Proceedings of the

30th International Workshop on Statistical Modelling

volume 2

July 6 – 10, 2015

Linz, Austria

Herwig Friedl, Helga Wagner

(editors)

Proceedings of the 30th International Workshop on Statistical Modelling,
volume 2,
Linz, July 6 – 10, 2015,
Herwig Friedl, Helga Wagner
(editors),
Linz, 2015.

Editors:
Herwig Friedl, hfriedl@tugraz.at

Institute of Statistics
Graz University of Technology
Kopernikusgasse 24/III
8020 Graz, Austria

Helga Wagner, Helga.Wagner@jku.at

Department of Applied Statistics
Johannes Kepler University Linz
Altenberger Straße 69
4040 Linz, Austria

Scientific Programme Committee

- Helga Wagner (Chair)
  Johannes Kepler University Linz, Austria
- Carlo-Giovanni Camarda
  INED, France
- Ciprian Crainiceanu
  Johns Hopkins University, USA
- Jean-Franǫis Dupuy
  INSA de Rennes, France
- Herwig Friedl
  Graz University of Technology, Austria
- Gillian Heller
  Macquarie University, Australia
- Thomas Kneib
  University of Göttingen, Germany
- Phillippe Lambert
  University of Liège, Belgium
- Stefan Lang
  University of Innsbruck, Austria
- Brian Marx
  Louisiana State University, USA
- Claire Miller
  University of Glasgow, UK
- Vicente Núñez-Antón
  Universidad del Pas Vasco UPV/EHU, Spain
- Jeff Simonoff
  New York University, USA

Local Organizing Committee

- Helga Wagner (Chair)
- Christine Duller
- Herwig Friedl
- Bettina Grün
- Gertraud Malsiner Walli
- Daniela Pauger
- Margarete Wolfesberger

Contents

Part III – Posters

George Agogo, Hilko van der Voet, Laura Trijsburg, Fred A. van Eeuwijk, Pieter van't Veer, Hendriek Boshuizen: Measurement error modelling for accelerometer activity data using Bayesian integrated nested Laplace approximation 3

Danilo Alvares, Carmen Armero, Anabel Forte, Luis Galipienso, Antonio Vicent, Luis Rubio: Bayesian multinomial logit model: an application to agriculture 7

Orasa Anan, Dankmar Böhning, Antonello Maruotti: Population size estimation in capture-recapture data based upon the Conway-Maxwell-Poisson distribution 11

Autcha Araveeporn: Estimating nonlinear autoregressive models using smoothing spline and penalized spline methods 15

Silvia Bacci, Francesco Bartolucci, Silvia Pandolfi: Longitudinal data with informative dropout: an approach based on an AR(1) latent process 19

Haakon Bakka: Log-Gaussian Cox processes with spatially varying second order properties 23

Somsri Banditvilai, Saowapa Mahakeeta: A receiving process simulation model of the consumer product distribution center ... 27

Edan Bar, Haim Bar: A bivariate index for stock classification .. 31

Eva Benková, Radoslav Harman: Barycentric algorithm for computing D-optimal size-and-cost constrained designs of experiments 35

Theophile Bigirumurame, Nolen Perualila-Tan, Adetayo Kasim, Ziv Shkedy: Integrated analysis of multi-source data in drug discovery experiments using structural equation models 39

Angela Bitto, Sylvia Frühwirth-Schnatter: Achieving shrinkage in the time-varying parameter models framework 43

Betsabé G. Blas: Skew-normal controlled calibration model 47

Marie Böhnstedt, Jutta Gampe: Density estimation from uncertain observations – Comparing parametric and nonparametric methods 51

Stella Bollmann, Andreas Hölzl, Helmut Küchenhoff, Moritz Heene, Markus Bühner: Evaluation of a new k-means approach for exploratory clustering of items 55

Robin Bruyndonckx, Niel Hens, Marc Aerts, Katrien Latour, Boudewijn Catry, Samuel Coenen: Using generalized estimating equations to study persistence of antimicrobial resistance in respiratory streptococci 59

S. Cakmak, C. Hebbern, J. Vanos, D. Crouse, C. Blanco: Using spatial and land use regression models in investigating the modifying effect of socioeconomic status on the interaction between traffic, air pollution and asthma 63

Mileno Cavalcante, Betsabé G. Blas: Beta calibration model with measurement errors 67

Marcella Corduas: Modelling correlated ordinal data by a copula approach 71

Kevin D. Dayaratna, Benjamin Kedem: A probabilistic examination of the efficacy of tort reform via Bayesian semiparametric density ratio modeling 75

Fernanda De Bastiani, Audrey Helen Mariz de Aquino Cysneiros, Miguel Angel Uribe-Opazo, Manuel Galea: Local influence on Gaussian spatial linear model with multiple replications 79

Vera Djordjilović, Monica Chiogna: An improved shrinkage procedure for the estimation of covariance matrices in graphical models 83

Michael Espendiller, Maria Kateri: Measures of association for 2×2 tables with a ϕ-divergence origin 87

Rosemeire L. Fiaccone, Robin Henderson: Dynamic analysis for event history data: Recurrent infant diarrhoea 91

Eder Lucio da Fonseca, Airlane Pereira Alencar, Pedro Alberto Morettin: Time-varying cointegration via wavelets . 95

Khuneswari Gopal Pillay, John H. McColl: Model selection and model averaging using restrictive strategies for imputation in linear models 99

Yann Guédon: Slope heuristics for multiple change-point models . 103

Jingyi Guo, Håvard Rue, Andrea Riebler: Making Bayesian bivariate meta-analysis practice friendly 107

Markus Hainy, Christopher C. Drovandi: Approximate Bayesian computation for spatial extremes using composite score functions 111

Philipp Hermann, Milan Stehlík: On some issues on statistical analysis for Wilms tumor 115

Freddy Hernández, Mabel Torres-Taborda, Lina Arteaga, Cristina Castro: GAMLSS applied to study bacterial cellulose yield 119

Manuel Higueras, David Moriña, Pedro Puig, Liz Ainsbury, Kai Rothkamm: A new inverse regression model and software for radiation biodosimetry 123

Abu Hossain, Robert A. Rigby, Dimitrios M. Stasinopoulos, Marco Enea: Modelling a proportion response variable using generalised additive models for location scale and shape 127

Maria Iannario, Domenico Piccolo: Analysing ordinal categorical data by means of a generalized mixture model 131

Amirhossein Jalali, Alberto Alvarez-Iglesias, John Newell: Dynamic nomogram in R 135

Gregor Kastner, Sylvia Frühwirth-Schnatter, Hedibert Freitas Lopes: Dynamic covariance estimation using sparse Bayesian factor stochastic volatility models 139

Lenka Komárková, Táňa Hajdíková, Arnošt Komárek: Model based segmentation of the Czech hospitals according to their longitudinal financial performance in 2007 – 2011 143

Altea Lorenzo-Arribas, Mark J. Brewer, Antony M. Overstall: A simulation study assessing the advantages of cumulative link models 147

Gertraud Malsiner-Walli, Bettina Grün, Paul Hofmarcher: Bayesian variable selection in semi-parameteric growth regression 151

Valentina Mameli, Monica Musio, Laura Ventura: Simulated adjustment of the signed scoring rule root statistic 155

Alessandra Marcelletti, Antonello Maruotti, Giovanni Trovato: A flexible bivariate location-scale finite mixture approach to economic growth 159

Kenan Matawie, Arshad Mehar, Anthony Maeder: An approach to determine clusters overlap for K-means clustering..... 163

Luís Meira-Machado, Marta Sestelo, Andreia Gonçalves: Nonparametric estimation of the survival function for ordered multivariate failure time data: a comparative study 167

Shirin Moghaddam, John Hinde, Milovan Krnjajić : Flexible models in survival analysis: an illustration 171

Annette Möller: Spatially adaptive probabilistic temperature forecasting using Markovian EMOS..... 175

Leacky Muchene, Ziv Shkedy, Tom Jacobs: Estimation of order-restricted mean structure for independent and correlated data using Bayesian variable selection models..... 179

Luiz R. Nakamura, Robert A. Rigby, Dimitrios M. Stasinopoulos, Roseli A. Leandro, Cristian Villegas: The Birnbaum-Saunders generalized-t distribution for positive skewed data..... 183

Mu Niu, Maurizio Filippone, Dirk Husmeier, Simon Rogers: Inference in nonlinear differential equations..... 187

Umberto Noè, Maurizio Filippone, Dirk Husmeier: Emulation of ODEs with Gaussian processes..... 191

Ruth O'Donnell, Claire Miller, Marian Scott: Within lake clustering of high resolution satellite retrievals - a functional data and clustering approach 195

Adrian O'Hagan: Bayesian model averaging for copula-based estimation of upper tail dependence in loss distributions..... 199

Thiago P. Oliveira, Rafael A. Moral, John Hinde, Clarice G.B. Demétrio, Silvio S. Zocchi, Ana B.R. Zanardo, Italo Delalibera Jr.: Generalized linear mixed models applied to overdispersed proportion data in a fungal occurrence study... 203

Willian Luís Oliveira, Carlos Alberto Ribeiro Diniz, María Durbán: A general class of bivariate regression models for mixed discrete and continuous responses 207

Johann Ospina-Galindez, Mercedes Andrade-Bejarano, Ramón Giraldo-Henao: Functional regression model of pentadal rainfall of Valle del Cauca (Colombia) in the period (1993 – 2011), integrating data from meteorological stations and satellite 211

Daniela Pauger, Helga Wagner: A comparison of different priors for sparse Bayesian modelling in regression models with ordinal predictors 215

Gilberto A. Paula, Carlos Eduardo M. Relvas: Partially linear models with autoregressive symmetric errors 219

Elisa Perrone, Werner Müller: D_s-optimality for discriminating between copula models: a first example 223

Eliane C. Pinheiro, Silvia L.P. Ferrari: A comparative review of generalizations of the Gumbel extreme value distribution 227

Hildete P. Pinheiro, Pranab K. Sen, Aluísio Pinheiro, Samara F. Kiihl: Constrained hypotheses testing via quasi U-statistics and its application to undergraduate performance assessment 231

Hauke Rennies, Thomas Kneib: Structural equation models for dealing with spatial confounding 235

Jose S. Romeo, Renate Meyer: Bayesian approach for modelling bivariate survival data through the PVF copula 239

Helene Roth, Stefan Lang: Posterior sensitivity of variance priors in Bayesian structured additive regression 243

Bruno Santos, Heleno Bolfarine: Analysis of Brazil's presidential election via Bayesian spatial quantile regression 247

Lukas M. Schäfer, Bruce J. Worton: Modelling wind direction with an application 251

Sabine K. Schnabel, Federico Torretta, Matthias Westhues: Quantifying LD decay by quantile regression – a case study 255

Max Schneider, Gilles Blanchard, Christian Levers, Tobias Kümmerle: Spatial variation of drivers of agricultural abandonment with spatially boosted models 259

Aliakbar Mastani Shirazi, Kalyan Das, Aluísio Pinheiro: Self-modeling ordinal model with time invariant covariates – an application to prostate cancer 263

Georgy Yu. Sofronov, Nikolay V. Glotov, Olga V. Zhukova: Statistical analysis of spatial distribution in populations of microspecies of Alchemilla L. 267

Katia Stefanova, Wallace Cowling, Arthur Gilmour : Use of genetic relationship matrices in the prediction of breeding values and their accuracy assessment 271

James Sweeney: A modified binomial likelihood model for zero and n-inflated count data 275

Fraser Tough, Charlotte M. Wright, John H. McColl: Conditional weight gain using an external reference 279

Diego Tovar-Rios, Rafael Tovar-Cuevas, Mercedes Andrade-Bejarano: Prior elicitation for estimation in a mixed effect logistic model 283

Nikolaus Umlauf, Reto Stauffer, Jakob W. Messner, Georg J. Mayr, Achim Zeileis: A conceptional Lego toolbox for Bayesian distributional regression models 287

Yannick Vandendijck, Christel Faes, Niel Hens: Geostatistical analysis using K-splines in the geoadditive model 291

Kim Van Kerckhove, Christel Faes, Philippe Beutels, Niel Hens: Do contacts over distance follow a power-law distribution? Estimation of the social contact distance kernel 295

Paul Wilson, Jochen Einbeck: A simple and intuitive test for number-inflation or number-deflation 299

Index 303

Part III - Posters

Measurement error modelling for accelerometer activity data using Bayesian integrated nested Laplace approximation

George Agogo[1], Hilko van der Voet[1], Laura Trijsburg[2], Fred A. van Eeuwijk[1], Pieter van't Veer[2], Hendriek Boshuizen[1]

[1] Biometris, Wageningen University, Netherlands
[2] Human Nutrition, Wageningen University, Netherlands

E-mail for correspondence: george.agogo@wur.nl

Abstract: Regular physical activity (PA) is associated with good health. PA cannot be measured exactly with an accelerometer, leading to measurement error (ME). ME results in loss of statistical power to detect an association between an exposure and an outcome of interest and bias in the association. We propose a ME model for a triaxial actigraph accelerometer and quantify loss of statistical power with a correlation coefficient between true and measured Total Energy Expenditure (validity coefficient) as derived from the accelerometer, and bias in the association with an attenuation factor. We estimated parameters in the ME model with Bayesian integrated nested Laplace approximation and compared the results with those of a Markov chain Monte Carlo and a Maximum Likelihood method. We applied the proposed method to the DuPLO validation study. Both the validity coefficient and the attenuation factor estimates were about 0.8; this implies a modest effect of ME.

Keywords: Accelerometer; Bayesian INLA; MCMC; Maximum likelihood; Measurement Error.

1 Introduction

Regular physical activity (PA) is associated with good health. Measuring PA with an accelerometer is difficult, leading to measurement error (ME). An accelerometer is a device that monitors body accelerations. ME leads to loss of statistical power to detect an association between an exposure (e.g., PA) and an outcome of interest (e.g., disease), and bias in the PA-outcome association. Loss of statistical power can be quantified with a correlation

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

coefficient between true and measured values (hereafter, validity coefficient) and the bias in the association with an attenuation factor (Carroll et al., 2006). In validation studies, doubly labelled water (DLW) is used as gold standard for total energy expenditure (TEE). TEE is composed of three components: energy expended due to PA, known as activity energy expenditure (AEE); energy expended at rest, known as basal energy expenditure (BEE) and thermic effect of food (TEF). To estimate TEE with accelerometer, AEE is estimated from accelerometer activity data using a prediction equation, BEE is estimated from anthropometric and demographic data using a prediction equation, and TEF is usually taken as 10% of TEE. The prediction equations are subject to error. The DLW measurements are used to validate the accelerometer data in the DuPLO study. The DuPLO study is a recently concluded validation study on PA and dietary assessment methods conducted in Wageningen and environs (in the Netherlands). In the DuPLO study, PA was assessed with a triaxial actigraph accelerometer that monitors body acceleration in three planes. Studies on validation of this accelerometer model are lacking. This study aims to quantify ME in daily TEE measurements from the accelerometer in the DuPLO study. We propose a ME model for the accelerometer. The proposed ME model contains latent true TEE, fixed bias terms and random error components (Ferrari et al., 2007). We estimated the ME model parameters with fast, non-sampling-based integrated nested Laplace approximation (INLA)(Rue et al., 2009), using data augmentation technique (Muff, 2015). We compare INLA results with those of a Markov Chain Monte Carlo (MCMC) with Gibbs sampler, and a Maximum Likelihood (ML) method with adaptive Gaussian quadrature.

2 Methods

2.1 Linear mixed measurement error (LMME) model

We denote the jth replicate of DLW measurement for the ith person by R_{ij}, accelerometer measurement by A_{ij} and the latent true daily TEE by T_i. We relate A_{ij} and R_{ij} with latent T_i by a linear additive ME model as

$$\begin{aligned} A_{ij} &= \beta_0 + \beta_A T_i + r_{A_i} + \epsilon_{A_{ij}} \\ R_{ij} &= T_i + \epsilon_{R_{ij}}, \end{aligned} \tag{1}$$

where $r_{A_i} \sim N(0, \sigma^2_{r_A})$, $\epsilon_{A_{ij}} \sim N(0, \sigma^2_{\epsilon_A})$, $\text{cov}(r_{A_i}, \epsilon_{A_{ij}}) = 0$, $\epsilon_{R_{ij}} \sim N(0, \sigma^2_R)$; $T_i \sim N(\alpha_0 + \alpha_Z^T \mathbf{Z}, \sigma^2_T)$ and $\mathbf{Z}$ is a vector of covariates.

2.2 Quantification of ME in the accelerometer

With the parameters from model (1), we quantify the loss of statistical power due to the ME in TEE with validity coefficient (ρ_{AT}) and bias in

the TEE-outcome association with attenuation factor (λ_A), defined as:

$$\begin{aligned}\rho_{AT} &= \frac{\text{cov}(A,T)}{\sqrt{\text{var}(A)\text{var}(T)}} = \frac{\beta_A\sigma_T}{\sqrt{(\beta_A^2\sigma_T^2+\sigma_{r_A}^2+\sigma_{\epsilon_A}^2)}}\\ \lambda_A &= \frac{\text{cov}(A,T)}{\text{var(A)}} = \frac{\beta_A\sigma_T^2}{\beta_A^2\sigma_T^2+\sigma_{r_A}^2+\sigma_{\epsilon_A}^2}.\end{aligned}$$

We use non-informative priors. For each of the 3 chains in MCMC, we discard 10^5 burn-in samples, save every 5th of the remaining 5×10^5 samples, resulting in 10^5 samples per chain. We assess mixing with trace plots. In ML, we use 10 quadrature points and bootstrapping to estimate the confidence interval(CI).

3 Results

Figure 1 plots the within-subject difference versus the subject average for the TEE measurements from the accelerometer (a) and the subject average TEE measurements from the accelerometer versus the subject average TEE measurements from the DLW (b). There is no discernible trend on the scatter plots, suggesting an additive within-subject random error (Figure 1(a)); a similar observation was made for scatter plots of TEE measurements from the DLW (figure not shown). Figure 1(b) shows that TEE for subjects with large mean DLW values are underestimated more with the accelerometer measurements than for subjects with small mean DLW values. This suggests a proportional scaling bias in the accelerometer measurements. These exploratory TEE data analyses support the choice of the LMME model.
Table 1 presents the posterior mean estimate (95% credible interval) for the validity coefficient and attenuation factor from MCMC and INLA, and the point estimate (95% confidence interval) from ML. TEE measurements from the accelerometer correlates strongly with true TEE with $\rho_{AT} \approx 0.8$.This implies modest loss of statistical power to detect significant TEE-outcome associations. Similarly, the attenuation factor is estimated as $\lambda_A \approx 0.8$ to imply minimal bias in TEE-outcome association. For instance, if the outcome model is linear and a true coefficient is Ω, then the observed coefficient from the use of the accelerometer will be 0.8Ω. INLA results are similar to those of MCMC and ML.

4 Conclusion

The accelerometer has satisfactory validity for measuring TEE in the DuPLO study population. There is minimal loss of statistical power to detect TEE-outcome association, when TEE is derived from accelerometer data. ME in the accelerometer causes modest bias in TEE-outcome association.

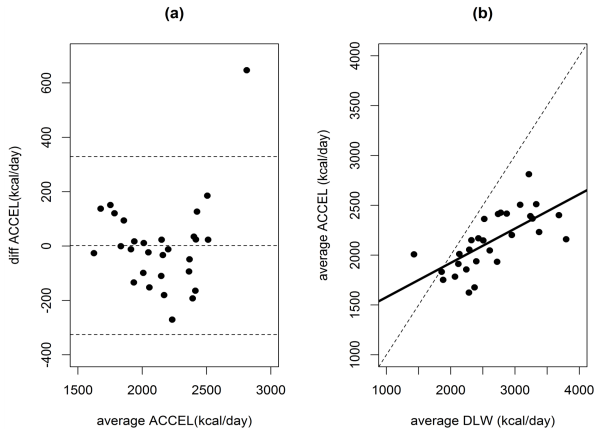


FIGURE 1. Within-subject difference versus subject average for accelerometer TEE, mean difference (middle dashed line) and 95% limits of agreement(extreme dashed lines) (a), and subject average accelerometer TEE versus subject average DLW TEE measurements(b).

TABLE 1. Parameter estimates (95 % CI) for validity coefficient (ρ_{AT}) and attenuation factor (λ_A) as estimated with INLA, MCMC and ML; CI is credible/confidence interval.

Quantity	INLA	MCMC	ML
ρ_{AT}	0.794(0.651; 0.893)	0.770(0.592; 0.886)	0.810(0.620;0.879)
λ_A	0.813(0.564; 1.025)	0.791(0.461; 1.083)	0.818(0.528;1.080)

References

Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models*. New York: Chapman & Hall/CRC.

Ferrari, P., Friedenreich, C., and Matthews, C.E. (2007). The role of measurement error in estimating levels of physical activity. *American Journal of Epidemiology*, **166**, 832 – 840.

Muff, S., Riebler, A., Held, L., Rue, H., and Saner, P. (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series C*, **64**, 231 – 252.

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

Bayesian multinomial logit model: an application to agriculture

Danilo Alvares[1], Carmen Armero[1], Anabel Forte[1], Luis Galipienso[2], Antonio Vicent[2], Luis Rubio[2]

[1] University of Valencia (UV), Burjasot, Spain
[2] Valencian Institute for Agricultural Research (IVIA), Moncada, Spain

E-mail for correspondence: daldasil@alumni.uv.es

Abstract: Chufa, *Cyperus sculentus*, also known as tiger nuts, is a herbaceous plant from which the *horchata*, a popular soft drink in Valencia (Spain), is obtained. Over the last four years, the cultivation of the chufa has suffered from a disease consisting in the appearing of black spots in some tubers, which must be discarded thus causing important economical losses to farmers.
This study deals with the statistical analysis of the seed transmission of the black spots to the harvest tubers. Bayesian generalized linear models with multinomial response are used to analyze data from an experiment in greenhouse containing asymptomatic and diseased seeds.

Keywords: Black spots in tiger nuts; Bayesian analysis; Posterior relative risk.

1 Introduction

Chufa, *Cyperus sculentus*, is a herbaceous plant that produces edible tubers. In Europe, its cultivation mainly occurs in Valencia (Spain), where is used for the preparation of *horchata*, a popular soft drink (Figure 1) with a long tradition.
Over the last four years, the chufa presents a pathology consisting in the appearance of black spots. This disease has unknown origin and tubers suffering from it must be discarded for precautionary reasons thus causing important economical losses to farmers.
The aim of this study is to statistically analyze the transmission of the black spots disease from the seed to the harvested tubers.

2 Experiment

The experiment was conducted on the premises of the Valencian Institute for Agricultural Research (Spain). The seeds were divided into two groups

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

FIGURE 1. Chufa and *horchata* (Source: www.lavanguardia.com).

of eight replicates each, the groups are classified as asymptomatic and with severe symptoms of the disease. After the harvest, the following information for each replicate was collected:

- `asymptomatic`: number of asymptomatic tubers;
- `mild`: number of tubers with mild symptoms;
- `severe`: number of tubers with severe symptoms

and our interest is focused on modeling the probability of harvesting an asymptomatic, mild or severe tuber from each group of seeds.

3 Bayesian multinomial logit model

Our variable of interest takes a finite number of values, which we refer as categories or classes (Congdon, 2005). A natural choice for analyzing this type of data is through generalized linear models with multinomial response (McCullagh and Nelder, 1989). We use logit links which connects each response category with the baseline category, asymptomatic seeds (Agresti, 2013).

$$\begin{aligned} (\boldsymbol{Y}_i|n_i,\boldsymbol{\theta}_i) &\sim \text{Multinomial}(n_i,\boldsymbol{\theta}_i) \\ \log\left(\frac{\theta_{ij}}{\theta_{i1}}\right) &= \alpha_j+\beta_i, \qquad j=2,3,\ i=1,2, \end{aligned}$$

where $Y_i=(Y_{i1},Y_{i2},Y_{i3})^{\mathrm{T}}$ represents the number of asymptomatic tubers, with mild and severe symptoms, respectively, coming from the ith group of seeds, $\boldsymbol{\theta}_i=(\theta_{i1},\theta_{i2},\theta_{i3})^{\mathrm{T}}$ are the subsequent probabilities for the group i, with $\sum_{j=1}^{3}\theta_{ij}=1$ for all i, and n_i is the total number of tubers from a type i seed. We assume prior independence among the parameters, α's and β's, in the model and consider non-informative normal distributions for the subsequent marginal priors.

The posterior distribution for the parameters of the model has been approximated by Markov chain Monte Carlo methods (Gelman et al, 2013). The

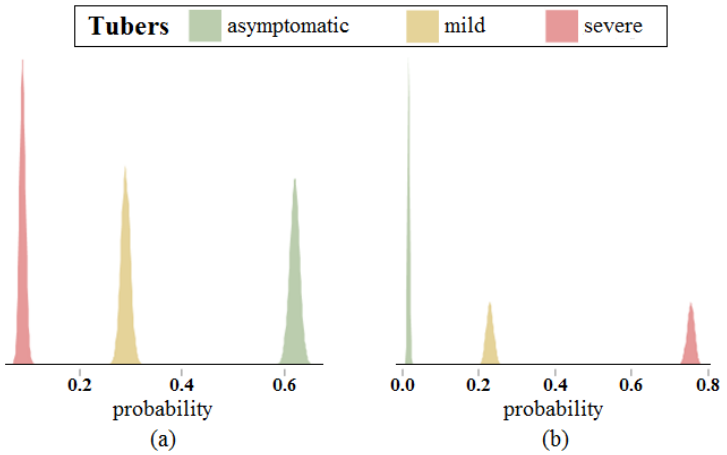


FIGURE 2. Posterior distributions for the probabilities θ_{ij} associated to tubers harvested from asymptomatic seeds (a) and seeds with severe symptoms (b).

MCMC algorithm has run for three Markov chains with 100000 iterations after a burn-in period with 1000 iterations. The effective iterations were thinned storing every 5th iteration in order to decrease auto-correlation in the sample. Figure 2 shows the posterior marginal distribution for the probabilities θ_{ij}, where remember that i refers to the group of seeds and j to the type of harvested tuber. For asymptomatic seeds, the probability of harvesting asymptomatic tubers (posterior mean 0.62) is greater than the probabilities corresponding to tubers with mild (posterior mean 0.29) and severe symptoms (posterior mean 0.09). In the case of seeds with severe symptoms, the probability of harvesting tubers with severe symptoms (posterior mean 0.75) is greater than the probabilities corresponding to tubers with mild symptoms (posterior mean 0.23) and asymptomatic tubers (posterior mean 0.02).
Probabilities of harvesting tubers with mild and severe symptoms with regard to the probability of harvesting asymptomatic tubers in each group of seeds are compared by means of the relative risks (RR).

$$RR_{ij} = \frac{\theta_{ij}}{\theta_{i1}}, \quad j = 2, 3, \ i = 1, 2. \tag{1}$$

Figure 3 shows the posterior distribution for the relevant RR. In the group of asymptomatic seeds, the posterior RR expectation of harvesting a tuber with mild and severe symptoms is 0.47 and 14.96, respectively. Posterior RR mean for mild and severe tubers harvested from seeds with severe symptoms is 0.14 and 49.43, respectively.

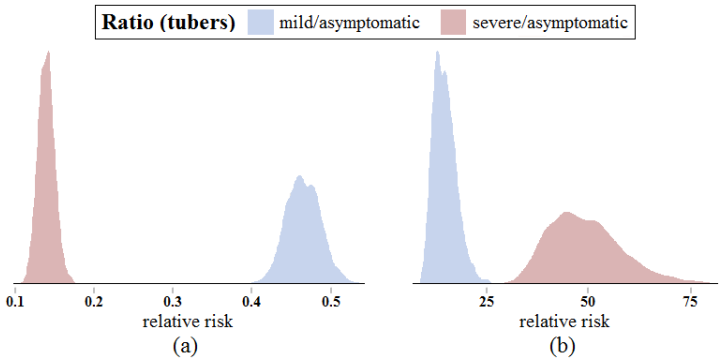


FIGURE 3. Posterior relative risk between the types of tubers in groups of asymptomatic seeds (a) and seeds with severe symptoms (b).

4 Conclusions

This study indicates a relevant mean reduction in the probability of harvesting asymptomatic tubers from seeds with severe symptoms of the disease. This interpretation is supplemented by the high relative risk of harvesting tubers with severe symptoms, compared with asymptomatic tubers, from seeds with severe symptoms. Furthermore, the results show that the selection of seeds seems to have a positive effect on the reduction of the prevalence of the disease. The fact that farmers can only market asymptomatic and mild symptoms tubers, combined with the information provided by the posterior distribution of the probability of harvesting tubers without severe symptoms, advise against the use of seeds with severe symptoms to obtain marketable tubers.

Acknowledgments: This paper has been partially supported by the Coordination for the Improvement of Higher Level Personnel (BEX 0047/13-9), Valencian Institute for Agricultural Research, and research grant MTM2 013-42323-P from the Spanish Ministry of Economy and Competitiveness.

References

Agresti, A. (2013). *Categorical Data Analysis*. NJ: Wiley.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. NY: Wiley.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*. London: Chapman & Hall/CRC.

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

Population size estimation in capture-recapture data based upon the Conway-Maxwell-Poisson distribution

Orasa Anan[1], Dankmar Böhning[1], Antonello Maruotti[12]

[1] University of Southampton, Southampton, UK
[2] Universita di Roma Tre, Rome, Italy

E-mail for correspondence: oa2e12@soton.ac.uk

Abstract: The purpose of the study is to estimate the population size under a truncated count model that accounts for heterogeneity. The proposed estimator is based on the Conway-Maxwell-Poisson (CMP) distribution. The benefit of using the CMP distribution is that it includes the Bernoulli, the geometric and the Poisson distributions as special cases and, furthermore, allows for heterogeneity. Parameter estimates can be obtained by exploiting the ratios of successive frequency counts in a weighted linear regression framework. The results of the comparisons with Turings, the maximum likelihood Poisson, Zeltermans and Chaos estimators reveal that our proposal can be beneficially used.

Keywords: Capture-recapture methods; Ratio plot; Heterogeneous populations.

1 Introduction

Capture-recapture (CR) methods have been adopted in a wide range of applications, including ecology (Alunni-Fegatelli and Tardella, 2013; Farcomeni, 2011), criminal activity (van der Heijden et al., 2003; Farcomeni and Scacciatelli, 2013), epidemiology (Böhning et al., 2005), official statistics (Rocchetti et al., 2011), in the estimation of the size of hidden populations. CR analyses are based on the repeated sampling from a population and, consequently, on the use of recapture information to infer the number of uncaptured units. Let X_i, $i = 1, \ldots, N$, denote the number of times unit i is captured over the m sampling occasions, and let $p_x = \Pr(X_i = x)$. Also let f_x denote the frequency of units captured exactly x times, $x = 0, 1, \ldots, m$. As $X_i = 0$ is not observed, the corresponding f_0 is unknown and might be replaced by its expected value Np_0. However, p_0 is

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

usually unknown and has to be estimated. Basically, the Poisson model with parameter λ may represent a natural starting point, however, this model is restrictive because it assumes a unit variance-to-mean ratio. Hence, even if the Poisson distribution can be recognized as an important tool to model count data, it may not be suitable for CR data, which are characterized by overdispersion/underdispersion, i.e. the variance is grater/lower than the corresponding sample mean, mainly due to unobserved heterogeneity.

2 Model inference

We wish to contribute extending a more general count distribution that captures a wider range of dispersion settings than existing distributions. In detail, we look at a two-parameter generalized form of the Poisson distribution, called the Conway-Maxwell-Poisson (CMP) distribution (Shmueli et al., 2005) to account for heterogeneity as it includes as special sub-models important distributions (i.e. Poisson, Bernoulli, and geometric distributions) and generalizes the Poisson distribution allowing for overdispersion as well as underdispersion.

2.1 The Conway-Maxwell-Poisson distribution

The CMP distribution is a flexible model for analyzing count data. Its probability mass function CMP(λ, ν) is given by $p_x = \frac{\lambda^x}{(x!)^\nu} \frac{1}{z(\lambda,\nu)}$, $x = 0, 1, 2, \ldots$; $\lambda \geq 0$; $\nu \geq 0$, where the normalizing constant $z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ is a generalization of well-known infinite sums. The CMP distribution has been overlooked for long time due to the complexity in dealing with the infinite sum $z(\lambda, \nu)$, that is often approximated.

2.2 The ratio-plot

In this work, we avoid classical approaches to estimation of population size and propose a method based on ratios of successive probability counts, namely, $r_x = (x+1)\frac{p_{x+1}}{p_x}$ which is a function of the observed count x. In CR studies, the zero counts are truncated and, hence, the observed sample frequencies $f_1, f_2, \ldots$ arise from the zero-truncated distribution $\frac{p_x}{1-p_0}$. However, the ratio r_x for the truncated and the untruncated distribution is identical. This is an important result as it makes the ratio applicable into a CR framework. The ratio for the CMP distribution has the form $r_x = (x+1)\frac{p_{x+1}}{p_x} = (x+1)\frac{\frac{\lambda^{x+1}}{\{(x+1)!\}^\nu}\frac{1}{z(\lambda,\nu)}}{\frac{\lambda^x}{(x!)^\nu}\frac{1}{z(\lambda,\nu)}} = \lambda(x+1)^{1-\nu}$ and does not depend on the complex normalizing constant term $z(\lambda, \nu)$. If we consider the ratio on the log-scale, we achieve a linear relationship. Accordingly, $\log(r_x) = \log\lambda + (1-\nu)\log(x+1) = \beta_0 + \beta_1 \log(x+1)$. We have that $\lambda = \exp(\beta_0)$ and $\nu = 1 - \beta_1$; however, due to $\nu \geq 0$ (or, equivalently,

$1-\nu \leq 1$), we must constrain $\beta_1 \leq 1$. In practice, we approximate capture probabilities by relative frequencies, therefore the ratio can be obtained by $r_x^* = (x+1)\frac{\hat{p}_{x+1}}{\hat{p}_x} = (x+1)\frac{f_{x+1}/N}{f_x/N} = (x+1)\frac{f_{x+1}}{f_x}$, as well as $\log(r_x^*) = \log\left\{(x+1)\frac{f_{x+1}}{f_x}\right\}$, where f_x is the frequency of count x and $N = \sum_{x=0}^{m} f_x$. By plotting $\log(r_x^*)$ against $\log(x+1)$, we derive a graphical diagnostic tool for detecting the validity of Conway-Maxwell-Poisson model. A log-ratio plot showing a positive slope indicates for the presence of overdispersion with respect to the Poisson distribution. On the other hand, in the case of underdispersion, the log-ratio plot displays a straight line with a negative slope. Finally, when the log-ratio plot displays a horizontal line, the equi-dispersion case is the Poisson distribution. Accordingly, the unknown f_0 can be then estimated by considering that $\log\left(\frac{f_1}{f_0}\right) = \hat{\beta}_0$, hence $\hat{f}_0 = f_1 \exp(-\hat{\beta}_0)$, where $\hat{f}_0$ is the unobserved frequency estimator. The linear regression estimator based on the Conway-Maxwell-Poisson distribution (LCMP) of target population size can be readily achieved as $\hat{N}_{LCMP} = n + \hat{f}_0 = n + f_1 \exp(-\hat{\beta}_0)$. We also obtain an estimated probability of the count to be zero (unobserved) as $\hat{p}_0 = \frac{\hat{f}_0}{\hat{N}_{LCMP}}$. The asymptotic variance estimator of $\hat{N}_{LCMP}$ is obtained as $\widehat{\mathrm{Var}}(\hat{N}_{LCMP}) = \frac{n f_1 e^{-\hat{\beta}^0}}{n+f_1 e^{-\hat{\beta}^0}} + (e^{-\hat{\beta}_0})^2 f_1[1 + f_1 \mathrm{Var}(\hat{\beta}_0)]$.

3 Simulation study

In the simulation study schemes with different population sizes $N = 100$, 1000, 10000 and levels of heterogeneity, and five estimators of population size were compared: the Turing's estimator, the maximum likelihood estimation under the zero-truncated Poisson model, Chao's lower bound estimator, Zelterman's estimator and (weighted) linear regression estimator under the zero-truncated Conway-Maxwell-Poisson model. We further test estimators performance by generating data from a CMP distribution. We expect that the LCMP estimator, as well as Chao and Zelterman estimators, outperforms Turing and MLEPoi estimators under *heterogeneous* schemes. Overall, the LCMP has the best performance when the population size is small or medium, whereas Turing and MLEPoi estimators underestimate the population size. Even the other heterogeneous estimators tend to underestimate the population size, providing reasonable results for $N = 10000$ only. Indeed, the CMP distribution is a very general one and accounts for many (possible) data features, that may not be captured by existing estimators.

4 Conclusion

A diversity of estimators in the capture-recapture field exists, being widely applied in many areas of interest. Here, we have introduced a new method of estimating the population size under a specific form of heterogeneity based on the Conway-Maxwell-Poisson distribution. We have also been able to see how accurate and precise the method is performing when it is compared to other frequently used estimators. Although the proposed estimator showed superior performance in terms of accuracy, it evidently gave also the largest variation; nonetheless, the variation of the new estimator considerably decreases for large population size (1000 and more) which is typically the case in real-world applications. We also provided a formula of variance approximation of the new estimator. This variance formula is not only useful to determine the efficiency of estimating, but it can be also used to construct confidence intervals. In short, the new estimator can be an alternative form of population size estimation especially for large populations and heterogeneous capturing probabilities.

References

Alunni-Fegatelli, D. and Tardella, L. (2013). Improved inference on capture recapture models with behavioural effects. *Statistical Methods & Applications*, **22**, 45 – 66.

Böhning, D., Dietz, E., Kuhnert, R., and Schön, D. (2005). Mixture models for capture-recapture count data. *Statistical Methods & Applications*, **14**, 29 – 43.

Farcomeni, A. (2011). Recapture models under equality constraint. *Biometrika*, **98**, 237 – 242.

Farcomeni, A. and Scacciatelli, D. (2013). Heterogeneity and behavioural response in continuous time capture-recapture, with application to street cannabis use in Italy. *Annals of Applied Statistics*, **7**, 2293 – 2314.

Rocchetti, I., Bunge, J., and Böhning, D. (2011). Population size estimation based upon ratios of recapture probabilities. *Annals of Applied Statistics*, **5**, 1512 – 1533.

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*, **54**, 127 – 142.

Van der Heijden, P.G., Bustami, R., Cruyff, M.J., Engbersen, G., and Van Houwelingen, H.C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling*, **3**, 305 – 322.

Estimating nonlinear autoregressive models using smoothing spline and penalized spline methods

Autcha Araveeporn[1]

[1] Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

E-mail for correspondence: kaautcha@hotmail.com

Abstract: The objective of this research is to compare the estimation of the Non-Linear AutoRegressive (NLAR) model with Smoothing Spline (SS) and Penalized Spline (PS) methods in a class of nonparametric regression method. NLAR model consists of a response variable and a function of predictor variable as a past of response variable. Moreover, the nonparametric regression method has been developed the smoothing technique which produces a smoother based on NLAR model. The SS and PS methods are computed to fit NLAR model with stationary and nonstationary time series data. For simulation study, the data is generated by the autoregressive process with several coefficient autocorrelations and sample sizes. The performance of SS and PS methods is used the criterion by minimizing the average mean square error values. The SS method exhibits a good power estimation in all cases of stationary and nonstationary data. For economic data, the gold price is an important factor for pretty much all of the world market. The gold price (US Dollars per Troy Ounce) is then applied by using SS and PS methods that collected in term of the monthly volume from January, 1984 to December 2013. The result is founded that the SS method performs better than PS method which is similar the result in case of simulation study.

Keywords: Nonlinear autoregressive model; Smoothing splines; Penalized splines.

1 Introduction

Parametric regression method is widely used for estimating regression function when dependent variable and independent variable are focused on the relationship. More specifically, parametric regression method requires several assumptions. To overcome these assumptions, nonparametric regression method is a choice for estimating function when the data may not

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

be available to use. The basic idea of nonparametric regression method is to let the dependent variable produce a smoothing function from independent variable. Popular nonparametric regression methods can be used by smoothing spline method (Wahba, 1990; Green and Silverman, 1994) and penalized spline method (Ruppert et al., 2003). The nonlinear autoregressive (NLAR) model is developed from the autoregressive model and mixed the smoothing function of nonparametric regression model in terms over the previous values. In this paper, we propose the smoothing spline method and penalized spline method to approximate smoothing function from NLAR model.

2 The nonlinear autoregressive model

The nonlinear autoregressive model is written as

$$y_t = \mu(y_{t-1}) + \varepsilon_t, \qquad t = 2, 3, \ldots, n, \tag{1}$$

where y_t are the dependent variables, y_{t-1} are the past of independent variables at lag 1, $\mu(y_{t-1})$ are the smoothing function of nonlinear autoregressive model, and ε_t denote the error terms.

