INLA; latent Gaussian model; survival.

1 Introduction

Time to event data is observed in its most simplest form, as the time when a particular event happens or when the monitoring process halts (censoring time). A competing risks model (Gooley et al. (1999)) arises in the case of multiple events being monitored. Each observational unit is at risk for each of the C events until one event occurs. After this, the risk of the other events are either zero or changed. For this purpose, competing risks should be dealt with accordingly. Early research in this field, suggested to treat all competing risks as censored observations when the focus is on only one risk. This could be troublesome due to the assumption that the competing events could still occur under the censored idea. In this framework, various approaches has been proposed like the subdistribution hazards (Dixon et al. (2011)), latent lifetimes and cause-specific hazards approaches (Prentice at al. (1978)). Critiques have been raised against the former two methods due to their restricting assumptions concerning the independence of cause and time. Therefore, we will adopt the latter approach.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

Features that originate from the data generating process, can pose difficulties in the modeling process. These features can involve clustering effects, spatial effects, treatment affects, to name a few. Most of the available software for competing risks are specifically formulated to handle a single feature. This is cumbersome for users since many different packages should be familiarized for every day use. We argue that most competings risks models are actually part of the group of latent Gaussian models and as such, can be handled with ease using R-INLA (Rue (2009)).

2 Competing risks

In this work we focus on the cause-specific hazard functions to characterize the risk of a particular event to a patient. Suppose we have C competing events and N patients, then the cause-specific hazard functions are defined as

$$\lambda_{c,i}(t) = \lambda_{c,0} \exp(\eta_{c,i}), \quad c = 1, ..., C, i = 1, .., N \tag{1}$$

with $\lambda_{c,0}$ the baseline hazard function for cause c and

$$\eta_{c,i} = \boldsymbol{\beta}^T \boldsymbol{X}_i + \boldsymbol{u}_i(\boldsymbol{z}_i) + \epsilon_i \tag{2}$$

where $\boldsymbol{\beta}$ represent the linear fixed effects of the covariates X, $\boldsymbol{\epsilon}$ is the unstructured random effects and $\boldsymbol{\gamma}$ represents the known weights of the unknown non-linear functions $\boldsymbol{u}$ of the covariates $\boldsymbol{z}$. The unknown non-linear functions, also known as structured random effects, $\boldsymbol{u}$ include spatial effects, temporal effects, non-seperable spatio-temporal effects, frailties, subject or group-specific intercepts and slopes etc.

If $\eta_{c,i}$ depends on time, then we do not have proportional hazards but an accelarated failure time model. The baseline hazard function can be specified parametrically or nonparameterically, and most well-known cases, including the Cox model, are accessible in *R-INLA*.

The hazard rates $\lambda_{c,i}$ are characterized, in part, by the corresponding linear predictors $\eta_{c,i}$. The INLA methodology can be applied since the data enters the model, exclusively, through the linear predictor. The linear predictor can be as complex as needed without much computational effort.

3 The INLA method

Hierarchical Bayesian additive models are widely used in various applications. A specific subset of Bayesian additive models is the class of latent Gaussian models (LGM). An LGM can be efficiently modelled using the INLA methodology implemented in the *R-INLA* package. (Rue (2009)) This class comprises of well-known models such as mixed models, temporal and spatial models. An LGM is defined as a model having a specific hierarchical structure, as follows: The likelihood is conditionally independent based on the likelihood parameters (hyperparameters),

$\boldsymbol{\theta}$ and the linear predictors, η_i, such that the complete likelihood can be expressed as

$$\pi(\boldsymbol{y}|\boldsymbol{\eta},\boldsymbol{\theta}) = \prod_{i=1}^{N} \pi(y_i|\eta_i(\boldsymbol{\mathcal{X}}),\boldsymbol{\theta}). \tag{3}$$

The linear predictor is formulated as follows:

$$\eta_i = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}_i + \boldsymbol{u}_i(\boldsymbol{z}_i) + \epsilon_i \tag{4}$$

where $\boldsymbol{\beta}$ represent the linear fixed effects of the covariates X, $\boldsymbol{\epsilon}$ is the unstructured random effects and $\boldsymbol{\gamma}$ represents the known weights of the unknown non-linear functions $\boldsymbol{u}$ of the covariates $\boldsymbol{z}$. The unknown non-linear functions, also known as structured random effects, $\boldsymbol{u}$ include spatial effects, temporal effects, non-seperable spatio-temporal effects, frailties, subject or group-specific intercepts and slopes etc. This class of models include most models used in practice since time series models, spline models and spatial models, amongst others, are all included within this class. The main assumption is that the data, $\boldsymbol{Y}$ is conditionally independent given the partially observed latent field, $\boldsymbol{\mathcal{X}}$ and some hyperparameters $\boldsymbol{\theta}_1$. The latent field $\boldsymbol{\mathcal{X}}$ is formed from the structured predictor as $(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{\eta})$ which forms a Gaussian Markov random field with sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_2)$, i.e. $\boldsymbol{\mathcal{X}} \sim N(\boldsymbol{0},\boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2))$. A prior, $\boldsymbol{\pi}(\boldsymbol{\theta})$ can then be formulated for the set of hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$. The joint posterior distribution is then given by:

$$\boldsymbol{\pi}(\boldsymbol{\mathcal{X}},\boldsymbol{\theta}) \propto \boldsymbol{\pi}(\boldsymbol{\theta})\boldsymbol{\pi}(\boldsymbol{\mathcal{X}}|\boldsymbol{\theta}) \prod_i \pi(Y_i|\boldsymbol{\mathcal{X}},\boldsymbol{\theta}) \tag{5}$$

The goal is to approximate the joint posterior density (5) and subsequently compute the marginal posterior densities, $\boldsymbol{\pi}(\mathcal{X}_i|\boldsymbol{Y}), i = 1...n$ and $\boldsymbol{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$. Due to the possibility of a non-Gaussian likelihood, the Laplace approximation to approximate this analytically intractable joint posterior density. The sparseness assumption on the precision of the latent Gaussian field ensures efficient computation.

4 Application

We show that a competing risks model with spatially clustered frailties is actually a latent Gaussian model and we can thus take advantage of the INLA method. We illustrate the applicabilty of our method to breast cancer data from the Surveillance Epidemiology and End Results database of the National Cancer Institute (SEER (2017)). This data presents competing risks and spatial random effects per region and per cause. Hesam et al. (2018) proposed a correlated spatial frailty model and implemented it using OpenBUGS software. They mention that if more than two competing risks are present, their computational framework is burdensome and inefficient. We, thus, aim to illustrate a computationally efficient approach that does not depend on the number of risks and has no difficulty handling spatial random effects.

References

Dixon, S.N., Darlington, G.A., Desmond, A.F. (2011). A competing risks model for correlated data based on the subdistribution hazard. *Lifetime Data Analysis* **17(4)**, 473 – 495.

Gooley, T.A., Leisenring, W., Crowley, J., Storer, B.E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine* **18(6)**, 695 – 706.

Hesam, S., Mahmoudi, M., Foroushani, A.R., Yaseri, M. and Mansournia, M.A. (2018). A cause-specific hazard spatial frailty model for competing risks data. *Spatial Statistics*, **26**, 101 – 124.

Prentice, R.L., Kalbfleisch, J.D. , Peterson Jr., A.V., Flournoy, N., Farewell, V.T., Breslow, N.E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 541 – 554.

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 (Statistical Methodology)*, **71(2)**, 319 – 392.

Surveillance, Epidemiology , and End Results (SEER) Program Populations (1969-2016) (2017). (www.seer.cancer.gov/popdata), National Cancer Institute, DCCPS, Surveillance Research Program.

Studying the Softplus Function as a Response Function in Regression Models

Paul Wiemann[1], Thomas Kneib[1]

[1] Chair of Statistics, University Göttingen, Germany

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

Abstract: The choice of the link function often depends only on the domain of the parameter in a regression model. We investigate the softplus function $\text{softplus}(x) = \log(\exp(x) + 1)$ in generalised linear models and their extension to distributional regression models as an alternative to the typically assumed exponential function as the response function for positive constrained parameters.

Keywords: generalised linear model; distributional regression; link function; response function; softplus

1 Introduction

In regression models, properties of the observed response variable y are related to the vector of available covariate information x. This is usually done via the linear predictor, a linear combination of the covariates $\eta = x'\beta$. Without constraining the vector of regression coefficients β, the domain of the linear predictor is the real numbers, but the modelled properties may be restricted to a subset of them. Suppose the expected value of the response variable is to be related to a linear combination of covariates, as done in GLMs. For a non-negative response variable, the expected value is also non-negative and thus the linear predictor should not be negative either. To achieve this, one would have to constrain the regression parameters, with the exact constraint depending on covariates. Instead, the linear predictor is usually mapped to the domain of the quantity to be modelled by means of a response function.
The same problem arises in the context of distributional regression, in which all parameters of a parametric distribution are linked to the covariate information (Umlauf and Kneib, 2018), since many distribution parameters can only have positive values. In both model classes the response functions are essential part of the model assumptions, but their choice is rarely questioned. In particular, if the modelled quantities are restricted to be greater than zero, the exponential

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

function is often used as the response function of choice. In general, there is no indication that this is the correct choice. Of course, the used response function should emulate the data generating process as closely as possible, but also other aspects as interpretability of the results and estimability might be important aspects to statisticians. Therefore, we propose to take the softplus function (Dugas et al., 2001) as response function into consideration.

2 Softplus Function

The Softplus function, mainly used in deep neural networks, is a continuous differentiable approximation of the ramp function $\max(0, x)$. We define the softplus function with an additional goodness of approximation (goa) parameter $a > 0$ as follows

$$\text{softplus}_a(x) = \log\left(1 + \exp(ax)\right)/a$$

with $x \in \mathbb{R}$. The additional parameter a allows for a better approximation and the approximation error can be kept arbitrarily small. Like the exponential function the softplus function is a smooth and bijective function mapping from the set of real numbers to the positive domain while having a positive first derivative.
As shown in Figure 1, the softplus function follows the identity function very closely in the positive domain and rapidly approaches zero in the negative domain for x towards minus infinity. This behaviour can be further accentuated by increasing the goodness of approximation parameter a.

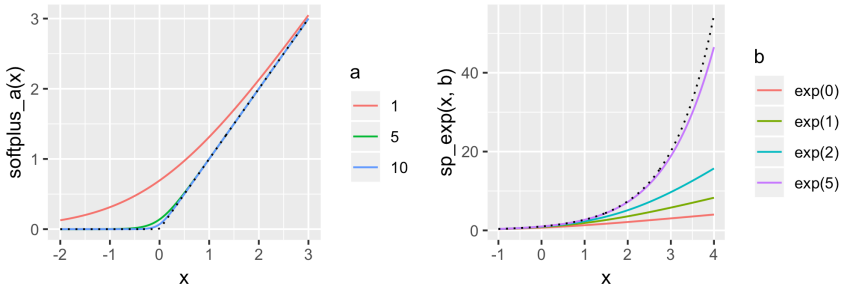


FIGURE 1. Plots of the softplus (left) and softplus exponential function (right) with the function be be approximated.