3 Nonparametric regression

3.1 Smoothing spline method

The smoothing spline was studied by Wahba (1990) that the smoothing spline estimator is estimated the natural polynomial spline $S_\lambda^{(K)}(\mu)$. Green and Silverman (1994) emphasized the natural cubic spline to fit the nonparametric regression function by minimizing

$$S_\lambda^{(K)}(\mu) = \sum_{t=1}^{n} \{y_t - \mu(x_t)\}^2 + \lambda \int_a^b \{\mu''(x_t)\}^2 dx_t.$$

In this case, we propose the NLAR model via smoothing spline method, and the natural cubic spline can be written as

$$S_\lambda^{(K)}(\mu) = \sum_{t=2}^{n} \{y_t - \mu(y_{t-1})\}^2 + \lambda \int_a^b \{\mu''(y_{t-1})\}^2 dy_t.$$

The smoothing spline estimator is approximated by minimizing $S_\lambda^{(K)}(\mu)$. Wahba (1977) suggested Generalized Cross-Validation (GCV) for choosing the smoothing parameter λ.

3.2 Penalized spline method

Penalized spline smoother is estimated using the truncated power function (Ruppert and Carroll, 2000). The natural cubic spline of called the low-rank thin-plate spline representation of $\mu(\cdot)$ is

$$\mu(x_t, \theta) = \alpha_0 + \alpha_1 x_t + \sum_{k=1}^{K} \beta_k |x_t - \tau_k|^3, \qquad t = 1, 2, \ldots, n,$$

where $\theta = (\alpha_0, \alpha_1, \beta_1, \ldots, \beta_K)^{\mathrm{T}}$ is the vector of regression coefficients, and $\tau_1 < \tau_2 < \cdots < \tau_K$ are fixed knots. In this case, we focus the NLAR model based on penalized spline method, then the natural cubic spline can be written as

$$\mu(y_{t-1}, \theta) = \alpha_0 + \alpha_1 y_{t-1} + \sum_{k=1}^{K} \beta_k |y_{t-1} - \tau_k|^3, \qquad t = 2, 3, \ldots, n.$$

To avoid overfitting, we minimize the natural cubic spline to estimate penalized spline estimator.

4 Simulation study

At the beginning, we generate data y_t, $t = 1, 2, \ldots, n$, from an AutoRegressive (AR) process at lag 1 by taking the coefficients $\rho = 0.1, 0.5, 0.7$, and 0.99 in the equation as

$$y_t = \rho y_{t-1} + \varepsilon_t, \qquad t = 1, 2, \ldots, n,$$

where ρ is the AR coefficients and ε_t is the error in term of normal distribution with $\mu = 0$ and $\sigma^2 = 1$. We simulate data with sample sizes $n = 50, 100, 200$, and 400 and repeat the generation at 500 times by R program. The criterion to assess the efficiency of the smoothing spline and the penalized spline method is the Mean Square Error (MSE).
From Table 1, the average MSE of smoothing spline method is less than penalized spline method for all cases. However, when the sample size n is large, the average MSE of smoothing spline method and penalized spline method is slightly different.

5 Application to real data

In this section, we consider the application of NLAR model using smoothing spline and penalized spline methods that we developed in the previous section. As financial data, we use the monthly volume of gold price (US Dollars per Troy Ounce) from January, 1984 to December 2013. The MSE values show that the MSE of smoothing spline is 685.0235 and penalized spline is 733.9173. Therefore it can be concluded that the smoothing spline method outperforms penalized spline method.

TABLE 1. The average MSE of Smoothing Spline (SS) method and Penalized Spline (PS) method based on NLAR model.

Method	n	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.99$
SS	50	3.0353	2.9404	2.9606	2.7581
	100	3.7652	3.7804	3.7402	3.6566
	200	3.8829	3.8596	3.8608	3.8906
	400	3.9451	3.9684	3.9544	3.9254
PS	50	3.8311	3.8084	3.7417	3.6084
	100	3.8414	3.8934	3.8404	3.7827
	200	3.9291	3.9130	3.9163	3.9378
	400	3.9684	3.9905	3.9809	3.9553

6 Conclusion

We have focused on smoothing spline method and penalized spline method for fitting NLAR model. Depending on the simulated data, and real data, the smoothing spline method works more efficient than the penalized spline method. Furthermore the smoothing parameter could be sensitive to fit mean function because the smoothing parameter with GCV can be converged to interpolating spline in a class of smoothing spline estimator. This finding suggests that the user can choose smoothing spline method that provided a good approximation for fitting mean function. the examples very tightly.

References

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Ruppert, D. and Carroll, R.J. (2000). Spatial-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, **42**, 205–224.

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

Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In: *Application of Statistics*, (P.R. Krisnaiah, ed.), Amsterdam: North Holland, 507–523.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

Longitudinal data with informative dropout: an approach based on an AR(1) latent process

Silvia Bacci[1], Francesco Bartolucci[1], Silvia Pandolfi[2]

[1] Department of Economics, University of Perugia, Perugia, Italy
[2] Department of Political Sciences, University of Perugia, Perugia, Italy

E-mail for correspondence: francesco.bartolucci@unipg.it

Abstract: A critical problem in repeated measurement studies is the occurrence of nonignorable missing observations. A common approach to deal with this problem is joint modelling the longitudinal and survival processes for each subject on the basis of a random effect that is usually assumed to be time constant. We relax this hypothesis allowing for time-varying subject-specific random effects, which follow a first-order autoregressive process. We also adopt a generalised linear model formulation to accommodate for different longitudinal outcomes (i.e., continuous, binary, counts). Estimation of the likelihood of the resulting joint model is based on quasi-Newton algorithms and on recursions developed in the hidden Markov literature. The properties of the proposed approach will be illustrated by a Monte Carlo simulation study.

Keywords: Generalised linear models; Nonignorable missing; Survival process.

1 Introduction

A typical situation encountered in the context of longitudinal data is the presence of informative dropout or nonignorable missing data (Little and Rubin, 2002), usually due to the manifestation of a certain event of interest, typically death. A possible approach to model these data is based on Joint Models (JMs; Wulfsohn and Tsiatis, 1997; Rizopolous, 2012), where an underlying latent process is introduced and it affects both the longitudinal and the missing (or survival) processes.
The standard formulation of JMs is characterised by a generalised linear mixed model for the longitudinal process, with subject-specific time-constant normal random effects, and by a proportional hazard model for

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

the survival process, where the risk of the event of interest at a given time depends on the expectation of the longitudinal response at the same time. In order to relax the restrictive assumption of time-constant random effects, Bartolucci and Farcomeni (2014) introduce a family of mixed latent Markov models, where the nonignorable missing process is accounted for through a discrete time-to-event history approach. Differently, Barrett et al. (2015) illustrate an approach for continuous longitudinal responses based on the discretisation of time-to-event and on a probit model for the dropout.
In our contribution, we propose to adopt a first-order autoregressive process, AR(1), instead of a discrete one, so that the resulting model is more parsimonious than that of Bartolucci and Farcomeni (2014), and we generalise the approach of Barrett et al. (2015) to different longitudinal outcomes, such as binary, ordinal, and count responses.

2 The proposed model

Our proposal consists in formulating a random intercept generalised linear model for the longitudinal process for observation j of individual i taken at time t_{ij} following in a certain period $s_{ij} = s(t_{ij})$:

$$g(\mu_{ij}) = \alpha_{is_{ij}} + \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta},$$

where $g(\cdot)$ is a suitable link function, μ_{ij} is the conditional expected value of the outcome $y_{ij} = y_i(t_{ij})$, and $\mathbf{x}_{ij} = \mathbf{x}_i(t_{ij})$ is the corresponding vector of covariates that may include time t_{ij} itself. Note that our proposal is completely general in the sense that any type of outcome is admitted (i.e., continuous, binary, counts). Besides, we assume that $\alpha_{is} = \alpha_{i,s-1}\rho + \eta_{is}\sqrt{1-\rho^2}$, with $s > 1$, $\alpha_{i1} = \eta_{i1}$, $\rho = \mathrm{cor}(\alpha_{i,s}, \alpha_{i,s-1})$, and independent errors $\eta_{is} \sim N(0, \sigma^2)$.
The formulation of the proposed JM is completed by a logit model for the survival process, that is, $\mathrm{logit}\,[p(S_i > s | S_i > s-1, \alpha_{is})] = \alpha_{is}\gamma + \mathbf{w}_{is}^{\mathrm{T}}\boldsymbol{\delta}$, with S_i denoting the number of periods that individual i survives, with $S_i \in \{1, \ldots, m\}$, and $\mathbf{w}_{is}$ a vector of covariates affecting the survival process.
Estimation of model parameters relies on the maximisation of the likelihood function, which is defined on the basis of a quadrature method using an equally spaced grid of nodes and a recursion developed in the hidden Markov literature (Baum et al., 1970). Moreover, as the exact likelihood approach of Barrett et al. (2015) cannot be implemented in our general model, we base the maximisation of the likelihood on alternating quasi-Newton algorithms, in which the observed information matrix is obtained by a numerical method based on finding the zeros of the score vector.

3 Simulation study

We simulated longitudinal data with dropout for $n = 1000, 2000$ individuals and different longitudinal sub-models (with continuous, binary and count

outcomes). The longitudinal measurements are randomly distributed over $m = 5, 10$ time intervals, with 1-10 or 1-20 repeated measurements per person. Each individual may have a varying number of visits in each time interval. We considered a uniform distribution for the visit time, and we assumed that the dropout may occur during any time interval. For each simulation study, we generated 500 datasets. Following Barrett et al. (2015), we included two covariates, in both longitudinal and survival models: age, with initial values generated uniformly from the interval (10-30), and a binary covariate (sex), which assumes value 1 with probability 0.5. We then considered two different values for the variance of the random effects, by letting $\sigma^2 = 1, 4$, and for the parameters $\gamma = 0.05, 0.5$, and $\rho = 0.7, 0.9$. For continuous data, we also let $\tau^2 = 1, 4$, with τ^2 denoting the variance of the repeated measures.

With reference to four different scenarios with continuous outcomes, Table 1 reports the values of the mean and related standard errors and root mean

TABLE 1. Mean parameter estimates (standard errors in parentheses) and root mean square errors (RMSE) under four different scenarios for continuous data.

	Scenario 1 (Benchmark)		Scenario 2 $\sigma^2 = 4$		Scenario 3 $m = 10$		Scenario 4 $n = 2000$	
	Mean	*RMSE*	*Mean*	*RMSE*	*Mean*	*RMSE*	*Mean*	*RMSE*
Longitudinal								
β_0	89.995	0.136	90.002	0.217	90.003	0.119	89.995	0.091
	(0.136)		(0.217)		(0.119)		(0.091)	
Time	-1.701	0.017	-1.701	0.024	-1.700	0.010	-1.700	0.012
	(0.017)		(0.024)		(0.010)		(0.012)	
Age t_0	-1.700	0.006	-1.700	0.010	-1.700	0.005	-1.700	0.004
	(0.006)		(0.010)		(0.005)		(0.004)	
Sex	2.005	0.067	1.998	0.117	1.997	0.064	2.005	0.047
	(0.067)		(0.117)		(0.064)		(0.046)	
Survival								
δ_0	2.000	0.225	2.007	0.237	1.997	0.178	1.992	0.155
	(0.225)		(0.237)		(0.178)		(0.154)	
Time	0.008	0.040	0.008	0.038	0.010	0.016	0.010	0.025
	(0.039)		(0.038)		(0.016)		(0.025)	
Age t_0	0.010	0.009	0.010	0.010	0.010	0.007	0.010	0.006
	(0.009)		(0.010)		(0.007)		(0.006)	
Sex	0.104	0.108	0.096	0.100	0.097	0.087	0.010	0.006
	(0.108)		(0.100)		(0.087)		(0.006)	
γ	0.049	0.077	0.049	0.033	0.050	0.076	0.051	0.057
	(0.077)		(0.033)		(0.076)		(0.057)	
Others								
τ^2	1.001	0.032	0.997	0.032	1.002	0.043	1.000	0.022
	(0.032)		(0.032)		(0.043)		(0.022)	
σ^2	0.992	0.055	3.993	0.148	0.995	0.062	1.000	0.038
	(0.054)		(0.148)		(0.062)		(0.038)	
ρ	0.698	0.026	0.699	0.018	0.699	0.030	0.700	0.019
	(0.026)		(0.018)		(0.030)		(0.019)	

square errors (RMSE) of the considered estimators. The first scenario assumes the same parameter setting as in Barrett et al. (2015), which is taken as benchmark design. More in detail, we have $\boldsymbol{\beta} = (90, -1.7, -1.7, 2)^{\mathrm{T}}$, $\boldsymbol{\delta} = (2, 0.01, 0.01, 0.1)^{\mathrm{T}}$, $\gamma = 0.05$, $\tau^2 = 1$, $\sigma^2 = 1$, $\rho = 0.7$, $n = 2000$, $m = 5$, and 10 repeated observations. Scenario 2 is based on a larger variance of the random effects, with $\sigma^2 = 4$, whereas Scenario 3 is based on a larger number of time intervals, $m = 10$. The last scenario evaluates the effect of an increase in the number of individuals, with $n = 2000$. The results lead us to conclude that, under all scenarios, the means of the parameter estimates are close to the true values. The first scenario confirms the results illustrated in Barrett et al. (2015). This allows us to validate the proposed estimation method. Under the second scenario, we observe that the standard errors and the RMSE of parameter estimates for the longitudinal model and those for σ^2 are larger that those obtained under Scenario 1. On the other hand, the standard error of the estimator of γ is smaller. Under Scenario 3, we observe an improving of the behavior of the estimators of the survival model parameters, with smaller standard errors and RMSE. Finally, as n increases, the behaviour of the proposed estimation method improves for all parameters.

Acknowledgments: Authors acknowledge the financial support from the grant FIRB ("Futuro in ricerca") 2012 on "Mixture and latent variable models for causal inference and analysis of socio-economic data," which is funded by the Italian Government (RBFR12SHVV).

References

Barrett, J., Diggle, P., Henderson, R., and Taylor-Robinson, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society, Series B*, **77**, 131 – 148.

Bartolucci, F. and Farcomeni, A. (2014). A discrete time event-history approach to informative drop-out in multivariate latent Markov models with covariates. *Biometrics*, doi: 10.1111/biom.12224.

Baum, L.E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41**, 164 – 171.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley.

Rizopolous, D. (2012). *Joint models for longitudinal and time-to-event data with applications in R*. Boca Raton, FL: Chapman&Hall/CRC Press.

Wulfsohn, M.S. and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330 – 339.

Log-Gaussian Cox processes with spatially varying second order properties

Haakon Bakka[1]

[1] Norwegian University of Science and Technology (NTNU), Norway

E-mail for correspondence: haakon.bakka@math.ntnu.no

Abstract: We consider log-Gaussian Cox processes, a class of point processes that accounts for spatial aggregation through a Gaussian random field. These models may be interpreted as latent Gaussian models in a Bayesian hierarchical modelling framework, and hence be fitted to point pattern data using integrated nested Laplace approximation (INLA).
While modelling approaches typically assume that second order properties of the latent spatial field are the same everywhere, this assumption may not be realistic in practice. This poster considers situations where the range of the Gaussian field varies in space, giving different cluster sizes in different regions, while the average number of points within a region remains constant.

Keywords: Spatial point process; Non-stationary; SPDE; Ecology.

1 Non-stationary point patterns

As an illustration, Figure 1 shows a simulated point pattern with a non-stationary structure. Specifically, the upper half of the pattern does not seem to exhibit much clustering, while the lower half of the pattern has a distinct clustering structure. We can set up a hierarchical model and fit it to the data using INLA. The model is

$$y_i \sim u(s_i) + m, \tag{1}$$
$$u(s) - \nabla h^{-2} \nabla u(s) = h^{-1} \tau \mathcal{W}(s), \tag{2}$$

where $u(s)$ is the Gaussian Field, h and τ are constants. Also, $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$, and $\mathcal{W}(s)$ denotes white noise. For more details, see Lindgren et al. (2011). This model's latent Gaussian field is a stationary Matérn field, and there

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

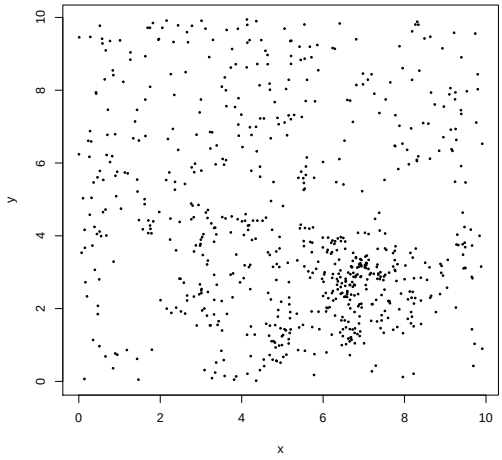


FIGURE 1. A simulated point pattern with non-stationary second order behaviour.

are no covariates. Inference yields the estimated log-intensity in Figure 2. At first sight, the estimated field seems to reflect the observed structure. However, when we use the posterior mean estimates of the hyper-parameters to simulate new intensity fields and point patterns, the simulations do not have the same structure as our original dataset. See Figure 3 for an example of such a simulated point pattern. This pattern does not have the same structure as the original data in Figure 1; the points on the bottom half of the plot are clearly not more clustered than the points on the top half. The reason this happens is that our stationary model does not allow for different clustering structure, but uses the same range in both parts of the field.

To remedy this, we propose to allow for non-stationarity in the field, using what we will refer to as the *Difficult Terrain model*, see Section 2 for details. The inferred field is shown in Figure 4. We see that the top half of the field is almost completely flat, representing the scattered structure of the data in that region. This is clearly non-stationary, because the field has a rapid transition from the flat part (in the top) to the varying part. In addition, point patterns simulated from this non-stationary model exhibit non-stationary behaviour. The data presented in Figure 1 were actually generated using the Difficult Terrain model.

In the poster, we will illustrate the relevance of non-stationary models in an ecological context including point patterns of rainforest trees in different habitats.

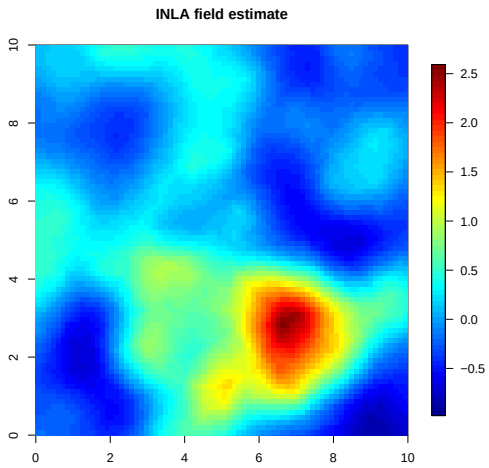


FIGURE 2. The estimated log-intensity field for the model in equation (1).

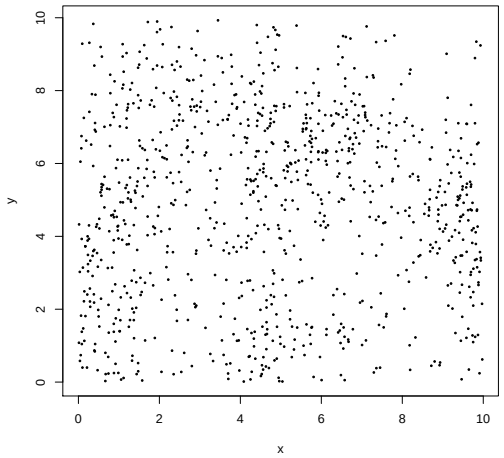


FIGURE 3. A point pattern simulation from the posterior INLA model. The log-intensity field underlying this pattern was simulated using the same range as for the inference in Figure 2.

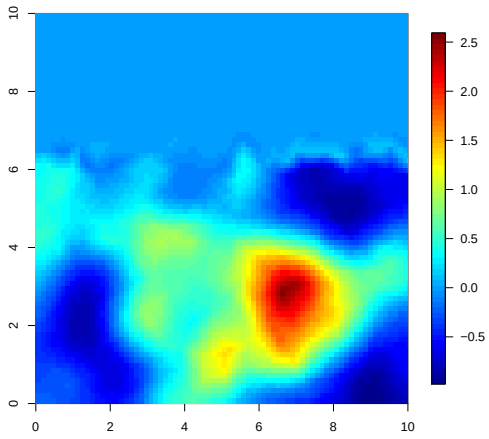


FIGURE 4. The posterior mean of the log-intensity field you get with a non-stationary difficult terrain model.

2 Details of the difficult terrain model

The Gaussian field in the non-stationary difficult terrain model used here is defined by the solution $u(s)$ to

$$u(s) - \nabla h^{-2}(s) \nabla u(s) = h^{-1}(s) \tau \mathcal{W}(s), \tag{3}$$

where $u(s), s \in \Omega \subseteq \mathbb{R}^2$ is the GF, $h(s)$ is a locally constant function, and τ is constant. Also, $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right)$, and $\mathcal{W}(s)$ denotes white noise. To see how this can be approximated with a GMRF, see Lindgren et al. (2011), and note that we have taken their equation (2), but re-parametrized it and fixed $\alpha = 2$.

Acknowledgments: Special thanks to my supervisors Janine Illian, Daniel Simpson and Håvard Rue.

References

Lindgren, F., Rue, H., and Lindstrøm, J. (2011). *An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society, Series B*, **73**, 423 – 498.

A receiving process simulation model of the consumer product distribution center

Somsri Banditvilai[1], Saowapa Mahakeeta[2]

[1] King Mongkut's Institute of Technology Ladkrabang, Thailand
[2] Cannon Pachinburi Ltd., Thailand

E-mail for correspondence: kbsomsri@kmitl.ac.th

Abstract: This paper presents a case study of the current state of the queueing system of trucks in the receiving process of the consumer product distribution center in Bangkok, Thailand. This study emphasizes methods of reducing the high cost of lost time expense that is paid for trucks that are waiting to transfer goods in the receiving process. There are 3 receiving gates. Gate 1 receives only home care products. Gate 2 receives only personal care products and Gate 3 receives only food products. Data collected from the database of truck usage time at each point of the distribution center and a survey form of the moving and storing time of products was analyzed. Arena 14.0 was used to create the simulation model, then verify and validate the model. The model represented the actual system used to test each policy to improve the receiving process. The findings from this study show that the first policy is to allow Gates 1 and 2 to receive both home care products and personal care products. This could reduce the average time of the receiving process by 35.07% and reduce the lost time expense by 3.7%. The second policy is to add one checker and one staff member for moving and storing products at each gate. This could reduce the average time of the receiving process by 25.99% and reduce lost time expense by 1.28%. The third policy is to change the sequence of the receiving process by removing the truck canvas before going to wait in the car park. This is found to reduce the average time of the receiving process by 48.09% and lost time expense by 11.23%.

Keywords: Simulation model; Receiving process; Distribution center.

1 Introduction

Distribution Center is a warehouse which is designed for specific purposes to make it more convenient for storing and moving products in and out. It is for collecting products from plant and distributing them to retailers,

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

wholesalers or customers (Lambert et al., 1998). The case study of the distribution center is of a consumer product distribution center. This research studies the process of trucks circulating in the receiving process of the distribution center. Each day trucks enter the distribution center and carry different types of products in different quantities. Each receiving gate is specific for each type of product, and in peak time, there are a lot of trucks entering the distribution center where manpower and receiving gates are limited. This causes the receiving process to be delayed. The distribution center needs to pay for lost time expense for trucks waiting over one hour, and this greatly increases operation costs of the distribution center.

2 The scope of this research

This research simulated the receiving process of a case study consumer product distribution center and does not concern about the distribution process. This distribution center does not have any problems with the distribution process. The distribution center staff of the receiving process were informed by the truck drivers in advance about the product type and quantity, staff used Exceed system to check and prepare the space for products that will arrive at the distribution center. Therefore, there was no problem with the space available.

3 Methodology

3.1 The study of the receiving process

The case study distribution center operates 24 hours per day and is divided into 3 shifts: The morning shift is from 6.30 a.m. to 2.30 p.m. The afternoon shift is from 2.30 p.m. to 10.30 p.m. and the night shift is from 10.30 p.m. to 6.30 a.m. It operates from Mondays to Saturdays and closes on Sundays. This distribution center has 3 receiving gates. Gate 1 receives only home care products. Gate 2 receives only personal care products and Gate 3 receives only food products. Gate 1 and Gate 2 have two staff members for moving and storing products and two checkers. Gate 3 has only one staff member for moving and storing products and one checker. The steps for the receiving process are as follows:

1. Trucks enter the distribution center and security staff at the gate will scan to record the plate no. and the truck arrival time, then check the type of product and the receiving gate of that type whether it is available. If the receiving gate is not available, trucks will process to a parking space.
2. When the receiving gate is available, the receiving staff will inform the security staff. The truck needs to be scanned again at the gate to record the time of entry inside the distribution center. Then proceed to remove canvas.

3. After removing canvas, trucks proceed to the receiving gate.
4. Moving and storing staff members move the products down and place them at the front of the receiving gate.
5. The checker checks products and matches with the documents. If so, (s)he returns the documents to the truck driver and prints labels to indicate where to place the products.
6. The truck travels from the receiving gate to the exit gate and scans again to record the departure time.
7. After placing products inside the receiving gate the receiving gate becomes available and then informs the security guard to call the next truck. Staff members move and store products on shelves.

3.2 Data collection and analysis

The distribution center system recorded the truck usage time, starting from arrival at the distribution center, and going through each point in the receiving process until departing from the distribution center. Data collected from the database of the truck usage time at each point of the distribution center and a survey form of moving and storing time of the products.

TABLE 1. The probability distribution of time related to the receiving process that was analyzed by Input Analyzer of Arena 14.0 and Chi-square test.

Time in the receiving process	The probability distribution
Trucks arrival time	2+EXPO(22,2)
From security gate to car park	U(0.86, 1.34)
From car park to security gate	0.75+0.6*BETA(1.93, 1.68)
From security gate to remove canvas	U(0.85, 1.45)
Moving products down and placing at the front of receiving gate	U(4.5, 6.5)
Removing canvas	24.5 + 17 * BETA(1.32, 1.5)
From removing canvas to receiving gate	N(1.15, 0.119)
Moving products down and placing at the front of receiving gate	U(4.5, 6.5)
Inspecting time for truck with 22 pallets	U(4.5, 6.5)
Inspecting time for truck with 44 pallets	U(10.5, 12.5)
Time of pushing products inside the gate	U(23.5, 25.5)
From receiving gate to security gate	U(2.5, 4.5)

3.3 Model building, verification and validation

Simulation is the imitation of the operation of a real-world process or system over time (Banks et al., 2005). By Arena 14.0, a discrete-event simulation model of receiving process was built. The model was verified and validated extensively in order to confirm that the model represents the current

receiving process and experiments with the policies to improve the receiving process. In a total of 1000 simulation runs, the average truck spending time in distribution center of the model is less than the actual system by 4.62%, the average truck waiting time in the car park of the model is less than the actual system by 4.25% and the average truck spending time in the receiving gate is less than the actual system by 2.78%, which is not beyond an acceptable range of 10% (Law, 2007). Therefore, the model represents the current receiving process.

3.4 Set the policies to improve the receiving process

From the study of the receiving process, we found that 42% are home care products, 23% are personal care products and 35% are food products. Since trucks need to remove canvas before going to the receiving gate and the time for removing canvas is quite long, we set three policies to improve the receiving process as follows. The first policy is to allow Gates 1 and 2 to receive both home care products and personal care products. The second policy is to add one checker and one staff member to move and store products at each gate. The third policy is to change the sequence of the receiving process by removing the truck canvas before going to wait in car park.

4 Results and conclusion

The indication of the efficiency of the system is the truck spending time in the distribution center. The first policy can reduce trucks spending time in the distribution center from 276.49 min to 179.52 min which decreased by 35.07% and reduced the lost time expense by 3.71%. The second policy can reduce trucks spending time in the distribution center from 276.49 min to 204.63 min which decreased by 25.99% and reduced the lost time expense by 2.33%. The third policy can reduce trucks spending time in the distribution center from 276.49 min to 43.52 min which decreased by 48.09% and reduced the lost time expense by 13.23%. This third policy should be implemented since it can greatly minimize the average time of the receiving process and the lost time expense. The drawback of this policy is that it does not work well in the rainy season because the products may get wet so we should build a roof to cover the car park.

References

Banks, J., Carson J.S., and Nelson, B.L. (2005). *Discrete-Event System Simulation*, Upper Saddle River, NJ: Pearson Education.

Law, A.M. (2007). *Simulation Modeling and Analysis*, Singapore: McGraw-Hill.

Lambert, D.M., Stock, J.R., and Ellram, L.M. (1998). *Fundamental of Logistics Management*, Singapore: McGraw-Hill.

A bivariate index for stock classification

Edan Bar[1], Haim Bar[2]

[1] Cornell University, USA
[2] University of Connecticut, USA

E-mail for correspondence: eb538@cornell.edu

Abstract: We develop a statistical model for classifying publicly-traded stocks in terms of their profitability and volatility. Our goal is to create an informative and robust method for evaluating securities in an investment portfolio.

Keywords: Mixture model; EM algorithm; Volatility; Technical indicators.

1 Introduction

Stock trading is often seen as a task for experts. Most people choose to invest in what they perceive as a simpler, safer approach such as trading in either mutual funds, where a basket of stocks is selected by analysts, or index funds, which include all stocks in an exchange or category. While diversification is a prudent strategy in general, many funds might over-diversify in the sense that they include highly correlated stocks, which do not significantly reduce risk as intended. There are a number of reasons why learning to invest in an informed way is not easy. First, performing research on each stock is time consuming. Second, comparing stocks or funds is hard. One may choose from a host of performance indexes, but it is not clear which index is most informative, or how to combine information from several indexes. Third, although several indexes provide intuition about the value or state of a company, there is usually no probabilistic model for investment choices. Our goal is to provide a novel, model-based approach which yields a probabilistic criterion for selecting stocks. While we do not claim that it is possible to predict future performance of stocks, we aim to create useful and robust technical indicators that reflect stocks' profitability and volatility.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

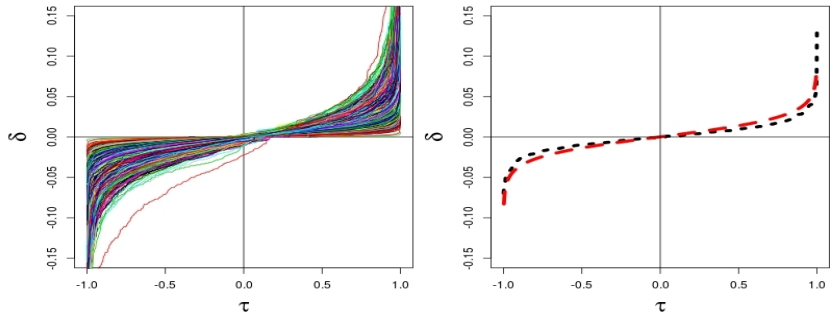


FIGURE 1. Left: each line represents the ordered daily changes, $\delta_i(\tau)$, of a stock, over a period of 754 days. The index of the sorted 754 observations is converted to be equally spaced values in (-1,1). Right: The dashed line represents the sorted daily change value for one stock, and the dotted line is obtained from fitting our model to the data.

1.1 The model

Let b_{it} and e_{it} be the open and close values, respectively, of stock $i = 1, \ldots, n$ at time $t = 1, \ldots, k$. Let the daily change in stock value be $\delta_{it} = (e_{it} - b_{it})/b_{it}$. We sort the values of δ_{it} in increasing order, and denote the sorted list by $\delta_{i(t)}$. We transform the index (t) by using $\tau = \frac{(t)-k/2-1/2}{k}$. Then, we consider the sorted daily changes as a function, $\delta_i(\tau)$, observed at equally spaced values in $[-1+\frac{1}{k}, 1-\frac{1}{k}]$. The left panel of Figure 1 shows the values of $\delta_i(\tau)$ for 3,112 stocks, over 754 days. This plot motivates our modeling approach. We assume that

$$\delta_i(\tau) = \eta_i \log\left(\frac{\gamma_i(1+\tau)}{1-\tau}\right). \tag{1}$$

A stock can be characterized in terms of the parameters $\eta_i > 0$ and $\gamma_i > 0$. The parameter η_i represents the slope of the curve $\delta_i(\tau)$. A large η_i is obtained if the variance of the daily changes is large (higher volatility). The parameter γ_i represents the x-intercept: when $\gamma_i = 1$, we have $\delta_i(0) = 0$, which means that exactly half the time the stock moved up, and half the time it went down. When $\gamma_i > 1$ the intercept is at $\tau < 0$ which means that the stock value increased more often than it decreased. We obtain the following parameter estimates:

$$\eta_i = \frac{\delta_i(\tau_0) - \delta_i(-\tau_0)}{4\tau_0} \tag{2}$$

$$\gamma_i = \exp\left(\frac{\sum_{\tau\in(-1,1)} \delta_i(\tau)}{k\eta_i}\right), \tag{3}$$

where τ_0 is close to 0. Note that η_i is obtained from the derivative of $\delta_i(\tau)$ at $\tau = 0$, where the function is approximately linear. In practice, we take

several (small) values of τ_0 and compute the average of the estimates we obtain from (3).
Next, we want to estimate the *distributions* of η_i and γ_i, across all companies. To do that, we assume that each parameter is drawn from a mixture of three log-normal distributions, representing low/medium/high levels. Specifically, we assume that

$$\log \eta_i \sim c_{0\eta} N(\mu_{0\eta}, \sigma^2_{0\eta}) + c_{1\eta} N(\mu_{0\eta} + \mu_{1\eta}, \sigma^2_{1\eta}) + c_{2\eta} N(\mu_{0\eta} - \mu_{1\eta}, \sigma^2_{1\eta}), \quad (4)$$

and similarly for $\log \gamma_i$, with parameters $c_{0\gamma}, c_{1\gamma}, c_{2\gamma}, \mu_{0\gamma}, \mu_{1\gamma}, \sigma^2_{0\gamma}, \sigma^2_{1\gamma}$. We estimate the parameters using the EM algorithm (Dempster et al., 1977), treating the indicator variables c_{jx} ($x = \eta, \gamma$, and $j = 0, 1, 2$), as 'missing data'. The expected values of c_{jx} are estimated using Bayes' rule, as in Bar et al. (2014). Thus, $\hat{c}_{jx}$ are conveniently interpreted as posterior probabilities. For example, if $\hat{c}_{2\gamma}(i) > 0.8$, then stock i increases significantly more often than it decreases. Furthermore, this modeling approach allows us to associate a bivariate index $\psi_i = (\eta_i, \gamma_i)$ with each company, as demonstrated in the next section.

2 Analyzing US stock data, 2012 – 2014

We obtained a list of 6,683 US stock symbols from NASDAQ. For each stock we obtained the 2012-2014 daily values from Yahoo Finance and key ratios from Morningstar. We used Excel, Python, and R to retrieve and parse the data. We excluded stocks with low median trading volume, and stocks that were not traded in all business days in the 2012-4 period. Greenblatt (2010) points out that utilities and financial stocks can be problematic due to their accounting methods so we removed 'income trust' companies. For each of the remaining 3,112 stocks we fitted model (1) and obtained estimates for η_i and γ_i, using equations 2 and 3. Then, we used the mixture model (4) to fit the distributions of $\log(\eta_i)$ and $\log(\gamma_i)$. The two distributions are shown in the left and middle panels of Figure (2). The dotted curves represent the fitted normal distributions for the three mixture components in (4), and the solid, purple curve is their weighted mixture. In both cases the EM algorithm yields an excellent fit. For $\log(\gamma_i)$ we obtain $\mu_{0\gamma} = 0.056$, and $\sigma_{0\gamma} = 0.072$, and the probability of being in the 'null' component of the mixture is 0.87. Thus, for 87% of the stocks, γ_i is approximately 1.06, and their daily change, $\delta_i(\tau)$, is expected to be positive only slightly more often than negative. The probability of the negative component of $\log \gamma$ is 0.12, while the probability of the positive component is 0.006. Thus, there is a very small number of stocks whose daily change is negative much more often than positive, and approximately 12% of the stocks are classified as having a significant posterior probability for $\delta_i(\tau)$ to be positive. The right panel in Figure (2) shows the bivariate distribution of $\log(\eta_i)$ vs. $\log(\gamma_i)$. The dashed lines represent $\hat{\mu}_{0\eta}$ and $\hat{\mu}_{0\gamma}$. According to model (1), points

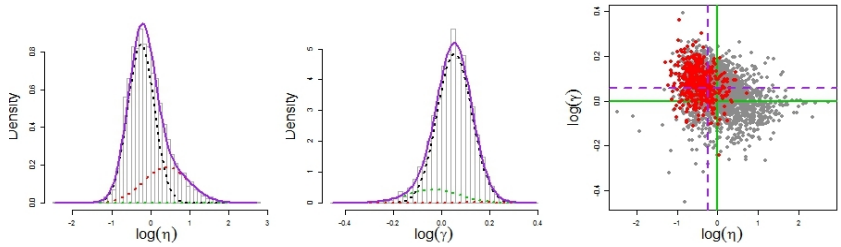


FIGURE 2. Left, Middle: histograms of $\log(\eta_i)$, $\log(\gamma_i)$, and the fitted three-component mixture distributions. Right: the bivariate distribution of $(\log(\eta_i), \log(\gamma_i))$. The dashed lines represent the means of the null components. The red dots represent the 500 companies with the largest market capitalization.

in the left-upper quadrant are 'good stocks', in the sense that their daily change is more likely to be positive than negative and the magnitude of the daily changes is relatively small (low volatility). In contrast, points in the lower-right quadrant are 'bad stocks'. The red dots represent the 500 companies with the largest market capitalization.
In summation, we find that our model fits daily trading data very well and provides intuitive classification of stocks. We plan to test how security selection using our model compares with other models, using historical data. For example, methods based on P/E ratios, the Capital Asset Pricing Model (CAPM), etc. (Malkiel, 1999). We also plan to check whether our approach is useful to detect market cycles, and how it performs in different market conditions (e.g., 'bull' vs. 'bear' markets).

Acknowledgments: We thank Mr. Dennis Merryfield, for useful conversations.

References

Bar, H.Y., and Booth, J.G., and Wells, M.T. (2014). A bivariate model for simultaneous testing in bioinformatics data. *Journal of the American Statistical Association*, **109**, 537 – 547.

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 – 38.

Greenblatt, J. (2010). *The Little Book that till Beats the Market*. Hoboken, N.J.: John Wiley & Sons.

Malkiel, B. (1999). *A Random Walk down Wall Street Including a Life-Cycle Guide to Personal Investing*. New York: Norton.

Barycentric algorithm for computing D-optimal size-and-cost constrained designs of experiments

Eva Benková[1,2], Radoslav Harman[2]

[1] Faculty of Social Sciences, Economics and Business, Johannes Kepler University Linz, Austria
[2] Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

E-mail for correspondence: eva.benkova@jku.at

Abstract: We study the problem of D-optimal experimental design under two linear constraints, which can be interpreted as simultaneous restrictions on the size and on the total cost of the experiment. For computing a size-and-cost constrained approximate D-optimal design, we propose a "barycentric" algorithm with sequential removal of redundant design points, whose convergence can be proved analytically.

Keywords: Optimal design of experiments; D-optimality; Barycentric algorithm.