The softplus function, used as a response function in a regression model, allows for a straightforward interpretation of the regression coefficients: as long as the predictor is large enough, in particular within the linear part of the function, the effects can be interpreted directly on the parameter, i.e. a change in the covariate by one unit causes a change by β units on the parameter, where β denotes corresponding the regression coefficient. In addition, the additivity of the effects in the predictor is transferred to the parameter space within the linear part of the function. By choosing a sufficiently large a, the linear part covers almost the entire positive domain.
In the negative domain and for a sufficiently large a a small change of the covariate does usually not cause a significant change on the parameter, since the

softplus function outputs values very close to zero. To ensure the validity of this interpretation, it is necessary to check the range of values of the linear predictor for the observations in the data set. Most of them should be located within the linear part of the softplus function. This additive interpretation is in contrast to the usual multiplicative interpretation for positively constraint parameters that raises form the use of the log-link.
The calculation of the softplus function does not involve numerical issues, since one can exploit $\log(1 + \exp(x)) = \max(0, x) + \log(1 + \exp(-|x|))$ and the calculation of $\log(1 + x)$ can be done very precisely even for $|x| \ll 1$ (Nielsen, F. and Sun, K., 2016; Abramowitz and Stegun, 1972, p. 68).
Another feature of the softplus function is that it can be used to define a function that initially follows the exponential function very closely, but then increases more slowly as its first derivative approaches a predefined upper limit. To achieve this, we define the softplus exponential function as

$$\text{sp\_exp}(x, b) = b\,\text{softplus}_1(x - \log(b)) = b\log(1 + b^{-1}\exp(x))$$

with $b > 0$ being the limit of the first derivative. We refer to Figure 1 for a plot of this function. We will not discuss the softplus exponential function here but will provide details during the talk, i.e. we believe that its use can lead to more numerical stability in some estimation algorithms.

3 Application: Doctor Visits

The objective of this application is the discussion of the application of the softplus function as a response function to data from the Australian Health Survey (AHS) 1977-1987 (for a detailed description of the data visit Cameron and Trivedi, 1986). With a regression model we relate the expected count of doctor consultations within two weeks to a set of covariates. The data consists of 5190 observations of which 4141 show a zero count of doctor visits. For the remaining 1049 observations an average of 1.49 consultations can be reported.
We consider the Poisson distribution and Negative Binomial distribution as possible response distribution for the outcome of interest and use the exponential function and the softplus function with goa parameter set to 10 as potential response functions to ensure positivity of the modelled quantity. The choice $a = 20$ for the softplus parameter is made since it keeps the expected count very close to zero when the predictor is negative and thus let the model better deal with the excess zero-count in the data. Furthermore, with this choice we can apply the additive interpretation almost on the hole positive domain.
To choose the best fitting model with employ the deviance information criterion (DIC, Spiegelhalter et al., 2002). As Table 1 shows, the model with negative binomial response distribution and softplus response function outperforms the other models in terms of DIC.
We refer to Table 2 for a listing of the posterior mean estimates and their 95% equal-tailed intervals. For both response functions, the signs of the posterior mean values of the regression coefficients are the same. However, some of the regression coefficients are significant with regard to the credibility intervals while using one of the two response function but not while using the other, i.e. the effect of age.

TABLE 1. DIC values of fitted models with assumed Poisson or Negative Binomial distributed responses for softplus ($a = 20$) response function respectively exponential response function.

response function	poisson	negbin
softplus	6443.05	6282.69
exponential	6735.44	6479.28

The difference in interpretation can be highlighted by considering the covariate days of reduced activities due to illness in the last two weeks. The posterior mean of the corresponding regression coefficient is 0.11 in both models. With the log-link each additional day of reduced activity would lead to a multiplicative change of 1.12 expected doctor consultations. For the softplus model, the same change in the covariate would lead to 0.11 additional expected doctor consultations. Of course this interpretation is only valid for the linear part of the softplus function. With the $a = 20$ this is basically the case on the whole positive domain of the predictor. For negative values of the predictor the softplus function (with goa parameter set to 20) outputs values close to zero so that a change of the covariate only affects the expected count if the threshold at 0 is exceeded.

4 Application: Average Rental Duration of Bicycles

In this section we demonstrate the applicability of the softplus function as a response function in a Bayesian distributional regression model. We employ data from Capital Bikeshare, a bicycle sharing service located in Washington D.C., to

TABLE 2. Posterior estimates of the regression coefficients on the expected value together with their 95% credibility intervals.

	softplus			exponential		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%
(Intercept)	-0.02	-0.07	0.03	-2.20	-2.45	-1.96
age in 100 years	0.12	0.02	0.22	0.30	-0.08	0.67
income in 1k dollars	-0.03	-0.07	0.01	-0.14	-0.34	0.06
female	0.05	0.02	0.08	0.17	0.04	0.29
number of illnesses	0.07	0.05	0.08	0.19	0.15	0.24
days of reducd actvty	0.11	0.09	0.12	0.11	0.10	0.12
health score	0.01	0.00	0.02	0.04	0.01	0.06
privateyes	0.05	0.01	0.08	0.20	0.03	0.36
freepooryes	-0.06	-0.13	0.01	-0.55	-1.03	-0.13
freerepatyes	0.10	0.04	0.15	0.20	0.00	0.40
nchronicyes	0.01	-0.02	0.04	0.13	0.00	0.28
lchronicyes	0.05	-0.01	0.11	0.17	-0.03	0.36

analyse the mean rental duration in minutes within each hour in the years 2016 - 2017. A raw descriptive analysis of this quantity gives an average of 10.7 and a standard derivation of 1.75. We assume it to be normally distributed and model both distribution parameters (mean, and standard deviation) with structured additive predictors. For both predictors we use the structure $\eta = f_1(\text{yday}) + f_2(\text{dhour}) + x'\beta$, where yday denotes the day of the year, dhour denotes the hour of the day and the term $x'\beta$ contains the intercept and linear effects. We employ cyclic P-splines (Eilers and Marx, 1996) for the smooth function f_1 and f_2 and a response function to transform the predictor for the standard deviation to the positive domain.

To illustrate the difference in interpretation between the softplus ($a = 10$) response function and the common log-link, we estimate the model for each of the two response functions with the BAMLSS software package via a MCMC algorithm (Umlauf et al., 2018).

DIC shows similar values for each model and does not clearly favour one of the response functions (exponential: 55529, softplus_10: 55389). Results regarding the parameters of the predictor for the mean are very similar for both models and are omitted here.

Instead, we focus on the predictor of the standard deviation of the reaction distribution and, as an example, on the effect of dhour. Figure 2 shows the effect of the time of the day on the predictor. Both models exhibit a similar pattern. For the softplus model, the values of the linear predictor are larger than 0.42 and thus covariate effects can be interpreted as effects on the parameter whereas in the other model the exponential function has to be applied and then the effect can be interpreted as multiplicative. In the early morning we observe an increased standard deviation for both models. The softplus model shows an increment by about 2.3 units while the exponential model outputs a multiplicative change by 3.5.

Due to the additive nature of its interpretation, the softplus function is even an alternative if both functions fit equally and it is up to the practitioner to decide which one to prefer.

References

Abramowitz, M. and Stegun, I. A. (1972). Handbook of Mathematical Functions. Number 55. In: *National Bereau of Standards: Applied Mathematics*. Wash-

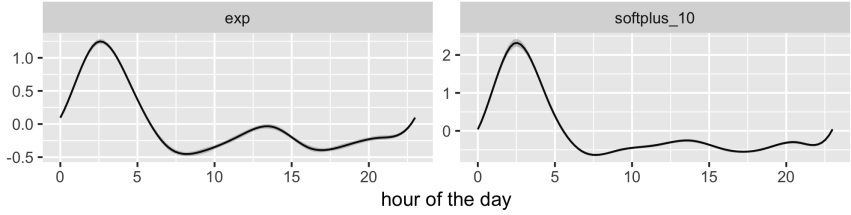


FIGURE 2. Posterior mean estimates on the predictor of the standard deviation together with 95% pointwise credible intervals for both response functions.

ington, D.C.: U.S. Government Printing Office, 10 ed.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data. Comparisons and applications of some estimators and tests. In *Journal of Applied Econometrics*, **1(1)**, 29 – 53.

Dugas, C., Bengio, Y., Belisle, F., Nadeau, C., and Garcia, R. (2001). Incorporating Second-Order Functional Knowledge for Better Option Pricing. In: *Advances in Neural Information Processing Systems*, **13(1)**, 451 – 457.

Eilers, P. H. C., and Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. In: *Statistical Science*, **11(2)**, 89 – 102.

Nielsen, F. and Sun, K. (2016). Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp inequalities. In: *Entropy*, **18(12)**, 442 – 467.

Spiegelhalter, D. J. , Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. In: *Journal of the Royal Statistical Society: Series B*, **64(4)**, 583 – 639.

Umlauf, N., Klein, N., and Zeileis, A. (2018). BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond). In: *Journal of Computational and Graphical Statistics*, **27(3)**, 612 – 627.

Umlauf, N. and Kneib, T. (2018). A Primer on Bayesian Distributional Regression. In: *Statistical Modelling*, **18(3-4)**, 219 – 247.

Non-crossing quantile regression via monotone B-spline varying coefficients

Gianluca Sottile[1], Vito MR Muggeo[1]

[1] University of Palermo, Department of Economics, Business and Statistics, Viale delle Scienze building 13, 90128, Palermo, Italy.

E-mail for correspondence: gianluca.sottile@unipa.it

Abstract: Quantile regression is often used to obtain nonparametric estimates of the conditional quantiles with respect to a continuous covariate. The presence of quantile crossing, however, leads to an invalid distribution of the response and makes it difficult to use the fitted model for prediction. In this paper, we show that crossing can be eliminated by estimating the multiple quantile curves jointly while modeling the regression coefficients via constrained B-splines. The estimating algorithm for such constrained optimization can be used to estimate quantile functions with the non-crossing property.

Keywords: quantile regression; non-crossing; monotone B-spline; fourth Dutch growth study

1 Introduction

Quantile regression (QR) was first developed by Koenker and Bassett (1978) to deal with estimation of quantiles of a continuous response variable as a function of multiple covariates. A well-known problem, coming from its nonparametric nature, when fitting several quantiles is represented by quantile crossing: e.g., the estimate of the 95th quantile, say, may be larger than that of the 99th quantile, at some covariate values. While this may not hinder the interpretation of the regression coefficients, quantile crossing can lead to unpleasant consequences when the fitted model is used for prediction or classification, e.g., in growth charts (Wei et al., 2006, Muggeo et al. 2013).
Crossing in quantile regression has been discussed by several authors, including He (1997), Chernozhukov et al.(2009), Bondell et al. (2010), Muggeo et al (2013) and Schnabel and Eilers (2013). More recently Frumento and Bottai (2016) propose the quantile regression coefficients modeling (QRCM) framework, wherein estimation is performed jointly by imposing a global structure for the quantile

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

curves, rather than estimating them individually, however noncrossing is just discouraged and not eliminated. Based on such idea, we develop a new approach within the L_1 optimization framework which does allow to constrain the quantile curves to fulfil the noncrossing property.