1 Introduction

The role of the optimal design of experiments as a statistical discipline is to study methods of conducting an experiment so that the maximum amount of the useful information is attained (e.g., Pukelsheim, 2006).
The research related to the optimal experimental designs is usually focused on situations with a single linear constraint on the "size" of the experiment. However, situations with multiple constraints do occur in practice; they can represent restrictions on various resources consumed by the experiment, see, e.g., Cook-Fedorov (1995). This extended abstract deals with an important special case of constrained designs and is based on the submitted paper Harman and Benková (2015).
We can represent the experimental design by an n-dimensional vector $\mathbf{w}$ of "design weights", with components $w_x \geq 0$, $x \in \mathcal{X} = \{1, \ldots, n\}$.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

Our goal is to propose a method of constructing a D-optimal design $\mathbf{w}^*$ in the set of all designs that satisfy both the size constraint (2) and the cost constraint (3)

$$\begin{aligned} \mathbf{w}^* &\in \operatorname{argmax}_{\mathbf{w}}\{\phi(\mathbf{w})\} && (1)\\ \text{subject to} \quad \sum_{x\in\mathcal{X}} w_x &= 1, && (2)\\ \sum_{x\in\mathcal{X}} c_x w_x &= 1. && (3) \end{aligned}$$

In above, c_x is a positive constant representing the "normalized" cost of the trial in the design point $x \in \mathcal{X}$ and the function $\phi : [0,\infty)^n \to [0,\infty)$ is the criterion of D-optimality defined by $\phi(\mathbf{w}) = \det^{1/m}(\mathbf{M}(\mathbf{w}))$, where $\mathbf{M}(\mathbf{w}) = \sum_{x\in\mathcal{X}} w_x \mathbf{f}(x)\mathbf{f}^{\mathrm{T}}(x)$ is the standardized information matrix of the size $m \times m$. For simplicity, we will assume regularity in the sense that the vectors $\mathbf{f}(1), ..., \mathbf{f}(n)$ span $\mathbb{R}^m$, and $\mathbf{f}(x) \neq \mathbf{0}_m$ for all $x \in \mathcal{X}$. The vectors $\mathbf{f}(x)$, $x \in \mathcal{X}$, can represent known regressors of the linear regression model

$$y(x) = \mathbf{f}(x)^{\mathrm{T}}\beta + \varepsilon(x),$$

where $\beta \in \mathbb{R}^m$ is a vector of unknown parameters and $\varepsilon(x), x \in \mathcal{X}$ are uncorrelated homoscedastic errors with mean values equal to zero.
It is possible to show that the criterion of D-optimality is continuous, concave, monotonic, and homogeneous on $[0,\infty)^n$. Due to the homogeneity of ϕ, a statistically natural definition of efficiency of a design $\mathbf{w}^a$ relative to a design $\mathbf{w}^b$ with $\phi(\mathbf{w}^b) > 0$ is given by $\mathrm{eff}(\mathbf{w}^a|\mathbf{w}^b) = \phi(\mathbf{w}^a)/\phi(\mathbf{w}^b)$.
It can be shown that having a proper method to find a solution to (1)-(3) is enough to solve the real-life motivated problem given by restrictions $\sum_{x\in\mathcal{X}} w_x \leq 1$ and $\sum_{x\in\mathcal{X}} c_x w_x \leq 1$ (see Harman and Benková (2015)). Note that it is important that we require both (2) and (3) to be satisfied; the single cost constraint (3) can be easily transformed to the size constraint (2) and solved by standard methods.

2 Theoretical results and the barycentric algorithm

Let $\mathcal{X}_+ = \{x \in \mathcal{X} : c_x > 1\}$, $\mathcal{X}_- = \{x \in \mathcal{X} : c_x < 1\}$, $\mathcal{X}_0 = \{x \in \mathcal{X} : c_x = 1\}$. In order to ensure feasibility of (1)-(3), we assume that $\mathcal{X}_+ \neq \emptyset$ and $\mathcal{X}_- \neq \emptyset$. We will also assume that $\mathcal{X}_0 \neq \emptyset$; all results can be modified in a straightforward way for the (simpler) case $\mathcal{X}_0 = \emptyset$.
Define $\delta_{x_+} = c_{x_+} - 1 > 0$ for $x_+ \in \mathcal{X}_+$ and $\delta_{x_-} = 1 - c_{x_-} > 0$ for $x_- \in \mathcal{X}_-$ and let $\mathbb{Q}_+^n$ denote the the polytope of all feasible designs of (1)-(3).
For any regular $\mathbf{w} \in \mathbb{Q}_+^n$, let $\mathbf{d}(\mathbf{w})$ denote the variance (sensitivity) function, which is, in our case, the n-dimensional vector with components $d_x(\mathbf{w}) = \mathbf{f}^{\mathrm{T}}(x)\mathbf{M}^{-1}(\mathbf{w})\mathbf{f}(x)$; $x \in \mathcal{X}$.

Theorem 1 *Let* $\mathbf{w} \in \mathbb{Q}_+^n$ *be a regular design. Then, the following three statements are equivalent: (i) The design* $\mathbf{w}$ *is D-optimal in* $\mathbb{Q}_+^n$*. (ii) There exists* $h \in \mathbb{R}$ *such that* $d_x(\mathbf{w}) \leq m + h(c_x - 1)$ *for all* $x \in \mathcal{X}$*. (iii) It holds that* $\max_{x_0 \in \mathcal{X}_0} d_{x_0}(\mathbf{w}) \leq m$ *and*

$$\max_{x_+ \in \mathcal{X}_+} \frac{d_{x_+}(\mathbf{w}) - m}{\delta_{x_+}} + \max_{x_- \in \mathcal{X}_-} \frac{d_{x_-}(\mathbf{w}) - m}{\delta_{x_-}} \leq 0.$$

The barycentric algorithm is a multiplicative method proposed in Harman (2014) for computing approximate D-optimal designs under linear constraints on the weights. The key component is a formula for barycentric coordinates of each $\mathbf{w} \in \mathbb{Q}_+^n$ in the set of all extreme vectors of $\mathbb{Q}_+^n$. This formula results in a proper form of the barycentric transformation (5)-(8). For any regular $\mathbf{w} \in \mathbb{Q}_+^n$ and $x_+ \in \mathcal{X}_+$, $x_- \in \mathcal{X}_-$ define the weighted variances $\tilde{d}_{x_+x_-}(\mathbf{w}) = (\delta_{x_+} + \delta_{x_-})^{-1}(\delta_{x_+} d_{x_-}(\mathbf{w}) + \delta_{x_-} d_{x_+}(\mathbf{w}))$.
The barycentric algorithm starts with a positive design $\mathbf{w}^{(0)} \in \mathbb{Q}_+^n$ and computes a sequence of designs $\{\mathbf{w}^{(t)}\}_{t=0}^{\infty}$ by the formula

$$\mathbf{w}^{(t+1)} = \mathbf{T}^B(\mathbf{w}^{(t)}) \text{ for all } t = 0, 1, \ldots, \tag{4}$$

where $\mathbf{T}^B(\cdot)$ is for any regular $\mathbf{w} \in \mathbb{Q}_+^n$ given by

$$\mathbf{T}^B(\mathbf{w}) = \mathbf{w} \odot \mathbf{d}^\pi(\mathbf{w}), \tag{5}$$

where $\odot$ is the componentwise multiplication and $\mathbf{d}^\pi(\mathbf{w})$ is given by

$$d^\pi_{x_+}(\mathbf{w}) = \frac{\sum_{x_- \in \mathcal{X}_-} w_{x_-} \delta_{x_-} \tilde{d}_{x_+x_-}(\mathbf{w})}{m s^\delta(\mathbf{w})};\ x_+ \in \mathcal{X}_+, \tag{6}$$

$$d^\pi_{x_-}(\mathbf{w}) = \frac{\sum_{x_+ \in \mathcal{X}_+} w_{x_+} \delta_{x_+} \tilde{d}_{x_+x_-}(\mathbf{w})}{m s^\delta(\mathbf{w})};\ x_- \in \mathcal{X}_-, \tag{7}$$

$$d^\pi_{x_0}(\mathbf{w}) = \frac{d_{x_0}(\mathbf{w})}{m};\ x_0 \in \mathcal{X}_0, \tag{8}$$

where $s^\delta(\mathbf{w}) = \sum_{x_+ \in \mathcal{X}_+} \delta_{x_+} w_{x_+} = \sum_{x_- \in \mathcal{X}_-} \delta_{x_-} w_{x_-}$, for all $\mathbf{w} \in \mathbb{Q}_+^n$.
The general theory in Harman (2014) implies that the sequence of information matrices $\{\mathbf{M}(\mathbf{w}^{(t)})\}_{t=0}^{\infty}$ corresponding to the sequence of designs generated by (4) converges to some non-singular matrix $\mathbf{M}^\infty$, but it is does not guarantee that $\mathbf{M}^\infty$ is the optimal information matrix. However, for the specific case of size-and-cost constrained designs and the specific barycentric coordinates used, the following theorem holds.
Theorem 2 *Let* $\mathbf{w}^{(0)} \in \mathbb{Q}_+^n$ *be a positive design and let* $\mathbf{w}^{(t+1)} = \mathbf{T}^B(\mathbf{w}^{(t)})$ *for* $t = 0, 1, \ldots$. *Let* $\liminf_{t \to \infty} s^\delta(\mathbf{w}^{(t)}) > 0$. *Then,* $\lim_{t \to \infty} \phi(\mathbf{w}^{(t)}) = \phi(\mathbf{w}^*)$*, where* $\mathbf{w}^*$ *is any solution of* (1)-(3).
The second part of the following theorem introduces a "deleting rule", which can be employed to remove redundant design points in order to increase the speed of the algorithm.

Theorem 3 *Let* $\mathbf{w} \in \mathbb{Q}_+^n$ *be a regular design, let* $\mathbf{w}^* \in \mathbb{Q}_+^n$ *be a D-optimal design, and let* $\epsilon = \max\left(\max_{x_+ \in \mathcal{X}_+, x_- \in \mathcal{X}_-} \tilde{d}_{x_+x_-}(\mathbf{w}), \max_{x_0 \in \mathcal{X}_0} d_{x_0}(\mathbf{w})\right) - m$. *Then,*

$$\mathrm{eff}(\mathbf{w}|\mathbf{w}^*) \geq \frac{m}{m+\epsilon}.$$

Let $h_m(\epsilon) = m\left(1 + \epsilon/2 - \sqrt{\epsilon(4+\epsilon-4/m)}/2\right)$. *Then,*
(i) $\max_{x_- \in \mathcal{X}_-} \tilde{d}_{x_+x_-}(\mathbf{w}) < h_m(\epsilon)$ *for some* $x_+ \in \mathcal{X}_+$ *implies* $w^*_{x_+} = 0$.
(ii) $\max_{x_+ \in \mathcal{X}_+} \tilde{d}_{x_+x_-}(\mathbf{w}) < h_m(\epsilon)$ *for some* $x_- \in \mathcal{X}_-$ *implies* $w^*_{x_-} = 0$.
(iii) $d_{x_0}(\mathbf{w}) < h_m(\epsilon)$ *for some* $x_0 \in \mathcal{X}_0$ *implies* $w^*_{x_0} = 0$.

Figure 1 displays times and efficiencies of 100 randomly generated problems computed by the barycentric algorithm and selected competing methods.

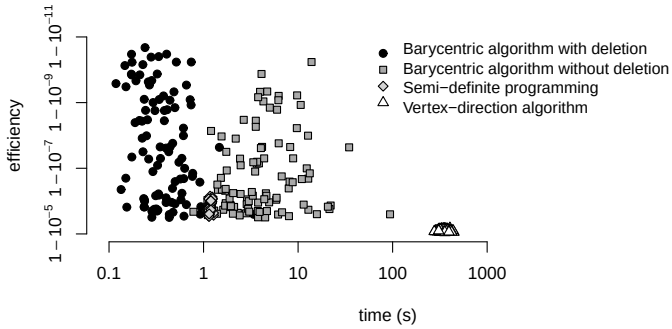


FIGURE 1. A comparison of selected algorithms for computing approximate *D*-optimal size-and-cost constrained designs.

References

Cook, D. and Fedorov, V. (1995). Constrained optimization of experimental design. *Statistics*, **26**, 129 – 178.

Harman, R. (2014). Multiplicative Methods for Computing D-Optimal Stratified Designs of Experiments. *Journal of Statistical Planning and Inference*, **146**, 82 – 94.

Harman, R. and Benková, E. (2015). Approximate D-optimal experimental design with simultaneous size and cost constraints. *arXiv:1408.2698 [stat.CO]*.

Pukelsheim, F. (2006). *Optimal design of experiments*. Classics in Applied Mathematics, SIAM.

Integrated analysis of multi-source data in drug discovery experiments using structural equation models

Theophile Bigirumurame[1], Nolen Perualila-Tan[1], Adetayo Kasim[2], Ziv Shkedy[1]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Hasselt, Belgium
[2] Wolfson Research Institute, Durham University Queen's Campus,Durham, U.K

E-mail for correspondence: theophile.bigirumurame@uhasselt.be

Abstract: The drug discovery and development processes are typically costly and time consuming. Hence, it is crucial to identify early failure of candidate compounds and thereby save time and investment in a later stage. We propose structural equation modeling (SEM) based approach for an integrated analysis which combines information from three data sources: (1) bioactivity variables, (2) variables representing the chemical structure of the compounds, and (3) gene expression data. The proposed model allows to estimate the effects of the gene expression on the biological activity variable and furthermore, it allows to decompose the effect of the chemical structure on the biological activity into direct and indirect (i.e. the effect via the gene expression) effects.

Keywords: Structural equation modeling; Microarray data; Bioassays.

1 Introduction

The drug discovery and development processes are typically costly and time consuming. Hence, it is crucial to identify early failure and thereby save time and investment in a later stage. The decision to continue/stop a development process in drug discovery should ideally be based on scientific parameters that are predictive of later outcomes, and which can be determined quickly and at relatively low cost.
Currently, microarray technology (Amaratunga et al., 2014) is used to monitor simultaneously the activity of thousands genes and their response to a certain drug. Microarrays are providing new insights into the molecular

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

pathology of human cancers and are helping to identify many new additional targets for drug discovery. By understanding gene expression patterns, researchers can gain information that can link sites of expression, biochemical pathways, and normal or pathological functions in organs and whole organisms.
However, relevant biologically data are acquired in a late stage of the research process. The use of biomarkers can reduce the costs and increase efficiency, and they should be incorporated early in the development process to gain information that can aid the development. In this paper we propose an approach based on structural equation modeling to combine information from the three data source, i.e., the bioactivity, the chemical structure of the compound, and the gene expression, and select a subset of genes which can be used as biomarkers.

2 Methodology: Structural equations modeling (SEM)

Three data sources were obtained from an oncology project, which focused on the inhibition of the fibroblast growth factor receptor (Verbist et al., 2015). The chemical structure (presence or absence of fingerprint feature, FFP, in a compound/molecule), the gene expression data and the bioactivity (IC50) outcome. Let X_{ij} be the jth gene expression ($j = 1, \ldots, 3595$), of the the ith compound ($i = 1, \ldots, 35$). The measurement for the bioactivity is denoted by Y_i. Let Z_i be an indicator variable, which takes value 1 if the fingerprint feature (FFP) is present in the ith compound, and 0 otherwise. The key idea behind a structural equation model (SEM) is that the causal relationships among the variables determine the expected pattern of correlation (Li et al., 2006). For the analysis presented in this paper, SEM with observed variables were considered (Bollen, 1989). The main advantage of our approach is that it allows to decompose the total effect of chemical substructure on the bioassay into the direct (effect of the Z on Y unmediated by X) and indirect effects (effect of the Z on Y, mediated by X). Our primary interest is to select genes which maximize the indirect effects. The indirect and direct effects are shown in Figure 1.
The SEM consists of a structural model (i.e., a path analysis model) which describes the casual relationship between the variables. The model is visualized in Figure 1 (right panel). Formally the model can be expressed as set of two models given by:

$$X = \gamma_1 Z + \varepsilon_1, \qquad Y = \gamma_2 Z + \beta X + \varepsilon_2, \tag{1}$$

where: γ_1 and γ_2 are the fingerprint effects on the gene expression and the bioassay respectively, β is the gene specific effect on the bioassay, ε_1 and ε_2 are the uncorrelated measurement errors. It is assumed that $(\varepsilon_1, \varepsilon_2) \sim$

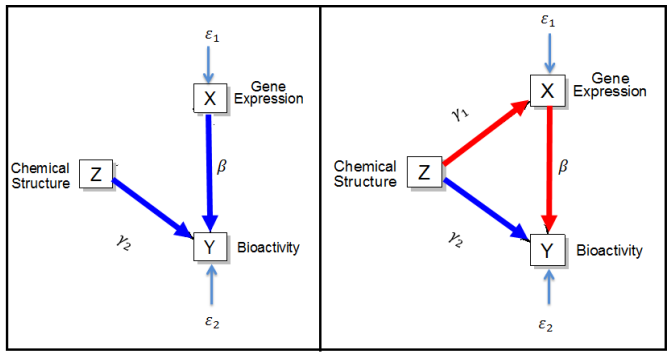


FIGURE 1. Fingerprint feature and gene expression direct effects on the bioassay (blue arrows in the left panel. Fingerprint feature indirect effect on the bioassay through the gene expression (red arrows in the right panel).

$N(0, \psi)$, $\text{var}(Z) = \phi$, and $\text{cov}(\varepsilon_i, Z) = 0$. The indirect effect of the FFP for a given gene j is equal to $\gamma_{1j}\beta$, whereas the direct effect is γ_2.
The unknown parameters can be estimated using the maximum likelihood estimation method. The model in Equation (1) is fitted gene by gene, and multiple testing adjustment using FDR (Benjamini and Hochberg, 1995) is performed to find significant parameters.

3 Results

In this section we present results for one of the genes that was selected using the SEM. This gene (gene 1) corresponds to the subset of genes which maximize the indirect effects. For these subset of genes, we expect to explain most of the FFP effects on the bioassay through the FFP effects on the gene expression. This type of genes is characterized by high FFP direct effect on the gene expression and high gene direct effect on the bioassay. These genes have relatively high correlation between the gene expression and the bioassay and they are differentially expressed. Figure 2 shows a typical gene. The indirect effect was equal to -0.56 and the direct effect was equal to -0.11. Note that the BH-FDR procedure was applied to correct for multiple testing

4 Discussion

There are many challenges in the drug discovery and development. Relevant biological data are acquired too late in the research processes and the use of biomarkers can reduce the cost and increase efficiency. The SEM presented in this paper can be used to select genetic biomarkers which can help in

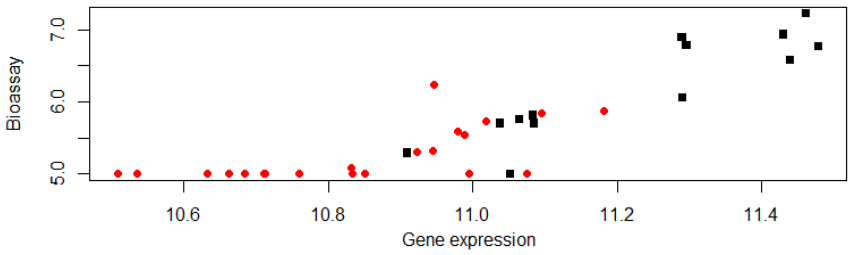


FIGURE 2. Typical gene which maximizes the indirect effect.

the development process. Genes maximizing the indirect effect could help in explaining the the effect of the FFP on the bioassay.
After detecting significant genes, one can find to which pathways they belong, in order to have an insight about the mechanism of action of given chemical. The gained information, thus could help in lead optimization. If toxicity related genes are identified, it could help in deciding to continue/or stop the development process with compounds having a given chemical substructure.

References

Amaratunga, D., Cabrera, J., and Shkedy, Z. (2014). *Exploration and Analysis of DNA Microarray and Other High Dimensional Data*. New York: John Wiley.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289 – 300.

Bollen, K.A. (1998). *Structural Equation Models*. New York: Wiley Series in Probability and Mathematical Statistics.

Li, R., Tsaih, S.W., Shockley, K., Stylianou, I.M., Wergedal, J., Paigen, B., and Churchill, G.A. (2006). Structural model analysis of multiple quantitative traits. *PLoS Genet*, **2**, e114.

Verbist, B., Klambauer, G., Vervoort, L., Talloen, W., QSTAR Consortium, Shkedy, Z., Thas, O., Bender, A., Hinrich, W., Göhlmann, H., and Hochreiter, S. (2015). Using transcriptomics to guide lead optimization in drug discovery projects: lessons learned from the QSTAR project. *Drug Discovery Today*, In Press.

Achieving shrinkage in the time-varying parameter models framework

Angela Bitto[1], Sylvia Frühwirth-Schnatter[2]

[1] WU Vienna University of Economics and Business, Institute for Statistics and Mathematics, Vienna
[2] WU Vienna University of Economics and Business, Institute for Statistics and Mathematics, Vienna

E-mail for correspondence: angela.bitto@wu.ac.at

Abstract: The present paper contributes to the literature in the following way. We investigate shrinkage for time-varying Parameter (TVP) models based on the normal-gamma prior which has been introduced by Griffin and Brown (2010) for standard regression models. Our approach extends Belmonte et al. (2014) who considered the Bayesian LASSO prior, which is a special case of the normal-gamma prior. We show how the normal-gamma prior can easily be extended to the TVP models. We present both a univariate and a multivariate application. First we choose EU area inflation modelling based on the generalized Phillips curve, then we draw our attention to a multivariate time series with a time-varying covariance matrix and analyse DAX-30 data. Our findings suggest, that the normal-gamma prior bears advantages over the Bayesian Lasso prior in terms of statistical efficiency and performs significantly better when drawing attention to the predictive performance.

Keywords: Time-varying parameter model; Hierarchical shrinkage; Normal-gamma prior.

1 Introduction and model

Time-varying parameter (TVP) models are a popular tool for handling data with smoothly changing parameters. The model reads:

$$\begin{aligned} y_t &= \mathbf{x}_t\boldsymbol{\beta} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q}), \end{aligned}$$

where $\boldsymbol{\beta}_t$ follows a random walk starting from the unknown initial values $\boldsymbol{\beta}_0$, which is assumed to be specified by a normal prior distribution

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

$\boldsymbol{\beta}_0 \sim \mathcal{N}_d(\boldsymbol{a_0}, \boldsymbol{P_0})$, further $\boldsymbol{\beta}_0$ is independent of (ε_t) and $(\boldsymbol{\omega}_t)$, which are independent and mutually independent Gaussian innovations. $\mathbf{x}_t$ is a d-dimensional vector $\mathbf{x}_t = (x_{t1}, \ldots, x_{td})$ with $x_{t1} = 1$ containing the regressors of this regression model. We assume $\mathbf{Q}$ to be a diagonal matrix. In general we allow σ_t^2 to be time-dependent which allows to introduce flexible error models such as stochastic volatility error models. However, in situations with many parameters the flexibility underlying these models may lead to overfitting models and, as a consequence, to a severe loss of statistical efficiency. This occurs, in particular, if only a few parameters are truely time-varying, while the remaining ones are constant or even insignificant. As a remedy, hierarchical shrinkage priors have been introduced. We make use of the normal-gamma prior and show how it can easily be extended to TVP models. In particular we use shrinkage priors for the initial values and the square root of the variances in a non-centered reparametrization (cf. Frühwirth-Schnatter and Wagner, 2010) of the model above:

$$\begin{aligned} \beta_j &\sim \mathcal{N}(0, \tau_j^2), \qquad & \tau_j^2 &\sim \mathcal{G}(a^\tau, a^\tau \lambda^2/2), \\ \sqrt{\theta}_j &\sim \mathcal{N}(0, \xi_j^2), \qquad & \xi_j^2 &\sim \mathcal{G}(a^\xi, a^\xi \kappa^2/2), \end{aligned}$$

Obviously the gamma distribution with $a^\tau = 1$ corresponds to the exponential distribution and subsequently the Bayesian Lasso prior results as a special case of the normal-gamma prior. We assume that a^τ and a^ξ are fixed. λ^2 and κ^2 follow a gamma distribution with fixed hyperparameters d_1, d_2, e_1, e_2. We discuss the critical choice of these hyperparameters and aim at giving a guideline for users.

2 Applications

In order to demonstrate the capability of our approach we apply it to a real-world dataset. We follow Belmonte et al. (2014) and analyse EU-area inflation based on the generalized Phillips curve. In this context inflation depends on lags of inflation and other predictors such as unemployment rate, money supply and changes in the oil price.

Due to the lack of formal classification rules, like posterior inclusion probabilities, one needs to find an alternative way to classify the predictors into three different categories of interest (time-varying, significant but static, insignificant). One possible way it to use simple visual inspection. Figure 1 displays the posterior paths of the regressors in the centered parametrization $\beta_{tj} = \beta_j + \sqrt{\theta_j}\tilde{\beta}_{tj}$ and Figure 2 shows the corresponding posterior densities of $\sqrt{\theta_j}$ of those four predictors for two different hyperparameters settings. The blue line indicates a normal-gamma prior setting with $a^\tau = a^\xi = 0.1$ and the red line the special case of a Lasso prior setting with $a^\tau = a^\xi = 1.0$. Neither *1-year Euribor* (Euro Interbank Offered Rate) nor *Loans* exhibit a signed square root of the variance which significantly differs

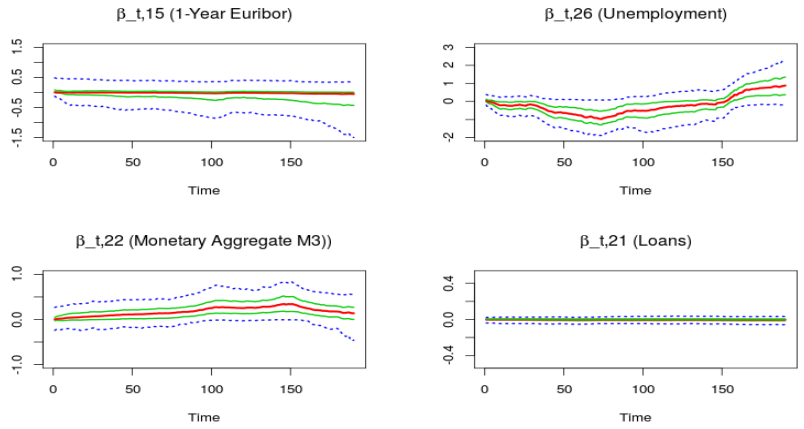


FIGURE 1. Posterior paths of four regressors $\beta_{tj} = \beta_j + \sqrt{\theta_j}\tilde{\beta}_{tj}$.

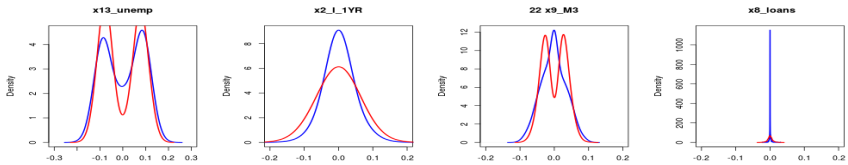


FIGURE 2. Posterior densities for various predictors; blue: $a^\tau = a^\xi = 0.1$, red: $a^\tau = a^\xi = 1.0$; $d_1 = d_2 = e_1 = e_2 = 0.001$.

from zero. Therefore the hypothesis of time-variation of those parameters can be rejected in both cases. For *Unemployment Rate* we can find a different pattern. It exhibits clearly time-varying behaviour as indicated by the bimodal structure of the posterior density plots, for *Monetary Aggregate M3* it is difficult to decide whether it can be regarded as significant or time-varying, depending on the choice of hyperparameters the outcome differs. In the interest of comparing the predictive performance of various shrinkage priors we evaluate the cumulative log predictive scores for various shrinkage priors as in Figure 3. We can see, that the normal-gamma prior significantly outperforms the Bayesian LASSO prior, which is a special case of the normal-gamma prior with $a^\xi = a^\tau = 1.0$.

Our approach can easily be extended to a multivariate time series $\mathbf{y}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{\Sigma}_t)$ with time-varying covariance matrix. We analyse German stock index data (DAX-30) between 2001 and 2012. Using a time-varying Cholesky decomposition in the spirit of Lopes et al. (2013), it is possible to present this model as a set of time-varying regressions. We estimate more than 400 dynamic coefficients $\beta_{ij,t}$ and find a very sparse pattern as shown in

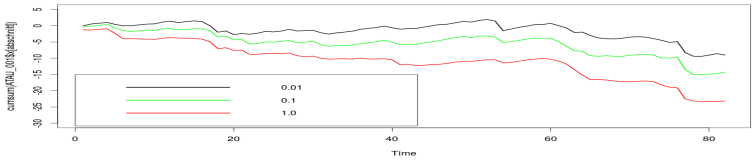


FIGURE 3. Predictive evaluation for various shrinkage priors over the last 80 months. Red: $a^\tau = a^\xi = 1.0$; green: $a^\tau = a^\xi = 0.1$, black: $a^\tau = a^\xi = 0.01$, $d_1 = d_2 = e_1 = e_2 = 0.001$.

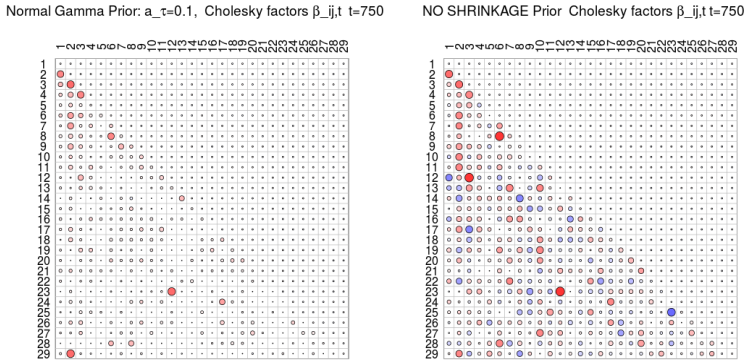


FIGURE 4. Comparing Cholesky factors $\beta_{ij,t}$ for $t = 750$. Left: Shrinkage via the normal-gamma prior, right: without a shrinkage prior setting.

Figure 4 for most of the time, here we show only one point in time: $t = 750$.

References

Belmonte, M.A.G., Koop, G., and Korobilis, D. (2014). Hierarchical shrinkage in time varying parameter models. *Journal of Forecasting*. **33**, 80 – 94.

Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partially non-Gaussian state space models. *Journal of Econometrics*, **154**, 85 – 100.

Griffin, J.E. and Brown, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 151 – 170.

Lopes, H.F., McCulloch, R.E, and Tsay, R.S. (2013). Cholesky Stochastic Volatility Models for High-Dimensional Time Series. *Working paper*.

Skew-normal controlled calibration model

Betsabé G. Blas[1]

[1] Universidade Federal de Pernambuco, Recife, PE, Brazil

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

Abstract: The so called normal linear calibration model (Usual-M) is used frequently in many areas. In analytical chemistry to determine the concentration of an element in samples, beforehand the analyst makes up a set of standards of know concentrations and assumes that the measurement error on this sample standards preparation process is negligible. Those measurement errors can lead to incorrect statistical inferences on the Usual-M. In the last two decades, the skew-normal distribution has been shown beneficial in dealing with asymmetric data in various theoretic and applied problems. In this work, we introduce a new calibration model considering skew-normal distributed error model, replicated measurement on the response variable and also assuming measurement errors on independent variable, which has the Usual-M as special case. Illustrative examples with chemical real data are reported.

Keywords: EM algorithm; Calibration model; Berkson variable.

1 Introduction

In analytical chemistry, a calibration model is a general method for determining concentration of a substance in a sample. The calibration models are defined on two stages, in the first stage, it was prepared standard solutions with known concentrations (say, X) chosen to cover a required range, and then it was measured the absorbance (say, Y) of each solution. In the second step, it was measured the absorbance (say, Y_0) of the unknown concentration (say, X_0) solution.

Linear calibration models with measurement errors can be found in Blas et al. (2007) and Blas and Sandoval (2010). In many applied problems, those calibration models may be not suitable when the data involves highly asymmetric observations. Figueiredo et al. (2010) studied the linear calibration model assuming that the errors have skew-normal distribution, as

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

one generalization of Figueiredo's work we propose the new linear calibration model by assuming measurement errors and skew-normal distribution for the error model. The rest of this extended abstract unfolds as follows. Section 2 briefly outlines some preliminaries of the skew-normal distribution. In Section 3 it is shown the hierarchical representation for the skew controlled calibration model (SCCM) by incorporating two latent variables. In Section 4 we apply to a real data set.

2 The skew-normal distribution

As developed by Azzalini (1985) a random variable Y follows a univariate skew-normal distribution with location parameter ξ, scale parameter σ^2 and skewness parameter $\lambda \in R$ if it has the density $\psi(y|\xi,\sigma^2,\lambda) = \frac{2}{\sigma}\phi\left(\frac{y-\xi}{\sigma}\right)\Phi\left(\frac{y-\xi}{\sigma}\right)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density function and cumulative distribution function, respectively; then, for brevity, we say that $Y \sim SN(\xi,\sigma^2,\lambda)$. Nota that if $\lambda = 0$, the density of Y reduces to the $N(\xi,\sigma^2)$ density.
Using the Henze (1986) reparametrization we write the variable $Y \sim SN(\xi, \sigma^2,\lambda)$ as $Y = \xi + \sigma\left(\frac{\lambda}{\sqrt{1+\lambda^2}}U_i + \frac{1}{\sqrt{1+\lambda^2}}V_i\right)$, with $U_i \sim HN(0,1)$ and $V_i \sim N(0,1)$. Hence, considering the reparametrization $\tau = \sigma\frac{\lambda}{\sqrt{1+\lambda^2}}$, $\eta = \sigma\frac{1}{\sqrt{1+\lambda^2}}$ we write $Y = \xi + \eta U_i + \tau V_i$.

3 New approach

The skew-normal controlled calibration model is defined by:

$$\begin{aligned} Y_{ij} &= \alpha + \beta x_i + \epsilon_{ij} \qquad j = 1,\ldots,m_i, \quad \text{and } i = 1,\ldots,n, && (1)\\ x_i &= X_i + u_i \qquad i = 1,\ldots,n, && (2)\\ Y_{0i} &= \alpha + \beta X_0 + \epsilon_{0i} \qquad i = 1,\ldots,r && (3)\end{aligned}$$

with the following assumptions on the random errors: $\epsilon_{ij},\epsilon_{0i} \overset{ind}{\sim} N(0,\sigma_\epsilon^2)$, $u_i \overset{ind}{\sim} SN(0,\sigma_u^2,\lambda)$, $\text{cov}(u_i,\epsilon_{ij}) = 0$ for all i,j and $\text{cov}(u_i,\epsilon_{0j}) = 0$ for all $j = 1,\ldots,m_i$ and $i = 1,\ldots,n$. The model parameters are $\alpha,\beta,X_0,\sigma_\epsilon^2$ and σ_u^2 and the main interest is the estimation of the unobserved quantity X_0. Some comments are in order here. The variable X_i in (1) is controlled, then $(\epsilon_{ij} + \beta u_i)$ is independent of X_i (Berkson-type errors).
Considering the reparametrization defined in the model (1) – (3) can be written as hierarchical form as

$$\begin{aligned} Y_{ij}|x_i &\sim N(\alpha+\beta x_i,\sigma_\epsilon^2), \qquad x_i|t_i \sim N(X_i+\triangle t_i,\Gamma), \qquad t_i \sim HN(0,1),\\ Y_{0j} &\sim N(\alpha+\beta X_0,\sigma_\epsilon^2), && (4)\end{aligned}$$

where $\triangle = \sigma_u\delta$, $\Gamma = \sigma_u^2(1-\delta^2)$ and $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$. An approximation for the variance of $\hat{X}_0$ is derived from the Fisher information matrix.

3.1 The EM algorithm

From the hierarchical model (4), it follows that the complete-data log-likelihood function, which is denoted by $l_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Y_0}, \mathbf{x}, \mathbf{t})$, with $\mathbf{t} = (t_1, \ldots, t_n)$, $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{Y} = (Y_1^\top, \ldots, Y_n^\top)^\top$, $\mathbf{Y}_0 = (Y_{01}, \ldots, Y_{0r})^\top$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})$, $i = 1, \ldots, n$.

For the current value $\widehat{\boldsymbol{\theta}}^{(k)} = (\widehat{\alpha}^{(k)}, \widehat{\beta}^{(k)}, \widehat{X_0}, \widehat{\sigma_\epsilon^2}^{(k)}, \widehat{\sigma_u^2}^{(k)})$ in the iteration k, the E-step of the EM-type algorithm requires the evaluation of $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = E[l_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Y}_0, \mathbf{x}, \mathbf{t})|\mathbf{Y}, \mathbf{Y}_0, \widehat{\boldsymbol{\theta}}^{(k)}]$, where the expectation is taken with respect to the joint conditional distribution of $\mathbf{x}$ given $\mathbf{Y}$ and $\mathbf{Y}_0$. Let the following quantities $\widehat{x_i}^{(k)} = E[x_i|\mathbf{Y}, \mathbf{Y}_0, \mathbf{t_i}, \widehat{\boldsymbol{\theta}}^{(k)}]$, $\widehat{x_i^2}^{(k)} = E[x_i^2|\mathbf{Y}, \mathbf{Y}_0, \mathbf{t_i}, \widehat{\boldsymbol{\theta}}^{(k)}]$, $\widehat{t_i}^{(k)} = E[t_i|\mathbf{Y}, \mathbf{Y}_0, \mathbf{x_i}, \widehat{\boldsymbol{\theta}}^{(k)}]$, $\widehat{t_i^2}^{(k)} = E[t_i^2|\mathbf{Y}, \mathbf{Y}_0, \mathbf{x_i}, \widehat{\boldsymbol{\theta}}^{(k)}]$, and $\widehat{t_i x_i}^{(k)} = E[t_i x_i|\mathbf{Y}, \mathbf{Y}_0, \widehat{\boldsymbol{\theta}}^{(k)}]$, then the EM algorithm is given by:

E-Step: Given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$, compute $\widehat{x_i}^{(k)}, \widehat{x_i^2}^{(k)}, \widehat{t_i}^{(k)}, \widehat{t_i^2}^{(k)}$ and $\widehat{t_i x_i}^{(k)}$ for $i = 1, \ldots, n$.

M-Step: Update $\widehat{\boldsymbol{\theta}}^{(k+1)}$ by maximizing $E[l_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Y}_0, \mathbf{x}, \mathbf{t})|\mathbf{Y}, \mathbf{Y}_0, \hat{\theta}^{(k)}]$ over $\theta^{(k)}$.

4 Application

An application is presented by using the data supplied by the chemical laboratory of the "Instituto de Pesquisas Tecnológicas (IPT)" - Brazil. The main interest is to estimate the unknown concentration value X_0 of the element barium.

Table 1 (a) presents the fixed values of concentration of the standard solutions and the corresponding observed intensities for the element barium, which are supplied by the plasma spectrometry method. This data set is referred to as the *first stage* of the calibration model. Table 1 (b) presents the observed intensities corresponding to the 3 sample solutions. This data set is referred to as the *second stage* of the calibration model. All estimates were computed considering the intensities divided by 10000 in order to achieve numerical stability.

Table 2 describes the ML estimates for $\alpha, \beta, X_0, \sigma_\epsilon^2, \sigma_\delta^2$ and the confidence interval amplitude $U(\hat{X}_0)$ for the normal controlled calibration model (N) and for the skew normal controlled calibration model (SN). The amplitude $U(\hat{X}_0)$ is given by the product of the squared root of the estimated variance of $\hat{X}_0$ and 1.96. We can observe that the estimated asymptotic standard errors from the new model are smaller than the values obtained from the usual model N.

TABLE 1. Left: Concentration (mg/g) and intensity of the standard solutions from barium element. Right: Intensity of the sample solutions of barium element.

X_i			
$X_1 = 0.1$	$X_2 = 0.2$	$X_3 = 0.5$	$X_4 = 1.06$
185082.2	373543.2	913829.3	1895804.6
184583.0	375166.9	894229.6	1926319.8
184906.3	369481.1	911759.2	1886632.8

Intensity (Y_0)
279034.2
279562.1
278462.2

TABLE 2. Parameter estimates of the proposed and usual models. Values in parenthesis are the estimated asymptotic standard errors.

distribution	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{X}_0$	$U(\hat{X}_0)$
SN	1.073	1.786e+02	1.502e-01	2.548e-05
	(0.184)	(0.518)	(1.3e-05)	-
N	1.138	1.786e+02	1.499e-01	7.031e-03
	(0.437)	(0.733)	(0.003)	-

5 Concluding remarks

In the application we can observe that estimated asymptotic standard errors from the new model is smaller than the usual model. In the error proposed model term, due to spectral interferences, stray light and slow response of the amplifier-recorder system, also due to the measurement errors from the standard sample preparation process seem to lead to a skew-normal error distribution. Thus, this new approach could fit better for chemical data.

References

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

Blas, B., Sandoval, M.C., and Satomi, O.Y. (2007). Homoscedastic controlled calibration model. *Journal of Chemometrics*, **21**, 145 – 155.

Blas B. Sandoval M.C. (2010). Heteroscedastic controlled calibration model applied to analytical chemistry. *Journal of Chemometrics*, **24**, 241 – 248.

Figueiredo C.C., Bolfarine H., Sandoval M.C., and Lima, C.R.O.P. (2010). On the skew-normal calibration model. *Journal of Applied Statistics*, **37**, 435 – 451.

Henze, N. (1986). A probabilistic representation of the skew- normal distribution. *Scandinavian Journal of Statistics*, **13**, 271 – 275.

Density estimation from uncertain observations – Comparing parametric and nonparametric methods

Marie Böhnstedt[1], Jutta Gampe[1]

[1] Max Planck Institute for Demographic Research, Rostock, Germany.

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

Abstract: Uncertainty of observations is modelled as a distribution over the potential values, and the underlying density of the (exact) data is to be estimated nonparametrically. We propose a smoothing technique based on a penalized likelihood and compare this approach to parametric procedures in a simulation study.

Keywords: Uncertain observations; Penalized likelihood; Smoothing; Generalized nonlinear models.

1 Introduction

When data are observed exactly, nonparametric estimation of the underlying density is an established alternative to parametric models. Several approaches to nonparametric density estimation are available, their properties have been thoroughly studied, and the additional flexibility mostly outweighs the loss in efficiency as compared to parametric methods.
If data are not observed exactly, but an uncertainty distribution over the potential values is given for each observation, the situation is less well studied. The aim of this work is to assess and compare the performance of a nonparametric and a corresponding parametric method for density estimation based on uncertain observations by means of a simulation study.

2 Model and methods

We denote by X the continuous random variable of interest with probability density function f and cumulative distribution function F. The support of the target density is partitioned into K (narrow) disjoint intervals I_k

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

of length Δ with midpoints u_k, $k = 1, \ldots, K$. This implies that $P(X \in I_k) = \int_{I_k} f(x)\,dx$ and we denote these probabilities by γ_k. To keep notation simple, we also denote this discretized version of the random variable by X and identify the interval I_k with the midpoint u_k, so that $P(X = u_k) = \gamma_k$. Estimation of the density f essentially is replaced by estimating the vector of probabilities $\gamma = (\gamma_1, \ldots, \gamma_K)^{\mathrm{T}}$. Instead of a sample of exact data $(x_1, \ldots, x_n)$ we only obtain uncertain observations. That is, for each observation i we have a distribution $g_i = (g_{i1}, \ldots, g_{iK})$ over the potential values u_k, with $g_{ik} \geq 0$ and $\sum_k g_{ik} = 1$. The properties of the individual g_i can vary across i but here we will consider the following scenario: The uncertainty distributions are centered around the true observation and have non-zero elements over an interval $[l_i, r_i]$ only. The interval ranges over $1 + 2\frac{\varepsilon}{\Delta}$ discretization intervals I_k, where the middle interval contains the true observation. The size of ε determines how wide the interval $[l_i, r_i]$ is; for $\varepsilon = 0$ we would obtain exact (discretized) observations. (Intervals at the boundaries are adapted so that they stay within the support of the distribution.) The probabilities g_{ik} are always such that the expected value of the uncertainty distribution is equal to the true value. The uncertainty distributions are collected in a $n \times K$ matrix $G = (g_{ik})_{i,k}$.

2.1 Nonparametric estimation

To estimate $\gamma = (\gamma_1, \ldots, \gamma_K)^{\mathrm{T}}$, and thereby the underlying density f, nonparametrically based on the uncertain data collected in the matrix G, we only assume that f (and consequently the elements of γ) is smooth. The element g_{ik} gives the contribution of observation i to the interval I_k. As proposed by Çetinyürek Yavuz and Lambert (2011), we can construct pseudocounts $y_k = \sum_{i=1}^{n} g_{ik}$ as column sums of G, which describe the contribution of the whole sample to the interval I_k, $k = 1, \ldots, K$. (Note that these pseudocounts in general are not integer numbers, but we could round y_k to the nearest integer value.) Then, $y = (y_1, \ldots, y_K)^{\mathrm{T}}$ has a multinomial distribution with probability vector γ. Equivalently, we can view the y_k as independently Poisson distributed with expectation $\mu_k = n\gamma_k$. We apply the discrete penalized likelihood smoothing introduced by Eilers and Borgdorff (2007) to derive a density estimate from the pseudocounts. That is, we maximize the penalized log-likelihood $l_p = \sum_{k=1}^{K}(y_k\eta_k - \mu_k) - \frac{\lambda}{2}\eta^{\mathrm{T}}D^{\mathrm{T}}D\eta$, where $\eta_k = \log \mu_k$ and $D\eta$ are second or third order differences of the elements of $\eta = (\eta_1, \ldots, \eta_K)^{\mathrm{T}}$. The smoothing parameter λ is chosen using AIC or BIC (Çetinyürek Yavuz and Lambert, 2011).

2.2 Parametric estimation

To estimate the density f in a parametric model, that is, under the assumption that it is known up to a finite-dimensional parameter vector θ, we modify the estimating strategy as follows. Again we use pseudocounts as

in Section 2.1, however, we replace the vector γ by an appropriate function of the parameters θ. The specific form of $\gamma(\theta)$ depends on the distribution that is to be estimated in the following way: The expected values μ_k of the pseudocounts y_k are given by $\mu_k = n\gamma_k = n\int_{I_k} f(x;\theta)\,dx$. If the length Δ of the discretization intervals I_k is sufficiently small, μ_k can be approximated by $n\Delta f(u_k;\theta)$. This parameterization of μ_k enables us to estimate θ by fitting a corresponding Poisson model with a predictor that is nonlinear in θ. The `gnm`-package in R offers algorithms to estimate such generalized nonlinear models (Turner and Firth, 2012).

3 Simulation study

A simulation study was conducted in order to examine and compare the two approaches. We generated samples of size $n = 100$ and $n = 1000$ according to two target distributions: a unimodal Gompertz distribution and a bimodal mixture of two Gompertz distributions. The target interval $[0, 110]$ was subdivided into $K = 110$ intervals of length $\Delta = 1$.
Several uncertainty distributions g were implemented by varying the interval width ε and the shape of g (truncated normal or uniform). Each setting was replicated 100 times. We assessed the quality of the estimates graphically and by a discrete approximation to the integrated squared error, $\mathrm{ISE}(\hat{f}) \approx \sum_{k=1}^{K} \Delta(\hat{f}(u_k) - f(u_k))^2$.

4 Discussion and outlook

Differences between the two approaches become noticeable only when the width of the uncertainty intervals is considerable. Figure 1 illustrates the results for $n = 100$ and interval width $\varepsilon = 8$. For the unimodal Gompertz distribution both methods are fairly equal, while the parametric approach naturally is advantaged in the bimodal case (if the model is specified correctly), if only slightly. The setting will be extended further to uncertainties that show some systematic pattern over the range of X.

References

Çetinyürek Yavuz, A. and Lambert, P. (2011). Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. *Statistics in Medicine*, **30**, 75 – 90.

Eilers, P.H.C. and Borgdorff, M.W. (2007). Non-parametric log-concave mixtures. *Computational Statistics & Data Analysis*, **51**, 5444 – 5451.

Turner, H. and Firth, D. (2012). Generalized nonlinear models in R: An overview of the gnm package. (R package version 1.0-7). (`http://CRAN.R-project.org/package=gnm`).

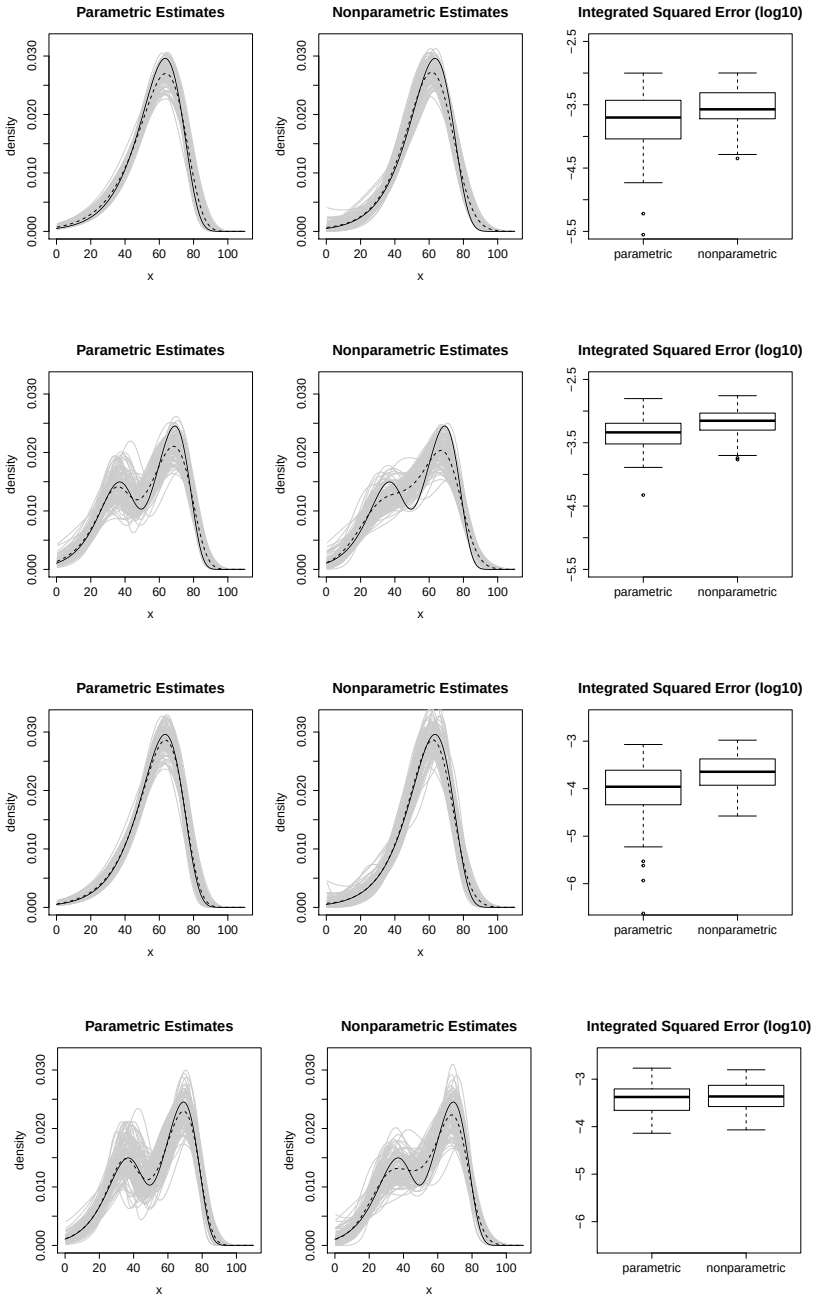


FIGURE 1. Uncertainty distribution with $\varepsilon = 8$. Top: uniform, bottom: truncated normal. Nonparametric estimates based on AIC and penalty order 3. Sample size $n = 100$. Solid black line: true distribution, dashed black line: mean of 100 replications.

Evaluation of a new k-means approach for exploratory clustering of items

Stella Bollmann[1], Andreas Hölzl[2], Helmut Küchenhoff[2], Moritz Heene[1], Markus Bühner[1]

[1] Department of Psychology, Ludwig-Maximilians University Munich, Germany
[2] Department of Statistics, Ludwig-Maximilians University Munich, Germany

E-mail for correspondence: `stella.bollmann@psy.lmu.de`

Abstract: In order to group variables of survey data into homogeneous subgroups, usually the exploratory factor analysis is used. This method is based on the common factor model. In the present study an alternative approach is suggested that does not make assumptions about an underlying model. We propose two new k-means approaches: *k-means scaled distance measure* where items are represented in a coordinate system in a way so that their distance is based on one minus their correlation; and *k-means cor* where items inter-correlations are directly taken as the coordinate points of the items. The comparative performance of both methods is tested with real data and simulated data. Advantages of these new methods are discussed.

Keywords: K-means; Clustering of items; Exploratory structure detection.

1 Introduction

The most frequently used method for exploratory structure detection in psychometric data is the exploratory factor analysis (EFA). The aim of EFA is the detection of the "'true"' factor model,

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{U}^2,$$

where $\mathbf{\Sigma}$ is the $p \times p$ correlation matrix of the p observed variables, $\mathbf{U}^2$ is the diagonal $p \times p$ matrix of the unique variances, $\mathbf{\Lambda}$ is the $p \times m$ matrix of the factor loadings on the $m < p$ factors, and $\mathbf{\Phi}$ stands for the $m \times m$ matrix of the factor inter-correlations.
However, EFA is known to exhibit some problems. The major mathematical drawback is the factor indeterminacy. Further problems are for example its

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

weak performance in small sample sizes ($n \leq 150$) and with high cross-loadings (e.g. Sass, 2010; Velicer and Fava, 1998), as well as the general issue of the underlying measurement model including uncorrelated residual variances, what may be difficult to justify (MacCallum and Tucker, 1991). We suggest two new k-means approaches as an alternative.

2 Non-hierarchical clustering of items

2.1 K-means clustering of items

The idea of k-means clustering of items is that the items are points in a coordinate system and the square distances between each item of a cluster and its centre are minimized iteratively. Therefore, a way has to be found in which the items can be represented in a coordinate system. Both of the methods suggested here are based on correlations.
In the first approach, *k-means scaled distance measure (sdm)*, the idea is to create coordinate points in a way so that the distance between these points are directly based on correlations. Therefore, the distance between the two variables X_1 and X_2 is described as

$$d_{sdm}(X_1, X_2) = \sqrt{0.5 - 0.5 \cdot \mathrm{Cor}(X_1, X_2)}$$

To represent the items in a coordinate system in a way so that the distance between two items is $d_{sdm}(X_1, X_2)$, a scaling procedure is used that is based on the idea of multidimensional scaling.
The idea of the second approach is that the correlations between the items are directly used as the coordinates of the items. The vector of correlations of one item to all the other items is represented as the coordinates of this item. The distance between two items is then

$$d_{cor}(X_1, X_2) = \sum_i \big(\mathrm{Cor}(X_1, X_i) - \mathrm{Cor}(X_2, X_i)\big)^2$$

The item assignment to clusters has to be preceded by a dimensionality assessment in which the best-fitting number of clusters is determined.

2.2 Silhouette width for assessment of dimensionality

The silhouette width (Brock et al., 2008) is based on each item's silhouette value, which is defined in the following manner:

$$S(j) = \frac{b_j - a_j}{\max(b_j, a_j)},$$

where a_j is the average distance between j and all the other observations in the same cluster, and b_j is the average distance between j and all the observations in the *nearest neighboring cluster*.
These values are then all summed up across items in order to get the silhouette width. The number of clusters is selected for which the silhouette width is highest.

3 Comaparison of methods

3.1 Design of the simulation study

The two new k-means approaches are compared to EFA and existing cluster analysis approaches: ClustOfVar and two hierarchical clustering methods, CAAL and CACL. Three different types of simulations are used for comparison of the accuracy of the used methods. First, we use the Real World simulation, a new approach using real data. Second, a traditional Monte-Carlo simulation is conducted. And third, we make a cross validation with confirmatory factor analysis (CFA).

The Real World simulation is a resampling technique in which the results of each method in a real, huge data set is regarded as the population model. Samples of different sizes are then drawn with replacement from this data set and the results of each method in the sub samples are compared to the population model. The two data sets we use are the NEO-PI-R, a widely used big-5 personality inventory ($n = 11724$) and the IST-2000-R, that measures intelligence in three major domain ($n = 1352$).

For dimensionality assessment, success rates are reported for each sample size, i.e. the percentage of identical number of factors as in the population data. And for item assignment, we set the number of factors to the theoretically assumed number. Similarity is then determined using the Rand Index (Rand, 1971), calculated by counting the number of correctly classified pairs of elements.

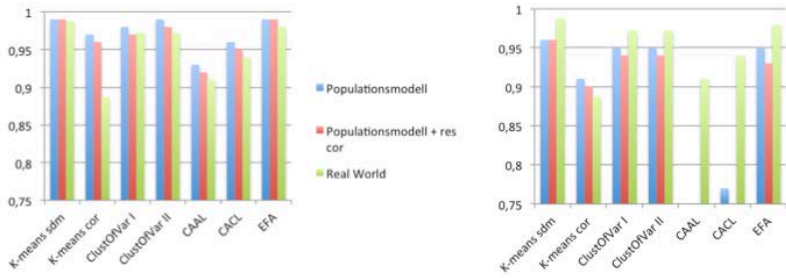
In the traditional Monte-Carlo simulation, we specify the factor model, the EFA has found in each of the population data sets and draw samples of different sizes. In the first simulation condition, all parameters are taken from the population model and in the second simulation condition, residual correlations are set to zero.

For the CFA cross-validation, we specify factor models with combinations of different dimensionality assessment methods and different item assignment methods on sub-samples of the data set. We subsequently test the specified model on the entire data set with CFA, reporting BIC values.

3.2 Results

In the Real World simulation the dimensionality assessment methods with the highest success rates are in both data sets the EFA methods and k-means cor with silhouette. The hierarchical CA methods with silhouette had the lowest success rates across data sets. For item assignment, the new k-means sdm shows the highest Rand Index of all methods in both data sets (see Figure 1). Especially for smaller sample sizes it outperforms EFA.

In the traditional Monte-Carlo simulation, when specifying the population model without residual correlations, EFA dimensionality assessment methods were the ones to achieve the highest success rates. When adding



(a) *NEO-PI-R* (b) *IST-2000-R*

FIGURE 1. Item assignment: Rand Indexes for Real World simulation and traditional simulation.

residual correlations though, their performance dropped considerably while other methods like k-means cor performed better. For item assignment in the traditional simulation, the highest Rand Indexes were achieved by k-means sdm followed by EFA.
In the CFA cross-validation, the combination of k-means sdm and EFA Parallel Analysis obtained the lowest BIC's in both data sets.

4 Discussion

The above results show that the main advantage of the new approaches are (a) that cluster scores are determinate and (b) for item assignment k-means sdm obtains better results than EFA and other cluster analysis approaches. We therefor suggest to use a combination of EFA Parallel Analysis methods for dimensionality assessment and k-means for item assignment.

References

Brock, G., Pihur, V., Datta, S., and Datta, S. (2008). clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, **25**.

MacCallum, R. and Tucker, L.R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, **109**, 502 – 511.

Sass, D.A. (2010). Factor loading estimation error and stability using exploratory factor analysis. *Educational and Psychological Measurement*, **70**, 557 – 577.

Velicer, W. and Fava, J.L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, **3**, 231.

Using generalized estimating equations to study persistence of antimicrobial resistance in respiratory streptococci

Robin Bruyndonckx[12], Niel Hens[13], Marc Aerts[1], Katrien Latour[4], Boudewijn Catry[4], Samuel Coenen[25]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BIOSTAT), Hasselt University, Hasselt, Belgium
[2] Laboratory of Medical Microbiology, Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium
[3] Centre for Health Economic Research and Modelling Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium
[4] Scientific Institute of Public Health, Brussels, Belgium
[5] Centre for General Practice, Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium

E-mail for correspondence: robin.bruyndonckx@uhasselt.be

Abstract: The use and misuse of antibiotics has over time lead to resistance of bacteria to several antibiotics. The aim of this study is to compare persistence of resistance to penicillin and erythromycin in respiratory streptococci. Based on a forward selection procedure for a multiple logistic regression model, a generalized estimating equations model was constructed. Because we assume that there is some residual resistance in the population, the link function was adjusted. Using this model we have shown that the evolution of resistance to penicillin and erythromycin over time does not differ significantly and that parameter estimates are influenced greatly by accounting for residual resistance in the population.

Keywords: Generalized estimating equations; Persistence of resistance.

1 Introduction

Antibiotic resistance poses a substantial threat to public health as it is related to treatment failure and increased mortality (French, 2005). One part of the solution to this problem is to study the relationship between antibiotic use and resistance in order to develop targeted interventions.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

The impact of antibiotic use on resistance in oropharyngeal streptococci has been assessed in two randomized placebo-controlled trials (RCTs) (Malhotra-Kumar et al., 2007; Chung et al., 2007). These studies have shown that persistence of resistance after exposure to macrolides lasts for more than six months, while it is estimated to be much shorter after exposure to amoxicillins. The objective of this paper is to compare the evolution of resistance to penicillin and erythromycin in respiratory streptococci based on routinely collected data with RCT results.

2 Methodology

Information on resistance of respiratory streptococcus isolates to penicillin or erythromycin (data for 2005) was coupled with information on oral consumption of penicillins and cephalosporins (CD) or macrolides and tetracyclines (AF), respectively (reimbursement data for July 2004 – December 2005) (Catry et al., 2008). For each isolate antibiotic use prior to sampling was considered and isolates for which time between use and sampling was at least four days were selected. The data contain information on the status of resistance for 358 isolates in 339 patients. Explanatory variables that were considered in the statistical analyses are treatment (CD or AF), bacteria (*S. pyogenes* (PY) or *S. pneumoniae* (PN)) and log(time).
A multiple logistic regression model using a logit link and the status of resistance as the binary response was constructed. A forward selection procedure was used to reach a model that includes all significant exploratory variables and two-way interactions. Included variables were treatment, bacteria and log(time). Since we were mainly interested in the difference in the evolution of resistance between the two treatment groups, the interaction between log(time) and treatment was added to this model. A generalized estimating equations (GEE) model including these four variables and an independent working correlation was constructed in order to account for the dependency between isolates from the same patient.
Both models implicitly assume that the resistance rate falls back to 0% when there is enough time between use and sampling. To relax this assumption, we acknowledged the residual resistance in the population (baseline resistance (BR)) and adjusted the link function accordingly:

$$\log\left(\frac{p}{1-p}\right) \quad \rightarrow \quad \log\left(\frac{p}{g-p}\right) \quad \text{with } g = 1 - BR.$$

3 Results

We plotted the evolution of BR over time, where BR was calculated as the percentage of resistant isolates with time between use and sampling between $t - 186$ and t (with $t = 186, \ldots, 372$) (Figure 1). Because this

estimate was rather unstable, we conducted a sensitivity analysis in which we fitted the GEE model using different BR estimates (for $t = 248$ (BR $3-8$ months), $t = 310$ (BR $5-10$ months) and $t = 372$ (BR $7-12$ months)) (Table 1).

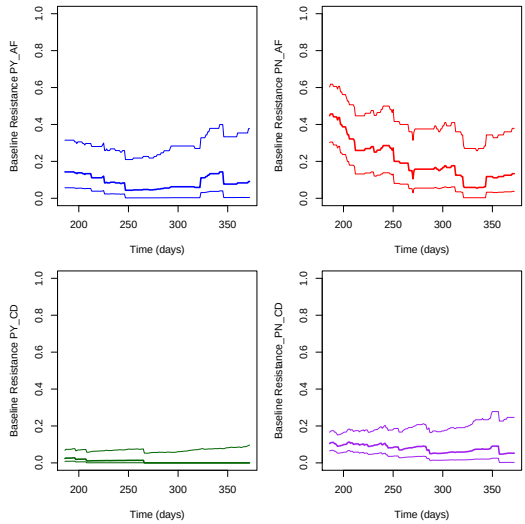


FIGURE 1. Evolution of baseline resistance over time for *S. pyogenes* (left) and *S. pneumoniae* isolates (right) after treatment with macrolides and tetracyclines (top) or penicillins and cephalosporins (bottom).

TABLE 1. Estimates (95% confidence intervals) of baseline resistance (BR) for *S. pyogenes* (PY) and *S. pneumoniae* (PN) isolates after treatment with macrolides and tetracyclines (AF) or penicillins and cephalosporins (CD).

	BR $3-8$ months	BR $5-10$ months	BR $7-12$ months
PY AF	0.044 (0.002; 0.210)	0.063 (0.003; 0.283)	0.091 (0.005; 0.377)
PY CD	0.014 (0.001; 0.073)	0.000 (0.000; 0.065)	0.000 (0.000; 0.096)
PN AF	0.273 (0.132; 0.482)	0.177 (0.062; 0.410)	0.133 (0.037; 0.379)
PN CD	0.082 (0.036; 0.178)	0.059 (0.016; 0.191)	0.053 (0.003; 0.246)

When comparing the four GEE models, it can be seen that parameter estimates altered by accouting for different BR estimates (Table 2). The interaction between log(time) and treatment was not significant for all four models. Further reduction of the models would result in a model including treatment, log(time) and bacteria (for BR $= 0$ and $7-12$ months) and a model including treatment and log(time) (for BR $3-8$ and $5-10$ months).

TABLE 2. Parameter estimates for the adjusted GEE model when using different values for baseline resistance (BR).

Parameter	BR = 0	BR 3 − 8 months	BR 5 − 10 months	BR 7 − 12 months
Intercept	-0.1011	-9.8544	-0.7523	-0.7255
Treat AF	-2.4276	6.5846	-2.4914	-2.4976
Bacteria PY	1.5141**	1.2608	1.1099	1.3522*
Log(time)	0.5970**	5.3333	0.9936	0.9408*
Log(time)*treat AF	0.0974	-4.1991	0.0836	0.0901

*: p-value < 0.05; **: p-value < 0.01.

4 Conclusions

Compared to the RCT results, in this study no significant difference in the evolution of resistance after treatment with penicillins or cephalosporins and macrolides or tetracyclines was found. Our recommendation for future studies on persistence of resistance is to acknowledge that baseline resistance might not be zero and to adjust the link function accordingly.

References

Catry, B., Hendrickx, E., Prael, R., and Mertens, R. (2008). Eindrapport IMA-BAPCOC-WIV: Verband tussen antibiotica consumptie en microbiële resistentie bij de individuele patiënt.

Chung, A., Perera, R., Brueggemann, A.B., Elamin, A.E., Harnden, A., Mayon-White, R., Smith, S., Crook, D., and Mant, D. (2007). Effect of antibiotic prescribing on antibiotic resistance in individual children in primary care: prospective cohort study. *British Medical Journal*, **335**, 429.

French, G.L. (2005). Clinical impact and relevance of antibiotic resistance. *Advanced Drug Delivery Reviews*, **57**, 1514 – 27.

Malhotra-Kumar, S., Lammens, C., Coenen, S., Van Herck, K., and Goossens, H. (2007). Impact of azithromycin and clarithromycin therapy on pharyngeal carriage of macrolide-resistant streptococci among healthy volunteers: a randomised, double-blind, placebo-controlled study. *Lancet*, **369**, 482 – 90.

Using spatial and land use regression models in investigating the modifying effect of socioeconomic status on the interaction between traffic, air pollution and asthma

S. Cakmak[1], C. Hebbern[1], J. Vanos[2], D. Crouse[1], C. Blanco[3]

[1] Population Studies Division, Health Canada, Canada
[2] University of Texas, USA
[3] University of San Sebastian, Chile

E-mail for correspondence: sabit.cakmak@hc-sc.gc.ca

Abstract: Urban air pollution exposure is associated with increased respiratory health effects. Traffic-related air pollution specifically has been linked to adverse respiratory health outcomes, and living near major roadways is associated with increased respiratory illness. The city of Windsor, Ontario, is located on the USA-Canada border and affected by trans-boundary air quality issues with high density traffic burdens from trucks, cars, and commercial vehicles. A land use regression (LUR) study (Wheeler et al., 2008) to predict seasonal multiple-source pollutant concentrations of NO2, SO2 and volatile organic compounds indicated that concentrations increased in the city with proximity to the international border, with strong inter-pollutant correlations. Air pollution exposure can also interact with socio-economic factors. Living in communities with lower household income and education levels has been shown to be associated with increased vulnerability to air pollution (Cakmak et al., 2006). Living near major roadways is related to increases in respiratory illness and asthma in children (Dockery DW, 2002). However, none of the previous studies linking ambient air pollution to the prevalence of asthma considered spatial clustering. There could be two types of clustering: first, clustering within neighbourhood. Clustering will induce a positive correlation of the response of subjects in the same location and thus suggest that there are inadequately modeled risk factors specific to the location itself. Failure to account for this clustering issue can lead to underestimation of the uncertainties associated with effect estimates. Second, health responses of subjects living in closer neighbourhoods may be more similar compared to health responses of subjects living in neighbourhoods further apart. Failure to account for this type of clustering can also lead to underestimation of the uncertainties associated with effect estimates. In this study, we present a spatial regression

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

model linking spatial variation in ambient air pollution to respiratory health. Our methods are illustrated with an analysis of respiratory health of elementary school children in Windsor, Canada to determine whether indicators of social status, such as family income and education, modify the respiratory health effects of gaseous and particulate air pollution, as well as the effect of roadway or traffic density on children's respiratory health.

Keywords: Spatial model; GLM; Land use regression; Population health.

1 Introduction

The study included children with and without asthma in grades 4 to 6 in the Windsor public school system. An estimated 7,200 children were approached for inclusion in 2005, with 2328 participating. Family socio-economic status and medical history information were collected from the Windsor Children's Respiratory Health Study questionnaires. The volume and type of traffic were collected for roadways in the vicinity of the subjects home by a trained observer using an electronic counter within the city of Windsor's Public Works Department and Geomatics Division. The Simple Traffic Count (STC) is the total volume of cars and trucks on each road segment, expressed as a daily average. It is measured for approximately one week by an automated counter and data are available for all arterial roads, and 75 percent of collector roads. The Turning Movement Count (TMC) is the volume of traffic on a roadway segment, determined by the volume and direction of traffic entering or leaving the segment of roadway at adjacent intersections. The distance of the child's home to various types of roadways is determined by creating a 200 m radius around each child's neighbourhood, centered on the postal code. This radius was chosen as traffic-related pollutants (e.g., nitrogen oxides, carbon monoxide, volatile organic compounds) peak close to roadways and fall to background levels approximately 200 m from the pollutant source. Exposure to traffic-related air pollution is calculated as the sum of the traffic counts on all roadways within this boundary. Yearly city wide levels of air pollution were estimated by averaging measurements from two fixed monitors within the city for hourly fine particulate matter (PM2.5), nitrogen dioxide (NO2, ppb), and sulphur dioxide (SO2, ppb), obtained from Environment Canadas National Air Pollution Monitoring System (NAPS) resolved to the participants neighborhood using a land use regression model. Subjects were categorised into three income levels based on questionnaire responses: $< \$35000$, $\$35000 - 80000$, and $> \$80000$, and into three education levels: less than high school, high school or community college, and university or higher. We then applied a statistical model for linking spatial variation in ambient air pollution to prevalence of asthma. Our methods are illustrated with an analysis of respiratory health of the elementary school children in Windsor, Canada to determine whether indicators of social status, such as

income and education, modify the respiratory health effects of gaseous and particulate air pollution, as well as the effect of roadway or traffic density on children's respiratory health.

2 Statistical model

The model is formulated in two stages. In "Stage One", health outcome data is modeled by covariates at the individual level and indicator functions for each community. Community-level covariates, such as air pollution, are not included at this stage. Three basic regression models are considered here: linear, logistic, and time to event. Estimates of the community-specific health responses are determined using standard computer software for linear, logistic and Cox proportional hazard survival models. We used SAS procedures because they can accommodate very large sample sizes (SAS, 1997). For linear and logistic regression models, we do not specify an intercept term. The estimates of the indicator functions can be interpreted as the average response in a specific community after adjusting for individual-level covariates. If these individual-level covariates are deviated from their mean value, then the community-specific outcomes can be interpreted as the predicted health response for a subject whose risk factors are all evaluated at the average for all respondents. Output from stage one is the community-specific adjusted health responses denoted by $\{\hat{\delta}(s),\, s = 1, \ldots, S\}$, where s denotes a zero dimensional point in Cartesian (x, y) space representing the location of one of communities under study. Additional output from this stage is the variance-covariance matrix of the $\hat{\delta}(s)$, denoted by $\hat{v}$, which describes the uncertainty in the estimates of the community-specific adjusted health response.

In "Stage Two", estimates of community health responses are related to risk factors defined at the community level using a linear random effects regression model as follows:

$$\hat{\delta}(s) = \zeta(s) + \beta^{\mathrm{T}} Z(s) + \eta(s) + \epsilon(s),$$

where $\epsilon(s)$ is a random process with zero expectation, variance-covariance matrix $\hat{v}$, independent of the spatial random effects process $\eta(s)$. $\zeta(s)$ is the two-dimensional trend term to account for residual spatial variability, and β is a vector of unknown regression coefficients linking the vector of spatial level risk factors, $Z(s)$, to the community-specific health responses. We assume that the spatial process $\eta(s)$ is stationary (i.e., expectation does not vary in space), has zero expectation, variance $\Theta > 0$, and correlation matrix Ω with dimension S. The correlation of the random effects between two areas can be modeled by their distance apart or some other characteristic of their locations. Here, $\hat{\delta}(S)$ has expectation $\mu(s) = \zeta(s) + \beta^{\mathrm{T}} Z(s)$ and variance-covariance matrix $\Sigma = \Theta\Omega + \hat{V}$. We consider non-parametric smoothed estimates of $\zeta(s)$ using the robust locally-weighted regression

(LOESS)(Green and Silverman, 1994) smoothers within the generalized additive model (GAM) framework (Hastie and Tibshirani, 1990). This method is implemented in the statistical computing software package S-Plus(6.2). The unknown parameter vector β is also estimated using GAM in S-plus.

3 Results and discussion

With regards to clustering within neighbourhoods, while our model gave similar air pollution asthma prevalence estimates as the standard logistic model, the standard errors of these estimates were somewhat higher than those from the standard logistic model. With regard to between neighbourhoods, we have observed a pattern of spatial autocorrelation in asthma prevalence that cannot be fully explained by ambient air pollution concentrations, even after controlling for a host of risk factors measured at the individual level. We also found that the association between air pollution and prevalence of asthma was somewhat sensitive to the specification of the complexity of the spatial surface, with more complex surface specifications resulting in lower estimates of the air pollution effect and higher standard errors. We have also found that asthma risk is higher in the lower education group compared to the university educated. These results suggest that there may be some confounding due to missing or systematically mis-measured risk factors that are also spatially correlated with pollution. We demonstrate that the spatial auto-correlation in community health outcomes can be accounted for through the inclusion of location in the deterministic component of the model assessing the effects of air pollution on prevalence of asthma.

References

Cakmak, S., Dales, R.E., and Judek, S. (2006). Respiratory health effects of air pollution gases: modification by education and income. *Arch. Environ. Occup. Health*, **61**, 5 – 10.

Dockery DW, P.C (2002). Outdoor particulates. In: *Environmental Epidemiology*, New York:Oxford University Press, 119 – 166.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Wheeler, A.J., Smith-Doiron, M., Xu, X., Gilbert, N.L., and Brook, J.R. (2008). Intra-urban variability of air pollution in Windsor, Ontario – Measurement and modeling for human exposure assessment. *Environmental Research*, **106**, 7 – 16.

Beta calibration model with measurement errors

Mileno Cavalcante[1], Betsabé G. Blas[2]

[1] Petrobras S.A., Rio de Janeiro, RJ, Brazil
[2] Universidade Federal de Pernambuco, Recife, PE, Brazil

E-mail for correspondence: milenoc@yahoo.com, betsabe.bg@gmail.com

Abstract: In this work we extend the beta calibration model proposed by Cavalcante and Blas (2014). This extension is suitable for a beta-distributed variable which depends on an explanatory variable that is subject to Berkson-type measurement errors in the first stage. An application is also presented.

Keywords: Calibration model; Beta regression; Berkson model.

1 Introduction

Measurement errors problem in calibration models can be of considerable interest for researchers whose main field of study covers regression models. The model we present in this work is an extension of the beta calibration model (BCM) proposed by Cavalcante and Blas (2014). Similarly to the original model, it is also suitable for a nonlinear response variable which lies in the open interval $(0, 1)$, as it is the case for beta-distributed variables. However, in the present case, x is an unobservable random variable, but a pre-fixed value of its surrogate is available. The relationship between x and its surrogate (W) is modelled through the Berkson model, which means that $x = W - \delta$, where δ is the measurement error variable such that $\delta \sim N(0, \sigma_\delta^2)$. In this case, we say that W is a *controlled* variable. Estimation of the parameters of the proposed model can be carried out considering two cases for the measurement error variance: known and unknown σ_δ^2.

2 Beta calibration model (BCM)

Cavalcante and Blas (2014) suppose for the BCM that the response variable y is beta-distributed, $y \in (0, 1)$, with $p > 0$ and $q > 0$ (i.e. $y \sim \mathcal{B}(p, q)$). A

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

regression structure for y is obtained following a reparameterization for its probability density function (pdf). Then, if $y \sim \mathcal{B}(p, q)$, with $p > 0$, $q > 0$, for the first stage (calibration stage), and $y_0 \sim \mathcal{B}(p_0, q_0)$, with $p_0 > 0$, $q_0 > 0$, $p \neq p_0$ and/or $q \neq q_0$ for the second stage, the reparameterization is: $\mu = \frac{p}{p+q}$, $\phi = p + q$, $\mu_0 = \frac{p_0}{p_0+q_0}$, and $\phi_0 = p_0 + q_0$.
Now, let $y_i \sim \mathcal{B}(\mu_i, \phi)$, with $i = 1, \ldots, n$, and $y_{i0} \sim \mathcal{B}(\mu_0, \phi_0)$, with $i = n+1, \ldots, n+k$, be a random sample. For a logit link function $g : (0, 1) \to \mathbb{R}$, strictly monotonic, and twice differentiable, with $g(\mu_i) = \eta_i = \alpha + \beta x_i$ and $g(\mu_0) = \eta_0 = \alpha + \beta x_0$, we get $\mu_i = \mathbb{E}(y_i) = h(\alpha + \beta x_i)$, $\mu_0 = \mathbb{E}(y_{i0}) = h(\alpha + \beta x_0)$, $Var(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi}$, and $Var(y_{i0}) = \frac{\mu_0(1-\mu_0)}{1+\phi_0}$.
Cavalcante and Blas (2014) obtained the maximum likelihood estimates (MLEs) for the regression parameters $(\alpha, \beta, x_0, \phi, \phi_0)$ by maximizing the joint likelihood function of (y_i, y_{i0}), with $\psi(x) = \frac{d \log(\Gamma(x))}{dx}$, $x > 0$, and $\Gamma(\cdot)$ as a gamma function. Since the ML estimators for these parameters do not have closed form, they used numerical methods to obtain them (e.g. L-BFGS-B). For estimation and simulation purposes, the Fisher's information matrix can be approximated by numerical methods.

3 Beta controlled calibration model (BCCM)

In real data applications, there are situations where it is impossible to observe the values of x_i directly, but its surrogate values (W_i) are available. If we assume that W_i has a pre-fixed value (i.e. it is a *controlled* variable), we can write $x_i = W_i - \delta_i$, where x_i is unobservable and δ_i is a random variable such that $\mathbb{E}(\delta_i) = 0$, and $\mathbb{E}(\delta_i, \delta_j) = 0$, $\forall i \neq j$. In this case, the measurement error process follows a Berkson model.
Building on the BCM assumptions, we suppose, for the BCCM's first stage, that the measurement error δ_i is non-differential, that is $v(y_i|x_i, W_i) = v(y_i|x_i)$ (Roy et al., 2005). Also, for the measurement error process, if $\delta_i \sim N(0, \sigma_\delta^2)$ and $\mathbb{E}(\delta_i, \delta_j) = 0$, $\forall i \neq j$,

$$x_i|W_i \sim N(\mu_{x|W}, \sigma^2_{x|W}) \tag{1}$$

with $\mu_{x|W} = \mathbb{E}(x|W) = W$, $\sigma^2_{x|W} = Var(\delta) = \sigma^2_\delta$, since $Cov(\delta_i, W_i) = 0$.
For $y \in (0, 1)$ with a non-differential δ_i (Roy et al., 2005, p. 271), we get

$$\mathbb{P}(0 < y < 1|x_i, W_i) = \int_{x|y\in(0,1)} h(\alpha^* + \beta^* x_i) s(x_i|W_i) dx_i \tag{2}$$

where $s(x_i|W_i)$ is the distribution in (1) and $h(\cdot)$ is a link function.
For a logit link function, the integral in (2) does not have a closed form, but it can be approximated (Monahan and Stefanski, 1992). Then,

$$h\left(\frac{\mathbf{B}^{\mathrm{T}}\mathbf{W}_i}{\sqrt{1 + \frac{\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B}}{k^2}}}\right), \tag{3}$$

where $\mathbf{B}^{\mathrm{T}} = [\alpha^*, \beta^*]$, $\mathbf{W_i^{\mathrm{T}}} = [1, W_i]$, $\mathbf{\Sigma} = \begin{bmatrix} Var(1|W_i) & 0 \\ 0 & Var(x_i|W_i) \end{bmatrix}$, and $h(u) = \frac{exp(u)}{1+exp(u)}$. Roy et al. (2005) assume that $k^2 = 1.70$.
Since $Var(1|W_i) = 0$, for the calibration stage, we have $\mu_i^* = \mathbb{E}(y_i) = h(\alpha^* + \beta^* W_i)$ and $Var(y_i) = \frac{\mu_i^*(1-\mu_i^*)}{1+\phi^*}$. Analogously, for the second stage, $\mu_0^* = \mathbb{E}(y_{i0}) = h(\alpha^* + \beta^* x_0)$ and $Var(y_{i0}) = \frac{\mu_0^*(1-\mu_0^*)}{1+\phi_0^*}$. Note that $\alpha^* = \frac{\alpha}{b}$ and $\beta^* = \frac{\beta}{b}$ are the same for both stages, where $b = \sqrt{1 - \frac{\beta^2 Var(x_i|W_i)}{k^2}}$, and α and β as the parameters for the naive model (BCM). Here, we follow Blas, et al. (2007) and assume that there is no measurement error related to the parameter x_0. It is easy to see that, when $\sigma_\delta^2 = 0$, this model becomes the BCM, with $x_i = W_i$. The joint likelihood function for the BCCM is

$$\ell(\mathbf{y}, \mathbf{y}_0, \mu^*, \mu_0^*, \phi^*, \phi_0^*) = \sum_{i=1}^{n} \ell_{i1}(\mu_i^*, \phi^*) + \sum_{i=n+1}^{n+k} \ell_{i0}(\mu_0^*, \phi_0^*), \qquad (4)$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{y_0} = (y_{(n+1)0}, \ldots, y_{(n+k)0})^{\mathrm{T}}$, $\mu^* = (\mu_1^*, \ldots, \mu_n^*)^{\mathrm{T}}$, $\ell_{i1}(\mu_i^*, \phi^*) = \log f_1(y_i|\mu_i^*, \phi^*)$, $\ell_{i0}(\mu_0^*, \phi_0^*) = \log f_0(y_{i0}|\mu_0^*, \phi_0^*)$, and $f_1(\cdot)$ and $f_0(\cdot)$ as beta pdfs for the first and second stages, respectively.
In order to obtain the MLEs for $(\alpha^*, \beta^*, x_0, \phi^*, \phi_0^*)$, we consider two cases: (i) known $Var(x_i|W_i) = \sigma_\delta^2$; and (ii) unknown $Var(x_i|W_i) = \sigma_\delta^2$. In case (i), the MLEs for the parameters of interest can be calculated in the same way as in the BCM after making the appropriate substitutions in α^* and β^* formulae. For case (ii), however, the set of score functions differs from that of Cavalcante and Blas (2014) by including the first derivative of (4) with respect to σ_δ^2. So, we need to obtain the MLEs for $(\alpha^*, \beta^*, x_0, \phi^*, \phi_0^*, \sigma_\delta^2)$.