2 Methods

To illustrate, let $Q(\tau \mid \boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_\tau$ the quantile regression equation at quantile τ with corresponding check function to be minimized $L(\boldsymbol{\beta}(\tau))$. The main idea is to assume a parametric model for the regression coefficients namely, $\boldsymbol{\beta}_\tau = \boldsymbol{\theta}\boldsymbol{b}(\tau)$, where $\boldsymbol{b}(\tau)$ is a set of known basis functions of τ (e.g., polynomials) and $\boldsymbol{\theta}$ is the new parameter vector to be estimated by minimizing the *integrated* loss function

$$\bar{L}(\boldsymbol{\theta}) = \int_0^1 L(\boldsymbol{\beta}(\tau \mid \boldsymbol{\theta}))\mathrm{d}\tau. \tag{1}$$

$\bar{L}(\boldsymbol{\theta})$ is smooth and therefore optimization via Newton-Raphson algorithms is straightforward, although some numerical procedure has to be used to solve (partially) the integral in (1). Minimization of such integrated loss objective (1) is implemented in the `qrcm` package in R and polynomials are typically exploited as basis functions. However joint estimation via minimization of (1) in its current and plain formulation, only discourages and does not guarantee noncrossing.
We propose to estimate jointly all the quantile curves using a discrete approximation of the objective (1). Let $\boldsymbol{X}_0 = (x_1, \ldots, x_p)^{\mathrm{T}}$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be respectively the model matrix and the vector of responses. Given a set of probability values $\tau_1, \ldots, \tau_K$, the K quantile regression equations may be written as

$$\underbrace{\begin{bmatrix} Q(\tau_1 \mid \boldsymbol{x}) \\ Q(\tau_2 \mid \boldsymbol{x}) \\ \vdots \\ Q(\tau_k \mid \boldsymbol{x}) \end{bmatrix}}_{\boldsymbol{Y}_{[kn\times 1]}} = \underbrace{\begin{bmatrix} \boldsymbol{X}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{X}_0 \end{bmatrix}}_{\boldsymbol{X}_{[kn\times kp]}} \underbrace{\begin{bmatrix} \boldsymbol{\beta}(\tau_1) \\ \boldsymbol{\beta}(\tau_2) \\ \vdots \\ \boldsymbol{\beta}(\tau_K) \end{bmatrix}}_{\boldsymbol{\beta}_{[kp\times pq]}} \tag{2}$$

where the first block, say, $\boldsymbol{\beta}(\tau_1)$ of the whole regression coefficient vector $\boldsymbol{\beta}$ represents the covariate effects at quantile τ_1. In order to express each coefficient as a smooth function of the probability values $\tau_1, \ldots, \tau_K$, we first permute $\boldsymbol{\beta}$ into $\tilde{\boldsymbol{\beta}}$ such that its first elements refer to the first regression coefficient relevant to $\tau_1, \tau_2, \ldots, \tau_K$. Namely $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_{\boldsymbol{1}}^{\mathrm{T}}, \ldots, \tilde{\boldsymbol{\beta}}_{\boldsymbol{p}}^{\mathrm{T}})^{\mathrm{T}}$ where each block $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}$ includes the coefficients of covariate X_j corresponding to $\tau_1, \tau_2, \ldots, \tau_K$. In order to obtain a smooth pattern, we propose to express each 'block' via B-splines, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = \boldsymbol{B}\boldsymbol{\theta}_j$. Overall we can write $\tilde{\boldsymbol{\beta}} = (\boldsymbol{I}_{\boldsymbol{p}} \otimes \boldsymbol{B})\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ collects all the covariate-specific $\boldsymbol{\theta}_j$s.
We 'back' permute $\tilde{\boldsymbol{\beta}}$ to express $\boldsymbol{\beta}$ as a function of the basis coefficients $\boldsymbol{\theta}$, and plugging in (2) we obtain the augmented design matrix. Hence, by building such aforementioned design matrix, the response vector $\mathbf{1}_k \otimes \boldsymbol{y}$ and the weights $\boldsymbol{\tau} \otimes \mathbf{1}_n$, model estimation is carried out via an usual L_1 optimization algorithms. However, like in the QRCM framework, simply re-parametrizing the coefficient vector in terms of spline coefficients does not guarantee noncrossing, and further

constraints are requested. Fortunately, the B-spline parametrization of each regression coefficient allows to set noncrossing constraints straightforwardly by enforcing positiveness of the first order differences of the spline coefficients. Positiveness is easily obtained by imposing a system of linear inequality constraints: i.e. $\boldsymbol{R}_{[p(q-1)\times pq]}\boldsymbol{\theta}_{[pq\times 1]} \geq \mathbf{0}_{[p(q-1)\times 1]}$, where $\boldsymbol{R}$ is the matrix of monotonicity constraints. These inequality constraints can be easily accounted by the Frisch-Newton algorithm for constrained optimization as explained in Koenker and Ng (2002). Computational efficiency is attained via managing sparse matrices as in the usual QR framework.

3 Application

We apply the proposed approach to a well known dataset referring to the Fourth Dutch Growth Study (Fredriks et al., 2000). It is a cross-sectional study that measures growth and development of the Dutch population between the ages 0 and 21 years. The study collects, among other variables, height, weight, head circumference and age for 7482 males and 7018 females. For illustrative purpose, here we consider only a subsample of $n = 1000$ observations and the BMI as main outcome.

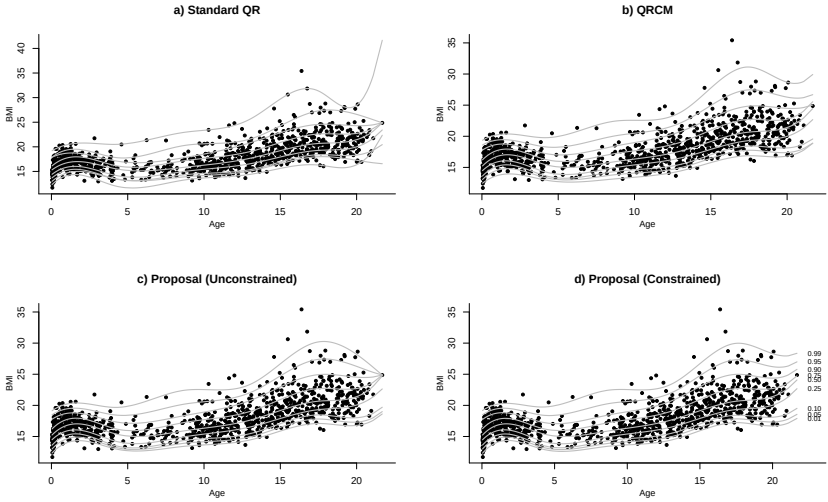


FIGURE 1. Estimated quantiles at $\tau = \{.01, .05, .10, .25, .50, .75, .90, .95, .99\}$ of the BMI with respect to AGE according different approaches. Panels **c** and **d** refer to our proposal, both unconstrained (i.e., without noncrossing constraints) and constrained (i.e., with noncrossing constraints), respectively. For the QRCM method and our proposal, a 3rd degree B-splines (rank 8) is used to model τ.

The relationship between age and BMI is well-known to be non-linear, hence we model the age effect via a 3rd degree B-spline basis with rank equal to 8. We

compared our proposal, both unconstrained and constrained, with the standard QR and the QRCM. Results are portrayed in Figure 1
The standard QR approach estimates each curve separately, leading to crossing curves several times, at about 3-4 years and beyond 20 years. The quantile curve at .99 has also a rise on the right side which is probably a model artefact related to the extreme quantile being estimated. On the other hand, joint estimation (panels b) and c)) with a smooth parameterization for the regression coefficients, strongly alleviates noncrossing, but without eliminating it completely (see quantile curves at higher ages). However inequality constraints easily included in the optimization algorithm, guarantee noncrossing curves as reported in panel d).

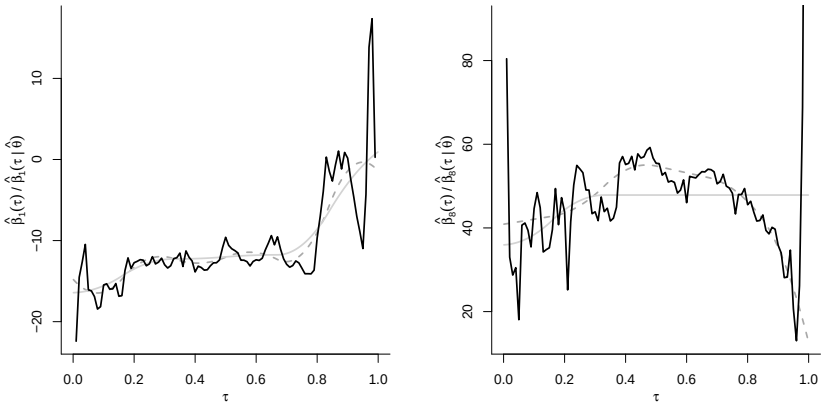


FIGURE 2. Two estimated regression coefficients. Black solid lines refer to standard QR; grey broken and solid lines refer to our proposal, both unconstrained and constrained, respectively.

Figure 2 compares the estimated regression coefficients corresponding to the 1st and the 8th basis function coming from the standard QR with the same coefficients coming from our approach with and without monotonicity constraints. The estimate coming from the simple QR, as a consequence of individual estimation, is clearly very wiggly and inefficient, with an important background noise.

4 Conclusion

In this paper we have proposed an efficient algorithm to estimate jointly multiple quantile curves as a function of several covariates, possibly modelled via B-splines. Estimation is performed via minimization of a naive L_1 loss function with proper augmented response, design matrix and weights. The non-crossing property is attained through simple inequality to perform constrained optimization. While estimation is based on a set of K fixed probability values, any desired quantile curve can be obtained by evaluating the B-spline basis at that fixed probability

value. Preliminary simulations have shown promising results. Current implementation has been discussed without any penalty on the coefficients but inclusion of any regularized criterion depending upon (a fixed) tuning parameters appears feasible and worth discussing. Also quantifying the estimates uncertainty, beyond any bootstrap solution, appears a crucial point to be investigated.

References

Koenker, R. and Bassett G. Jr. (1978). Regression Quantiles. *Econometrica*, **46**, 33 – 50.

Wei, Y., Pere, A., Koenker, R., and He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, **25**, 1369 – 1382.

Muggeo, V.M.R., Sciandra, M., Tomasello, A., and Calvo, S. (2013). Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental and Ecological Statistics*, **20**, 519 – 531.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, **96**, 559 – 575.

Bondell, H.D., Reich, B.J., and Wang, H. (2010). Non-crossing quantile regression curve estimation. *Biometrika*, **97**, 825 – 838.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, **51**, 186 – 192.

Frumento P. and Bottai, M. (2016). Parametric modeling of quantile regression coefficient functions. *Biometrics*, **72**, 74 – 84.

Koenker, R. and Ng, P. (2005). A Frisch-Newton Algorithm for Sparse Quantile Regression. *Acta Mathematicae Applicatae Sinica*, **21**, 225 – 236.

Fredriks, A.M. and van Buuren, S. and Wit, J.M. and Verloove-Vanhorick, S.P. (2000). Body index measurments in 1996-7 compared with 1980. *Arch Dis Child*, **82**, 107 – 112.

Schnabel, S.K., and Eilers, P.H.C. (2013). Simultaneous estimation of quantile curves using quantile sheets. *AStA Adv Stat Anal*, **97**, 77 – 87.

Flexible multivariate point processes with applications to modelling football matches

Santhosh Narayanan[1], Ioannis Kosmidis[1], Petros Dellaportas[2]

[1] University of Warwick, UK
[2] University College London, UK

E-mail for correspondence: S.Narayanan.1@warwick.ac.uk

Abstract: We consider the modelling of sequences of multivariate point processes where the occurrence rate depends on past occurrences within the process. Building on a traditional model for point processes, the Hawkes process, the main idea is to take advantage of the decomposition that motivated partial likelihood to separate the modelling of the event types and the occurrence times. We present an application on the modelling of event-sequences in football, where match events can be treated as a multivariate spatio-temporal point process. The aim is to provide inferences about previously unquantified measures governing the dynamics of the game as well as predicting the occurrence of events of interest, such as goals, in a specified interval of time.

Keywords: Point Processes; Partial Likelihood; Bayesian Inference.