4 Application to MnO_4^- concentration

We considered a real data set presented by Barros Neto et al. (2010) to test this model. This sample was obtained from a set of standard solutions that were submitted to ultraviolet-visible spectrophotometric analysis to determine their permanganate (MnO_4^-) concentration for one wavelength ($\lambda = 440$ nanometers (nm)), yielding readings in absorbance units (A). Permanganate proportion was in milliliters (mL).
The present analysis considers that because measurement errors might arise during the preparation of these solutions, MnO_4^- concentration (W_i) which was pre-fixed (*controlled* for each experiment) might not be the real concentration for each solution (x_i), which leads to a Berkson-type error.
We applied the Beer-Lambert law's logarithmic transformation to all readings in absorbance (i.e. $A = -\log_{10} T$, where T is transmittance, $T \in (0, 1)$). Our sample had $n = 6$ and $k = 3$ data points for the first and second stages, with y_0 assumed to be associated to an unknown x_0, despite the fact that x_0's true value was 5.0. This procedure allowed us to measure

how far apart was $\hat{x}_0$ from its true value. We had no prior information regarding the true σ_δ^2, so we treated it as unknown (case (ii)).
For BCCM, ML parameter estimation was performed using the `R` package `optim` (algorithm L-BFGS-B), with $(x_0, \phi^*, \phi_0^*, \sigma_\delta^2) \in \mathbb{R}_4^+$. The same applies to BCM estimation, excluding σ_δ^2. Table 1 presents the results.

TABLE 1. MLEs results for BCCM and BCM.

Model	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\phi}$	$\hat{\phi}_0$	$\hat{x}_0$	$\hat{\sigma}_\delta^2$	$\widehat{Var}(\hat{x}_0)$
BCCM	−0.1307	−0.0846	346.05	409.48	5.3436	1.7619	0.8686
BCM	−0.1307	−0.0846	346.05	409.47	5.3437	-	0.8687

Table 1 shows that ML parameter point estimates for both models are pretty close for all parameters. For x_0, in particular, their estimates are not very different from its true value (5.0) for BCCM and BCM. Their variance estimates differ only by 0.0001. These results suggest that the presence of a Berkson-type error, under the assumption made for k^2, appear to have no or very little effect on MLEs results for the BCC model.

References

Barros Neto, B., Scarminio, I.S., and Bruns, R.E. (2010). *Como Fazer Experimentos: Pesquisa e Desenvolvimento na Ciência e na Indústria*, 4th ed., Porto Alegre: Bookman, 258 – 259.

Blas, B.G., Sandoval, M.C., and Yoshida, O.S. (2007). Homoscedastic controlled calibration model. *Journal of Chemometrics*, **21**, 145 – 155.

Cavalcante, M.T. and Blas, B.G. (2014). Beta calibration model. In: *Proceedings of the 29th International Workshop on Statistical Modelling*, Göttingen, Germany, Kneib, T., Sobotka, F., Fahrenholz, J., and Irmer, H., (Eds.), pp. 19 – 22.

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

Monanhan, H. and Stefanski, L.A. (1992). Normal scale mixture approximations to $F^*(z)$ and computation of the logistic-normal integral. *Handbook of the Logistic Distribution*. Balakhrisnan, N. (Ed.), New York: Marcel Dekker, 592 – 540.

Roy, S., Banerjee, T., and Maiti, T. (2005). Measurement error models for misclassified binary responses. *Statistics in Medicine*, **24**, 269 – 283.

Modelling correlated ordinal data by a copula approach

Marcella Corduas[1]

[1] Department of Political Sciences, University of Naples Federico II, Italy

E-mail for correspondence: marcella.corduas@unina.it

Abstract: In this article, we investigate the problem of modelling multivariate ordinal data with CUB margins. This is a convex combination of a discrete Uniform and a shifted Binomial distribution useful for representing evaluation data. We discuss how the Plackett's distribution can be used in the D-vine algorithm in order to derive a multivariate distribution having CUB margins.

Keywords: CUB models; Plackett's distribution; D-vine; Ordinal data; Food preferences.

1 Introduction

Evaluation data originated by rating one or more items can be modeled by means of CUB distributions (Piccolo, 2003). This is the convex combination of a discrete Uniform and a shifted Binomial distribution:

$$P(x;\boldsymbol{\theta}_x) = \pi_x \left(\begin{array}{c} m-1 \\ x-1 \end{array} \right) (1-\xi_x)^{x-1}\xi_x^{m-x} + (1-\pi_x)\frac{1}{m}, \qquad x = 1,\ldots,m.$$

We will refer to this probability mass distribution (pmf) as $X \sim CUB(\pi,\xi)$. The model mimics a simplified choice mechanism which is supposed to underly the moulding of the judgements when a rater is requested to express preferences, degree of satisfaction about a certain item or, generally speaking, his agreement with a given statement by means of a Likert scale (Iannario and Piccolo, 2012 and therein references). The model is statistically identifiable when $m > 3$ (Iannario, 2010). The interest for CUB model relies on its flexibility in representing observed data by means of a parsimonious formulation and on the fact that the interpretation of the estimated parameters can be easily found. In particular, $(1-\pi_x)$ is interpreted as a measure of *uncertainty* that affect the rater's judgements whereas $(1-\xi_x)$

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

describes the strength of attraction (*feeling*) that the rater feels towards the object under evaluation. Parameters may be related to explanatory variables characterizing raters by means of a logistic link function, but this aspect will not be considered in the present work.
Some recent contributions have investigated the problem of modelling multivariate ordinal data with CUB margins (Corduas, 2014; Andreis and Ferrari, 2013; Giordano, 2014). In the same line, in this article, we propose to apply the approach developed by Panagiotelis et al. (2012) in order to build a multivariate distribution using pair copulas. For this aim, we will consider the Plackett distribution as a possible copula to produce the distributions involved by the various steps of the algorithm based on the discrete vine representation.

2 The joint distribution

Firstly, we briefly recall the method introduced by Plackett (1965) for constructing a one parameter bivariate distribution from given margins. A bivariate Plackett random variable (X, Y) is characterized by the following joint cumulative distribution function:

$$C(F(x), G(y); \psi) = \frac{M(x,y) - [M^2(x,y) - 4\psi(\psi-1)F(x)G(y)]^{1/2}}{2(\psi-1)},$$

where $\psi \in (0, \infty)$. Here, $F(x)$ and $G(y)$ are the pre-defined marginal distribution functions. Moreover, $M(x,y) = 1 + (F(x) + G(y))(\psi - 1)$. The parameter ψ is a measure of association between X and Y; specifically, $\psi = 1$ implies that X and Y are independent, whereas $\psi < 1$ and $\psi > 1$ refer to negative and positive association, respectively.
Molenberghs et al. (1994) introduced the multivariate Plackett's distribution but its construction is in general rather demanding from a computational point of view. However, the bivariate Plackett's distribution may become the building block for constructing a multivariate discrete distribution where the scalar variables follow a CUB distribution.
Without loosing in generality, we consider the case of a three dimensional variable: (Y_1, Y_2, Y_3) where each scalar random variable Y_i takes values $y_i \in S(Y_i) = \{1, \ldots, m\}$, m is given and the ordinal scale is such that 1 is associated to the worst judgement and m to the best one.
Following the approach by Panagiotelis et al. (2012), given an observed sample of ordinal data, (y_{1i}, y_{2i}, y_{3i}), for $i = 1, \ldots, n$, the estimation algorithm is summarized as follows. In order to simplify the notation, whenever possible we drop the reference to the argument of the function.

- The CUB model is fitted to each sample of observed ordinal data (y_{hi}), for $i = 1, \ldots, n$, $h = 1, 2, 3$ obtaining the marginal models: $CUB_1(\pi_1, \xi_1)$, $CUB_2(\pi_2, \xi_2)$, $CUB_3(\pi_3, \xi_3)$ and the corresponding cumulative distribution functions: F_1, F_2, F_3;

- Estimate the joint distributions using the Plackett bivariate copula C: $F_{12} = C(F_1, F_2; \psi_{12})$ and $F_{32} = C(F_3, F_2; \psi_{32})$. These yield the evaluation of the corresponding joint pmf P_{12} and P_{32};
- Compute the conditional distributions $F_{1|2}(y_1|y_2 = i; \psi_{1|2=i})$ and $F_{3|2}(y_3|y_2 = i; \psi_{3|2=i})$, $i = 1, \ldots, m$;
- Estimate the joint conditional distribution functions by means of the copula: $F_{13|2=i} = C(F_{1|2}(y_1|y_2 = i), F_{3|2}(y_3|y_2 = i); \psi_{13|2=i})$, and then, from those, the conditional pmf $P_{13|2}$ for $i = 1, \ldots, m$;
- Compute the multivariate joint distribution $P_{123} = P_{13|2}P_2$.

Note that in each step at most a bivariate copula is needed. The estimation can be performed either by means of the IFM (Inference For the Margins) method (Joe, 1997), or by full maximum likelihood.

3 An empirical application: extra-virgin olive oil data

The proposed method is applied to the ratings that 1000 Italian consumers from AC Nielsen panel gave about on extra virgin olive (EVO) oil (Corduas, 2014). Each interviewee was asked to rate the importance of three EVO oil attributes (colour, taste, clarity) in determining his/her purchase decision on a 7 point Likert scale (where 1 denoted "not important at all" and 7 "extremely important"). In Table 1. we summarize the results of the best fitted model for the mentioned attributes (in parenthesis the jackknife estimated standard error is reported).

TABLE 1. Extravirgin olive oil.

	Colour	Taste	Clarity
π	0.873	0.469	0.736
s.e.	0.024	0.007	0.039
ξ	0.308	0.353	0.307
s.e.	0.015	0.031	0.009
(Colour,Taste)	(Clarity, Taste)		
$\psi_{12} = 2.616\,(0.112)$	$\psi_{32} = 2.210\,(0.107)$		
(Colour,Clarity|Taste)			
$\psi_{13\|2=1} = 10.644\,(4.876)$	$\psi_{13\|2=2} = 6.853\ (3.043)$		
$\psi_{13\|2=3} = 20.825\,(7.467)$	$\psi_{13\|2=4} = 14.604\,(3.588)$		
$\psi_{13\|2=5} = 8.022\,(2.219)$	$\psi_{13\|2=6} = 13.328\,(3.705)$		
$\psi_{13\|2=1} = 16.120\,(5.392)$			
$R^2_{CU} = 0.753$			

The adequacy of the model has been assessed by means of the pseudo-R^2: $R^2_{CU} = (1-\exp(LR_{no}/n))/(1-\exp(LR_{max}/n))$ where $LR_{no} = 2(L_M - L_0)$

and $LR_{max} = 2(L_{max} - L_0)$, being: L_M the maximized log-likelihood value of the considered model, L_0 is the value of the log-likelihood of the null model where independence among the scalar random variables is assumed, L_{max} is the log-likelihood value of the model with a perfect fit (Cragg and Uhler, 1970).

Acknowledgments: This work was supported by Univ. Federico II and Compagnia di San Paolo grant CUP/E68C130000200003.

References

Andreis, F. and Ferrari, P. (2013). On a copula model with CUB margins, *QdS – Journal of Methodological and Applied Statistics*, **15**, 33 – 51.

Corduas, M. (2014). Analyzing bivariate ordinal data with CUB margins. *Statistical Modelling*, doi: 10.1177/1471082X14558770.

Cragg, J.G. and Uhler, R.S. (1970). The demand for automobiles. *Canadian Journal of Economics*, **3**, 386 – 406.

Giordano, S. (2014). A multivariate CUB model. *ERCIM Conference Abstract Book*, Pisa.

Iannario, M. (2010). On the identifiability of a mixture model for ordinal data. *METRON*, **LXVIII**, 87 – 94.

Iannario, M. and Piccolo, D. (2012). CUB models: Statistical methods and empirical evidence. In: Kenett, R.S. and Salini S. (eds.), *Modern Analysis of Customer Surveys: with applications using R*. Chichester: J. Wiley, 231 – 258.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using multivariate Plackett distribution. *Journal of the American Statistical Association*, **89**, 633 - 644.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*. **107**, 1063 – 1072.

Piccolo, D. (2003). On the moments of a mixture of Uniform and shifted Binomial random variables. *QdS – Journal of Methodological and Applied Statistics*, **5**, 85 – 104.

Plackett, R.L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516 – 522.

A probabilistic examination of the efficacy of tort reform via Bayesian semiparametric density ratio modeling

Kevin D. Dayaratna[1], Benjamin Kedem[2]

[1] The Heritage Foundation, United States
[2] University of Maryland, United States

E-mail for correspondence: kevin.dayaratna@heritage.org

Abstract: This study presents a novel adaptation of existing semiparametric density ratio methods to model individual level heterogeneity by uniting the approach with Bayesian methods. We apply this approach to medical malpractice loss data from the previous decade to compute the probabilities of extreme losses. Our results illustrate the effectiveness of recently implemented medical malpractice reforms.

Keywords: Tort reform; Medical malpractice reform; Bayesian computation; Density ratio estimation; Semiparametric modeling.

1 Introduction

1.1 Tort reform

Throughout the United States, medical doctors face the consistent risk of unnecessary litigation. These concerns often manifest themselves in increased health care costs by impelling doctors to order unnecessary tests (Kessler and McClellan 1996; Studdert et al 2005, Crain et al 2009). In some situations, these risks prevent doctors from even pursuing fields of medicine that they are interested in. Reforms to the civil justice system to ameliorate these issues are known as tort, or in this specific context, medical malpractice reforms. In this paper, we improve upon existing semiparametic statistical methodologies to quantify the impact of recently instituted tort reforms reforms to the medical profession.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

1.2 Density ratio estimation

Density ratio estimation (DRE) methods have been used since the late 1970s. The general idea behind DRE methods is instead of assuming a strict parametric distribution for a statistical model, to assume that the distribution in question as well as another "reference" distribution have a particular ratio (Prentice and Pyke 1979; Owen 1988; Qin and Zhang 1997; Qin and Zhang 2005). This semiparametric methodology has seen many uses in applied statistical research ranging from biostatistics to time series analysis to public health among many others (Fokianos 2004; Kedem et al., 2009, 2014).

However, despite the fact that a few studies have utilized Bayesian methods in conjunction with empirical likelihood applications, no studies, to our knoweldge have done so joining the semiparametric DRE method with Bayesian statistics to model heterogeneity (e.g. Yang 2012; Mengersen et al., 2013). In this study, we address this issue, incorporating state-level heterogeneity regarding litigiousness as there is no reason to assume that all states respond to tort reforms in an identical manner.

1.3 Bayesian inference using density ratio estimation

Define $G(\mathbf{x}) = G_{I+1}(\mathbf{x})$ as our reference cumulative distribution function and let $p_{ij} = dG(\mathbf{x}_{i,j}) = dG_{I+1}(\mathbf{x}_{i,j})$. The empirical likelihood function, based on our amalgamated data $\mathbf{x}_{ij}$ is:

$$L(\boldsymbol{\theta}, G_M) = \prod_{i=1}^{M}\prod_{j=1}^{n_i} p_{ij} \prod_{i=1}^{I}\prod_{j=1}^{n_i} e^{\alpha_i + \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{h}(x_{ij})},$$

where $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_I, \beta_{11}, \ldots, \beta_{IP})$ and $\mathbf{h}(x_{ij})$ is a mapping specified a priori by the researcher. We assume that $n_i = 1\ \forall i = 1, \ldots, I$ and $n_{I+1} = n_M = N = I$. Without loss of generality, we can marginalize the empirical likelihood function across normally distributed heterogeneity distributions as follows:

$$\begin{aligned}
\mathrm{ML}(\mu_\alpha, \boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, G_M) &= \int_{\mathbb{R}}\cdots\int_{\mathbb{R}} \prod_{i=1}^{M}\prod_{j=1}^{n_i} p_{ij} \prod_{i=1}^{I}\prod_{j=1}^{n_i} e^{\alpha_i + \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{h}(x_{ij})} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\alpha_i - \mu_\alpha)^2}{2}} \\
&\quad \times \prod_{p=1}^{P} \frac{1}{\sqrt{2\pi\sigma_{\beta_p}^2}} e^{-\frac{(\beta_{ip} - \mu_{\beta_p})^2}{2\sigma_{\beta_p}^2}}\, d\alpha_i d\beta_{ip} \\
&= \prod_{i=1}^{M}\prod_{j=1}^{n_i} p_{ij} \prod_{i=1}^{I}\prod_{j=1}^{n_i} e^{\mu_\alpha + \frac{1}{2}} e^{\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{h}(x_{ij}) + \frac{1}{2}\mathbf{h}(x_{ij})^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{h}(x_{ij})}.
\end{aligned}$$

This marginalized likelihood can be estimated subject to constraints similar to those used in Voulgaraki et al. (2012), i.e., $p_{ij} \geq 0$, $\sum_{i=1}^{M}\sum_{j=1}^{n_i} p_{ij} = 1$,

and

$$\sum_{i=1}^{M}\sum_{j=1}^{n_i} p_{ij} e^{\mu_\alpha + \frac{1}{2} + \boldsymbol{\mu_\beta}^{\mathrm{T}} \mathbf{h}(x_{ij}) + \frac{1}{2}\mathbf{h}(x_{ij})^{\mathrm{T}} \boldsymbol{\Sigma_\beta} \mathbf{h}(x_{ij})} = 1.$$

This constraint is easy to see after integration of both sides of the constraint imposed in Voulgaraki et al. (2012): $\sum_{i=1}^{M}\sum_{j=1}^{n_i} p_{ij} e^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}} \mathbf{h}(x_{ij})} = 1$ over the heterogeneity distribution F, providing us with

$$\int \sum_{i=1}^{M}\sum_{j=1}^{n_i} p_{ij} e^{\alpha_k + \boldsymbol{\beta_k^{\mathrm{T}}} \boldsymbol{h(x_{i,j})}} dF = \int 1 dF \; \forall k = 1, \ldots, I.$$

The marginalized empirical likelihood function can be optimized numerically to estimate μ_α, $\boldsymbol{\mu_\beta}$, and $\boldsymbol{\Sigma_\beta}$. Distribution functions for computation of probabilities can also be obtained as discussed in Dayaratna (2014). The advantage of this approach is that the marginalization dramatically reduces the dimensionality of the problem involving a vast number of densities (and hence parameters) to just two.

2 An application: tort reform

Our dataset was identical to that used in the Crain et al. (2009) study, which also examined per capita tort losses. We used the model specification described above allowing $\mathbf{h}(x_{ij})$ to be the identity mapping. Our results computing the probabilities of extreme medical malpractice losses are below. Confidence intervals were estimated via a bootstrap approach:

TABLE 1. Analysis of 2004 tort loss data, using Bayesian DRE approach.

Probability	Estimate	Lower 95% Limit	Upper 95% Limit
P(Tort Losses > 35000)	0.100	0.010	0.148
P(Tort Losses > 45000)	0.085	0.005	0.104
P(Tort Losses > 55000)	0.019	0.000	0.060

TABLE 2. Analysis of 2006 tort loss data, using Bayesian DRE approach.

Probability	Estimate	Lower 95% Limit	Upper 95% Limit
P(Tort Losses > 35000)	0.068	0.010	0.144
P(Tort Losses > 45000)	0.045	0.005	0.102
P(Tort Losses > 55000)	0.019	0.000	0.060

The above results indicate a notable reduction in probabilities of extreme tort losses in 2006 compared to 2004, indicating the efficacy of state-based malpractice reforms instituted around that time period. These results and their implications, along with model comparisons, are discussed in detail in Dayaratna (2014).

References

Crain, N., Crain, M., McQuillan, L.J., and Abramyan, H. (2009). Tort law tally: How state tort reforms aect tort losses and tort insurance premiums. *Pacic Research Institute*, San Francisco, CA.

Dayaratna, K.D. (2014). Contributions to Bayesian Statistical Modeling in Public Policy Research. *Dissertation*, University of Maryland.

Fokianos, K. (2004). Merging information for semiparametric density estimation. *Journal of the Royal Statistical Society, Series B*, **66**, 941 – 958.

Kedem, B., Qin, J., and Short, D.A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, **43**, 1.

Kedem, B., Kim, E.Y., Voulgaraki, A., and Graubard, B.I. (2009). Two-dimensional semiparametric density ratio modeling of testicular germ cell data. *Statistics in Medicine*, **28**, 2147 – 2159.

Kedem, B., Pan, L., Zhou, W., and Coelho, C. (2014). Interval estimation of small tail probabilities - applications in food safety.

Kessler, D. and McClellan, M. (1996). Do doctors practice defensive medicine? *The Quarterly Journal of Economics*, **111**, 353 – 390.

Mengersen, K.L., Pudlo, P., and Robert, C.P. (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, **110**, 1321 – 1326.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237 – 249.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403 – 411.

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, **84**, 609 – 618.

Qin, J. and Zhang, B. (2005). Density estimation under a two-sample semiparametric model. *Nonparametric Statistics*, **17**, 665 – 683.

Studdert, D.M., et al. (2005). Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. *Journal of the American Medical Association*, **293**, 2609 – 2617.

Voulgaraki, A., Kedem, B., and Graubard, B.I. (2012). Semiparametric regression in testicular germ cell data. *The Annals of Applied Statistics*, **6**, 1185 – 1208.

Yang, Y. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, **40**, 1102 – 1131.

Local influence on Gaussian spatial linear model with multiple replications

Fernanda De Bastiani[1], Audrey Helen Mariz de Aquino Cysneiros[1], Miguel Angel Uribe-Opazo[2], Manuel Galea[3]

[1] Universidade Federal de Pernambuco, Recife 50740-540, Brazil
[2] Universidade Estadual do Oeste do Paraná, Cascavel 85819-110, Brazil
[3] Pontificia Universidad Católica de Chile, Santiago 782-0436, Chile

E-mail for correspondence: fernandadebastiani@gmail.com

Abstract: We present local diagnostics techniques to assess the influence of observations on Gaussian spatial linear models with multiple replications, considering appropriated perturbation in the response variable and in the scale matrix.

Keywords: Geostatistics; Maximum likelihood; Spatial variability.

1 Introduction

Geostatistical data are collected at known locations in space, from a process that has a value at every location in a certain domain. Given a model for the trend, and under some stationary assumptions, geostatistical modeling involves the estimation of the spatial correlation. For this study the observations are taken from different experimental units, which is different geographical location, where each variable is observed more than once. To assess the effect of small perturbations in the model (or data) on the parameter estimates, Cook (1986) has proposed the local influence method. Uribe-Opazo et al. (2012) used diagnostic techniques to assess the sensitivity of the maximum likelihood estimators, covariance functions and linear predictor to small perturbations in the data and/or in the Gaussian spatial linear model assumptions. We consider appropriate perturbation in the response variable and in the scale matrix proposed by Zhu et al. (2007).

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

2 Gaussian spatial linear model with multiple replications

Let $\mathbf{Y} = \mathbf{Y}(s) = \text{vec}(\mathbf{Y}_1(s), \ldots, \mathbf{Y}_r(s))$ be an $nr \times 1$ random vector of r independent stochastic processes of n elements each, that depend on the position $s \in S \subset \mathbb{R}^2$. It is assumed the ith stochastic process $\mathbf{Y}_i(s) = \text{vec}(Y_i(s_1), \ldots, Y_i(s_n))$, represents the $n \times 1$ vector, for $i = 1, \ldots, r$, which can be expressed as a linear model by $\mathbf{Y}_i(s) = \boldsymbol{\mu}_i(s) + \boldsymbol{\epsilon}_i(s)$, where $\boldsymbol{\mu}_i(s)$ is an $n \times 1$ vector, the means of the process $\mathbf{Y}_i(s)$, and $\boldsymbol{\epsilon}_i(s)$ is an $n \times 1$ vector of a stationary process, with $\text{E}[\boldsymbol{\epsilon}_i(s)] = \mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. As assumed by Smith (2001), the mean vector $\boldsymbol{\mu}_i(s)$ can be written as a spatial linear model by $\boldsymbol{\mu}_i(s) = \mathbf{X}(s)\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\text{T}}$ is a $p \times 1$ vector of unknown parameters, and $\mathbf{X} = \mathbf{X}(s) = [\mathbf{X}_{i1}(s), \ldots, \mathbf{X}_{ip}(s)]$ is an $n \times p$ matrix of p explanatory variables.
Let $\boldsymbol{\Sigma}_i = [C_i(s_u, s_v)]$ be the $n \times n$ covariance matrix of $\mathbf{Y}_i(s)$ for ith repetition, $i = 1, \ldots, r$. The matrix $\boldsymbol{\Sigma}_i$ is non-singular, symmetric and positive defined, associated to the vector $\mathbf{Y}_i(s)$, where for the stationary and isotropic process, the elements $C_i(s_u, s_v)$ depend on the Euclidean between points s_u and s_v. We considered an homogeneous process where $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, has a structure which depends on the vector of parameters $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)^{\text{T}}$ or $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3, \phi_4)^{\text{T}}$, depending on the form of the covariance structure $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} = \phi_1 \mathbf{I}_n + \phi_2 \mathbf{R}$, where $\phi_1 \geq 0$ is known as nugget effect; $\phi_2 \geq 0$ is known as sill; $\mathbf{R} = \mathbf{R}(\phi_3, \phi_4) = [(r_{uv})]$ or $\mathbf{R} = \mathbf{R}(\phi_3) = [(r_{uv})]$ is an $n \times n$ symmetric matrix, which is function of $\phi_3 > 0$, and sometimes also a function of $\phi_4 > 0$; ϕ_3 is a function of the model range (a), ϕ_4 when exists is known as the smoothness parameter, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. For each repetition of the random vector, $\mathbf{Y}_i \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$.

3 Local influence

The local influence method investigates the role of observations on the parameters estimation under small perturbations introduced by a vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_s)^{\text{T}} \in \mathbb{R}^s$. We assume that there is a point $\omega_0 \in \Omega$, where there is no perturbation. The influence of the perturbation $\boldsymbol{\omega}$ on the ML estimator proposed by Cook (1986) is evaluated by the likelihood displacement given by $LD(\boldsymbol{\omega}) = 2[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\hat{\boldsymbol{\theta}}|\boldsymbol{\omega})]$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood (ML) estimator of $\boldsymbol{\theta}$ in the postulated model and $\hat{\boldsymbol{\theta}}_\omega$ is the ML estimator of $\boldsymbol{\theta}$ in the model perturbed by $\boldsymbol{\omega}$. Cook (1986) proposed to study the local behavior of $LD(\boldsymbol{\omega})$ around $\boldsymbol{\omega}_0$ and showed that the normal curvature C_l of $LD(\boldsymbol{\omega})$ at $\boldsymbol{\omega}_0$, in direction of some unit vector $\boldsymbol{l}$, is given by $C_l = C_l(\boldsymbol{\theta}) = 2|\boldsymbol{l}^{\text{T}}\boldsymbol{\Delta}^{\text{T}}\mathbf{L}^{-1}\boldsymbol{\Delta}\boldsymbol{l}|$, with $||\boldsymbol{l}|| = 1$, where $-\mathbf{L}$ is the observed information matrix evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\boldsymbol{\Delta} = (\boldsymbol{\Delta}^{(1)}, \ldots, \boldsymbol{\Delta}^{(r)})$, $\boldsymbol{\Delta}^{(i)} = (\boldsymbol{\Delta}_\beta^{(i)\top}, \boldsymbol{\Delta}_\phi^{(i)\top})^{\text{T}}$, where $\boldsymbol{\Delta}_\beta^{(i)} = \partial^2 \mathcal{L}_i(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial\boldsymbol{\beta}\partial\boldsymbol{\omega}^{\text{T}}$ and $\boldsymbol{\Delta}_\phi^{(i)} = \partial^2 \mathcal{L}_i(\boldsymbol{\theta}|\boldsymbol{\omega})/\partial\boldsymbol{\phi}\partial\boldsymbol{\omega}^{\text{T}}$, evaluated

at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Let define $\mathbf{F} = \boldsymbol{\Delta}^{\top}\mathbf{L}^{-1}\boldsymbol{\Delta}$. Since C_l is not invariant under uniform change of scale, Poon and Poon (1999) proposed the conformal normal curvature $B_l = C_l/tr(2\mathbf{F})$. We denote by $B_i = 2|f_{ii}|/tr(2\mathbf{F})$ the conformal curvature in the unit direction with ith entry 1 and all other entries 0. To check appropriate choice of a perturbation vector and to calculate influence measures we followed Zhu et al. (2007).

Perturbation on the scale matrix: Let us assume $\omega_i^{-1}\boldsymbol{\Sigma}$ instead of $\boldsymbol{\Sigma}$ for this perturbation scheme, with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_r)^{\top}$ the perturbation vector, $\omega_i > 0$, and $\boldsymbol{\omega}_0 = (1, \ldots, 1)^{\top}$ being the vector of non-perturbation. The log-likelihood for each repetition is given by $\mathcal{L}_i(\boldsymbol{\theta}|\omega_i) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{n}{2}\log\omega_i - \frac{1}{2}\omega_i\boldsymbol{\epsilon}_i^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}_i$, $\boldsymbol{\Delta}_{\beta}^{(i)} = \mathbf{X}^{\top}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\epsilon}_i$ and $\boldsymbol{\Delta}_{\phi}^{(i)} = \frac{\partial \mathrm{vec}^{\top}(\boldsymbol{\Sigma})}{\partial \boldsymbol{\phi}}\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\top}\boldsymbol{\Sigma}^{-1})$, where $\boldsymbol{\epsilon}_i = (\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta})$, evaluated in $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Perturbation on the response variable: We can write $\mathbf{Y}_{i\omega} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, with $\mathbf{Y} = \mathbf{Y}_i + (-1)\mathbf{A}\boldsymbol{\omega}_i$. We have that $\mathcal{L}_i(\boldsymbol{\theta}|\boldsymbol{\omega}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\delta_{i\omega}$, where $\delta_{i\omega} = [\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\omega}_i)]^{\top}\boldsymbol{\Sigma}^{-1}[\mathbf{Y}_i - \boldsymbol{\mu}(\boldsymbol{\omega}_i)]$, $\boldsymbol{\mu}(\boldsymbol{\omega}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\omega}$. So $\boldsymbol{\mu}(\boldsymbol{\omega}_i) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\omega}_i$ is a perturbation scheme appropriate for $i = 1, \ldots, r$, as shown in De Bastiani et al. (2014) and $\boldsymbol{\Delta}_{\beta}^{(i)} = \mathbf{X}^{\top}\boldsymbol{\Sigma}^{-1/2}$ and $\boldsymbol{\Delta}_{\phi}^{(i)} = \boldsymbol{\epsilon}_i^{\top}(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}^{1/2}}{\partial\phi_j} - \boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\phi_j}\boldsymbol{\Sigma}^{-1/2})$, where $\boldsymbol{\epsilon}_i = (\mathbf{Y}_{i\omega} - \mathbf{X}\boldsymbol{\beta})$, evaluated in $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

4 Application

The data set was collected in a grid of 7.20×7.20 m an experimental area with 1.33 ha, for the harvest years from 1998/1999 to 2002/2003 with 255 observations each. Soybean productivity data and four chemical contents of soil considered as explanatory variables were collected: phosphorus (P), potassium (K), calcium (Ca) and magnesium (Mg).

In Table 1 we note that P and Mg has an inverse proportional relationship with the mean of the productivity, and the opposite happens with K and Ca. The Gaussian geostatistical model was chosen by the maximum log-likelihood value and cross validation criteria.

In Figure 1 (left side) we can see that the repetition 5, which correspond to the year 2002/2003, has more influence of the estimation process. Figure 1

TABLE 1. Parameters estimates for Coodetec data - `GeoR` - and the asymptotic standard errors (in parenthesis) considering the Gaussian geostatistical model.

$\hat{\beta}_0$	$\hat{\beta}_1$ (P)	$\hat{\beta}_2$ (K)	$\hat{\beta}_3$ (Ca)	$\hat{\beta}_4$ (Mg)	$\hat{\phi}_1$ (nugget)	$\hat{\phi}_2$ (sill)	$\hat{\phi}_3$ f(range)
2.401	-0.002	0.327	0.012	-0.059	0.193	0.058	40.684
(0.046)	(0.005)	(0.078)	(0.018)	(0.023)	(0.018)	(0.022)	(26.842)

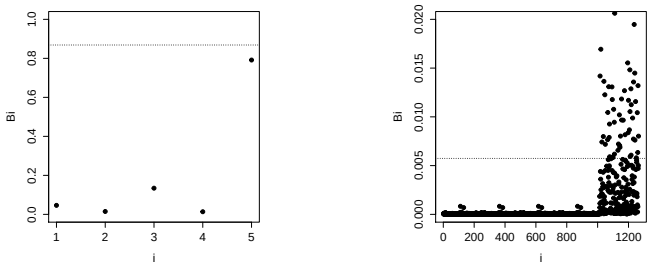


FIGURE 1. The data for five repetitions (five years) from 1998/1999 until 2002/2003 B_i vs i considering perturbation on the scale matrix (left side) and perturbation on the response variable (right side).

(right side) shows that observations belonging to the data collected on the latest year has more influence.

5 Conclusions

The local influence technique allowed us more than just identify influential observations, allow us to understand better the modeling process.

Acknowledgments: Fundação Araucária of Paraná State, Capes, CNPq and FACEPE, Brazil and project FONDECYT 1110318, Chile.

References

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133 – 169.

De Bastiani, F., Cysneiros, A.H.M.A., Uribe-Opazo, M.A., and Galea, M. (2014). Local influence on elliptical spatial linear models. *TEST*, DOI: 10.1007/s11749-014-0409-z.

Poon, W. and Poon, Y.S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **61**, 51 – 61.

Smith, R.L. (2001). *Environmental Statistics*. Notes at the conference Board of the Mathematical Sciences course at University of Washington.

Uribe-Opazo, M.A., Borssoi, J.A., and Galea, M. (2012). Influence diagnostics in Gaussian spatial linear models. *Journal of Applied Statistics*, **39**, 615 – 630.

Zhu, H.T., Ibrahim, J.G., Lee, S., and Zhang, H. (2007). Perturbation selection and influence measures in local influence analysis. *Annals of Statistics*, **35**, 2565 – 2588.

An improved shrinkage procedure for the estimation of covariance matrices in graphical models

Vera Djordjilović[1], Monica Chiogna[1]

[1] Department of the Statistical Sciences, University of Padova, Italy

E-mail for correspondence: djordjilovic@stat.unipd.it

Abstract: Estimation of high dimensional covariance matrices is a commonly encountered problem in the analysis of genomics data. The widely used sample covariance matrix is usually not appropriate. We propose an improved shrinkage estimator, that builds upon the work of Ledoit and Wolf (2004), adapting it to the graphical models setting.

Keywords: Shrinkage; Covariance matrix; "Small n large p", Gaussian graphical models.

1 Introduction

The starting point of many statistical procedures in graphical modelling is the estimation of the structured covariance matrix of a normally distributed random vector. We tackle the estimation of the covariance matrix when the number of variables under consideration is at least of the same order of magnitude as the number of available statistical units, a situation encountered particularly often in genomics setting.
To obtain an invertible and well conditioned estimate of a covariance matrix in high dimensional settings, Ledoit and Wolf (2004) proposed a shrinkage approach. Their shrinkage estimator is a weighted average of the sample covariance matrix and the identity matrix (sometimes referred to as *the target*). We adapt the approach of Ledoit and Wolf (2004) to the graphical models setting by replacing the identity matrix with a target that encodes the presumed graphical structure.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

2 The proposal

Consider a p-variate normal random vector X and assume that there is an undirected graph $G = (V, E)$, such that the distribution of X is Markov with respect to G. We thus have

$$(X_1, \ldots, X_p)^{\mathrm{T}} \sim \mathrm{N}_p(\mu, \Sigma), \quad \mu \in \mathbb{R}^p, \Sigma^{-1} \in S^+(G),$$

where $S^+(G)$ is the set of all $p \times p$ symmetric positive definite matrices with null elements corresponding to the missing edges of G.
To estimate Σ, we propose an estimator of the following form

$$U = \lambda T + (1 - \lambda)S, \tag{1}$$

where S is the sample covariance matrix, T is a target matrix and λ is the shrinkage parameter. In Ledoit and Wolf (2004) T is taken to be an identity matrix I.
Our choice of matrix T is guided by two criteria: it should be positive definite, so that U, being a convex combination of a positive semidefinite and a positive definite matrix, is also positive definite; and it should reflect the graphical structure of the considered distribution. We propose one target and three different specifications of it, each characterized by a different number of parameters. The target is obtained in two steps. We first start from the following working matrix W

$$W = \begin{pmatrix} v_1 & r\sqrt{v_1 v_2} & \cdots & r\sqrt{v_1 v_p} \\ r\sqrt{v_2 v_1} & v_2 & \cdots & r\sqrt{v_2 v_p} \\ \vdots & \vdots & \ddots & \vdots \\ r\sqrt{v_p v_1} & r\sqrt{v_p v_2} & \cdots & v_p \end{pmatrix},$$

for which we consider three different specifications. In the second step, we impose constraints on the inverse of W, such that $W^{-1} \in S^+(G)$. The resulting matrix is our target matrix.
As far as specifications are concerned, we consider W with equal correlation and different variances, of equal correlation and equal variances ($v_1 = \cdots = v_p$), and of fixed equal correlation and equal variances (for example, $r = 0.1, v_1 = \cdots = v_p = 1$). We refer to the target resulting from the first, the second and the third specification as T_1, T_2 and T_3, respectively. With the first two specifications, we estimate the unknown parameters from the data and then pass the resulting working matrices to the IPS (Iterative Proportional Scaling) algorithm to ensure that their inverses have the right zero structure.
The matrix U resulting from (1) will be invertible, but its inverse, in general, will not have the desired zero structure corresponding to the missing edges of G. Therefore in the last step, we apply the IPS algorithm to U to ensure that the model constraints are satisfied.

Clearly, choosing the right weight (shrinkage intensity) to give to the target is essential. We follow the asymptotic approach of Ledoit and Wolf and choose an asymptotically optimal λ. The optimal value minimizing the expected loss $\mathrm{E}\|U - \Sigma\|_F/p$, where $\|A - B\|_F = \sqrt{\mathrm{tr}[(\mathrm{A} - \mathrm{B})^{\mathrm{T}}(\mathrm{A} - \mathrm{B})]}$ is:

$$\lambda^* = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\left[\mathrm{var}(s_{ij}) - \mathrm{cov}(t_{ij}, s_{ij})\right]}{\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left(t_{ij} - s_{ij}\right)^2}.$$

To compute the quantities featured in the above expression, we propose to adopt a bootstrap approach. When the cost of resampling is prohibitive, we consider the third specification, containing no parameters. Since the target matrix is now deterministic, the previous expression simplifies to

$$\lambda^* = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\left[\mathrm{var}(s_{ij})\right]}{\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{E}\left(t_{ij} - s_{ij}\right)^2}.$$

To estimate the quantities in the numerator, we follow Schafer and Strimmer (2005).

3 Simulation studies

Performances of the proposed approach are studied via simulation. We compare three novel shrinkage estimators based on targets T_1, T_2 and T_3 with the standard shrinkage estimator (where the target is the identity matrix I). For purpose of comparison, we also include the sample covariance matrix S. We consider a directed acyclic graph (DAG) describing a particular biological pathway, the B cell pathway. The DAG derived from this pathway contains 35 nodes and is shown in Figure 1. This graph, alongside measurements of the expression levels of the participating genes, is an example featured in the R package `topologyGSA` (Massa and Sales 2013). We use these expression measurements to estimate the parameters of the moralized graph, and then use the estimated model to simulate 10000 datasets for each of the considered sample sizes. The results are shown in Table 1.
The shrinkage estimators based on T_1 and T_2 give the best results in this simulation study. An interesting observation is that for the small sample sizes the standard shrinkage estimator (I) outperforms the competing shrinkage estimator (T_3), a result probably due to the chosen value for the common correlation coefficient (0.1).
This shrinkage method seems to be a simple and affordable way to insert apriori information in the shrinking procedure about the structure of relations among variables. We are currently investigating the effects of misspecification of the graphical structure on the expected loss of the estimators (results not reported here). When misspecification foresees more edges than the truly existing ones, the performance appears to be not affected.

TABLE 1. The B cell pathway model: root mean square error (and standard deviation) of different covariance estimators, multiplied by 10^2.

n	T_1	T_2	T_3	I	S
10	4.74	4.37	5.30	4.91	12.51
	(0.81)	(0.54)	(0.72)	(0.47)	(1.71)
20	3.47	3.50	4.33	4.12	8.73
	(0.54)	(0.34)	(0.47)	(0.33)	(0.86)
30	3.01	3.12	3.82	3.76	7.03
	(0.41)	(0.38)	(0.38)	(0.25)	(0.63)
50	2.57	2.62	3.34	3.36	5.43
	(0.30)	(0.38)	(0.28)	(0.22)	(0.37)
100	1.99	1.95	2.36	2.79	3.83
	(0.24)	(0.30)	(0.23)	(0.17)	(0.26)

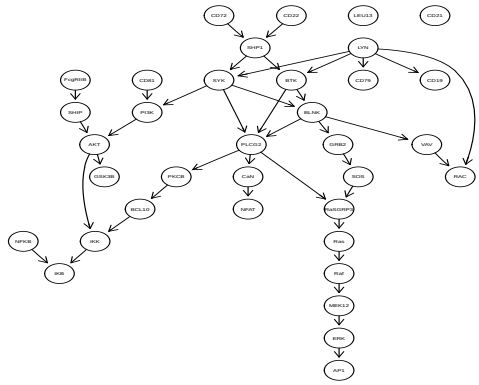


FIGURE 1. DAG representing the B cell pathway.

References

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, **88**, 365 – 411.

Massa, S. and Sales, G. (2013). topologyGSA: Gene Set Analysis Exploiting Pathway Topology. R package version 1.4.2.

Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**.

Measures of association for 2×2 tables with a ϕ-divergence origin

Michael Espendiller[1], Maria Kateri[1]

[1] Institute of Statistics, RWTH Aachen University, Germany

E-mail for correspondence: Espendiller@stochastik.rwth-aachen.de

Abstract: The odds ratio is the predominant measure of association in 2×2 contingency tables, which, for inferential purposes, is usually considered on the log-scale. A drawback however is that in case of a sampling zero, the log odds ratio is estimated as infinite or minus infinite. Under an information theoretic view, it is connected to the Kullback-Leibler divergence. Considering a generalized family of divergences, the ϕ divergence, alternative association measures are derived. Their properties are studied and asymptotic inference is developed. For some members of this family, the estimated association measures remain finite in the presence of a zero cell and have finite variance, allowing thus the construction of confidence intervals. Special attention is given to the power divergence, which is a parametric family. The role of its parameter, in terms of the asymptotic confidence intervals' coverage probability and average relative length, is lightened via an evaluation study.

Keywords: Odds ratio; Kullback-Leibler divergence; Power divergence; Delta method; Asymptotic confidence intervals.

1 Introduction

One of the most fundamental measures of association in contingency tables analysis is the odds ratio, playing an important role in modeling. Focusing on 2×2 contingency tables, the bibliography on inference for the odds ratio is enormous rich, with part on the study of the behavior of corresponding confidence intervals when a zero cell count occurs. In such cases, the odds ratio is estimated either as 0 or ∞. To ensure that the estimate exists, usually a small constant c $(0 < c < 1)$ is added on all cells, when one of them is zero. For large samples, approximate confidence intervals (CIs) for the odds ratio are obtained, based on the asymptotic normality of the log odds ratio, derived by the delta method (cf. Agresti, 2013). The presence

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

of zero cells leads furthermore to infinite asymptotic variance estimates, which is avoided by the c-corrected estimator. Over the years, there have been proposed alternative asymptotic CIs, targeting to a better behavior for relative small samples. For a short overview of the literature on the odds ratio, we refer to Kateri (2014, Section 2.5.2).
In this work, we propose a generalization of the log odds ratio by replacing the log-scale through a family of scales, based on a generalized family of divergences, the ϕ divergence. In Section 2, the ϕ-scaled odds ratio is introduced and its properties are discussed. Some members of this family do not face the problem of infinite estimation of the log-scale while others have finite variance as well. In case of the Kullback-Leibler (KL) divergence, the ϕ-scaled odds ratio turns out to be the classical log odds ratio. Special attention is given to the power divergence (Cressie and Read, 1984), which is a broad family itself, controlled by a link parameter, λ. The Wald asymptotic CI for the ϕ-scaled odds ratio are derived and an indicative example is provided. For the power divergence, the role of λ in the behavior of CIs is investigated in an extensive evaluation study for various table structure scenarios. These results are shortly commented in Section 3.

2 The ϕ-scaled odds ratio

Let $\mathbf{n} = (n_{ij})$ be a 2×2 contingency table of counts, observed under a multinomial sampling scheme with $\sum_{i,j} n_{ij} = n$ and probability matrix $\boldsymbol{\pi} = (\pi_{ij})$, satisfying $\pi_{ij} > 0$, $i, j = 1, 2$. Let further X, Y denote the binary row and column classification variables, respectively. The log odds ratio is then defined as $\log\theta = \log\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$ and is estimated by $\log\hat{\theta} = \log\frac{n_{11}n_{22}}{n_{12}n_{21}}$. The estimator $\log\hat{\theta}$ of $\log\theta$ is approximately normal distributed with asymptotic variance estimated by $\hat{\sigma}^2 = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$. For given marginal probabilities, i.e. for given π_{1+} and π_{+1}, the value of $\log\theta$ specifies uniquely the underlying probability table, with $\log\theta = 0$ corresponding to the probability table under independence of X and Y. The information actually required for specifying a 2×2 probability table, additional to $(\pi_{1+},\ \pi_{+1})$, is a measure of the underlying association of the form $\theta^g = g\left(\frac{\pi_{11}}{\pi_{1+}\pi_{+1}}\right) + g\left(\frac{\pi_{22}}{\pi_{2+}\pi_{+2}}\right) - g\left(\frac{\pi_{12}}{\pi_{1+}\pi_{+2}}\right) - g\left(\frac{\pi_{21}}{\pi_{2+}\pi_{+1}}\right)$, which compares the departure of the cell probabilities from the independence, scaled through g. When this departure is measured by the ϕ divergence, for a strictly convex ϕ-function, then it can be proved that $g = \phi'$ (through the ϕ-association models for $I \times J$ tables (Kateri and Papaioannou, 1995); for $I = J = 2$, the model is saturated).
The ϕ divergence between two discrete finite bivariate distributions $\mathbf{p} = (p_{ij})$ and $\mathbf{q} = (q_{ij})$ is defined as $\mathcal{D}^{\phi}(\mathbf{p}, \mathbf{q}) = \sum_{i,j} q_{ij}\phi\left(\frac{p_{ij}}{q_{ij}}\right)$, where ϕ is a convex function on $[0, \infty)$, such that $\phi(1) = \phi'(1) = 0$, $0 \cdot \phi(0/0) = 0$, and $0 \cdot \phi(x/0) = x \cdot \lim_{u\to\infty}\phi(u)/u$. It was introduced in 1963 by Csiszár.

Well-known members of the ϕ-divergence family are the KL divergence $[\phi(x) = x \log x - x + 1]$, the Pearson divergence $[\phi(x) = \frac{1}{2}(x-1)^2]$ and the power divergence of Cressie and Read (1984) for $\phi_\lambda(x) = \frac{x^{\lambda+1} - x - \lambda(x-1)}{\lambda(1+\lambda)}$ $(\lambda \neq -1, 0)$. The last is a parametric family, including the Pearson $(\lambda = 1)$ and the KL $(\lambda \to 0)$ divergences.
Hence, for strictly convex ϕ-functions, we define the ϕ-scaled odds ratio as

$$\theta^\phi = \theta^\phi(\boldsymbol{\pi}) = \sum_{i=j} \phi'\left(\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}\right) - \sum_{i \neq j} \phi'\left(\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}\right), \qquad i, j = 1, 2, \quad (1)$$

which is estimated by $\hat{\theta}^\phi = \theta^\phi(\mathbf{n})$, obtained by replacing in (1) the probabilities with the corresponding sample frequencies. Using the ϕ-function for KL divergence, θ^ϕ in (1) becomes the log odds ratio. The θ^ϕ corresponding to the power divergence (θ^{ϕ_λ}) was introduced by Rom and Sarkar (1992). However, their approach was model based and their focus different.
The ϕ-scaled odds ratio is sampling scheme invariant, in the sense that the ϕ-scaled odds ratio under product binomial sampling turns out to be equivalent to (1).
Properties:
(P1) $\theta^\phi = 0$ if and only if X and Y are independent.
(P2) For fixed row and column marginals, θ^ϕ is increasing in π_{11}.
(P3) For fixed row (column) marginals, θ^ϕ is increasing in the column (row) marginals.
(P4) θ^ϕ is invariant under table rotation.
(P5) θ^ϕ changes sign when rows or columns are interchanged.

Although $\log\theta$ is unbounded, θ^ϕ is bounded if ϕ is twice differentiable and strictly convex with $\phi'(0) = \lim_{t \searrow 0} \phi'(t) > -\infty$. In this case, $\hat{\theta}^\phi$ remains finite in the presence of a sampling zero. The asymptotic distribution of $\hat{\theta}^\phi$ is derived by the delta method. Note that although $\hat{\theta}^\phi$ is invariant of the sampling scheme, its variance is not. For the KL divergence the variances under multinomial and product binomial sampling coincide. Furthermore, the estimated standard error of $\hat{\theta}^\phi$ (SE_ϕ) is finite if $\phi''(0) = \lim_{t \searrow 0} \phi''(t)$ exists finitely.
The $\hat{\theta}^{\phi_\lambda}$ for $\lambda = 0, 1/3, 1$ are illustrated for the data of Table 1 (left). They are provided in Table 1 (right). The underlying sampling scheme is the product binomial with $n_1 = n_2 = 15$.

3 Discussion

Focusing on the θ^{ϕ_λ}, evaluation studies were performed with regard to the coverage probabilities of the associated 95% asymptotic CIs and for various scenarios for the sample size of the tables and the balancing among the cell probabilities. The log odds ratio CIs tend to underweight extreme situations, which explains the low coverage when $|\log\theta| > 4$ as Agresti

TABLE 1. Prednisolone data (Source: Kristensen et al., 1992; *Journal of Int. Med.*, **232**, 237 – 245). Estimates, bounds, standard errors and asymptotic 95% Wald CIs for θ^{ϕ_λ}, for $c = 0$ and $c = 0.5$ in the upper and lower part, respectively.

	Normalization	
	Yes	No
Prednsl.	7	8
Placebo	0	15

λ	$\hat{\theta}^{\phi}$	Range	SE	95 % CI
0	∞	$(-\infty, \infty)$	-	-
1/3	4.40	[-7.56, 7.56]	-	-
1	2.61	[-4.00, 4.00]	0.22	[2.18, 3.04]
0	3.31	[-6.87, 6.87]	1.52	[0.33, 6.29]
1/3	2.79	[-5.10, 5.10]	0.88	[1.06, 4.53]
1	2.33	[-3.75, 3.75]	0.46	[1.43, 3.24]

(1999) also pointed out. In such cases, we suggest to use $\lambda = 1/3$, since the corresponding CI improves the coverage when approaching the boarders of the parameter space. Overall, the $\lambda = 1/3$-odds ratio CI is less conservative than the classical log odds ratio CI. Comparisons among CIs for different ϕ-scaled odds ratios with respect to average length are not straight forward, due to the scale difference. For this, they are compared in terms of their average relative length. For example, for the data of Table 1, the relative lengths of the CIs for $\lambda = 0$ and $\lambda = 1/3$, when $c = 0.5$, are $(6.29 - 0.33)/(2 \cdot 6.87) = 0.434$ and $(4.53 - 1.06)/(2 \cdot 5.10) = 0.340$, respectively.

References

Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics*, **55**, 597 – 602.

Agresti, A. (2013). *Categorical Data Analysis (3rd ed.)*. Hoboken: Wiley & Sons, Inc.

Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *Royal Statistical Society, Series B*, **46**, 440 – 464.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. New York: Birkäuser.

Kateri, M. and Papaioannou, T. (1995). f-divergence association models. *International Journal of Mathematical and Statistical Science*, **3**, 179 – 203.

Rom, D. and Sarkar, S.K. (1992). A generalized model for the analysis of association in ordinal contingency tables. *Journal of Statistical Planning and Inference*, **33**, 205 – 212.

Dynamic analysis for event history data: Recurrent infant diarrhoea

Rosemeire L. Fiaccone[1], Robin Henderson[2]

[1] Statistics Department, Universidade Federal da Bahia, Brazil
[2] School of Mathematics and Statistics, Newcastle University, UK

E-mail for correspondence: rose.fiaccone@gmail.com

Abstract: In many situations, individual subjects or units may experience events that occur repeatedly. We illustrate the use of modern event-history analysis in the analysis of recurrent diarrhoea episodes in three cohorts of infants. The data are complicated by time-dependent covariates, time-dependent effects, intermittent missingness and dropout. In our approach the conditional mean based on the history is modelled as a function of possibly time-varying covariates through an additive regression model for longitudinal binary data subject to both intermittent missingness and dropout. The idea is to show how the array of additive intensity modelling techniques can be applied to longitudinal binary data, providing valuable inferences without highly computationally intensive procedures. In addition to this, it is easy to assess the effect of covariates on the event of interest, even for complex time varying effects of time varying covariates. On the other hand the more natural logistic regression approach can be unstable when events are rare, and the use of the Firth correction may lead to biased predictions, especially when, as here, we sum over a large number of ages.

Keywords: Additive model; Counting process; Recurrent event.

1 Introduction

Recent research has focussed on complex recurrent event settings which include large number of recurrent events, time-dependent covariates, time-dependent effects and dependent censoring among other features. One way to analyse data of this type is to use longitudinal techniques for count or binary outcomes, either the number of episodes over a period of time, or the presence/absence of diarrhoea each day. The conditional mean based on the history is modelled as a function of possibly time-varying covariates through an additive regression model for longitudinal binary data subject

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

to both intermittent missingness and dropout. We illustrate the use of modern event-history analysis in the analysis of recurrent diarrhoea episodes in three cohorts of infants. In particular, we will be interested in recurrent events, which are multiple occurrences of the same event for an individual, in our case repeated episodes of diarrhoea. The aim is to incorporate evolution over time, measured on three scales: time in study, calendar time, and age of subject through dynamic models. These models include time dependent covariates representing individual-specific histories not known at outsent of study (Fosen et al., 2006; Borgan et al., 2007). This is a real issue since it is hoped that incidence of diarrhoea would decline in calendar time as sanitation and health awareness improves, but at the same time it is known that incidence declines with increasing age. The analysis is further complicated by the presence of heterogeneity between children and missing data.

2 Methods

The present approach can be viewed as a flexible alternative in which the conditional mean based on the history is modelled as a function of possibly time-varying covariates. Consider $N_i(t)$ to be a process that counts the number of events observed for child i up to age (or time) t. If a diarrhoea episode is observed in a small interval of age (or time), then $N_i(t)$ jumps one unit. Denote by $\mathcal{F}_{it}$ be the all information available to the researcher at age (or day) t (on diarrhoea episodes, covariates, missing observations, etc for individual i. At each event time a vector of 0s and 1s (for people at risk) at age (or time) is formed ($dN(t)$). Then $dNi(t)$ is a binary random variable indicating the number of events that are seen to occur for individual i in the short interval $[t; t + dt)$. It is possible to model its conditional mean by the observed intensity function

$$E[dN_i(t)|\mathcal{F}_t] = P(dN_i(t) = 1|\mathcal{F}_t) = Y_i(t)\alpha_i(t|\mathcal{F}_t)dt,$$

where $Y_i(t)$ is a risk indicator (0 if cannot observe any change in $N_i(t)$ at t, 1 otherwise), and $\alpha_i(t|\mathcal{F}_{t^-})$ is the true underlying intensity function. In other words, $N_i(t)$ will have a jump at time t if an event occurs and is observed to occur. The at-risk process can be allowed to depend on the observed past, though this is suppressed in the notation. Usually $\alpha_i(t|\mathcal{F}_{t^-})$ is of most interest and is the object of statistical modelling.
According to Andersen et al. (1993), the counting process $N_i(t)$ has a *compensator* $\Lambda_i(t)$ such that $M_i(t) = N_i(t) - \Lambda_i(t)$ is a *martingale*, which can be thought of as a type of residual or noise process. In our case $\Lambda_i(t)$ is the *cumulative intensity process*. General martingale theory can be used to derive asymptotic properties of estimators (via the martingale central limit theorem) and as the basis of variance estimation, with few assumptions and almost no additional work (Andersen et al., 1993). We will assume

that individuals are mutually independent and associated with individual i is a collection of possibly time-varying covariates $X_{i1}(t), \ldots, X_{ip}(t)$. The Aalen additive model that underlies our work is of the form

$$\alpha_i(t|\mathcal{F}_{t^-}) = \beta_0(t) + \beta_1(t)X_{i1}(t) + \cdots + \beta_p(t)X_{ip}(t), \tag{1}$$

and, loosely, $dN_i(t) = Y_i(t)dB_0(t) + \sum_{k=1}^{p} Y_i(t)dB_k(t)X_{ik}(t) + dM_i(t)$, where $\beta_0(t)$ is the baseline intensity, $\beta_j(t)$ are regression functions ($j = 1, \ldots, p$), and $B_j(t) = \int_0^t \beta_j(u)du$ are *cumulative coefficients*. More complex versions of (1) might incorporate covariate trajectories rather than their instantaneous values, or summaries of previous event patterns as in the dynamic models of Fosen et al. (2006) or Borgan et al. (2007). For each age t, this model has the form of a standard linear regression model with uncorrelated errors. Therefore, we may estimate informally the regression functions $\beta(t)$ by least squares estimation. Technically this is because $\hat{B}(t)$ is a consistent estimator whereas $\hat{\beta}(t)$ is not, and the martingale theory touched on above applies to $\hat{B}(t)$: subject to mild conditions $\sqrt{n}(\hat{B}(t) - B(t))$ converges in distribution to a mean zero $(p+1)$-dimensional Gaussian martingale with a variance matrix that can be estimated easily from the data.

3 Analysis and results

This work arose out of issues encountered in a large-scale sanitation programme carried out from 1997 to 2004 in Salvador, Brazil, which involved three longitudinal studies designed to evaluate the impact of environmental sanitation measures on the health of the population. For the analysis, we used 21 areas (similar socioeconomic and sanitary conditions, called sentinal areas) common to all three studies, and also the same covariates. A day with diarrhoea was defined as the occurrence of three or more liquid or loose stools starting when the child woke in the morning (Morris et al., 1994). For brevity in this report we will concentrate on diarrhoea prevalence as the outcome variable of interest. Prevalence is defined as the probability that a child has diarrhoea on a given day, and prevalent days form the recurrent events. To investigate the effect of the various socioeconomic and demographic factors, particularly sanitation arrangements, and as the three studies varied by calendar and follow up periods, we used an age rather than calendar time scale to adjust all three cohorts in the same structure on scale: 6 months to 36 months (Age). The advantage of the additive regression model is that the plots of cumulative regression functions have a direct interpretation despite possibly complex effects. Periods when a cumulative regression function decreases implies that the covariate reduces the risk of the event happening over those ages, and an increasing cumulative regression function implies the opposite. We have assumed in this analysis that covariate effects are common across phases. Separate analyses of each phase indicate that this is a reasonable assumption. Our

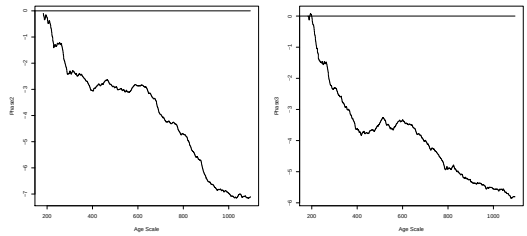


FIGURE 1. Cumulative regression function estimates.

results showed an overall reduction in diarrhoea as age increases (not included) and also across the phases (Figure 1).

4 Conclusion

In the reported analysis we focused on the additive regression model for recurrent events data. This method presents several advantages when dealing with this kind of data. First, the model is very simple as it is of linear form and therefore the computations are almost instant, as we can use least squares estimation. Second, it is easy to assess the effect of covariates on the event of interest, even for complex time varying effects of time varying covariates. Third, powerful martingale theory supports inference and model checking. A disadvantage is that there is no constraint for intensity estimates to be non-negative. On the other hand the more natural logistic regression approach can be unstable when events are rare, and the use of the Firth correction may lead to biased predictions, especially when, as here, we sum over a large number of ages.

References

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New-York.

Borgan, O., Fiaccone, R.L., Henderson, R., and Barreto, M.L. (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. *Scandinavian Journal of Statistics*, **34**, 53 – 69.

Fosen, J., Borgan, O., Weedon-Fekaer, H., and Aalen, O.O. (2006). Dynamic analysis of recurrent event data using the additive model. *Biometrical Journa*, **48**, 381 – 398.

Morris, S.S., Cousens, S.N., Laneda, C.F., and Kirkwood, B.R. (1994). Diarrhoea-defining the episode. *Scandinavian Journal of Statistics*, **23**, 617 – 623.

Time-varying cointegration via wavelets

Eder Lucio da Fonseca[1], Airlane Pereira Alencar[1], Pedro Alberto Morettin[1]

[1] Institute of Mathematics and Statistics – IME-USP, Sao Paulo, Brazil

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

Abstract: In recent decades, interest in literature on the subject of cointegration increased expressively. Traditional models that address this issue assumes that the cointegration vector does not vary over time. However, there is evidence in literature that this assumption may be considered too restrictive. Using the concept of wavelets, we propose a vector error correction model wherein is allowed to the cointegration vector vary over time.

Keywords: Time-varying VEC models; Wavelets; Dynamic cointegration.

1 Introduction

Since the seminal studies of Engle and Granger (1987) and Johansen (1988), the interest regarding cointegration increased significantly in literature. However, traditional models consider the assumption that the cointegration vector does not change over time (Lütkepohl, 2005). As pointed by Bierens and Martins (2010), this assumption may be considered very restrictive. Hence, the main objective of the article will be, unlike traditional cointegration models, analyze the economic assumption of cointegration varying over time. Similar to Bierens and Martins (2010), we propose to model the cointegration vector temporally, via decomposition by wavelets.

2 Time-varying vector error correction models representation with wavelets

Assuming that $\mathbf{\Pi}_t = \boldsymbol{\alpha}\boldsymbol{\beta}_t^{\mathrm{T}}$ with fixed $\boldsymbol{\alpha}$, consider a time-varying vector error correction model of order p (TV-VECM(p)) with Gaussian errors, without intercepts or time trends

$$\Delta \mathbf{Y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{Y}_{t-1} + \mathbf{\Gamma}\mathbf{X}_t + \mathbf{u}_t, \tag{1}$$

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

where T is the number of observations, $t = 1, \ldots, T$ and $\mathbf{u}_t = (u_{1t}, \ldots, u_{nt})^{\mathrm{T}}$, with $\mathbf{u}_t \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_u)$. For the $n \times 1$ vector time series $\mathbf{Y}_t$, we supposed that there are fixed $r < n$ linearly independent columns of the time-varying $n \times r$ matrix $\boldsymbol{\beta}_t = (\beta_{1t}, \ldots, \beta_{nt})^{\mathrm{T}}$ of cointegrating vectors. We can represent the components β_{it} of $\boldsymbol{\beta}_t$, $i = 1, \ldots, n$, by a linear combination of wavelet (mother) and scaling (father) functions as (Nason, 2008)

$$\beta_{it} = \sum_{k \in Z} c_{j_0,k} \phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty} \sum_{k \in Z} d_{j,k} \psi_{j,k}(t), \tag{2}$$

where $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ respectively represents the functions of wavelet and scale functions of the chosen base. In turn, c and d terms represents the coefficients of the linear combination. Using (2), it is possible to rearrange Equation 1, such that

$$\Delta \mathbf{Y}_t = \boldsymbol{\alpha} \mathbf{W}_J^{\mathrm{T}} \mathbf{Y}_{t-1}^{(J)} + \boldsymbol{\Gamma} \mathbf{X}_t + \mathbf{u}_t,$$

where

$$\begin{aligned}
\boldsymbol{\Gamma} &= [\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_{p-1}], \\
\boldsymbol{X}_t &= [\Delta \mathbf{Y}_{t-1}^{\mathrm{T}}, \ldots, \Delta \mathbf{Y}_{t-p+1}^{\mathrm{T}}]^{\mathrm{T}}, \\
\mathbf{W}_J^{\mathrm{T}} &= [\mathbf{c}_{0,0}^{\mathrm{T}}, \ldots, \mathbf{d}_{J,2^J-1}^{\mathrm{T}}], \\
\mathbf{Y}_{t-1}^{(J)} &= [\phi(t)\mathbf{Y}_{t-1}^{\mathrm{T}}, \ldots, \psi_{J,(2^J-1)}(t)\mathbf{Y}_{t-1}^{\mathrm{T}}]^{\mathrm{T}}.
\end{aligned}$$

3 Numerical results

Here we present the estimation of a specific simulated TV-VECM(p) model, namely

$$\begin{aligned}
Y_{1t} &= \beta_{2t} Y_{2,t-1} + u_{1t} \\
Y_{2t} &= Y_{2,t-1} + u_{2t}
\end{aligned}$$

with $\boldsymbol{\beta_t} = [1, \beta_{2t}]$. Note that the $\boldsymbol{\beta_t}$ component varies with time only through the component β_{2t}.

3.1 Considering just one change on $\boldsymbol{\beta_{2t}}$ over time

In this case, the component β_{2t} receives only two distinct values over time. In other words, we have two regimes for the cointegration relations, i.e.

$$\boldsymbol{\beta}_{2t} = \begin{cases} \beta_{21} & \text{if } 1 \leq t \leq \frac{T}{2} \\ \beta_{22} & \text{if } \frac{T}{2} < t \leq T. \end{cases}$$

We estimate the components of the β_{2t} vector via conditional maximum likelihood in three situations, as can be seen in the Table 1. Note that, in each case, we satisfactorily estimate the correct value of β_{2t}, even in the case where this vector does not vary with time (traditional cointegration).

TABLE 1. Considered cases to simulate the components of the β_{2t} vector.

β_{2t}	β_{21}	β_{22}	Description
Case 01	-1	+1	Change in signal
Case 02	-1	-2	Change in scale
Case 03	-1	-1	Traditional cointegration

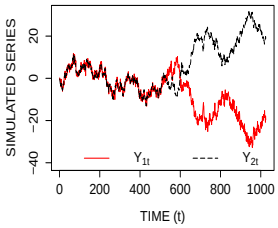


FIGURE 1. Case 01.

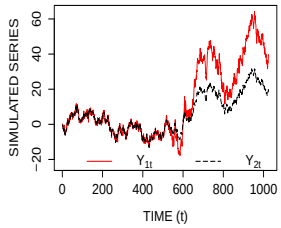


FIGURE 2. Case 02.

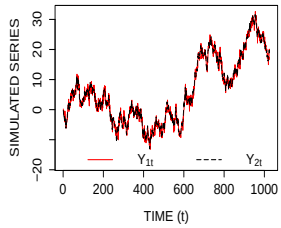


FIGURE 3. Case 03.

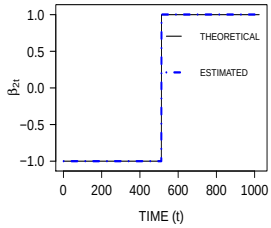


FIGURE 4. Case 01.

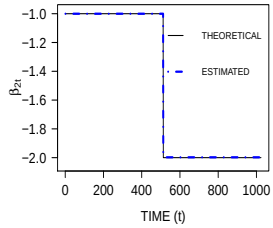


FIGURE 5. Case 02.

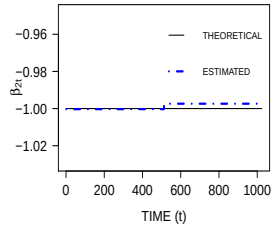


FIGURE 6. Case 03.

3.2 Allowing vector β_{2t} freely vary with time

In this case, we consider the component β_{2t} receiving distinct values over time. Thus, we can envisage other forms to the cointegration vector varying with time, as can be seen in the Table 2. Like the first case, we estimate the components of β_{2t} vector via conditional maximum likelihood.

TABLE 2. Considered cases to simulate and estimate the β_{2t} vector.

β_{2t}	Description
Case 04	Three distinct values over time
Case 05	Seasonal behaviour
Case 06	Linear trend
Case 07	Logarithmic trend

Note that, we also satisfactorily estimate the correct value of β_{2t}.

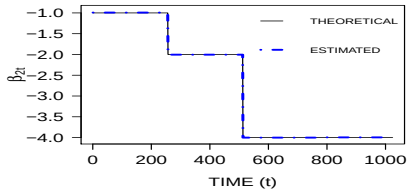


FIGURE 7. Case 04.

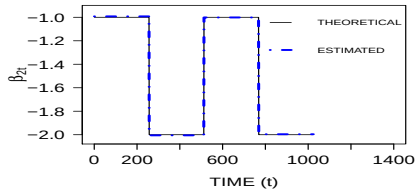


FIGURE 8. Case 05.

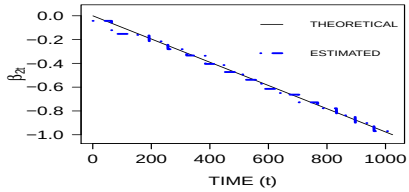


FIGURE 9. Case 06.

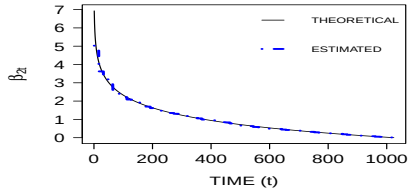


FIGURE 10. Case 07.

4 Concluding remarks

Considering $\boldsymbol{\beta}_t$ which varies with time permits a natural generalization of traditional cointegrated models. Combining wavelets and conditional maximum likelihood estimation proved to be a powerful tool to model the time-varying cointegration vector, especially when it assumes exotic behaviour over time. Using this method, we also can estimate other parameters of the model along with likelihood ratio tests.

Acknowledgments: The first author is especially grateful to CAPES and IME-USP for financial support.

References

Bierens, H.J. and Martins, L.F. (2010). Time-varying cointegration. *Econometric Theory*, **26**, 1453 – 1490.

Engle, R.F. and Granger, C.W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, **55**, 251 – 276.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, **12**, 231 – 254.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, New York.

Nason, G. (2008). *Wavelet Methods in Statistics with R*. Springer, New York.

Model selection and model averaging using restrictive strategies for imputation in linear models

Khuneswari Gopal Pillay[1], John H. McColl[1]

[1] University of Glasgow, United Kingdom

E-mail for correspondence: k.gopal-pillay.1@research.gla.ac.uk

Abstract: Model selection introduces additional uncertainty into the model-building process, but the standard errors of parameter estimators obtained from the selected model by standard statistical procedures will under-estimate the true variability. Model averaging aims to incorporate the uncertainty associated with model selection into parameter estimation, by combining estimates over a set of possible models. We have carried out a series of Monte Carlo experiments to compare the theoretical properties of model selection and model averaging procedures for multiple regression models when some covariate values are missing and have to be imputed. A restrictive strategy (where minimal use is made of auxiliary variables in both prediction and imputation models) and a strategy using non-overlapping variable sets (where the auxiliary variable is only used in the imputation model) were investigated. The mean square error of prediction at a standard grid of points in covariate space was used to compare the predictive ability of model selection and model averaging procedures. The results showed that model averaging often provides better prediction than the best model obtained through model selection. It is advisable to use model averaging with a restrictive strategy, as opposed to non-overlapping variable sets, to make predictions in the presence of missing data.

Keywords: Restrictive; On-overlapping; Auxiliary variable; Single imputation.

1 Introduction

Model selection is well-known for introducing additional uncertainty into the model-building process (Buckland et al., 1997) but the properties of standard parameter estimates obtained from the selected model do not reflect the stochastic nature of model selection. This problem can be more

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

severe in the presence of missing data (Wood et al., 2008). In the literature, model averaging has been proposed as an alternative to model selection which is intended to overcome the under-estimation of standard errors that is a consequence of model selection. The main objective of this paper is to compare model selection and model averaging in imputed data sets in the context of missing data, in terms of prediction, using a restrictive strategy and non-overlapping variable sets. Collins et al. (2001) defined a restrictive strategy as including few or no auxiliary variables in both the imputation and prediction models, where auxiliary variables are defined as variables that are included in an analysis solely to improve the performance of missing data procedures. A strategy of using non-overlapping variable sets (an extremely restrictive strategy) is defined as not including auxiliary variables in the prediction model. In contrast, an inclusive strategy includes numerous auxiliary variables and overlapping variable sets in both the imputation and prediction models.

2 Design of simulation

All data were simulated from the following multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where Y is the response variable, X the explanatory variables, β the parameters of the model/coefficients, ε a normal error term and n the number of observations. $\mathbf{X}$ (X_1, X_2, and later X_3) values, where X_3 is an auxiliary variable, were simulated from a multivariate normal distribution with fixed zero means and a specified covariance matrix. ρ_{23} denotes the correlation between X_2 and X_3, $\rho_{23} = -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75$, and $\rho_{12} = \rho_{13} = 0$. The Y values were created based on (1), the simulated X_1 and X_2 values and error terms simulated from $N\left(0, \sigma_\varepsilon^2\right)$ where $\sigma_\varepsilon^2 = \frac{1}{16}, 1, 16$. In all simulations, $\beta_0 = \beta_1 = \beta_2 = 1$. Some X_2 values were deleted from the simulated data set completely at random. Simulations were carried out for every combination of sample size ($n = 50, 100, 200, 400$), missing percentage ($m = 0, 25$ and 50) and covariance matrix. The "norm.nob" imputation method in the MICE package (van Buuren and Groothuis-Oudshoorn, 2011) was used to impute any missing observations of X_2 using the auxiliary variable X_3. The imputation model used in both the restrictive strategy and non-overlapping variable sets was

$$X_{2i} = \hat{\varphi_0} + \hat{\varphi_3} X_{3i} + h_i. \tag{2}$$

$\hat{\varphi_0}$ and $\hat{\varphi_3}$ were estimated from the complete cases using least squares, and h_i was a random error from $N\left(0, \hat{\sigma}_h^2\right)$. For the restrictive strategy, eight possible prediction models were considered based on all possible subsets of variables X_1, X_2 and X_3. Whereas, for the non-overlapping strategy, four

possible prediction models were considered based on all possible subsets of variables X_1 and X_2 only. Model averaging was carried out across these sets of models using AICc based weights (Buckland et al., 1997). Model selection was carried out using AICc as the model-selection criterion (Claeskens and Hjort, 2008).

3 Results and discussion

Figure 1 shows the comparison between the MSE(P) values obtained using the restrictive strategy and non-overlapping variable sets on model averaging and model selection for each ρ_{23}, missing percentage, sample size ($n = 50$ and $n = 400$) and $\sigma_\varepsilon = 0.25, 1, 4$. In both the restrictive strategy and using non-overlapping variable sets, the MSE(P) for the selected best model and model averaging decreases as sample size increases and, generally, as $|\rho_{23}|$ increases. The negative and positive correlations of same magnitude showed similar results of MSE(P) for model selection and model averaging using the restrictive strategy and using non-overlapping variable sets. As the error variance increases, the MSE(P) for the selected best model and model averaging increases. MSE(P) for model averaging is lower than for model selection both in complete data sets and after imputation of missing values, and the difference is bigger for smaller sample sizes. However, the MSE(P) for model selection using non-overlapping variable sets is lower than using a restrictive strategy. A similar simulation study was carried out using an inclusive strategy in order to compare it with the restrictive strategy. However, the results showed that there is no clear difference between the restrictive and inclusive strategies in these simple contexts.

4 Conclusion

In conclusion, researchers need to clearly define the purpose of their model-building process with missing data and decide whether they are most interested in identifying which variables to include in the model for the response or in making the best predictions. If the interest of the research with missing data is to identify which variables to be included when making predictions, one should use model selection with non-overlapping variable sets (using the auxiliary variable only in the imputation model). On the other hand, it is advisable to use model averaging with restrictive strategies (using the auxiliary variable in both the imputation and prediction models) to make predictions. Researchers can use an auxiliary variable with high positive correlation to improve imputation. However, an auxiliary variable with a high negative correlation can degrade the performance of model selection, especially when there is relatively large error variance in the response.

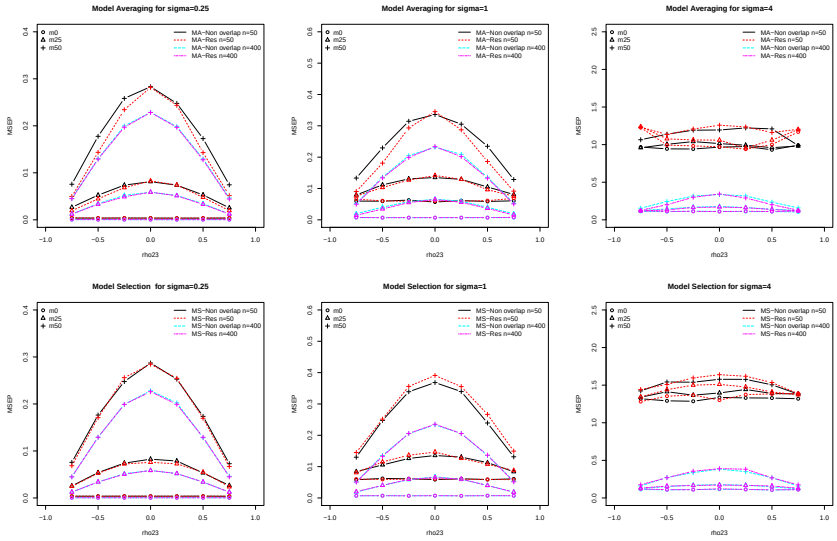


FIGURE 1. Comparison between the restrictive strategy and non-overlapping variable sets on model averaging and model selection for σ_ε.

References

Buckland, S.T., Burham, K.P., and Augustin, N.H. (1997). Model selection: An integral part of inference. *Biometrics*, **53**, 603 – 618.

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Collins, L.M., Schafer, J.L., and Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, **6**, 330 – 351.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, **45**, 1 – 31.

Wood, A.M., White, I.R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, **27**, 3227 – 3246.

Slope heuristics for multiple change-point models

Yann Guédon[1]

[1] CIRAD, UMR AGAP and Inria, Virtual Plants, Montpellier, France

E-mail for correspondence: guedon@cirad.fr

Abstract: With regard to multiple change-point models, much effort has been devoted to the selection of the number of change points. But, the proposed approaches are either dedicated to specific segment models or give unsatisfactory results for short or medium length sequences. We propose to apply the slope heuristic, a recently proposed non-asymptotic penalized likelihood criterion, for selecting the number of change points. In particular we apply the data-driven slope estimation method, the key point being to define a relevant penalty shape. The proposed approach is illustrated using two benchmark data sets.

Keywords: Data-driven slope estimation; Latent structure model; Model selection; Multiple change-point detection.