1 Introduction

A point process is a probabilistic model for a random collection of points on some space often used to describe the occurrence of random events over time. The Hawkes process (HP), introduced in Hawkes (1971), is a model for *self-exciting* point processes, where the chance of a subsequent occurrence is increased for some time period after the initial occurrence. We are developing multivariate point processes suitable for a wide range of applications that overcome the limitations of the HP model. In this paper we will present the challenging case of modelling in-game events in football. Specifically, we model all touch-ball events, where a player acts on the ball with some part of their body, as a mutually-exciting point process that depends on the past history of events and occurrence times.

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

2 Marked point processes

A marked point process consists of occurrence times $\boldsymbol{t} = \{t_i : t_i \in \Re, i = 1, \ldots, n\}$ and marks (event types) $\boldsymbol{m} = \{m_i : m_i \in \{1, \ldots, M\}, i = 1, \ldots, n\}$ with $t_1 < \ldots < t_n$. We define the history or filtration $\mathcal{F}_t$ at time t of the process as $\mathcal{F}_t = \{(t_j, m_j) : t_j \in \boldsymbol{t}, m_j \in \boldsymbol{m}, t_j \leq t\}$. We shall work under the setting where we observe a process from its beginning say at time $t = 0$ and $\mathcal{F}_0 = \emptyset$. The task is then to model each event, the pair of (t_i, m_i), given $\mathcal{F}_{t_{i-1}} (i = 1, \ldots, n)$.

3 Separating times and marks

The characteristic property of the HP is its self-exciting intensity, which leads to clustering of events in both the time and mark spaces. For applications like event sequences in team sports that is studied in this paper, the occurrence times of events appear to be uniformly spaced, and using the HP model may not be appropriate. In the mark space, self- or cross-excitation is the increase in the chance of a mark caused by the occurrence of another mark and we definitely want to capture these effects between events in our model. To restrict the self-exciting property of the process to the mark space, we take advantage of the decomposition of a multivariate model in Cox (1975). Specifically, the full likelihood of a marked point process can always be factorised as

$$\prod_{i=1}^{n} g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta}) \prod_{i=1}^{n} f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}), \tag{1}$$

where g and f are the probability density and mass functions for times and marks respectively, and $\boldsymbol{\zeta}, \boldsymbol{\theta}$ are the unknown parameter vectors. The second product in (1) is a *partial likelihood* based on the mark sequence $\boldsymbol{m}$.

4 Model specification

We specify the probability mass function for marks as

$$f(m_i \mid t_i, \mathcal{F}_{t_{i-1}}; \boldsymbol{\theta}) = \frac{\delta_{m_i} + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta_{m_j}(t_i - t_j)} \gamma_{m_j \to m_i}}{1 + \sum_{t_j < t_i} \mathrm{e}^{\alpha - \beta_{m_j}(t_i - t_j)}},$$

where $\delta_{m_i} \in [0, 1]$ is the the background intensity of mark m_i. The parameter $\alpha \in \mathbb{R}$ controls the magnitude of excitation, $\beta_{m_j} > 0$ are the event dependent rates at which the excitation decays over time and $\gamma_{m_j \to m_i} \in [0, 1]$ is the probability a parent event of mark m_j generates an off-spring event of mark m_i. By definition, $\sum_{m=1}^{M} \delta_m = 1$ and $\sum_{m=1}^{M} \gamma_{m_j \to m} = 1$ for every $m_j \in \{1, \ldots, M\}$. The probability density function for the occurrence times is set to

$$g(t_i \mid \mathcal{F}_{t_{i-1}}; \boldsymbol{\zeta}) \sim \mathbf{Gamma}[a(m_{i-1}), b(m_{i-1})],$$

where, as the notation indicates, the shape and rate parameters of the gamma distribution depend on the mark of the last observed event.

5 Application to football matches

For each touch-ball event we have the event type, time-stamp, (x, y) co-ordinates of its location in the field, team and player ids, game half, event outcome (successful/unsuccessful) and the end (x, y) co-ordinates if the event is a pass. In total we have approximately 1.1 million events recorded over 760 games from the 2013/14 and 2014/15 English Premier League seasons. A snapshot of the data is shown in Table 1.

TABLE 1. Dataset snapshot

second	minute	team_id	player_id	type	outcome	x	y	end_x	end_y
0	0	665	68312	Pass	Successful	49.1	51.0	52.5	44.8
2	0	665	14036	Pass	Successful	52.2	44.5	36.7	60.6
3	0	665	79050	Pass	Successful	36.7	60.6	24.9	39.1
5	0	665	14107	Pass	Unsuccessful	25.0	37.9	97.0	22.9
11	0	660	73379	Win	Successful	1.9	73.7	1.9	73.7
15	0	660	73379	Pass	Successful	5.5	65.3	20.9	21.5
17	0	660	6292	Pass	Successful	20.9	21.5	29.0	38.5
19	0	660	26820	Foul	Successful	25.8	37.4	25.8	37.4

6 Estimation and inference

For a total of S games, the likelihood is

$$\prod_{s=1}^{S}\left[\prod_{i=1}^{n_s} g(t_{s,i} \mid \mathcal{F}_{t_{s,i-1}};\boldsymbol{\zeta})\prod_{i=1}^{n_s} f(m_{s,i} \mid t_{s,i},\mathcal{F}_{t_{s,i-1}};\boldsymbol{\theta})\right],$$

where n_s is the number of events in game s, $t_{s,i}$ and $m_{s,i}$ are the occurrence time and mark of the i-th event in game s respectively.

6.1 Prior specification

We specify exponential priors for the vector of decay rates, $\boldsymbol{\beta} \sim \mathbf{Exp}(0.01)$ and for the Gamma shape and rate parameters, $\boldsymbol{a} \sim \mathbf{Exp}(0.01)\,, \boldsymbol{b} \sim \mathbf{Exp}(0.01)$. A Normal shrinkage prior on the unbounded parameter $\alpha \sim \mathbf{N}(0, \sigma_\alpha)$ with a hyper-prior $\sigma_\alpha \sim \textbf{half-Cauchy}(0, 5)$. We use non-informative priors on the constrained parameter vectors $\boldsymbol{\delta} \sim \textbf{Dirichlet}(1)$ and each row of the $\boldsymbol{\gamma}$ matrix, $\boldsymbol{\gamma}_r \sim \textbf{Dirichlet}(1)$ for every $r \in \{1, \ldots, M\}$.
The first 16 games of the 2013/14 season is used as training data. We obtain parameter samples by running 3 parallel MCMC chains of 1000 iterations each after burn-in. From the results of fitting our model, we highlight below the elements of event conversion matrix $\boldsymbol{\gamma}$ that can provide insights towards understanding the dynamics of the game of football.

6.2 Event conversion rates

The posterior means of the event conversion rate parameters in Table 2 highlight the advantage a team has over the opposition when playing at home. The higher

conversion rates between Home Passes (row 3, col 1) as compared to Away passes (row 6, col 3) indicate that the home team is more likely to keep possession of the ball. The conversion rates to a Home Shot (col 2) are also consistently higher compared to an Away Shot (col 4), meaning the home team is also more likely to take advantage of their possession and make goal scoring attempts.

TABLE 2. Posterior means of the conversion rate parameters for selected events. The suffix Pass_S refers to a successfully completed Pass event.

$\gamma_{m_j \to m_i}$	Home_Pass_S	Home_Shot	Away_Pass_S	Away_Shot
Home_Win	0.34	0.04	0.07	0.00
Home_Dribble	0.20	0.06	0.00	0.00
Home_Pass_S	0.75	0.02	0.00	0.00
Away_Win	0.09	0.00	0.35	0.02
Away_Dribble	0.04	0.00	0.23	0.01
Away_Pass_S	0.00	0.00	0.71	0.02

7 Prediction framework

We have N samples of the posterior parameter vector, $\boldsymbol{p}_k = \{\boldsymbol{\zeta}_k, \boldsymbol{\theta}_k\}$ for $k = 1, \ldots, N$. For a single game, for each $\boldsymbol{p}_k$, we generate M simulations of the game in the interval $(T, T+d)$, where is T is the game time at which prediction is made and d is the duration of the prediction interval. Each simulation is carried out iteratively as follows; we first simulate the occurrence time of next event given history and then its mark given time and history. This generated pair of (time, mark) is then added to the history as the most recent event. The simulation is stopped when the time exceeds $T + d$.

7.1 Validation

For each game in the test set, we get events counts from the $N \times M$ simulations and validate against the true counts observed in the game during the prediction interval. For comparison, we train a homogeneous Poisson model for each mark on the training data to use as a baseline. To evaluate the performance of the predictors, we calculate the sum of the log predictive probabilities over all marks. Table 3 shows the results from validation for $d = 2$ minute and $d = 5$ minute intervals with a prediction start time of $T = 10$ minutes in both cases. For each game we used $N = 100$ samples from the posterior and generated $M = 500$ simulations of the game in the interval for each sample.

In the case of $d = 2$, the model performs better than the baseline in 8 out of 10 games. In particular, the model performs exceptionally better in games 1, 5 and 10, which appear to be instances of relatively unpredictable event sequences indicated by their larger magnitudes. In the case of $d = 5$, the model outperforms the baseline in 6 out of 10 games indicating that the differences between the model and baseline reduce with the size of interval.

TABLE 3. Sums of log predicted probabilities for the first 10 games in the test set from predictions made on 2 minute and 5 minute intervals.

$T = 10$ min	$d = 2$ min		$d = 5$ min	
	Model	Baseline	Model	Baseline
Test set game 1	-29.31	-34.52	-40.19	-43.22
Test set game 2	-27.30	-27.93	-39.30	-35.95
Test set game 3	-24.11	-25.50	-36.74	-38.86
Test set game 4	-24.95	-24.87	-41.90	-46.45
Test set game 5	-30.75	-35.84	-46.46	-44.92
Test set game 6	-25.63	-24.46	-43.77	-42.35
Test set game 7	-22.24	-23.22	-39.18	-40.87
Test set game 8	-24.95	-27.36	-40.10	-38.36
Test set game 9	-23.76	-23.80	-34.93	-35.78
Test set game 10	-30.59	-35.48	-41.28	-44.64

8 Discussion

Our initial results indicate that the flexible specification of multivariate point processes that has been introduced here is suitable for the application of modelling in-game events in football. We are able to provide inferences about previously unquantified measures governing the dynamics of the game as well as predicting the occurrence of events in a specified interval of time.

8.1 Future work

Future work includes the following

Game states as covariates Our aim here is to capture the state of the game using a set of quantities (e.g. teams, location, score) measurable from the data available to us. Covariates can be incorporated in the conversion rate parameters using a baseline logit specification in Agresti and Kateri (2011). For example, in the application for this talk, team information can be incorporated as

$$\log\left(\frac{\gamma_{m_j\to m}(t_1,t_2)}{\gamma_{m_j\to M}(t_1,t_2)}\right) = \kappa_{m_j\to m} + \mu_{t_1,m} - \nu_{t_2,m} \qquad m \in \{1,\ldots,M-1\}\,,$$

where κ is the baseline conversion parameter, t_1 is the team in possession of the ball (attacking) and t_2 is the defending team. μ and ν are the team abilities to make or stop a conversion to mark m for the attacking and defending teams respectively.

Alternative application Another application, for the statistical methodology developed as a part of this project, which we wish to explore is in cybersecurity. Internet network companies are prone to incur huge losses due to malicious server attacks. These attacks can result in a breach of data security, reduced

bandwidth for data transfer and failures in user-network connections. Typically, the modelling task in such a scenario would involve predicting the time to the next attack and its type.

References

Agresti, Alan and Kateri, Maria (2011). *Categorical data analysis*. New York: Springer.

Cox, David R. (1975). Partial likelihood. *Biometrika*, **62**, 269 – 276.