1 Introduction

The slope heuristics were introduced by Birgé and Massart (2001) as a new non-asymptotic penalized likelihood criterion for model selection. They showed that there exists a minimal penalty such that the dimension of models (and the associated estimator risk) selected with lighter penalties becomes very large. Moreover, they proved that considering a penalty equal to twice this minimal penalty allows to select a model close to the best possible (or oracle) model in terms of estimator risk. This approach has been recently popularized by the introduction of the data-driven slope estimation method by Baudry et al. (2012) which is a practical method for implementing slope heuristics. In the maximum likelihood estimation framework, this practical method is based on the expectation of a linear relation between the penalty shape (a function of the model dimension) and the maximized log-likelihoods for overparameterized models. We focus here on the application of the slope heuristics for selecting the number of change points in

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

multiple change-point models.

2 Defining the log-likelihood function and the penalty shape for multiple change-point models

For multiple change-point models, the two possible log-likelihood functions are:

- $\log f(\mathbf{s}^*, \mathbf{x}; J)$, the log-likelihood of the most probable segmentation $\mathbf{s}^*$ in J segments (the number of change points is therefore $J-1$) of the observed sequence $\mathbf{x}$. Lebarbier (2005) used this log-likelihood function for defining a slope heuristic for Gaussian models.
- $\log f(\mathbf{x}; J)$, the log-likelihood of all the possible segmentations in J segments of the observed sequence $\mathbf{x}$ with $f(\mathbf{x}; J) = \sum_{\mathbf{s}} f(\mathbf{s}, \mathbf{x}; J)$.

These log-likelihoods for $K = 2, \ldots, J$ can be exactly computed by a single application of a dynamic programming algorithm for $\log f(\mathbf{s}^*, \mathbf{x}; J)$ and a smoothing algorithm for $\log f(\mathbf{x}; J)$ (Guédon, 2013). The slope estimation relies on maximized log-likelihoods computed for overparameterized models and, as shown in Guédon (2013, 2015), the most probable segmentation is often meaningless for overparameterized models. Consistently with our view of multiple change-point models as latent structure models (Guédon, 2013; 2015), we will thus focus on the log-likelihood of all the possible segmentations $\log f(\mathbf{x}; J)$ for defining a slope heuristic. We investigated on several data sets this log-likelihood over the range of J values corresponding to overparameterized models and noted that this log-likelihood function is markedly concave for overparameterized models if $J < T$ (e.g. $10 < T/J < 100$), where T is the sequence length, but far less if $J \ll T$.
To apply the slope heuristics, it is required that (Baudry et al., 2012):
(C1) The log-likelihood increases with J.
(C2) The penalty shape $\text{pen}_{\text{shape}}(J)$ increases with J.
To these two standard requirements, we add the two following specific requirements for multiple change-point models:
(C3) The penalty shape $\text{pen}_{\text{shape}}(J)$ depends on the sequence length T. Adding a segment for say $J = T/10$ entails a smaller increase of the penalty shape than adding a segment for $J \ll T$.
(C4) The first-order differenced penalty shape $\text{pen}_{\text{shape}}(J) - \text{pen}_{\text{shape}}(J-1)$ decreases with J. This decrease is not a linear function of J.
For the definition of the penalty shape, our starting point was $\text{pen}_{\text{shape}}(J) = \log n_J$ where $n_J = \binom{T-1}{J-1}$ is the number of possible segmentations in J segments. Consider the limiting case where all the segmentations are equally probable for a fixed J, then $\log f(\mathbf{x}; J) = \log \gamma_J + \log n_J$. In fact for overparameterized models, as shown in Guédon (2013, 2015) using different examples, $\log f(\mathbf{x}; J)$ decomposes into a structural part corresponding to

true change points and a noise part that increases with J. To respect the monotonicity of $\text{pen}_{\text{shape}}(J)$ as a function of J, we finally propose

$$\text{pen}_{\text{shape}}(J) = \log\left\{\frac{T^{J-1}}{(J-1)!}\right\},$$

with

$$\text{pen}_{\text{shape}}(J) - \text{pen}_{\text{shape}}(J-1) = \log\left(\frac{T}{J-1}\right).$$

3 Illustrations on benchmark data sets

The use of the proposed slope heuristic for multiple change-point models is illustrated using two benchmark data sets corresponding to different segment models and sequence lengths. The slope heuristic is compared with the "exact"ICL criterion proposed by Rigaill et al. (2012).

3.1 British coal mining disasters

The data consist of the dates of 191 coal mining disasters between 1851 and 1962 summarized as annual counts during the 112-year period. We assume that the number of disasters in any year has a Poisson distribution, and the underlying Poisson distribution parameter is piecewise constant through time.

TABLE 1. British coal mining disaster data: comparison of the exact ICL criterion and the slope heuristics (SH) with J as the penalty shape ($\text{pen}_{\text{shape}}0$) and the proposed penalty shape ($\text{pen}_{\text{shape}}1$). The criterion value and the corresponding posterior model probability $P(\mathcal{M}_J|\mathbf{x})$ are given for each J.

J	1	2	3	4	5
ICL_J	−413.14	−358.70	−362.31	−369.34	−375.74
$P(\mathcal{M}_J\|\mathbf{x})$	0	0.855	0.141	0.004	0
SH_J ($\text{pen}_{\text{shape}}0$)	−419.68	−358.81	−357.04	−358.55	−361.01
$P(\mathcal{M}_J\|\mathbf{x})$	0	0.2	0.49	0.23	0.07
SH_J ($\text{pen}_{\text{shape}}1$)	−407.72	−360.19	−368.05	−377.00	−385.38
$P(\mathcal{M}_J\|\mathbf{x})$	0	0.98	0.02	0	0

The slopes were estimated over the range $J = 6, \ldots, 20$ and, we obtained a residual standard deviation of 1.05 with the naive penalty shape J instead of 0.04 with the proposed penalty shape. Both the exact ICL criterion and the slope heuristic with the proposed penalty shape favour 2 segments while the slope heuristic with the naive penalty shape J favours 3 segments and puts weight on 4 segments which is not consistent with the outputs of the different validation approaches shown in Guédon (2013, 2015); see Table 1.

3.2 Well-log data

The data consist of 4050 measurements of the nuclear-magnetic response of underground rocks. The underlying signal is roughly piecewise constant, with each segment relating to a single rock type that has constant physical properties. We estimated Gaussian change in the mean and variance models on the basis of these data.

TABLE 2. Well-log data: comparison of the exact ICL criterion and the slope heuristic (SH) with the proposed penalty shape. The criterion value and the corresponding posterior model probability $P(\mathcal{M}_J|\mathbf{x})$ are given for each J.

J	15	16	17	18	19	20
ICL_J	−69355.4	−69330.0	−69316.3	−69309.7	−69309.6	−69313.3
$P(\mathcal{M}_J\|\mathbf{x})$	0	0	0.014	0.378	0.403	0.063
SH_J	−69479.6	−69461.8	−69466.0	−69478.6	−69482.8	−69495.4
$P(\mathcal{M}_J\|\mathbf{x})$	0	0.89	0.11	0	0	0

The slope was estimated over the range $J = 30, \ldots, 80$. The exact ICL criterion favours 18 and 19 segments while the slope heuristic with the proposed penalty shape favours mainly 16 segments; see Table 2. The 16-segment model selected by the slope heuristic is far more consistent with the analysis of the segmentation space presented in Guédon (2013) than the 18- or 19-segment models selected by the exact ICL criterion.

References

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, **22**, 455 – 470.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, **3**, 203 – 268.

Guédon, Y. (2013). Exploring the latent segmentation space for the assessment of multiple change-point models. *Computational Statistics*, **28**, 2641 – 2678.

Guédon, Y. (2015). Segmentation uncertainty in multiple change-point models. *Statistics and Computing*, **25**, 303 – 320.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, **85**, 717 – 736.

Rigaill, G., Lebarbier, E., and Robin, S. (2012). Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, **22** 917 – 929.

Making Bayesian bivariate meta-analysis practice friendly

Jingyi Guo[1], Håvard Rue[1], Andrea Riebler[1]

[1] Norwegian University of Science and Technology, Department of Mathematical Sciences, Trondheim, Norway

E-mail for correspondence: jingyi.guo@math.ntnu.no

Abstract: The use of priors can stabilise inference in bivariate meta-analysis, so that Bayesian inference has recently become attractive. However, Bayesian analysis is often computationally demanding and a well-motivated prior for the covariance matrix of the bivariate random effects is crucial. Integrated nested Laplace approximations provide an efficient solution to the computational issues, but the important question about prior elicitation remains. We apply the new penalised complexity prior framework to derive priors for the variance parameters and the correlation parameter. This allows us an intuitive specification of the hyperpriors based on interpretable contrasts. Using a simulation study we show that the new priors perform better than previously suggested priors in terms of sharpness, MSE and the proper Dawid-Sebastiani score. All methodology is implemented in the new user-friendly R-package `meta4diag` which provides a GUI for easy model specification and can be downloaded from R-forge `http://meta4diag.r-forge.r-project.org`.

Keywords: Bayesian analysis; Bivariate random effects; Integrated nested Laplace approximation; Meta-analysis; Penalized complexity priors.

1 Introduction

A diagnostic test study usually presents a two-by-two table from which pairs of sensitivity and specificity can be computed. A bivariate meta-analysis summarises the results from separately performed studies while keeping the two-dimensionality of the data (Chu and Cole, 2006). Since the number of studies is often small and data may be sparse, maximum likelihood estimation can be challenging. Paul et al. (2010) proposed to perform full Bayesian inference using integrated nested Laplace approximations (INLA) (Rue et al., 2009). Harbord (2011) noted that INLA has

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

considerable promise for use in routine analysis as no MCMC sampling is involved, but that it is hard to specify suitable prior distributions.
Recently penalised complexity (PC) priors have been proposed (Simpson et al., 2014), which allow the user to specify hyperparameters using intuitive contrasts about the parameter the prior is defined for. Here, we apply this approach to derive interpretable prior distributions for the precision and, in particular, the correlation parameter. The methodology is available to the applied scientist in the R-package `meta4diag` providing a GUI.

2 Bayesian bivariate meta-analysis

Let TP, FP, TN and TN denote the number of true positives, false positives, true negatives, and false negatives, respectively. Further, let $\text{Se} = \text{TP}/(\text{TP}+\text{FN})$ be sensitivity and $\text{Sp} = \text{TN}/(\text{TN}+\text{FP})$ specificity. A bivariate model summarises the results of several diagnostic studies $i = 1, \ldots, I$ by modelling sensitivity and specificity jointly:

$$\begin{aligned} &\text{TP}_i|\text{Se}_i \sim \text{Binomial}(\text{TP}_i + \text{FN}_i, \text{Se}_i), \quad \text{logit}(\text{Se}_i) = \mu + \mathbf{U}_i\alpha + \phi_i, \\ &\text{TN}_i|\text{Sp}_i \sim \text{Binomial}(\text{TN}_i + \text{FP}_i, \text{Sp}_i), \quad \text{logit}(\text{Sp}_i) = \nu + \mathbf{V}_i\beta + \psi_i, \\ &\begin{pmatrix}\phi_i \\ \psi_i\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_\phi^2 & \rho\sigma_\phi\sigma_\psi \\ \rho\sigma_\phi\sigma_\psi & \sigma_\psi^2\end{pmatrix}\right], \end{aligned} \tag{1}$$

where μ, ν are intercepts for $\text{logit}(\text{Se}_i)$ and $\text{logit}(\text{Sp}_i)$, respectively, and $\mathbf{U}_i$, $\mathbf{V}_i$ are possibly available covariates vectors. The covariance matrix of the random effects parameters ϕ_i and ψ_i is parameterised using between-study variances σ_ϕ^2, σ_ψ^2 and correlation ρ (Chu and Cole, 2006).
Usually vague or mildly informative priors are used for σ_ϕ^2, σ_ψ^2 and ρ. Harbord (2011) proposed to use a stronger prior for ρ that is possibly not symmetric around zero, but defined around a (negative) constant ρ_0. Penalised complexity (PC) priors allow for such a specification. Thinking about ρ, let

$$\boldsymbol{\Sigma}_{base} = \begin{pmatrix}1 & \rho_0 \\ \rho_0 & 1\end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma}_{flexible} = \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}$$

denote the covariance matrices of the base model and the flexible model, respectively. The increased complexity introduced by $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{flexible})$ compared to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{base})$ can be measured by the Kullback-Leibler discrepancy (KLD). We assume a constant rate penalisation of the distance $d(\rho) = \sqrt{2\text{KLD}(\rho)}$ to both sides of $\rho_0 = -0.2$, say, leading to a two-sided exponential prior with parameter λ. Motivated by real studies we would like to distribute the density mass unequally to both sides leading to two rate parameters λ_{left} and λ_{right} which can be inferred by the user using interpretable and intuitive contrasts motivated from findings in comparable meta-analyses. Here, we use $\text{P}(\rho \geqslant -0.2) = 0.6$ and $\text{P}(\rho < -0.95) = 0.05$,

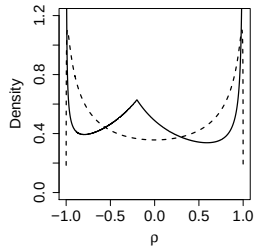


FIGURE 1. PC-prior (solid) for ρ compared to a $\mathcal{N}(0, 5)$ prior for the Fisher's z-transformed correlation (Paul et al., 2010) (dashed).

leading to $\lambda_{\text{left}} = 1.48$ and $\lambda_{\text{right}} = 0.99$. Figure 1 shows the obtained PC-prior for ρ and compares it with the prior used in Paul et al. (2010). Priors for σ^2_ϕ and σ^2_ψ are constructed analogously.

3 Simulation study

We performed a simulation study to compare the Bayesian approach with either PC-priors or the priors of Paul et al. (2010) to the frequentist approach `metandi` (Harbord and Whiting, 2009). For both Bayesian settings, we report summary statistics on the variance stabilising transformed parameters, which represent a more natural scale. Confidence intervals (CI) obtained by `metandi` were consequently transformed to the same scale. Table 1 shows the results for the transformed correlation parameter ρ in five different simulation settings, where the true correlation is either -0.6, -0.4, -0.2, 0 or 0.2. The Bayesian approach shows higher coverage probabilities compared to the frequentist approach and using the PC-prior the shortest average CI width is obtained over 1000 simulated datasets. Of note, when the true ρ is close to $\rho_0 = -0.2$, which corresponds to our base model, the empirical coverage is slightly higher than the nominal level but this is expected due to the concentration of the prior around -0.2. Although `metandi` gives a smaller bias, the MSE is larger than in both Bayesian approaches. To combine the measures of calibration and sharpness, we report the proper Dawid-Sebastiani score, where smaller values are preferred. Based on the mean DSS the PC-prior outperforms both other approaches. We conclude that the Bayesian approach is to be preferred over the frequentist approach. Further, the novel PC-priors show better behaviour than the prior proposed by Paul et al. (2009). Most importantly, the PC-prior approach allows the applied user to incorporate prior knowledge in a straightforward and intuitive way by specifying probability contrasts.

TABLE 1. Results for the Fisher's z-transformed correlation parameter. The 95%-coverage, the average 95% interval length (given in squared brackets), mean bias, mean squared error (MSE) and the mean Dawid-Sebastiani score (DSS) for 1000 simulated independent datasets from the bivariate model with μ =logit(0.8), ν =logit(0.7), $\sigma_\phi^2 = \sigma_\psi^2 = 1$ and different values of ρ are given.

True Value	Model	Coverage [CI-Length]	Bias	MSE	DSS
$\rho = -0.6$	Paul et al.	0.946 [2.25]	0.026	0.358	-0.142
	PC-prior	0.944 [2.23]	0.129	0.365	-0.084
	`metandi`	0.946 [2.30]	-0.101	0.440	-0.015
$\rho = -0.4$	Paul et al.	0.951 [2.12]	0.026	0.315	-0.245
	PC-prior	0.962 [2.01]	0.087	0.256	-0.479
	`metandi`	0.932 [2.15]	-0.052	0.379	-0.080
$\rho = -0.2$	Paul et al.	0.947 [2.06]	0.028	0.294	-0.297
	PC-prior	0.975 [1.94]	0.031	0.222	-0.657
	`metandi`	0.930 [2.09]	-0.008	0.350	-0.114
$\rho = 0$	Paul et al.	0.958 [2.05]	-0.003	0.264	-0.377
	PC-prior	0.974 [1.94]	-0.049	0.209	-0.627
	`metandi`	0.929 [2.07]	-0.003	0.316	-0.193
$\rho = 0.2$	Paul et al.	0.955 [2.06]	-0.025	0.288	-0.340
	PC-prior	0.955 [2.02]	-0.101	0.278	-0.373
	`metandi`	0.942 [2.09]	0.009	0.336	-0.170

References

Chu, H. and Cole S.R. (2006). Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology*, **59**, 1331 – 1333.

Harbord, R.M. (2011). "Commentary on: Multivariate meta-analysis: potential and promise". *Statistics in Medicine*, **30**, 2507 – 2508.

Harbord, R.M. and Whiting P. (2009). metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression *Stata Journal*, **9**, 211.

Paul, M., Riebler, A., Bachmann, L.M., Rue, H., and Held, L. (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Statistics in Medicine*, **29**, 1325 – 1339.

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

Simpson, D.P., Martins, T.G., Riebler, A., Fuglstad, G., Rue, H., and Sørbye, S.H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv:1403.4630*.

Approximate Bayesian computation for spatial extremes using composite score functions

Markus Hainy[1], Christopher C. Drovandi[2]

[1] Department of Applied Statistics, Johannes Kepler University, Linz, Austria
[2] Mathematical Sciences School, Queensland University of Technology, Brisbane, Australia

E-mail for correspondence: markus.hainy@jku.at

Abstract: Bayesian analysis of max-stable process models for spatial extremes is complicated by the fact that the likelihood function is usually intractable for more than a few locations. A potential solution to this problem is to resort to approximate Bayesian computation (ABC) techniques. However, using ABC immediately raises the question as to which summary statistic should be employed to compare the simulated pseudo-data to the observed data. Previous attempts have used the estimated extremal coefficients, often resulting in high-dimensional summary statistics. We instead propose to use the composite score vector derived from the pairwise likelihood representation of the intractable full likelihood. This reduces the dimensionality of the summary statistic to the number of unknown parameters. We perform an extensive simulation study to assess the utility of the composite score summary statistic compared to using the extremal coefficients.

Keywords: Max-stable processes; Approximate Bayesian computation; Composite score functions; Indirect inference.

1 Max-stable processes

Max-stable processes are a popular class of models for spatial extremes data. If it exists, the limiting distribution of the maximum of a suitably normalized sequence of independent and identically distributed (multivariate) random variables is in the family of multivariate extreme value distributions (MEVDs). Max-stable processes are the infinite-dimensional generalization of MEVDs. The D-dimensional marginal distribution of a max-stable process observed at D locations follows a D-dimensional MEVD. The uni-

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

variate margins are members of the univariate Generalized Extreme Value (GEV) family. To reduce complexity, it is common to consider max-stable processes with unit-Fréchet margins ($\Pr(Z \leq z) = \exp(-1/z)$). The analysis then focuses on the dependence structure of the max-stable process. The univariate GEV parameters can be estimated separately for each location in a first step, and the data can be transformed to unit-Fréchet scale using these estimates. For an overview of max-stable process models see e.g. Davison et al. (2012).

There are many ways to construct max-stable processes with unit-Fréchet margins, leading to different models such as the so-called Smith, Schlather, Brown-Resnick, or extremal-t model (see Davison et al., 2012). Unfortunately, for most of these models, it is infeasible to compute the likelihood function for more than a few locations. Therefore, classical or Bayesian inference based on the likelihood function would be impossible. However, the pairwise likelihoods can be used to construct a composite likelihood. The maximum composite likelihood estimator (MCLE) is a consistent estimator of the model parameters, see Ribatet et al. (2012).

2 ABC estimation

Several strategies have been proposed up to this point to perform Bayesian inference for max-stable process models. Ribatet et al. (2012) adjust the magnitude and curvature of the overly concentrated composite likelihood and use it instead of the true likelihood in a standard Bayesian estimation framework. Erhardt and Smith (2012), on the other hand, perform approximate Bayesian computation (ABC).

The basic idea of ABC is to simulate pairs of parameter values, $\{\theta_i\}_{i=1}^N$, and pseudo-data, $\{z_i\}_{i=1}^N$, from the model and to give higher weight to parameter samples with a small distance between the corresponding pseudo-data and the observed data, z_{obs}. In this way, a sample from an approximate posterior distribution is obtained. The tolerance level governs which discrepancies between pseudo-data and observed data, denoted by $d(z_i, z_{\text{obs}})$, are considered as "close" enough to warrant non-negligible weights. However, it is usually not advisable to compare the whole data set. Due to the curse of dimensionality, it would require a prohibitive amount of computing resources to maintain a desired approximation accuracy when the data are high-dimensional. Therefore, one usually seeks to find low-dimensional summary statistics of the data and to compare these, so $d(z_i, z_{\text{obs}}) = d(s(z_i), s(z_{\text{obs}}))$. For the max-stable process model, Erhardt and Smith (2012) define and estimate the tripletwise extremal coefficients between all triplets of locations and use them as summary statistics. The set of extremal coefficients is not a sufficient statistic for this model, but it nonetheless provides a very good summary of the dependence structure. However, since there are $\binom{D}{3}$ tripletwise extremal coefficients, Erhardt and Smith (2012) group them

into K clusters according to the shapes of the triangles formed by the location triplets (only depending on the locations, not on the data). Then they compute the means of the extremal coefficients within the K clusters, which results in a K-dimensional summary statistic. In their simulations with $D = 20$ locations, they use $K = 100$ clusters, which is still quite a large dimension for the summary statistic. We will denote the approach of using the extremal coefficients as summary statistics by ABC-ec.

3 Composite score function

If a composite likelihood function is available, Ruli et al. (2015) propose to use the score vector of the composite likelihood function as a summary statistic for ABC and call it ABC-cs (where cs denotes composite score). For models in the exponential family, the score vector is a sufficient statistic. Thus, the score vector is likely to provide informative summary statistics also for more general models, even in the case of composite likelihoods. The composite score vector is evaluated at the MCLE of the observed data, $\tilde{\theta}_{\text{obs}}$, so $s(z_{\text{obs}}, \tilde{\theta}_{\text{obs}}) = 0$. Ruli et al. (2015) show that using the discrepancy function $d(s(z_i, \tilde{\theta}_{\text{obs}}), s(z_{\text{obs}}, \tilde{\theta}_{\text{obs}})) = d(s(z_i, \tilde{\theta}_{\text{obs}}), 0) = s(z_i, \tilde{\theta}_{\text{obs}})^{\text{T}} J(\tilde{\theta}_{\text{obs}})^{-1} s(z_i, \tilde{\theta}_{\text{obs}})$, where $J(\tilde{\theta}_{\text{obs}})$ is the variance-covariance matrix of the composite score vector evaluated at the MCLE of the observed data, leads to approximate posterior distributions with the correct curvature. Furthermore, this discrepancy function is invariant to reparameterizations.

One major advantage of using the composite score vector as summary statistic is that the dimension of the statistic equals the dimension of the parameter space. Hence, if the model is parsimonious, the summary statistic will be of a low dimension. If the composite score function is available analytically, a second advantage is that the summary statistics are usually easy and fast to compute.

4 Simulation study

Our goal is to assess the utility of using the composite score vector as summary statistic for max-stable process models compared to using the extremal coefficients. In particular, we consider the Schlather (extremal Gaussian) model with a Whittle-Matérn correlation function. We performed our simulation studies for various settings, varying the number of locations, the number of repeated observations per location, and the true correlation parameters to produce strongly or weakly correlated observations. For each setting, 300 data sets were generated and ABC-cs and ABC-ec were applied to obtain approximate posterior samples for each data set. We used a slightly modified version of the adaptive population Monte Carlo (PMC)

algorithm developed by Lenormand et al. (2013). This simple but flexible PMC algorithm automatically reduces the tolerance level until the acceptance rate of newly simulated samples falls below some predefined threshold.

Mean squared errors of the posterior parameter samples and of the posterior means and posterior modes were computed and graphical analyses were made. These analyses often gave mixed results, with slight advantages for ABC-cs in most cases. However, the ultimate goal is to accurately estimate the dependence structure. For the Schlather model, the correlation function captures all information about the dependence structure. Furthermore, it is known that widely differing parameter values can lead to similarly shaped correlation functions. Therefore, we computed the integrated squared correlation difference (ISCD) for each sampled posterior parameter $\theta_i \in \{\theta_j\}_{j=1}^N$:

$$\mathrm{ISCD}_i = \int_0^\infty (\rho(h;\theta_i) - \rho(h;\theta_{\mathrm{true}}))^2 \mathrm{d}h,$$

where $\rho(h;\theta)$ denotes the correlation function with correlation parameters θ evaluated at distance h. This gives us a posterior sample of ISCDs for each of the 300 data sets, for which we can calculate summary statistics like the posterior mean or the posterior median. We observe that ABC-cs yields lower average posterior mean ISCDs (averaged over the 300 data sets) than ABC-ec in all settings that we considered. Using the integrated absolute correlation differences instead leads to the same conclusions.

References

Davison, A.C., Padoan, S.A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, **27**, 161 – 186.

Erhardt, R.J. and Smith, R.L. (2012). Approximate Bayesian computing for spatial extremes. *Computational Statistics and Data Analysis*, **56**, 1468 – 1481.

Lenormand, M., Jabot, F., and Deffuant, G. (2013). Adaptive approximate Bayesian computation for complex models. *Computational Statistics*, **28**, 2777 – 2796.

Ribatet, M., Cooley, D., and Davison, A.C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, **22**, 813 – 845.

Ruli, E., Sartori, N., and Ventura, L. (2015). Approximate Bayesian computation with composite score functions. *Statistics and Computing*, to appear.

On some issues on statistical analysis for Wilms tumor

Philipp Hermann[1], Milan Stehlík[12]

[1] Department of Applied Statistics, Johannes Kepler University Linz, Austria
[2] Departamento de Matemática, Universidad Técnica Federico Santa María, Chile

E-mail for correspondence: philipp.hermann@jku.at

Abstract: Standard mathematical procedures in the area of shape analysis are applied on Wilms tumor. A platonic body C60 is constructed for renal tumors with the aid of landmarks on the basis of 3D objects, computed from 2D magnetic resonance images. Differentiation between Wilms and non-Wilms groups is performed in addition to a pre and post clinical trial for chemotherapy patients.

Keywords: Shape analysis; Wilms tumor; Chemotherapy.

1 Introduction

Wilms tumor, also named nephroblastoma, is a cancer type which mainly affects the kidney of children under 7 years old. It is one of the three most frequent types of embryonal tumors in children (Ward et al., 2014). Genetic predisposition is suspected to increase the risk of nephroblastomas. Therefore, Wilms tumor is said to be one of the flagship-examples for abnormal developments which can be assigned to cancer predisposition in an organ (Schwab, 2001). The chance of curing cancer has been increasing in the last years, however, it is highly dependent on age, histological type, postoperative acuteness or volume of the tumor itself. Screenings every three months is advisable for risk groups in order to detect possible tumors as soon as possible (Ward et al., 2014).
Magnetic resonance images (MRI) deliver 2D images after being prescribed due to suspicion of Wilms tumor. Images of these screenings are basis for constructing a three dimensional object of the renal tumor as visible in Figure 1 taken from Giebel et al., (2012) and Hermann et al., (2015). Researchers have a huge interest to find markers for a good differentiation to avoid misclassifications (Giebel et al., 2012). Shape analysis allows to form

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

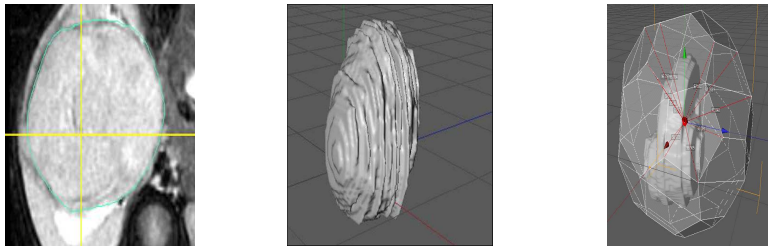


FIGURE 1. Left: transversal image of a renal tumor; middle: 3D Image of a renal tumor; right: surface of tumor and platonic body.

more-dimensional objects with the aid of mathematical procedures, characterizing objects on the basis of key points, called landmarks. However, standardization and centralization regarding size and position of the object allow comparing objects and differentiating between stages of tumors. In total 60 landmarks are taken as the cut-points between the surface of the tumor and the vector of the edge of the platonic body C60 (see Giebel et al., 2010, 2012). This platonic body, plotted in Figure 1, is formed on an explorative approach around the renal tumor with respect to minimization of Euclidean volume of the object.

2 Data analysis

The sample size consists of 40 patients compound of 30 patients affected with Wilms tumor, 7 with Non-Wilms and 3 missing values. Phrasing "Non-Wilms" represents patients with Wilms tumor of lower malignancy group, i.e. group I of four groups. The number of measured points in order to get the exact location of the landmarks varies to a greater extent (between 186 and 6638) due to the explorative approach based on geometric methods. Wilcoxon- and t-test yield evidence for statistical significance of recognized differences in descriptive statistics (mean, median) between the two groups. As a next step distributional differences between Wilms and non-Wilms patients are tested with the aid of a two-sample KS-test. In addition to p-values, density estimators with Gaussian kernel are given for all coordinates in Figure 2 indicating that significant differences concerning the distributions between the groups are observed.
Furthermore, we transform the data according to the three dimensional center point. Therefore, landmarks are reduced by their corresponding center value for every observation and coordinate. Due to this centralization approach, mean as well as median are very close to zero. Testing for differences between the means with the Wilcoxon-test does not deliver significant p-values. In contrast to that, the t-test would suggest that within x-coordinate different mean values are observed. Analogously to the

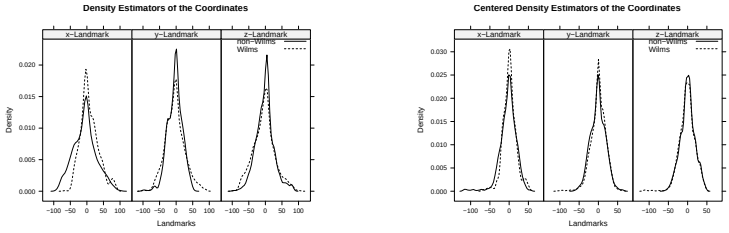


FIGURE 2. Left: density estimators of the coordinates and groups; right: density estimators of the centered coordinates and groups.

non-centered data, KS-test has been used to test for equal distribution, where only for x-coordinate the null-hypothesis of equal distributions was rejected. Centering the data leads to a loss in distributional differences between the groups. Former significant differences in the landmarks between the groups are removed due to this geometrical centralization, comparable to standardizations which are used e.g., to transform normal distributions to standard normal distributions.

3 Impact of chemotherapy

A pre and post clinical trial has been performed with 10 patients in order to test for developments of the landmarks before and after chemotherapy, resulting in 20 observations. For allowance of comparison of the size of the object, each landmark was centered according to its center value in advance. A decrease in the mean of the standard deviation can be observed for y- and z-coordinate after therapy. Moreover, mean and median are close to zero, however, minima and maxima are smaller for all coordinates after therapy, which allows the assumption that therapy could lead to a decrease of the volume of the object. In order to have a graphical comparison of the data before and after chemotherapy, density estimators were given in Figure 2(b) in Hermann et al. (2015). These figures show that therapy leads to less variance in the distribution for every coordinate, which strengthens the previous arising assumption. Testing for normality of the distributions before and after therapy has been performed in Hermann et al. (2015). The results of Shapiro test do not allow to reject normality for the y-coordinate before therapy as well as y- and z-coordinate after therapy. Although normality is still rejected for the x-coordinate, its p-value at least increases and a development in direction of normality can be assumed, among others because the shape of the density estimator is similar to the normal distribution.

Figure 3 contains 3D-scatterplots in order to compare the size of the object for three of the patients before and after therapy. Plots of the landmarks before therapy can be found in the first line and the corresponding plots af-

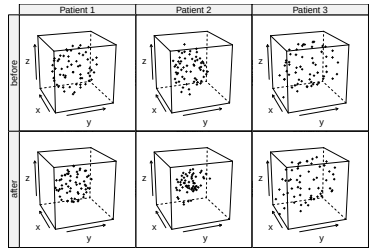


FIGURE 3. Scatterplot for patients before and after therapy.

ter therapy are beneath. Apparently, the volume of the object decreases for all displayed patients (all axes have same scale), but in different magnitude.

4 Conclusion

The standardly used mathematical procedure of constructing the platonic body C60 for the renal tumor allows to differentiate on the basis of distributions and means between the groups. Medical staff can be supported in the diagnosis decision process with this geometrical approach built on MRI, however, in its current stage the analysis is preliminary.

Acknowledgments: Milan Stehlík acknowledges Proyecto Interno 2015, REGUL. MAT 12.15.33, Modelaciòn del crecimiento de tejidos con aplicaciones a la Investigaciòn del càncer. Philipp Hermann acknowledges ANR project Desire FWF I 833-N18.

References

Giebel, S.M., Schiltz, J., Graf N., Nourkami, N., Leuscher, I., and Schenk, J.-P. (2012). Application of shape analysis on 3D images – MRI of renal tumors. *Journal of Iranian Statistical Society*, **11**, 131 – 146.

Giebel, S.M., Schiltz, J., and Schenk, J.-P. (2010). Application of shape analysis on renal tumors in 3D. In: *5th International Symposium on Health Informatics and Bioinformatics (HIBIT)*, 149 – 152.

Hermann, P., Giebel, S.M., Schenk, J.-P., and Stehlík, M. (2015). Dynamic shape analysis – before and after chemotherapy. Submitted to *International Conference on Risk Analysis (ICRA6/RISK 2015)*.

Schwab, M. (2001). *Encyclopedic Reference of Cancer*. Berlin-Heidelberg: Springer-Verlag.

Ward, E., DeSantis, C., Robbins, A., Kohler, B., and Jemal, A. (2014). Childhood and adolescent cancer statistics. *CA: A Cancer Journal for Clinicians*, **64**, 83 – 103.

GAMLSS applied to study bacterial cellulose yield

Freddy Hernández[1], Mabel Torres-Taborda[2], Lina Arteaga[2], Cristina Castro[2]

[1] Universidad Nacional de Colombia, Colombia
[2] Universidad Pontificia Bolivariana, Colombia

E-mail for correspondence: fhernanb@unal.edu.co

Abstract: The purpose of this research was to evaluate the effect of pH and time over mean and variance of bacterial cellulose yield using GAMLSS models. It was conducted a experiment with two factors and five levels per factor.

Keywords: GAMLSS; Bacterial cellulose yield.

1 Introduction

Cellulose can be produced by some bacterial species and it is expected to be a new commodity with diverse applications if production conditions could be improved. In this study, an experiment was conducted to analyze the effect of pH and cultivation time on the production of bacterial cellulose using the microorganism *Gluconacetobacter medellinensis*. In order to describe the behavior of the experimental units, different models were considered using GAMLSS models (Generalized Additive Model for Location Scale and Shape).

2 GAMLSS

Rigby and Stasinopoulos (2005) proposed GAMLSS models (Generalized Additive Model for Location Scale and Shape) that assume the response variable y_i, with $i = 1, \ldots, n$, are independent with probability density function $f(y_i \mid \boldsymbol{\theta}_i)$ in which $\boldsymbol{\theta}_i = (\mu_i, \sigma_i, \nu_i, \tau_i)^{\mathrm{T}}$ corresponds the to parameter vector. The first two elements μ_i and σ_i are location and scale parameters while the remaining are shape parameters. GAMLSS allow that each

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

parameter could be a function of a set of explanatory variables and the distribution of random variable y_i is not limited to the exponential family (Stasinopoulos and Rigby, 2007). The GAMLSS models consider continuous as well as discrete distributions with different parameterizations for the same distribution of the response variable. Details of distributions and parameterizations used in GAMLSS can be found in (Rigby and Stasinopoulos, 2010, page 199).
Another advantage of GAMLSS models is that the models allow the use of fixed effects, random effects and (semi) parametric univariate in the specification regression models, where all the parameters of the assumed distribution for the response can be modeled as additive functions of explanatory variables.

3 Results

Figure 1 shows the density plot and boxplot for cellulose variable, from these plots we can observe that response variable is right-skewed with minimum 0.0181, median 0.0787, maximum 0.5707 and five points from 32 that seem like outliers. For these reasons it seems reasonable to use a skewed distribution to model the cellulose yield. Figure 2 shows the scatterplot for bacterial cellulose yield, pH and time. We observed that maximum bacterial cellulose yield was obtained with pH 3.5 and 13 days, it was noted that the yield tends to decreases as the pH increases and it tends to increases as the time increases too.

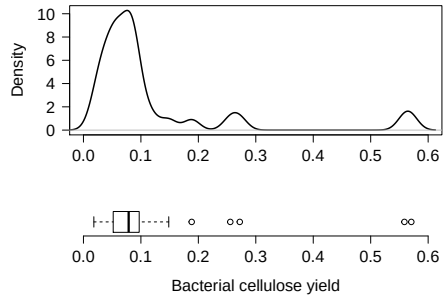


FIGURE 1. Density and boxplot for bacterial cellulose yield (g).

According to Figure 1, several models with gamma, log-normal and inverse Gaussian distribution were fitted for the response variable. For each model were considered polynomials up to grade two with pH and Time as covariates in the structure of μ and σ parameters. The preferred model was the one with the minimum AIC value (Akaike information criterion proposed by Akaike, 1973) that assumed gamma distribution for the response variable. The structure, covariates and results for this model are shown in

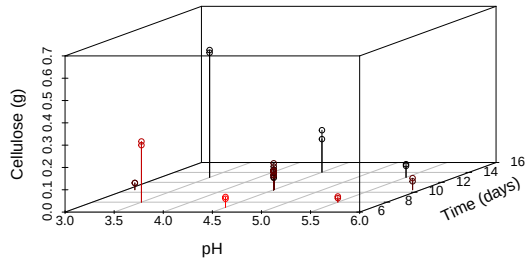


FIGURE 2. Scatterplot for bacterial cellulose yield (g), pH and time (days).

Table 1. From Table 1 we obtained the estimated expressions for μ and σ parameters as:

$$\log(\hat{\mu}) = -1.45 - 0.65\text{pH} + 0.18\text{Time} \tag{1}$$

$$\log(\hat{\sigma}) = 1.58 - 0.56\text{pH}. \tag{2}$$

Due to the fact that parameterization considered for gamma distribution, the mean and variance of cellulose yield can be expressed as:

$$\hat{E}(Y) = \hat{\mu} = e^{-1.45-0.65\text{pH}+0.18\text{Time}} \tag{3}$$

$$\hat{Var}(Y) = \hat{\mu}^2\hat{\sigma}^2 = e^{0.26-2.42\text{pH}+0.36\text{Time}}. \tag{4}$$

From the above expressions it was observed that for each additional day, at a fixed value of pH, the mean of bacterial cellulose yield increases in 19.72% (obtained from $e^{0.18} = 1.1972$); similarly, if we fixed the time, then for each additional unit of pH, the variance decreases in 91.11% (obtained from $e^{-2.42} = 0.0889$). Figure 3 shows a plot of estimated mean and variance for several values of time. From this figure we observed that mean and variance for cellulose yield decrease as pH increases, the opposite occurs for mean and variance when time increases.

TABLE 1. Parameter estimates.

$\log(\mu)$ model	Estimate	Std. Error	t value	P-value
Intercept	-1.45	0.58	-2.52	1.785e-02
pH	-0.65	0.09	-7.43	5.468e-08
Time	0.18	0.03	5.36	1.173e-05
$\log(\sigma)$ model	Estimate	Std. Error	t value	P-value
Intercept	1.58	0.57	2.79	0.0095142
pH	-0.56	0.12	-4.54	0.0001058

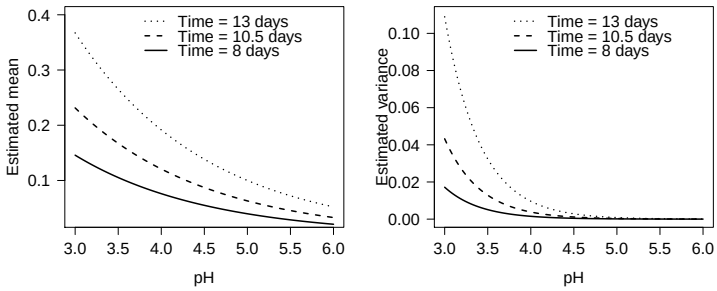


FIGURE 3. Estimated mean and variance for three time values.

4 Conclusions

The results found here agree with Castro et al. (2012) that conclude that optimal bacterial cellulose yield is found near pH 3.5. The two explanatory variables used in the model were significant to explain the mean and variance of bacterial cellulose yield, the equations 3 and 4 could be used to model the system behavior to those conditions and allow describing the variability of bacterial cellulose yield.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory*, Budapest: Akademia Kiado, 267 – 281.