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

Diggle, Peter J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. London: Chapman & Hall.

Hawkes, Alan G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58**, 83 – 90.

Rasmussen, Jakob Gulddahl (2013). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, **15**, 623 – 642.

Longitudinal models with informative time measurements

Inês Sousa[1], Adriana Vieira[2], Luis Castro[3]

[1] Centre of Molecular and Environmental Biology and Department of Mathematics, University of Minho, Portugal
[2] Department of Mathematics, University of Minho, Portugal
[3] Hospital de Braga, Portugal

E-mail for correspondence: isousa@math.uminho.pt

Abstract: In longitudinal studies individuals are measured repeatedly over a period of time for a response variable of interest. In classical longitudinal models the longitudinal observed process is considered independent of the times when measurements are taken. However, in medical context it is common that patients in worst health condition are more often observed, whereas patients under control do not need to be seen so many times. Therefore, longitudinal models for data with this characteristics should allow for an association between longitudinal and time measurements processes. In this work we propose a joint model for the distribution of longitudinal response and time measurement using maximum likelihood methodology to make inference on the model parameters. A simulation study is conducted and the model proposed is fitted to a data set on progression of oncological biomarkers in breast cancer patients.

Keywords: longitudinal; follow-up times; biomarkers

1 Introduction

Longitudinal data analysis plays a key role in a multiplicity of distinct areas, including medicine. One of the great difficulties in this type of study is related to different observation times for different individuals us in unbalanced studies, times that are usually treated as independent of the response variable. An even greater difficulty occurs when the different observation times are related with the response variable. For example, the doctor decides to mark more, or fewer, appointments according to the patient's state of health. That is, patients are usually measured according to their clinical condition. In cases where observation times and response variables are related, a simple longitudinal analysis will produce

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

biased estimators (Lin et al, 2004). The general linear model (Diggle et al. 2002) described for longitudinal data analysis, assumes a deterministic follow-up time process that is noninformative about the outcome longitudinal response. Therefore, it is necessary to develop new methodologies that allow the inclusion of this characteristic. We intend to present here some alternative model that fits into this problematic.

Others have been proposed models for situations where the longitudinal response variable and the time measurements are related. More lately, Fang et al (2016) proposed a joint model for longitudinal and informative observation using two random effects with additive mixed effect model for observation time. Cheng et al (2015) proposed a model where the probability structure of the observation time process is unspecified. Lipsitz et al (2002) consider a model where assumptions regarding the time measurements process result in the likelihood function separated in the two components. Lin et al (2004) approach is base on missing data and proposed a class of inverse intensity-of-visit process-weighted estimators in marginal regression models. Fitzmaurice et al (2006) consider the same problem when the longitudinal response is binary.

In this work we consider a response longitudinal variable with Gaussian distribution. We propose a model where the follow-up time process is stochastic. The model is described through the joint distribution of the observed process and the follow-up time process. Estimation of model parameters is through maximum likelihood. We conducted a simulation study of longitudinal data where model parameter estimates are compared, when using the model proposed and ignoring the association between processes. Finally, the model proposed is applied to a real data set when monitoring for biomarkers CEA and CA15.3 on breast cancer progression. In these cases the follow-up time process should be considered dependent on the longitudinal outcome process.

2 Model Proposal

Consider data observed for m individuals, where $\mathbf{Y}_i$ is the vector of longitudinal responses and $\mathbf{T}_i$ is the vector of time measurements, both for subject $i = 1, ..., m$. It is assumed a model for the joint distribution of the longitudinal outcome process $\mathbf{Y}$ and the time measurement process $\mathbf{T}$ through an unobserved stationary Guassian process $\mathbf{W}(s)$. Therefore, we propose the following model

$$[\mathbf{Y}_i|\mathbf{W}(s), \mathbf{T}_i] \sim \text{Normal}(\mu + \mathbf{W}(t_{ij}), \tau^2)$$

and intensity function for the time measurement process at time t_{ij}, $j = 1, ..., n_i$

$$\lambda(t_{ij})|\mathbf{W}_{history}(s) \sim \exp\left\{\mathcal{F}(\mathbf{W}_{history}(t_{ij}))\right\},$$

where, μ is the expected value that can include regression parameters and $\mathcal{F}(.)$ is any defined function. For example, to describe a time measurement process dependent on the progression of the patients unobserved health condition, we might define

$$\lambda(t_{ij})|\mathbf{W}_{history}(s) = \exp\left(\alpha + \beta \sum_{s=(t_{ij}-4)}^{t_{ij}} W(s)w(t_{ij} - s)ds\right)$$

where $\sum_{s=(t_{ij}-4)}^{t_{ij}} w(t_{ij}-s)=1$.
Notice that, process $\mathbf{W}(\mathbf{s})$ is continuous in time, though only a discrete version of it is observed at t_{ij}.
For inference we consider a likelihood approach, where the likelihood function is

$$\begin{aligned}[\mathbf{Y},\mathbf{T}] &= \prod_{i=1}^{m}[\mathbf{Y}_i,\mathbf{T}_i] \\ &= \prod_{i=1}^{m}\int_W [\mathbf{Y}_i|\mathbf{W}][\mathbf{T}_i|\mathbf{W}][\mathbf{W}]dW \\ &= \prod_{i=1}^{m} E_{\mathbf{W}|\mathbf{Y}_i}\left([\mathbf{T}_i|\mathbf{W}][\mathbf{Y}_i|\mathbf{W}_0]\frac{[\mathbf{W}_0]}{[\mathbf{W}_0|\mathbf{Y}_i]}\right)\end{aligned}$$

where, $\mathbf{W}_0$ is the subset with observed time points and $\mathbf{W}_1$ is the subset with unobserved time points.,
We then generate g samples from $[\mathbf{W}|\mathbf{Y}_i]$ and approximate the expectation by its Monte Carlo version

$$L_{MC}(\theta)=\prod_{i=1}^{m}\frac{1}{g}\sum_{j=1}^{g}\left(f(\mathbf{T}_i|\mathbf{W}_j)f(\mathbf{Y}_i|\mathbf{W}_{0j})\frac{f(\mathbf{W}_{0j})}{f(\mathbf{W}_{0j}|\mathbf{Y}_i)}\right)$$

3 Results

A simulation study is conducted and results are presented when fitting both, the model proposed and the general linear longitudinal model (Diggle et al, 2002).
A data set on oncological biomarkers, CEA and CA15.3, for breast cancer patients is available. There are data available on 550 patients, with a mean number of measurements per subject of 7.6 (median=7 and sd=4.1), with a total number of observations for CEA of 4166 and 5166 for CA15.3. In Figure 1 longitudinal profiles of CEA and CA15.3 (logarithm scale) of a random sample of 10 patients is shown, with black dots representing the location of the time measurements and the solid black line is the respective smooth spline for all data.
The proposed model is fitted to this data and results are compared with the classical longitudinal model.

Acknowledgments: This work was supported by project 028248/SAICT/2017 funded by COMPETE2020 (Programa Operacional Competitividade e Internacionalização) in its component FEDER (Fundo Europeu de Desenvolvimento Regional) and by FCT (Fundação para a Ciência e a Tecnologia, I.P.). This work was also supported by the PhD grant 128191/2016 funded by FCT I.P., by the Ministério da Ciência, Tenologia e Ensino Superior (MCTES) to the second author.

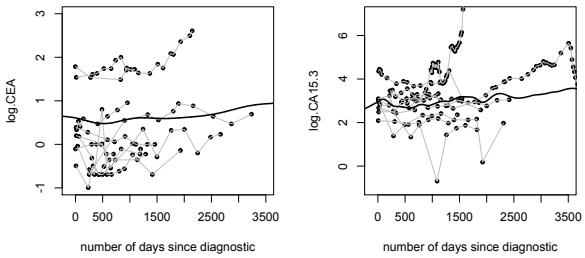


FIGURE 1. Longitudinal profiles of a random sample of 10 patients measured for CEA and CA15.3.

References

Chen, Y., Ning, J. and Cai, C.Y. (2015). Regression analysis of longitudinal data with irregular and informative observation times. *Biostatistics*, **16**, 727–739.

Diggle, P.J., Heagerty, P., Liang, K-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Oxford University Press (2nd edition).

Fang, S., Zhang, H.X. and Sun, L.Q. (2016). Joint analysis of longitudinal data with additive mixed effect model for informative observation times. *Journal of Statistical Planning and Inference*, **169**, 43–55.

Fitzmaurice, G., Lipsitz, S., Ibrahim, J., Gelber, R. and Lipshultz, S. (2006). Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. *Biostatistics*, **7**, 469–485.

Lin, H., Scharfstein, D. and Rosenheck, R. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society, Series B*, **66**, 791–813.

Lipsitz, S., Fitzmaurice, G., Ibrahim, J., Gelber, R. and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, **58**, 621–630.

Estimation of the probability of a giant 'doomsday' solar geomagnetic storm

Pedro Puig [12], David Moriña[12], Isabel Serra[234], Álvaro Corral[1235]

[1] Barcelona Graduate School of Mathematics (BGSMath), Spain
[2] Departament de Matemàtiques, Univ. Autònoma de Barcelona (UAB), Spain
[3] Centre de Recerca Matemàtica, Campus Bellaterra, Spain
[4] Barcelona Supercomputing Center, Barcelona, Spain
[5] Complexity Science Hub Vienna, Austria

E-mail for correspondence: ppuig@mat.uab.cat

Abstract: Intense solar geomagnetic storms can cause severe damage to worldwide electrical systems and communications. In this work, a counting process with Weibull inter-occurrence times is used in order to estimate the probability of extreme geomagnetic events.

Keywords: Arrival process; Carrington event; Weibull regression.

1 Introduction

A geomagnetic storm is a disturbance in the magnetosphere quantified by changes in the Dst (disturbance-storm time) index. This index measures the globally averaged change of the horizontal component of the Earth's magnetic field at the magnetic equator and it is recorded once per hour. During quiescent times, the Dst index varies between -20 and +20 nT (nanotesla). The Carrington event is the largest known example of geomagnetic storm, occurred by the end of August and early September 1859 and associated to a minimum Dst under -850 nT. Richard C. Carrington was observing sunspots on the solar disk and saw a large solar flare (Figure 1) with optical brightness lasting several minutes and equaling that of the background sun, due to the destabilization of a large region of the sun causing an extremely fast coronal mass ejection towards Earth. Nowadays, a Carrington-like geomagnetic storm would be catastrophic for electrical systems and communications.
The Dst index has been traditionally modelled by means of its physical properties (Riley 2012, Kataoka, 2013) although some work has also focused on exploring its statistical properties (Yokoyama and Kamide, 1997). As far as we know, all efforts in statistical modelling have been based on the assumption that the occurrence of a geomagnetic storm follows an homogeneous Poisson counting process

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

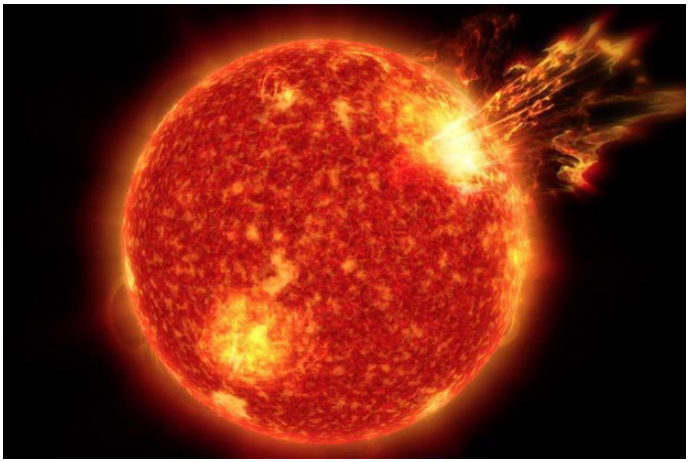


FIGURE 1. Huge solar flare recorded by NASA

(see for instance Riley, 2012). To analyse the process of temporal occurrence of geomagnetic storms we use the Dst index, recorded hourly from 1957-01-01 to 2017-12-31 and available from the World Data Center for Geomagnetism in Kyoto:

`http://wdc.kugi.kyoto-u.ac.jp/`

2 Statistical Modelling

When the Dst signal crosses a fixed negative threshold from above this defines the occurrence time or starting time of a geomagnetic storm with an intensity limited by the threshold. The inter-occurrence time is the time between two consecutive storms below the threshold, that is just the difference of their occurrence times. We have found that the distributions of inter-occurrence times seem to be well fitted by Weibull distributions. The choice of the Weibull distribution is based on purely empirical grounds, as a common generalization of the homogeneous Poisson process, which is recovered as a particular case. In terms of the complementary cumulative distribution function, the Weibull distribution takes the form $S(t) = P(X > t) = e^{(-t/\tau)^\gamma}$, where X is the random variable representing inter-occurrence times and γ, τ are respectively the parameters of shape and scale. The details of this research is fully described in Moriña et al. (2019).
It is found that the scale parameter of the inter-occurrence times distribution grows exponentially with the absolute value of the intensity threshold defining the storm, whereas the shape parameter keeps rather constant (see Figure 2).
Therefore, the inter-occurrence times were fitted using a Weibull regression model where the scale parameter changes with the threshold of the storm, T, according to $\log(\tau) = \beta_0 + \beta_1 T$ and the shape parameter γ is constant. The estimates are $\log(\hat{\gamma}) = -0.39$ (SE $= 0.023$), $\beta_0 = 2.96$ (SE $= 0.17$) and $\beta_1 = -0.0121$ (SE $=$ 0.0008). Because the shape parameter is below one, these Weibull distributions

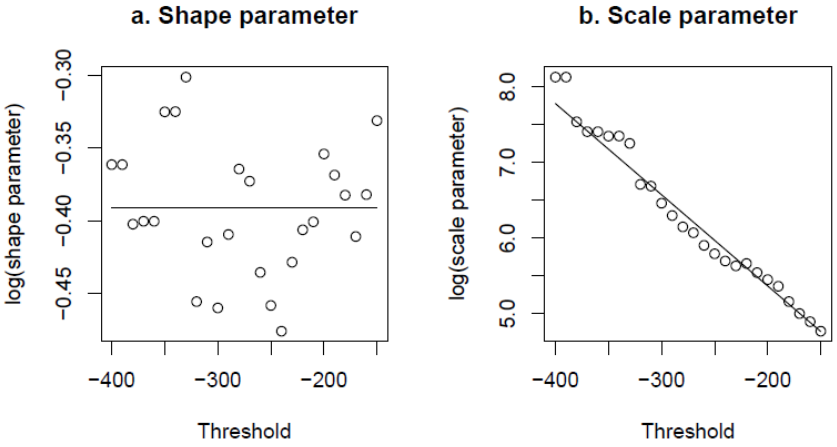


FIGURE 2. Relationship between Dst threshold (in nT) and Weibull shape (a.) and scale (b.) parameters, in log-scale, with scale parameter in days. Intensity thresholds range from -400 nT to -150 nT. The points correspond to maximum–likelihood estimates of the shape and scale parameters for fixed threshold values.

have a decreasing failure rate (DFR) or decreasing hazard. Therefore the associated count distributions (number of storms within this threshold in a period of time) should be overdispersed, a result confirmed in practice.

Knowing that the original Carrington event happened in 1859, about 58000 days ago, one can compute the probability of having a Carrington or more intense event during the next decade (2019-2028) conditioned to the fact that no event like this has happened since 1859, in this way,

$$P(X \leq t_c + t_d \mid X \geq t_c) = \frac{S(t_c) - S(t_c + t_d)}{S(t_c)} = 1 - \exp\left[\left(\frac{t_c}{\tau}\right)^{\gamma} - \left(\frac{t_c + t_d}{\tau}\right)^{\gamma}\right] = 0.0092,$$

with $t_c = 58000$ days and $t_d = 3652$ days (10 years). According to this model, the estimated probability is 0.92%, with a 95% confidence interval equal to $[0.46\%, 1.88\%]$. The value reported by Riley (2012) was about 12%, in sharp contrast with our result.

We can also estimate the expected number of geomagnetic storms for a period of time t, $E(N(t))$, with different thresholds. It can be done using the asymptotic approximation, $E(N(t)) = t/\mu$, where m is the average inter-occurrence time, in this case coming from the Weibull distribution, given by $\mu = \tau\Gamma(1 + 1/\gamma)$. For instance, for thresholds of -400 nT and -800 nT the estimated expected number of geomagnetic storms are 1.63 per 10 years and 1.37 per $1,000$ years, respectively.

Anyway, the estimated probability of a Carrington-type event, 0.92% for the next 10 years, is not insignificant. Public authorities should have a protocol of action for coping with this kind of disaster. In 2013 Lloyd's of London and Atmospheric

and Environmental Research (AER) published a report estimating the cost of a Carrington-like event to the U.S.: "The total U.S. population at risk of extended power outage from a Carrington-level storm is between 20-40 million, with durations of 16 days to 1-2 years. The duration of outages will depend largely on the availability of spare replacement transformers. If new transformers need to be ordered, the lead-time is likely to be a minimum of five months. The total economic cost for such a scenario is estimated at 0.6-2.6 trillion USD".

Acknowledgments: This work was partially supported by grants from ISCIII cofunded by FEDER funds /ERDF: RD12/0036/0056, PI11/02090, from AGAUR: 2014SGR 756, 2014SGR1307 and from MINECO: FIS2015-71851-P, MTM2015-69493-R. D. Moriña acknowledges support through M. de Maeztu Progr. for Units of Excellence in R&D: MDM-2014-0445 and from Fundación Santander Universidades.

References

Kataoka, R. (2013). Probability of occurrence of extreme magnetic storms. *Space Weather*, **11**, 214 – 218.

Moriña, D., Serra, I., Puig, P. and Corral, A. (2019). Probability estimation of a Carrington-like geomagnetic storm. *Scientific Reports*, **10**, 1 – 9. DOI: 10.1038/s41598-019-38918-8

`https://www.nature.com/articles/s41598-019-38918-8`

Riley, P. (2012). On the probability of occurrence of extreme space weather events. *Space Weather*, **10**, 1 – 12.

Yokoyama, N. and Kamide, Y. (1997). Statistical nature of geomagnetic storms. *Journal of Geophysical Research: Space Physics*, **102**, 14215 – 14222.

Diagnostic accuracy of ultrasound measures on large for gestational age: a Bayesian regression model for ROC curves with constraints

Zhen Chen[1], Soutik Ghosal[1]

[1] National Institutes of Health, United States

E-mail for correspondence: zhen.chen@nih.gov

Abstract: Predicting large fetuses at birth is of great interest for obstetricians. Using an NICHD Scandinavian study that collected longitudinal ultrasound examinations during pregnancy, we estimate diagnostic accuracy parameters of estimated fetal weight (EFW) at various times during pregnancy in predicting large for gestational age. We propose a placement value based Bayesian Beta regression model with random effects to ROC curves. The use of placement values allows us to model covariate effects directly on the ROC curves and the adoption of Bayesian approach accommodates *a priori* information and constraints. The proposed methodology is shown to perform better than a standard approach and its application to the Scandinavian study data suggests that diagnostic accuracy of EFW can improve almost 75% from week 17 to 37 of gestation age.

Keywords: AUC; Placement values; Macrosomia.

1 Introduction

Predicting large fetuses at birth is of great interest for obstetricians, as these newborns are usually at higher risk for perinatal morbidity and potentially long term metabolic complications. Using an NICHD Scandinavian study that collected longitudinal ultrasound examinations during pregnancy (Bakketeig et al. 1993), we seek to estimate diagnostic accuracy parameters of estimated fetal weight (EFW) at some pre-specified gestational age (GA) in discriminating large for gestational age (LGA). Here EFW is derived using biparital diameter, middle abdominal diameter and femur length (Hadlock et al. 1985), and LGA is defined as birth weight greater than 90^{th} percentile at a given gestational week. Several challenges arise in the analysis: 1) not all fetuses underwent ultrasound at the

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

same time, and certainly not at the pre-specified GA's; 2) each fetus underwent multiple ultrasound examinations, generating correlated data; and 3) there is a *a priori* belief that ultrasound examinations closer to delivery have higher predictive power. To address these issues, we propose a placement value (PV, Pepe and Cai 2004) based Bayesian Beta regression modeling framework with random effect to ROC curves. The use of PV-based regression to ROC curves allows us to model the effect any covariate directly on the ROC curves rather than on the distributions of the LGA and non-LGA EFWs, and the adoption of Bayesian approach accommodates *a priori* information and constraints.

2 Methods

Let y_{0ij} and y_{1ij} be the EFWs of a non-LGA and LGA fetus i from the jth ultrasound examination and x_{0ij} and x_{1ij} be the corresponding GAs. The proposed approach involves the estimation of the PVs before applying a beta regression models. As the placement value of y_{1ij} is defined as $z_{ij} = S_{0,x_{1ij}}(y_{1ij})$, where $S_{0,x}(\cdot)$ is the covariate x-specific survival function for the non-LGA population, the key step in estimating the PVs is to estimate $S_{0,x}(\cdot)$. Although more complex approach, such as quantile regression, can be used here, a simple parametric normal model is adequate in many situations. Once the PVs are estimated, we can model them in a Beta regression model as follows

$$
\begin{aligned}
&z_{ij} \sim Beta(a_{ij},\ b_{ij}) \\
&\mu_{ij} = \frac{a_{ij}}{a_{ij}+b_{ij}},\ \phi_{ij} = a_{ij} + b_{ij} \\
&logit(\mu_{ij}) = \beta_0 + \beta_1 x_{1ij}
\end{aligned}
$$

where $Beta(a,\ b)$ is a beta distribution with mean $\frac{a}{a+b}$ and $logit(u) = \frac{u}{1-u}$. We take a Bayesian approach to inference and adopt standard proper yet vague priors for model parameters. The Bayesian approach is preferred here as it allows the incorporation of *a priori* knowledge when available and makes it easy to accommodate variability associated with estimating PVs. To model correlation EFWs, we introduce a random intercept in the mean structure $logit(\mu_{ij}) = \beta_{i0} + \beta_1 x_{1ij}$, where $\beta_{i0} \sim N(\beta_0, \sigma^2)$ is a random effect term with $N(\mu,\ \sigma^2)$ denoting a normal distribution with mean μ and variance σ^2. Under the proposed model, the covariate-specific ROC curves are simply the covariate-specific Beta CDFs, and the covariate-specific AUCs are simply given by $AUC(x) = a_{ij}(x) + b_{ij}(x)$. With a pre-specified GA x^*, we can estimate the corresponding ROC curves and AUC. The accommodation of the *a priori* constraint can be achieved by specifying the prior $\beta_1 \sim N(\beta_{10}, \sigma^2_{\beta_1})I(\beta_1 > 0)$, where $I(c)$ is the usual indicator function.

3 Application results

We use 2072 participants from the NICHD Scandinavian study who have both ultrasound examination and birth weight data. Figure 1 provides an overall picture of EFWs over time separately for LGA and non-LGA fetuses. In general, LGA fetuses tend to have higher EFWs than non-LGA ones. We fitted the proposed Bayesian Beta regression model with random effect to the data and estimated

ROC curves and AUC measures at 17, 25, 33, and 37 weeks of gestational age. Table 1 provides the posterior estimates and their corresponding 95% credible intervals and Figure 2 provides the estimated ROC curves under the proposed approach at various gestational age. Overall, ultrasound examinations at larger GA have higher AUC, indicative of better discriminative power. For comparison, we also fit a naïve model where we group examinations into 4 groups according to their closeness to GA of 17, 25, 33, and 37 weeks. The estimated AUCs are biased upwards, possibly due to inappropriate grouping of examinations, and suffer from some statistical efficiency loss. Considering the *a priori* constraint does not impact AUC point estimates much but leads to some statistical efficiency gain (data not shown).

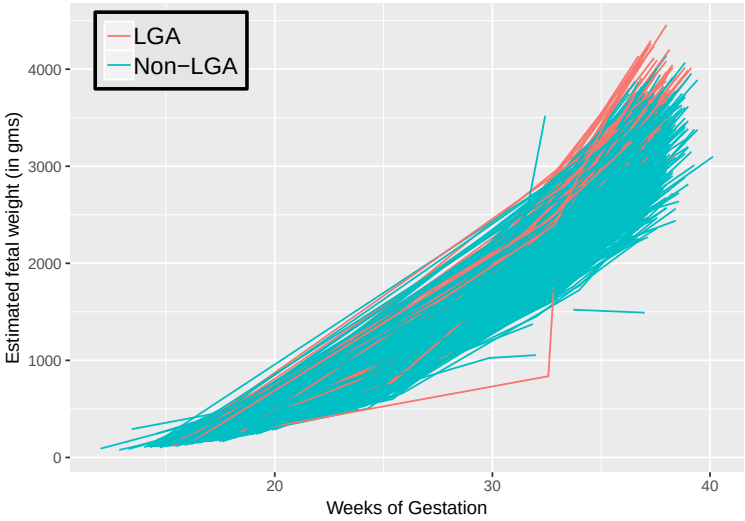


FIGURE 1. Spaghetti plot of estimated fetal weight (EFW) overtime during pregnancy stratified by large for gestational age (LGA) status from the NICHD Scandinavian study.

TABLE 1. Posterior estimates of area under ROC curves (AUC) for the NICHD Scandinavia study data.

GA (weeks)	Proposed			Naïve		
	Mean	95% CI		Mean	95% CI	
17	0.4454	0.4065	0.4844	0.5451	0.5048	0.5849
25	0.5960	0.5718	0.6198	0.6652	0.6240	0.7022
33	0.7305	0.7091	0.7508	0.7710	0.7386	0.8028
37	0.7859	0.7627	0.8078	0.8450	0.8180	0.8707

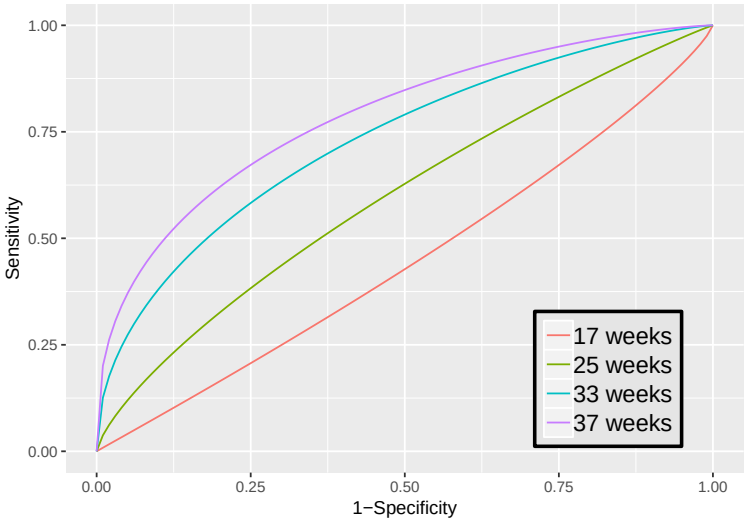


FIGURE 2. Posterior mean ROC curves at various age of gestation (in weeks) from the NICHD Scandinavian study.

References

Bakketeig, L.S. , Jacobsen, G., Hoffman, H.J., Lindmark, G., Bergsjø, P., Molney, K. and Rødsten, J. (1993). Pre-pregnancy risk factors of small-for-gestational age births among parous women in scandinavia. *Acta Obstetricia et Gynecologica Scandinavica*, **72**, 273 – 279.

Hadlock, F.P., Harrist, R.B., Sharman, R.S., Deter, R.L. and Park, S.K. (1985). Estimation of fetal weight with the use of head, body, and femur measurements – a prospective study. *American Journal of Obstetrics and Gyncology*, **151**, 333 – 337.

Pepe, M. and Cai, T. (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics*, **60**, 528 – 535.

A discrete competing risks mixed model with masked causes: a cow longevity study

Rafael Pimentel Maia[1], Clarice Garcia Borges Demétrio[2], Rodrigo Labouriau[3]

[1] Department of Statistics, University of Campinas, Campinas, Brazil
[2] Department of Exact Science, University of Sao Paulo, Piracicaba, Brazil
[3] Department of Mathematics, Aarhus University, Aarhus, Denmark

E-mail for correspondence: rpmaia@unicamp.br

Abstract: In longevity studies often the interest lies in modeling the time until death of a group of individuals that might die of different specific causes. The time to death of an individual is said to be cause-masked when the time to death of this individual is observed but not the specific cause of death, characterizing a competing risk problem with masked cause. This work will study some techniques, based on suitable variants of the EM-algorithm, to perform statistical inference in a competing risk scenario with partial masking and right censoring. The goal is to extent a class of multivariate proportional hazard models for competing risks containing suitably gaussian random components to characterize the quantitative genetic aspects of longevity in large scale animal production systems. The methods will be applied on real data of Danish dairy cattle.

Keywords: censored data; mixed model; multivariate model.

1 Introduction

In longevity studies often the interest lies in modeling the time until death of a group of individuals that might die of different specific causes. The time to death of an individual is said to be *cause-masked*, or simply *masked*, if it is observed the time but not the cause of death.
The competing risks with masked causes has been treated in the literature by using a finite mixture sub-model to represent the masked individuals and then performing inference via the EM algorithm for finite mixtures (see Flehinger et al. 1998, 2002 and Craiu and Duchesne 2004a, 2004b). On the other hand, Maia et al. (2014a,b) used multivariate proportional hazard models for competing risks containing suitably defined gaussian random components to characterize the

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

quantitative genetic determination of longevity in large scale animal production systems. Here we will propose a methodology to combine these two techniques in order to characterize quantitative genetic aspects of traits involving right censure, competitive risks and partial masking.

1.1 Data set

The dataset comprises records of 82,871 Danish dairy cows of the breed Jersey calving from 2000 to 2006. Table 1 displays the distribution of the cows according to their status. Is is presented two general culling reasons: death and slaugther. Death was treat as one competing risk with no masked individuals. In the other hand, slaughtering was split in to 4 distinct causes: low milk production (performance), infertility, udder problems and a group of other causes.

TABLE 1. Sample distribution according to the status of the cows.

Status	**Specific Cause**	**Label**	**n**	**%**
Dead	-	Cause 1	9,114	11.0
Slaughtered	Performance	Cause 2	4,126	5.0
	Infertility	Cause 3	2,571	3.1
	Udder problems	Cause 4	6,071	7.3
	Other	Cause 5	4,448	5.4
	Unknown	**Masked***	**30,878**	**37.3**
Censured	-	Censured	25,663	31.0
Total	-	-	82,871	100.0

$*$ The masked individuals are masked only among the specific causes 2 to 5 (slaughtered specific causes)

2 Methods

Define the following r.v.: T as the observed survival time ($T \in \mathbb{Z}_+$); D as the cause of death indicator, $D \in \{1, \ldots, J\}$; δ being the not censure indicator; and a γ the not masked cause indicator. Then, define the *cause-specific hazard probabilities functions*, for $j = 1, \ldots, J$, by

$$\lambda_j(t) = P\left[T = t, \delta = 1, D = j \,|\, T \geq t\right] . \tag{1}$$

The probability the cause of death of a given individual is masked is

$$\rho_j = P\left[\gamma = 0 \,|\, T, D = j, \delta = 1\right] . \tag{2}$$

The probability the true cause of death of a individual is j given it has a masked cause can be obtained by

$$\pi_j(t) = P\left[D = j \,|\, T = t, \gamma = 0, \delta = 1\right] = \frac{\rho_j \lambda_j(t)}{\sum_{k=1}^{J} \rho_k \lambda_k(t)} . \tag{3}$$

2.1 The proportional hazard model

Suppose, we have a sample o n individuals and it is also observed a range of explanatory variables (possibly time dependent) represented by the vectors $\mathbf{X}_i(t)$ and a matrix of gaussian random components, say $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_J)$ with a k dimensional component ($k \geq 1$) for each cause of death. The conditional cause specific hazard function for the j^{th} cause ($j = 1, 2, \ldots, J$), conditional on $\mathbf{U}_j = \mathbf{u}_j$, for the i^{th} individual ($i = 1, 2, \ldots, n$) at the time t, $t \in \mathbb{Z}_+$, is given by (see Maia et. al. (2014a))

$$\lambda_{ij}(t|\mathbf{u}_j) = \lambda_j(t) \exp\left(\mathbf{X}_i^t(t)\boldsymbol{\beta}_j + \mathbf{Z}_i^t\mathbf{u}_j\right) , \tag{4}$$

where the $\lambda_j(\cdot)$s are the baseline specific hazard functions and $\boldsymbol{\beta}_j$ are the vectors of fixed effects. $\mathbf{X}(\cdot)$ and $\mathbf{Z}$ are incidence matrices. It is assumed that $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_J)$ follows a multivariate normal distribution with mean equal to zero and covariance matrix given by $A \otimes \boldsymbol{\Sigma}$, where A is a known matrix (usually an identity matrix or a relationship matrix build from de pedigree data) and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1J} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1J} & \sigma_{2J} & \ldots & \sigma_J^2 \end{bmatrix} .$$

The marginal likelihood function for the respective model is given by

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n}\prod_{j=1}^{J}\left\{(1-\rho_j)^{\gamma_i\delta_{ij}}\rho_j^{(1-\gamma_i)\delta_{ij}}\right. \\ &\times \left.\int\left\{\prod_{i=1}^{n}[1-\lambda_{i\cdot}(t_i)]^{1-\delta_i}\, S_i(t_i-1)\prod_{j=1}^{J}\lambda_{ij}(t_i)\phi(\mathbf{u};\boldsymbol{\Sigma})^{\delta_{\mathbf{ij}}}\right\}d\mathbf{u}\right\} \\ &= \mathcal{L}_1(\boldsymbol{\rho}) \times \mathcal{L}_2(\boldsymbol{\beta}, \boldsymbol{\Sigma}) , \end{aligned} \tag{5}$$

where $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J\}$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_J)$, t_i is the observed survived time for the i individual, $\lambda_{i\cdot}(t) = \sum_{j=1}^{J}\lambda_{ij}(t)$, $S_i(t) = \prod_{s<t}[1-\lambda_{i\cdot}(s)]$ and $\phi(\cdot)$ is the multivariate normal probability density function. Note that the multiple integral above is typically of very high dimension.

2.2 EM algorithm

The E-step consists on calculate, at the l^{th} interaction, the expected values of δ_{ij} conditionally on previous estimates values of the parameters in the model, say $\boldsymbol{\rho}^{l-1}$ and $\boldsymbol{\beta}^{l-1}$. $\mathbb{E}\left[\delta_{ij}|\boldsymbol{\rho}^{l-1}, \boldsymbol{\beta}^{l-1}\right]$ is equal to : 1 if $\gamma_i = 1$ and $D_i = j$; 0 if $\gamma_i = 1$ and $D_i \neq j$ or if $\delta_i = 0$; $\pi_{ij}^l(t_i)$ if $\gamma_i = 0$ and $\delta_i = 1$. Where $\gamma_i = 1$ means not masked, $\delta_i = 0$ means censored and t_i is the observed survival time.
At the M-step we have to maximize

$$\mathcal{Q}_1\left(\boldsymbol{\rho}\,|\,\boldsymbol{\rho}^{l-1}\right) = \mathbb{E}\left[\log\mathcal{L}_1(\boldsymbol{\rho})\,|\,\boldsymbol{\rho}^{l-1}\right] \tag{6}$$

and

$$\mathcal{Q}_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\,|\,\boldsymbol{\beta}^{\mathbf{l-1}}, \boldsymbol{\Sigma}^{\mathbf{l-1}}\right) = \mathbb{E}\left[\log\mathcal{L}_2(\boldsymbol{\beta}, \boldsymbol{\Sigma})\,|\,\boldsymbol{\beta}^{\mathbf{l-1}}, \boldsymbol{\Sigma}^{\mathbf{l-1}}\right] \tag{7}$$

The maximum likelihood estimate for $\boldsymbol{\rho}$ at the l^{th} interaction can be easily obtained. The maximization of $\mathcal{Q}_1$ is equivalent to problem with no masking causes (See Maia et al. (2014a,b) for more details).,

2.3 Simulation Study

This study was based on a range of simulated data of competing risks with three specific risks based on a proportional hazard model with a binary fixed effect and a multivariate gaussian random components. Four scenarios were simulated representing different choices about the masked probabilities and the variance-covariance structure of the random component. The models involve three different causes, there are no censored observations and the baseline specific hazard probability function was assumed to be constant.

- The cause specific hazard probability, for the j$^{\text{th}}$ risk, is modeled by $\lambda_{ij}(t|\mathbf{u}) = \lambda_j \exp(\beta_j X_i + Z_i \mathbf{u}_j)$
- The set values for the fixed parameter are $\lambda_1 = 0.10$, $\lambda_2 = 0.11$, $\lambda_3 = 0.12$, $\beta_1 = -0.12$, $\beta_2 = -020$ and $\beta_3 = -015$
- The sample sizes are $10,000$ individuals.

2.4 Results

The set of explanatory variables (fixed effects) included in the ajusted model was: age at the first parity (categorized as: 1st quartile, 2nd and 3rd quartile, and 4ft quartile), herd sized (categorized as: 1st quartile, 2nd and 3rd quartile, and 4th quartile) and calving year (as a factor). It was included two random components: a sire random component (with the relationship matrix) representing the sire additive genetic effect, and a Herd-Year random component representing the environment effect.
Table 2 presents the sum among all individuals in the sample of the expected values of δ_{ij} for $j \in \{$ Performance, Udder Problems, Infertility, Other $\}$. We cand see that form masked individuals 5.540 was assiged to performance problems, 4.485 to infertility problems, 7.383 to udder problems and 13.480 to other problems.

TABLE 2. Predicted number of events for the specific slaughtering causes.

Specific Cause	$\sum_i E[\delta_{ij}\|\theta^l]$
Performance	9,666.4
Infertility	7,056.2
Udder problems	13,454.7
Other	17,928.6
Total	48,094.0

Table 3 presents the estimates of the variance components and correlations for the sire and herd-year effect, and the estimates of the dispersion parameters and heritability for each specifica cause. We see a large variance of the sire effect for

low performance (0.282 - sd = 0.020), infertility (0.155 - sd = 0.016) and udder problems (0.124, sd = 0.0111). The larger correlations among sire effects was between low performance and other causes (0.426 - sd = 0.057); udder problemas and other causes (0.528 - sd = 0.057); and death and low performance (-0.376 - sd = 0.094).

3 Conclusion

In general we conclude that the finite mixture model approach via EM algorithm is able to detect the variance of the random components. Woth the presented model it was possible to estimated some genetic parameters like variance components and heritablitues for a specific cause even with a larger presence of masked causes. A simulation study also showed that, in general the finite mixture model approach via EM algorithm is able to detect part of variance of the random components but tended to under estimate the variances specially when the probability of masking were high..

Acknowledgments: This work was financed by CNPq process 301323/2014-3.

References

Craiu, R.V. and Duchesne, T. (2004). Inference based on the EM algorithm for the competing risks model masked causes of failure. *Biometrika*, **91**, 543 – 558.

Craiu, R.V. & Reiser, B. (2006). Inference for the dependent competing risks model with masked causes of failure. *Lifetime Data Analysis*, **12**, 21 – 53.

Flehinger, B.J., Reiser, B. & Yashchin, E. (1998). Survival with competing risks and masked causes of failures. *Biometrika*, **85**, 151 – 164.

Flehinger, B.J., Reiser, B. & Yashchin, E. (2002). Parametric modeling for survival with competing risks and masked failure causes. *Lifetime Data Analysis*, **8**, 177 – 203.

Maia, R.P., Madsen, P. & Labouriau, R. (2014a).
Multivariate survival mixed models for genetic analysis of longevity trait. *Journal of Applied Statistics*, **42**, 1286 – 1306.

Maia, R.P., Madsen, P. & Labouriau, R. (2014b). Genetic determination of mortality rate in Danish dairy cows: A multivariate competing risks analysis based on the number of survived lactations. *Journal of Dairy Science*, **97**, 1753 – 1761.

TABLE 3. Estimates of the variance components (diagonal) and of the correlations (under diagonal) for the Sire and Herd-year random components, estimates of the dispersion parameter (ϕ_j) and marginal heritability.

	Death	*Perf.*	*Udder probl.*
Sire			
Death	0.052 (0.011)		
Perf.	-0.376 (0.094)	0.282 (0.020)	
Udder probl.	-0.136 (0.104)	-0.053 (0.059)	0.124 (0.011)
Infert.	-0.017 (0.112)	-0.026 (0.069)	-0.072 (0.069)
Other	0.074 (0.111)	0.426 (0.057)	0.528 (0.057)
Herd-Year			
Death	0.254 (0.014)		
Perf.	0.014 (0.036)	0.380 (0.016)	
Udder probl.	-0.044 (0.036)	-0.028 (0.031)	0.221 (0.001)
Infert.	-0.001 (0.038)	-0.027 (0.032)	-0.073 (0.030)
Other	0.014 (0.036)	0.003 (0.031)	-0.048 (0.029)
ϕ_j	0.845 (0.003)	0.360 (0.001)	0.414 (0.001)
$h^2_{\lambda_j}$	0.013	0.130	0.078

	Infert.	*Other*
Sire		
Death		
Perf.		
Udder probl.		
Infert.	-0.155 (0.016)	
Other	0.046 (0.075)	0.060 (0.006)
Herd-Year		
Death		
Perf.		
Udder probl.		
Infert.	0.253 (0.012)	
Other	0.011 (0.032)	0.250 (0.011)
ϕ_j	0.347 (0.001)	0.464 (0.002)
$h^2_{\lambda_j}$	0.055	0.053

Author Index

Ötting, M., 57, 125

Adam, T., 135
Aitkin, M., 37
Amoros, R., 165
Arima, S., 79

Bakka, H., 112, 291
Bao, Y., 130
Barone, R., 84
Bermudez, P. D., 147
Bernardi, M., 175, 283
Bird, R. K., 165
Bird, T. G., 165
Borges, C. G., 101, 324
Bowman, A. W., 29

Camarda, C. G., 89
Campos, L., 175
Campos, P., 49
Carvalho, M., 20
Castro, L., 312
Chakraborty, A., 279
Chen, Z., 320
Clarke, P. S., 130
Coatti, G. C., 14
Constans, M., 236
Cribari-Neto, F., 117
Currie, I., 95

de Bastiani, F., 73, 204
de Carvalho, M., 121, 147, 287
de Carvalho, V. I., 232, 287
de Uña-Álvarez, J., 210
de Zea Bermudez, P., 20
Dellaportas, P., 306
Diggle, P.J., 3
Durante, D., 283
Dutta, S., 279
Duyck, J., 62

Economou, T., 107
Espinheira, P. L., 117

Feijoo-Cid, M., 254
Friedl, H., 151

George, M., 267
Ghosal, S., 320
Gonçalves, A. M., 49
Groll, A., 67
Grundy, E., 180
Guedes. A. C., 117

Heller, G. Z., 204
Hinde, J., 101
Hohberg, M., 67
Hothorn, T., 226

Inácio de Carvalho, V., 184

Jácome, M. A., 200
Janzen, I., 169
Jensen, F. H., 57
Johnson, P. J., 165

Karlis, D., 263
Katina, S., 29
Kauermann, G., 141, 220
King, R., 169, 232
Klein, N., 157, 184, 214
Kneib, T., 214, 295
Kosmidis, I., 306
Kumada, T., 165

López-de-Ullibarri, I., 200
Labouriau, R., 324
Lampert, P., 8
Lang, M., 226
Lang, S., 214
Langrock, R., 57, 125, 135, 169
Lebacher, M., 220
Leonelli, M., 121
Lesaffre, E., 62, 259
Leyva-Moral, J. M., 254
Li, B., 194
Lourenço, V. M., 287
Lovas, A., 236

Maia, R. P., 324
Margaritella, N., 232
Maruotti, A., 125
Marx, B. D., 194
Mayr, A., 242
Mayr, G. J., 226
Menary, M. B., 107
Mews, I., 169
Morettin, P. A., 248
Mori, D., 316
Moriña, D., 254
Muggeo, V. M.R., 301

Narayanan, S., 306

Oliveira, L. M., 73
Omerovic, S., 151

Pedeli, X., 263
Pedroso-de-Lima, A. C., 14
Peng, C., 161
Pereira, P., 20, 147
Pereira, S., 20, 147
Pettitt A. N, 267
Pina, S., 49
Pohle, J., 57
Polettini, S., 79
Puig, P., 254, 316

Quick, N., 169
Quinter, L. A., 259

Ribeiro, E. E., 101
Ridall, P. G, 267
Rigby, R. A., 73, 204
Rocha, F., 14
Rodríguez–Álvarez, M. X., 184
Rodrigues, I., 45
Roos, M., 112
Rubio, R., 121
Rue, H., 112
Rue, H.Rue, 291

Sótonyi, P., 236
Safari, W. C., 200
Sampaio, J. M., 248
Schemm, J., 169
Schlosser, L., 226
Schmid, M., 242
Schneble, M., 141
Serra, I., 316
Siemiginowska, A., 175
Silva, G. L., 14
Simon, T., 157
Simonoff, J. S., 273
Singer, J. M., 14
Sottile, G., 301
Sottosanti, A., 175
Sousa, I., 312
Stasinopoulos, D. M., 73, 204
Stauffer, R., 226
Steele, F., 180
Stolf, P., 283
Szilágyi, B., 236

Tancredi, A., 84
Titze, S., 242
Toyoda, H., 165
Tran, T. D., 62
Turkman, K. F., 20
Turkman, M. A., 20

Umlauf, N., 157, 214

van den Hout, A., 189
van Niekerk, J., 291
Verbeke, G., 62, 259
Vieira, A., 312

Wang, W., 189
Weinhold, L., 242
Weiß, C. H., 135
Wiemann, P., 295

Zatz, M., 14
Zeileis, A., 226
Zhang, N., 273