Castro, C. et al. (2012). Bacterial cellulose produced by a new acid-resistant strain of Gluconacetobacter genus. In: *Carbohydrate Polymers*, **89**, 1033 – 1037.

Rigby, R. and Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**, 507 – 554.

Rigby, R. and Stasinopoulos, D. (2010). Instructions on how to use the gamlss package in R. http://gamlss.org/images/stories/papers/gamlss-manual.pdf.

Stasinopoulos, D. and Rigby, R. (2007). Generalized additive models for location, scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1 – 46.

A new inverse regression model and software for radiation biodosimetry

Manuel Higueras[1 2], David Moriña[3], Pedro Puig[2], Liz Ainsbury[1], Kai Rothkamm[1]

[1] Public Health England, UK
[2] Universitat Autònoma de Barcelona, Spain
[3] Centre for Research in Environmental Epidemiology (CREAL), Spain

E-mail for correspondence: Manuel.Higueras-Hernaez@phe.gov.uk

Abstract: New Bayesian-type count data inverse regression methods are introduced, with applications in cytogenetic radiation biodosimetry. A new R package to work with these models has been developed.

Keywords: Bayesian calibration; Biological dosimetry; Radiotherapy.

1 State of the art

Biological dosimetry is essential for the timely determination of the radiation dose received by an exposed individual in the event of a radiation accident or unplanned exposure. Biodosimetry relies on the relation between the amount of damage induced by radiation at a cellular level, e.g. counting micronuclei. The frequency of chromosome aberrations is a established biological indicator of radiation dose received.

1.1 Classical methodology and Bayesian alternative

Calibration curves are produced by irradiating n blood samples from a healthy donor with several doses x_i, $i = 1, \ldots, n$. Then, for each dose, n_i cells are analysed and the counts of observed chromosomal aberrations y_{ij}, $j = 1, \ldots, n_i$ are recorded. For the dicentric assay it is usually assumed that the counts y_{ij} follow a Poisson distribution whose population mean is a function of x_i and a set of parameters β, i.e. $\mathrm{E}(y_{ij}) = f(x_i; \beta)$. The parameters of this regression model are usually estimated by maximum likelihood, and the MLE and its estimated variance-covariance matrix are

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

calculated and recorded. Therefore, in a case of suspected overexposure, given a blood sample from an irradiated patient, m lymphocytes are scored obtaining the counts $\tilde{y}_1, \ldots, \tilde{y}_m$. The classical approach to estimate the received dose x and its confidence limits is to use the inverse regression method described as a standard procedure in the IAEA manual (2011). Groer and Pereira were the first to investigate the use of Bayesian models in chromosome dosimetry neutron exposure, and since then several researchers have used Bayesian methods in radiation biodosimetry. They presented a method for the Poisson model for the simple case where $\mathrm{f}(x;\beta) = \beta x$ using a Gamma distribution as a prior for β. A review of Bayesian methods in biodosimetry can be found in Ainsbury et al. (2013).

2 Original contributions

In Higueras et al. (2015), given a Poisson regression model the posterior of the population mean is approximated to a Normal distribution, using the asymptotic normality of the posterior distribution for large samples and the delta-method, i.e.

$$\mu|x \sim \mathrm{N}\left((\mathrm{f}(x,\hat{\beta}),\hat{\delta}), \nabla \cdot \hat{\Sigma} \cdot \nabla^{\mathrm{T}}\right),$$

where $\hat{\Sigma}$ is the estimated covariance matrix and ∇ is the gradient of $\mathrm{f}(x;\beta)$. This asymptotic Normal posterior distribution of the population mean can be approximated to a Gamma distribution, and the resulting calibrative density for solving the inverse problem is expressible in terms of a mixture of a Poisson distribution with a Normal (or Gamma) distribution,

$$\mathrm{f}(x|\tilde{y}) \propto \mathrm{p}(x) \int \mathrm{L}(\tilde{y}|\mu)\mathrm{P}(\mu|x)\mathrm{d}\mu,$$

i.e., the Hermite (or Negative Binomial) distribution. A new `R` package entitled `radir` (Moriña et al., 2015) for the models presented here is available at CRAN repository, see http://cran.r-project.org/web/packages/radir/.

2.1 Example: analysis of doses in thyroid cancer patients

Serna et al. (2008) studied chromosomal damage in lymphocytes of thyroid cancer patients after radioiodine treatment. The authors performed a micronucleus assay in binucleated cells of blood samples from 25 patients 3 days after internal Iodine-131 exposure. The *in vitro* calibration curve was fitted by a linear-quadratic model with intercept, $\mathrm{f}(x;\beta) = G\beta_2 x^2 + \beta_1 x + \beta_0$ according to Poisson's law, and the estimate of β_0 was not taken into account, because the authors in Serna et al. (2008) argued that the intercept could change for each patient. The constant G is the Lea-Catcheside generalized dose-protraction factor, being $G = 1$ for the *in vitro* assay. Taking

into account the characteristics of the Iodine-131 treatment, the authors in Serna et al. (2008) found the factor G to be close to 0.1. Then β_0, the background of each patient, can be estimated counting the micronuclei of the patient from a blood sample taken before the treatment. This leads to the fitted regression model $\mathrm{f}(x;\hat{\beta})$ with a covariance matrix that incorporates the variance of $\hat{\beta}_0$ without correlation with $\hat{\beta}_1$ and $\hat{\beta}_0$. To illustrate the novel techniques presented here, the dose received by Patient 1 in Serna et al. (2008) is estimated. The patient presented 13 cells with just one micronucleus each in a total of 500 cells.

TABLE 1. Statistical summary of the calibrative dose densities for both Uniform dose priors.

Prior dose distribution	Mode	Mean	SD	95% HPD
$\mathcal{U}(0,2)$	1.140	1.141	0.481	(0.319, 2.000)
$\mathcal{U}(0,4.5)$	1.140	1.561	0.858	(0.055, 3.281)

Therefore, $\mu|x$ will be considered to follow a distribution with $\mathrm{f}(x,\hat{\beta}) = G\hat{\beta}_2 x2 + \hat{\beta}_1 x + \hat{\beta}_0$ and variance $v(x,\hat{\beta}) = \nabla \cdot \hat{\Sigma} \cdot \nabla^{\mathrm{T}}$, where $\nabla = (1, x, Gx^2)$. In this case a Gamma mean prior is preferred instead of a Normal, due to constraint reasons. For a Gamma mean prior, the predictive posterior distribution represents the probability of a Negative Binomial random variable taking a value of 13 counts, with mean $4.810 \cdot 10^3 x^2 + 0.177x + 0.130$ and variance $4.326 \cdot 10^6 x^4 + 2.036 \cdot 10^4 x^3 + 9.922 \cdot 10^3 x^2 + 0.177x + 0.133$. Two calibrative densities have been calculated applying two different Uniform prior dose distributions, from 0 to 2 and 4.5 Gy. Figure 1 shows the plot of both densities of the estimated dose for the test data. Their statistics are indicated in Table 1. These results agree with those displayed in Serna et al. (2008), where the dose estimate for Patient 1 was 1.14 Gy.
This result can be reproduced with very simple `R` commands using `radir`.

```
> f <- expression(b0+b1*x+0.1*b2*x^2); pars <- c("b0","b1","b2")
> beta <- c(0.01, .0136, .0037)
> cov <- matrix(c(1.98e-05,0,0,0,.3121*10^(-4),-.0798*10^(-4),0,
+ -.0798*10^(-4), .0256*10^(-4)),nrow=3)
> ex.u1 <- dose.distr(f, pars, beta, cov, cells=500, dics=13,
+ prior.param=c(0, 2)); summary(ex.u1)
> ex.u2 <- dose.distr(f, pars, beta, cov, cells=500, dics=13,
+ prior.param=c(0, 4.5)); summary(ex.u2)
```

The Bayesian approach has a number of obvious advantages for cytogenetic radiation dose estimation, where high quality prior information is generally available and where the estimated dose is more correctly represented by a distribution of possible values, the calibrative density presented here. The above results demonstrate that the approach described is accurate and informative for practical cytogenetic dosimetry.

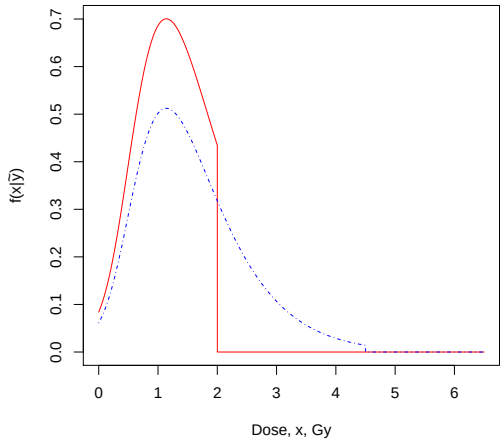


FIGURE 1. Calibrative dose densities from a $\mathcal{U}(0,2)$ (solid line) and a $\mathcal{U}(0,4.5)$ (dashed) prior dose.

References

Ainsbury, E.A., Vinnikov, V.A., Puig, P., Higueras, et al. (2013). Review of Bayesian statistical analysis methods for cytogenetic radiation biodosimetry, with a practical example. In: *Radiation Protection Dosimetry*. DOI: 10.1093/rpd/nct301.

Higueras, M., Puig, P., Ainsbury, E.A., and Rothkamm, K. (2015). A new inverse regression model applied to radiation biodosimetry. In: *Proceedings of the Royal Society A*. DOI: 10.1098/rspa.2014.0588.

I.A.E.A. (2011). *Cytogenetic Dosimetry: Applications in Preparedness for and Response to Radiation Emergencies*. Vienna: International Atomic Energy Agency.

Moriña, D., Higueras, M., Puig P., Ainsbury, E.A., and Rothkamm, K. `radir` package: An `R` implementation for cytogenetic biodosimetry dose estimation. *Journal of Radiological Protection*, (submitted).

Serna, A., Alcaraz, M., Navarro, J.L., Acevedo, C., Vicente, V., and Canteras, M. (2008). Biological dosimetry and Bayesian analysis of chromosomal damage in thyroid cancer patients. *Radiation Protection Dosimetry*, **129**, 372 – 380.

Modelling a proportion response variable using generalised additive models for location scale and shape

Abu Hossain[1], Robert A. Rigby[1], Dimitrios M. Stasinopoulos[1], Marco Enea[2]

[1] STORM, London Metropolitan University,UK
[2] University of Palermo, Italy

E-mail for correspondence: hossaina@londonmet.ac.uk

Abstract: In this paper two alternative approaches are proposed to model a response variable Y measured on the interval from zero to one, including both zero and one. The first proposed model employs a flexible four parameter distribution for $0 < Y < 1$, for example a logit skew exponential power distribution, inflated by including point probabilities at 0 and 1. The second proposed model is a generalised Tobit model, obtained from a flexible four parameter distribution on $(-\infty, \infty)$, for example the skew exponential power distribution, by censoring below 0 and above 1. The proposed models are applied to a real data set and compared with current popular models.

Keywords: GAMLSS; Generalised Tobit model; Logit skew exponential power distribution.

1 Introduction

The purpose of this paper is to provide two flexible modelling approaches for a proportion response variable measured on the interval from 0 to 1, including both 0 and 1, i.e. range $[0, 1]$. In the first approach a flexible distribution for Z with range $(-\infty, \infty)$ is transformed to Y with range $(0, 1)$, using an inverse logit transformation $Y = 1/(1+e^{-Z})$, which is then inflated by including point probabilities for Y at 0 and 1. Any available distribution on $(-\infty, \infty)$ within the `gamlss` package, Stasinopoulos and Rigby (2007), can be used for Z, for example the flexible four parameter skew exponential power (SEP) and skew student t (SST) distributions.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

The second approach is a generalised Tobit model, in which a flexible distribution on $(-\infty, \infty)$ is truncated below 0 and above 1 to provide range $0 \leq Y \leq 1$ with probabilities at 0 and 1. Again any available distribution on $(-\infty, \infty)$ including the SEP and the SST can be used for this purpose.

2 Inflated logitSEP distribution

2.1 Logit skew exponential power (logitSEP) distribution

Any distribution on range $-\infty < Z < \infty$ can be transformed to a restrictive range $0 < Y < 1$ by using an inverse logit transformation $Y = 1/(1+e^{-Z})$. The reason for the proposed model is that the usual beta distribution for $0 < Y < 1$ is often inadequate for modelling a proportion response variable. The inverse logit transformation of the skew exponential power distribution, Fernandez et al. (1995) called the logitSEP distribution, is introduced here to provide an improved model on the interval $(0, 1)$. Note that if $Z \sim \text{SEP}(\mu, \sigma, \nu, \tau)$ for $-\infty < Z < \infty$, then $Y = 1/(1 + e^{-Z}) \sim \text{logitSEP}(\mu, \sigma, \nu, \tau)$ for $0 < Y < 1$. The logitSEP distribution is created using the `gamlss` function `gen.Family()`, which allows any `gamlss` distribution with range $(-\infty, \infty)$, (e.g. SEP), to be transformed to a new `gamlss` distribution, (e.g. logitSEP), with range $(0, 1)$.

2.2 The logitSEP distribution, inflated at 0 and 1

The inflated logitSEP distribution is suitable for a proportion response variable on $0 \leq Y \leq 1$ that includes both 0 and 1. The inflated logitSEP distribution is a mixture of a logitSEP distribution for $0 < Y < 1$ and a Bernoulli distribution for Y at 0 or 1. The model includes three components: a discrete value 0 with probability p_0, a discrete value 1 with probability p_1 and a $\text{logitSEP}(\mu, \sigma, \nu, \tau)$ distribution on the unit interval $(0, 1)$ with probability $(1 - p_0 - p_1)$. The mixed (continuous-discrete) probability (density) function of $Y \sim \text{Inf.logitSEP}(\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$ is given by

$$f_Y(y|\mu, \sigma, \nu, \tau, \xi_0, \xi_1) = \begin{cases} p_0 & \text{if } y = 0 \\ p_1 & \text{if } y = 1 \\ (1 - p_0 - p_1) f_W(y|\mu, \sigma, \nu, \tau) & \text{if } 0 < y < 1 \end{cases}$$

for $0 \leq y \leq 1$, where $W \sim \text{logitSEP}(\mu, \sigma, \nu, \tau)$ has a logitSEP distribution with $-\infty < \mu < \infty$ and $\sigma > 0$, $\nu > 0$, $\tau > 0$ and $0 < p_0 < 1$, $0 < p_1 < 1$ and $0 < p_0 + p_1 < 1$. The parameters ξ_0 and ξ_1 are related to p_0 and p_1 by $\xi_0 = p_0/p_2$, $\xi_1 = p_1/p_2$, where $p_2 = 1 - p_0 - p_1$, so $\xi_0 > 0$ and $\xi_1 > 0$. Hence $p_0 = \xi_0/(1 + \xi_0 + \xi_1)$ and $p_1 = \xi_1/(1 + \xi_0 + \xi_1)$.

The default link functions relate the parameters $(\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$ to the predictors $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$, i.e.

$$\begin{aligned} \mu &= \eta_1 \\ \log \sigma &= \eta_2 \\ \log \nu &= \eta_3 \\ \log \tau &= \eta_4 \\ \log(p_0/p_2) = \log(\xi_0) &= \eta_5 \\ \log(p_1/p_2) = \log(\xi_1) &= \eta_6. \end{aligned}$$

The dependence of the predictors of the parameters (i.e. η_1 to η_6) on explanatory variables may be linear, nonlinear, non-parametric smooth, regression trees or neural network models. The parameter sets (μ, σ, ν, τ) and (ξ_0, ξ_1) are 'information' orthogonal. The inflated (inverse logit) transformed distributions (e.g. Inf.logitSEP) have the advantage of extra flexibility, in that the probabilities of Y at 0 and 1 are modelled independently of the distribution on $(0, 1)$, (e.g. logitSEP), but with the cost of introducing extra parameters (ξ_0, ξ_1) into the model. Note that the logit transformation is sensitive to values of Y very close to 0 or 1.

3 Generalised Tobit model

The original Tobit model for a response variable Y on $[0, 1]$ assumes that the response follows a normal distribution censored below 0 and above 1, Tobin (1958).

The generalised Tobit model on $[0, 1]$ requires censoring below 0 and above 1 of a flexible model response variable distribution on $(-\infty, \infty)$ for its positive probabilities at 0 and 1. Censoring refers to the transformation of observations outside the limiting interval to the border values, Hoff (2007). Here the values in the model distribution below 0 and above 1 are transformed to 0 and 1 respectively.

Let $Y_1 \sim D(\mu, \sigma, \nu, \tau)$ be a flexible uncensored distribution on $(-\infty, \infty)$. Let $Y \sim D_c(\mu, \sigma, \nu, \tau)$ be the corresponding distribution left censored below 0 and right censored above 1 with resulting range $[0, 1]$. Then

$$Y = \begin{cases} 0 & \text{if } Y_1 \leq 0 \\ Y_1 & \text{if } 0 \leq Y_1 \leq 1 \\ 1 & \text{if } Y_1 \geq 1. \end{cases}$$

Hence the (mixed continuous-discrete) probability (density) function of Y is given by

$$f_Y(y) = \begin{cases} P(Y_1 \leq 0) & \text{if } y = 0 \\ f_{Y_1}(y) & \text{if } 0 < y < 1 \\ P(Y_1 \geq 1) & \text{if } y = 1 \end{cases}$$

for $0 \leq y \leq 1$. In principle D can be any distribution on $(-\infty, \infty)$, for example the four parameter SEP or SST distribution given in Section 2.1. In the generalised Tobit models the probabilities of Y at 0 and 1 are directly related to the distribution between 0 and 1 and so are less flexible, but the model is more concise (i.e. parsimonious) in that it has less parameters. Also the Tobit model is not so sensitive to values of Y very close to 0 or 1.

4 Conclusion

This paper proposed two models for a proportion response variable measured on the interval $[0, 1]$ including 0 and 1. The first proposed model transforms a flexible four parameters distribution on $(\infty, -\infty)$, (e.g. SEP), to range $(0, 1)$, using an inverse logit transformation, and then extends the range to $[0, 1]$ by including point probabilities at 0 and 1, giving an inflated distribution (e.g. Inf.logitSEP). The second proposed model is a generalised Tobit model, obtained by censoring below 0 and above 1 a flexible four parameter distribution on $(\infty, -\infty)$, (e.g. SEP), giving range $[0, 1]$. The dependence of each of the parameters of the two proposed models on explanatory variables can be linear, nonlinear, non parametric smooth function, regression trees or neural network models. The models can be fitted in `R` using the packages `gamlss`, `gamlss.cens`, and `gamlss.inf`. The models are applied to a proportion response variable loss given default.

References

Fernandez, C., Osiewalski, J., and Steel, M.F. (1995). Modeling and inference with spherical distributions. *Journal of the American Statistical Association*, **90**, 1331 – 1340.

Hoff, A. (2007). Second stage dea: Comparison of approaches for modelling the DEA score. *European Journal of Operational Research*, **181**, 425 – 435.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics*, **54**, 507 – 554.

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location, scale and shape (gamlss) in r. *Journal of Statistical Software*, **23**, 1 – 46.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24 – 36.

Analysing ordinal categorical data by means of a generalized mixture model

Maria Iannario, Domenico Piccolo

[1] Department of Political Sciences, University of Naples Federico II, Italy

E-mail for correspondence: maria.iannario@unina.it

Abstract: The use of statistical methods for analyzing ordinal categorical data has increased dramatically, particularly in biomedical, social sciences, and marketing research. Several statistical methods for univariate and correlated multivariate categorical responses have been introduced. They are based on two main approaches: methods concerning latent variables linked to observed categorical responses via threshold models and methods which directly derive from a probability distribution. For the second approach a close relationship with the data generating process is required whereas for the first class of models more stringent fitting aptitudes are pursued. In this proposal we present a comprehensive mixture model which includes both paradigms and allows to unify most of the current proposals. The generalized mixture which we present is an useful framework to compare models, to discover unexpected similarities and to introduce new interesting distributions.

Keywords: Ordinal data; Mixture models; Data generating process.

1 Introduction

The starting point for devising a framework which encompasses some of the different proposals nowadays available in the literature for the analysis of univariate ordinal categorical data is the consideration that any model is or may be considered as a finite mixture of two components which we call *feeling* and *uncertainty*. The first is a combination of attractiveness, satisfaction, awareness whereas the second one is the summary of indecision, fuzziness, blurriness concerning the stochastic mechanism behind the discrete choices. These components have to be conveniently weighted by means of a mixture which is a convex combinations of different probability mass distributions.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

The detailed specification of the two components allows to introduce a comprehensive mixture. It includes the main frameworks for the analysis of ordinal responses as the case study shows.

2 Components of the generalized mixture model

Focusing on the notation, with respect to a given item $\mathcal{I}$, we denote as R_i the final response expressed by the i-th subject and by Y_i and V_i, $i = 1, \ldots, n$, the corresponding random variables for feeling and uncertainty, respectively. All these random variables are defined over the discrete support $\{1, \ldots, m\}$, for a known m. Let $\theta = (\beta, \mathbf{\Psi})$ the set of all parameters characterizing the distribution of $(R_1, \ldots, R_n)$ and $\mathbf{\Psi} = (\eta, \mathcal{T}_m)$ the parameters for the feeling component.
Available information on the subjects' characteristics (covariates) are collected in a matrix $\mathbf{T}$, and $\mathbf{T}^{(\pi)}$ and $\mathbf{T}^{(\Psi)}$ are submatrices of $\mathbf{T}$ containing subjects' covariates for uncertainty and feeling components, respectively.
Then, the basic formulation of the mixture approach is:

$$\Pr(R_i = j|\theta) = \pi_i \Pr(Y_i = j|\mathbf{t}_i^{(\Psi)}, \mathbf{\Psi}) + (1 - \pi_i) \Pr(V_i = j) \qquad (1)$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where $\pi_i = \pi(\mathbf{t}_i^{(\pi)}, \beta) \in (0, 1]$ has been introduced to weight the two components and $\mathbf{t}_i^{(\pi)} \in \mathbf{T}^{(\pi)}$. Here, Y_i and V_i are the random variables for the feeling and uncertainty, respectively. The case $\pi_i \equiv 0$ is logically possible if we are assuming that the selection process is controlled solely by indecision; however, this assumption rules out the identifiability of parameters concerning the feeling. Hereafter we prefer to set $\pi_i > 0$, $\forall i$. The probability distribution of uncertainty is assumed known on the basis of *a priori* assumptions and it does not require parameters. An increase of π_i implies a reduced impact of the uncertainty component; thus, the quantity $(1 - \pi_i)$ is a (normalized) measure of the uncertainty implied by the model.
If we concentrate the attention on the feeling Y, ordinal phenomena are genuinely observed or are derived by a continuous variable Y^* (a latent variable) which for convenience or necessity is examined in a discrete version by means of Y. In the first case the correspondence with integers is immediate (although the scale may be not metric) and the assumptions about a latent variable are not so relevant since we begin the statistical procedures with a direct consideration of Y (Piccolo, 2003). In the second case the discretization of Y is obtained by means of *cutpoints* which transform the continuous support of Y^* into a sequence of ordered bins concerning Y (McCullagh and Nelder, 1989).
Then, a classification of models for ordinal data commonly used in the literature derives from the information available on the set of cutpoints which we denote as $\mathcal{T}_m = \{-\infty = \tau_0, \tau_1, \ldots, \tau_{m-1}, \tau_m = +\infty\}$, where

$\tau_{j-1} < \tau_j$ for $j = 1, \ldots, m$. In this regard, we distinguish between unsupervised (classes I and II) and supervised discretization (class III). In the class I there are no cutpoints to define the ordered sequence, thus $\mathcal{T}_m \equiv \emptyset$. In this case, we directly formalize the probability mass distribution of Y. In the class II the cutpoints are defined on *a priori* basis since they are known or conventionally defined. Specifically, a density function for Y^* is transformed into a discrete version Y by means of some arbitrary convention. In the class III cutpoints are unknown and arbitrarily located on the support of Y^*. In this case, the mapping from Y^* to Y is obtained by an *ad hoc* specification of cutpoints which are determined by data during the estimation process.
In all situations, Y_i is a discrete random variable characterized by a structured probability distribution mass $Pr(Y_i = j \mid \mathbf{\Psi})$. However, this distribution may be reached in a direct way (case I), or by means of the probability distribution function $F_{Y_i^*}(.)$ of a continuous latent variable Y_i^* characterized by the η parameters and the knowledge (or the estimation) of the cutpoints in $\mathcal{T}_m$ necessary to discretize Y_i^* into Y_i (cases II or III).
In cases II and III, the cutpoints $(\tau_1, \ldots, \tau_{m-1})$ contained in $\mathcal{T}_m$ produce an exhaustive splitting of an ordinal variable with m categories.
Let p_j^V a fully specified probability distribution for the uncertainty component. Then, the *Generalized mixture model with uncertainty* (GEM) which summarizes the main models is defined as follows:

$$\Pr(R_i = j|\theta) = \pi_i \Delta G_j(\mathbf{t}_i^{(\eta)}; \mathcal{T}_m, \eta) + (1 - \pi_i) \underset{j}{\overset{V}{\Pr}}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, m$, where $\theta = (\beta, \mathbf{\Psi})$ and $\mathbf{\Psi} = (\mathcal{T}_m, \eta)$.
Then, the probability mass function for the feeling Y_i is

$$\Delta G_j(\mathbf{t}_i; \mathcal{T}_m, \eta) = \begin{cases} \Pr(Y_i = j|\eta) & \text{for a discrete distribution;} \\ F_{Y_i^*}(\tau_j; \eta) - F_{Y_i^*}(\tau_{j-1}; \eta) & \text{for a latent distribution.} \end{cases}$$

Table 1 shows an overview of models for ordinal data encompassed by the generalized mixture. Possible new candidates for the latent variable approach for both components characterized by few parameters of location, variability and/or shape which can be related to subjects' covariates by adequate link functions may be included. This framework may be used to compare models, discover similarities and introduce new interesting distributions that sometimes improve the fitting results. The following case study shortly shows the main results obtained by the implementation of the presented approach. Data concern the evaluation of Obama's political attitudes scored on a Likert scale, with $m = 7$ categories (where 1=completely unsatisfactory and 7=completely satisfactory). Sample data ($n = 710$) collected in 2012 include several subjects' covariates: *age*, which we consider with log transformation in the analysis, among them. Observe that the

TABLE 1. Main models encompassed by the generalized mixture.

Class	*DGP*	*Models*	*Variations*
I (*no* cutpoints)	`Discrete`	IHG	
Unsupervised discretization	`random`	SB	
	`variables`	CUB	VCUB
			HCUB
			LC-CUB
		CUBE	VCUBE
		CUB+*shelter*	GECUB
			CUB-DK
II (*known* cutpoints)	`Continuous`	CUN	
Unsupervised discretization	`variables`	D-BETA	
III (*estimable* cutpoints)	`Latent`	CUMULATIVE	*Logit*/*Probit*
Supervised discretization	`continuous`		*C-log-log*
	`variables`	CUP	(*idem*)

TABLE 2. Models fitted to the `ordinal` response (with covariates).

Models	$1-\hat{\pi}$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\log(\hat{\sigma})$	$\hat{\phi}$	*Log-lik*	*BIC*
IHG		−2.703	−0.414			−1180.3	2373.7
CUB	0.129	−1.548	−0.502			−1080.1	2179.9
D-BETA		0.247	0.127	−0.040		−1210.7	2447.7
POM			0.633			−1065.4	2176.7
CUP	0.090		0.713			*−1065.2*	2183.0
CUBE	0.088	−1.477	−0.486		0.066	−1071.4	*2169.1*

maximum of log-likelihood is obtained for a CUP model (with 8 parameters) whereas a comparable fit is obtained by a CUBE model (with just 4 parameters) which turns out to achieve the minimum BIC.

Acknowledgments: This research has been supported by FIRB 2012 project (code RBFR12SHVV) and the frame of SHAPE project in STAR Programme (CUP E68C13000020003).

References

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, (2nd ed.)*. London: Chapman & Hall.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85 – 104.

Dynamic nomogram in R

Amirhossein Jalali[1,2], Alberto Alvarez-Iglesias[2], John Newell[1,2]

[1] School of Mathematics, Statistics and Applied Mathematics, NUI Galway, Ireland
[2] HRB Clinical Research Faculty, NUI Galway, Ireland

E-mail for correspondence: a.jalali2@nuigalway.ie

Abstract: Translational Medicine promotes the convergence of basic and clinical research disciplines and the transfer of knowledge on the benefits and risks of therapies. In an analogous fashion we propose the concept of translational statistics (Newell and Hinde, 2014) to facilitate the integration of Biostatistics within clinical research and enhance communication of research findings in an accurate and accessible manner to diverse audiences (e.g. policy makers, patients and the media). As nomograms are a useful tool to achieve this aim, we present a new R package, "DynNom", to generate dynamic nomograms to aid this translational process.

Keywords: Translational statistics; Dynamic nomograms; Shiny package in R.

1 Introduction

As statistical inferential methods become more computational the models arising are increasingly complex and difficult to interpret. Translational Statistics is useful in helping to avoid the common mistakes which often arise when some statistical results are reported by non-statisticians. One example of this knowledge transfer occurs when modelling a binary outcome where the usual summary is an odds ratio. It has been argued that, when possible, a summary quoting the underlying probabilities is more informative than one based on ratios of odds or indeed of probabilities.

2 Nomograms

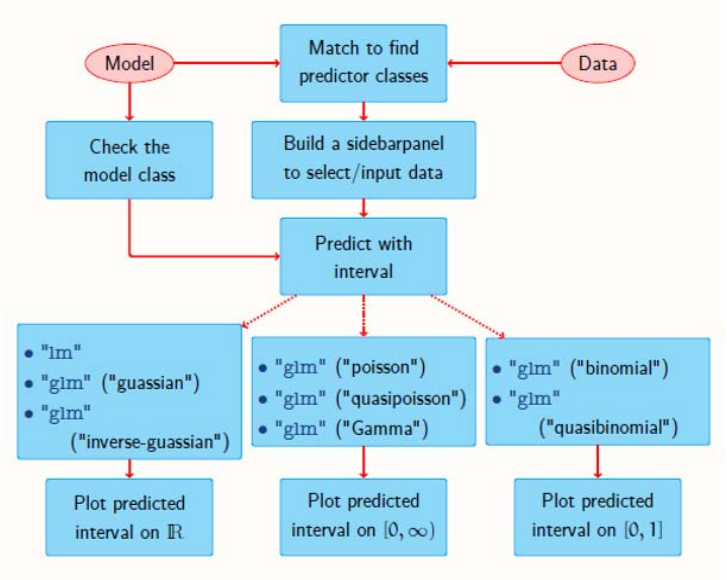
Nomograms are particularly popular in biomedical research to inform clinical decision making because of their ability to allow an individual calculate

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

a point estimate of a response variable, for a particular set of values for the explanatory variables, in the model of interest. A nomogram also provides a graphical summary of the relative importance of each predictor variable. Frank Harrell's rms library contains a function called nomogram for creating a (static) nomogram based on models supported by his rms library (Harrell 2001). The emergence of the rpanel (Bowman, 2007) and Shiny packages in R allow the creation of user-friendly graphical interfaces for R code to be displayed either in a standalone window (rpanel) or delivered as a webpage (Shiny).

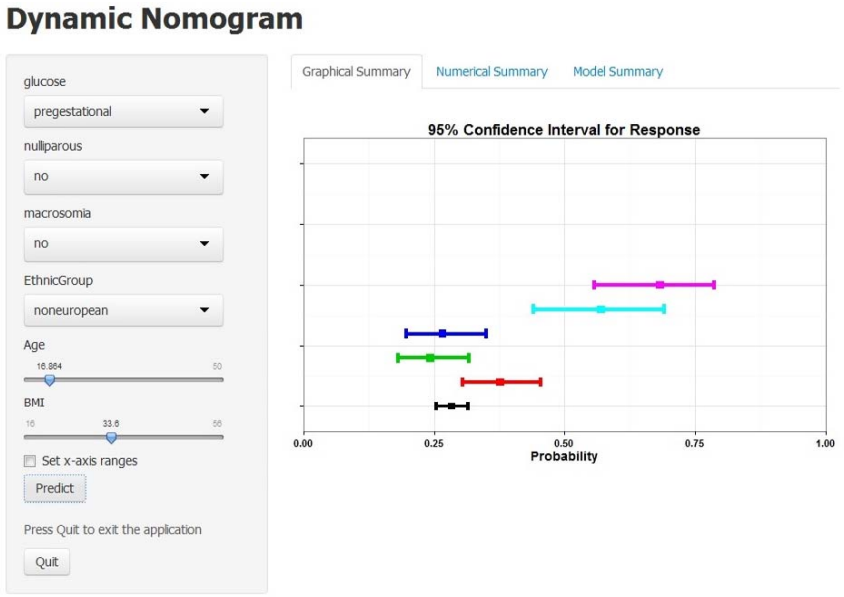
3 Dynamic nomograms (DynNom package)

In this paper I will introduce my DynNom R package which makes it possible to present the results of a lm or glm model object as a dynamic nomogram using Shiny that can be interacted with in a web browser. The package allows an investigation into the results of the proposed model, the relative importance of each explanatory variable (e.g. modifiable risk factors) and an assessment of model assumptions through accompanying model diagnostics.
The following flowchart shows the steps taken to create the DynNom package:



A Shiny application consists of two scripts, the user-interface, which controls the look and feel of the application, and the server script, which specifies the R actions. A reactive function is an R expression that uses widget

inputs and returns R outputs thereby updating the output whenever the original input changes.
The following screen shot depicts a DynNom object used to model a binomial outcome, namely the probability of a woman having a caesarian section:



It is conceivable that such a model, appearing in a scientific publication in medicine for example, could be accompanied with a URL directing the reader to the dynamic nomogram.
The next step in this project is to extend the package to include model objects created by linear and non-linear mixed models (e.g. lme4 and nlme) and models for time to event data (e.g. parametric survival and Cox proportional hazards models).

References

Bowman, A., Crawford, E., Alexander, G., and Bowman, R.W. (2007). rpanel, simple interactive controls for R functions using the tcltk package. *Journal of Statistical Software*, **17**.

Harrell, F.E. (2001). *Regression Modeling Strategies: With Application to Linear Models, Logistic Regression, and Survival Analysis*. Springer.

Kattan, M.W., Eastham, J.A., Wheller, T.M., Maru, N., Scardina, P.T., Erbersdobler, A., Graefen, M., Huland, H., Koh, H., Shariat, S.F.,

Slawin, K.M., and Ohori, M. (2003). Counseling men with prostate cancer: A nomogram for prediction presence of small, moderately differentiated, confined tumors. *The Journal of Urology*, **170**, 1792-1797.

Newell, J. and Hinde J. (2014). Translational statistics and dynamic Nnmograms. *Proceedings of the Conference on Applied Statistics in Ireland (CASI).*

Dynamic covariance estimation using sparse Bayesian factor stochastic volatility models

Gregor Kastner[1], Sylvia Frühwirth-Schnatter[1], Hedibert Freitas Lopes[2]

[1] WU Vienna University of Economics and Business, Austria
[2] Insper, São Paulo, Brazil

E-mail for correspondence: gregor.kastner@wu.ac.at

Abstract: We address the "curse of dimensionality" arising in time-varying covariance estimation by modeling the underlying volatility dynamics of a time series vector through a lower dimensional collection of latent dynamic factors. Furthermore, we apply a Normal-Gamma shrinkage prior to the elements of the factor loadings matrix, thereby increasing parsimony even more. Estimation is carried out via MCMC in order to obtain draws from the high-dimensional posterior and predictive distributions. To guarantee efficiency of the samplers, we utilize several ancillarity-sufficiency interweaving strategies (ASIS) for sampling the factor loadings. Estimation and forecasting performance is evaluated for simulated and real-world data.

Keywords: Shrinkage; Normal-gamma prior; Curse of dimensionality; Predictive distribution; Ancillarity-sufficiency interweaving strategy.

1 Introduction

We extend the standard factor stochastic volatility (SV) model (see, e.g., Chib et al., 2006, and the references therein) to allow for shrinkage of the loadings matrix towards zero. Within a Bayesian framework this can be achieved by employing the Normal-Gamma prior (Griffin and Brown, 2010) for the factor loadings. This hierarchical prior features excellent properties in the sense that factors which are irrelevant (exhibit little or no contribution to the likelihood) for certain components are effectively shrunk towards zero a posteriori. At the same time, the Normal-Gamma prior is flexible enough to cater for informative factors without "overshrinking". The model reads

$$\boldsymbol{y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\epsilon}_t, \qquad \boldsymbol{f}_t = \boldsymbol{V}_t^{1/2}\boldsymbol{u}_t,$$

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

where $\boldsymbol{y}_t$ denotes the vector of (potentially demeaned) log-returns of m observed time series at time t for $t = 1, \ldots, T$, $\boldsymbol{\Lambda}$ is an unknown $m \times r$ factor loadings matrix with elements Λ_{ij}, and $\boldsymbol{f}_t = (f_{1t}, \ldots, f_{rt})^{\mathrm{T}}$ represents the common latent factors at time t. $\boldsymbol{V}_t^{1/2} = \mathrm{Diag}(\exp(h_{1t}/2), \ldots, \exp(h_{mt}/2))$ denotes the latent idiosyncratic (i.e., component-specific) volatilities, and $\boldsymbol{V}_t^{1/2} = \mathrm{Diag}(\exp(h_{m+1,t}/2), \ldots, \exp(h_{m+r,t}/2))$ denotes the latent factor (i.e., common) volatilities. The errors $\boldsymbol{\epsilon}_t \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{I}_m)$ and $\boldsymbol{u}_t \sim \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{I}_r)$ represent i.i.d. m- respectively r-variate normal innovations with zero means and unit covariance matrices, where $\boldsymbol{\epsilon}_t$ and $\boldsymbol{u}_s$ are assumed to be pairwise independent for all $t, s \in \{1, \ldots, T\}$. Both the latent factors and the idiosyncratic shocks are allowed to follow independent SV processes, i.e.

$$h_{it} = \mu_i + \phi_i(h_{i,t-1} - \mu_i) + \sigma_i \eta_{it}, \quad \eta_{it} \sim \mathcal{N}(0, 1).$$

Following Griffin and Brown (2010), we substitute the usual factor loadings prior, $\Lambda_{ij} \sim \mathcal{N}(0, \tau^2)$ with τ^2 fixed, by a hierarchical Normal-Gamma prior,

$$\Lambda_{ij} | \tau_{ij}^2 \sim \mathcal{N}(0, \tau_{ij}^2), \quad \tau_{ij}^2 \sim \mathcal{G}(a_i, a_i \lambda_i^2/2), \quad \lambda_i^2 \sim \mathcal{G}(c_i, d_i),$$

with a_i, c_i, and d_i fixed. Choosing a_i small enforces strong shrinkage towards zero, while choosing a_i large imposes little shrinkage. Note that the Bayesian Lasso prior arises as a special case when $a_i = 1$. Univariate SV process priors are the same as in Kastner and Frühwirth-Schnatter (2014).

2 Identifiability and sampling efficiency

Without identifying the scaling of either the jth column of $\boldsymbol{\Lambda}$ or the variance of f_{jt}, the model is not identified. Aguilar and West (2000) assume that $\Lambda_{jj} = 1$, while the level μ_{m+j} of $h_{m+j,t}$ (which corresponds to the scaling of f_{jt}) is modeled to be unknown. Alternatively, one can fix the level μ_{m+j} at zero and leave the diagonal elements Λ_{jj} unrestricted. This is the baseline approach adopted in this paper.

It is fruitful to notice that by letting $\boldsymbol{\Lambda}^* := \boldsymbol{\Lambda} \times \mathrm{Diag}(\Lambda_{11}^{-1}, \ldots, \Lambda_{rr}^{-1})$ denote the restricted factor loadings matrix and $\boldsymbol{f}_{i\cdot}^* := \Lambda_{ii} \boldsymbol{f}_{i\cdot}$ denote the correspondingly transformed factor for $i = 1, \ldots, r$, one can easily move from one identification scheme to the other. This transformation can be exploited to substantially improve the usual Gibbs-sampler by utilizing ASIS (Yu and Meng, 2011) to redraw the factor loadings.

To illustrate the effectiveness of these simple reparameterizations, we consider simulated data. The top panel of Figure 1 exemplifies the output of the sampler for the first series' loading on the first latent factor without using any form of interweaving on the factor loadings. It stands out that even after a thinning of 100, posterior draws show extremely high autocorrelation. In fact, the extent of autocorrelation in these draws is so high that the dependence on the starting values is non-negligible, also after the long

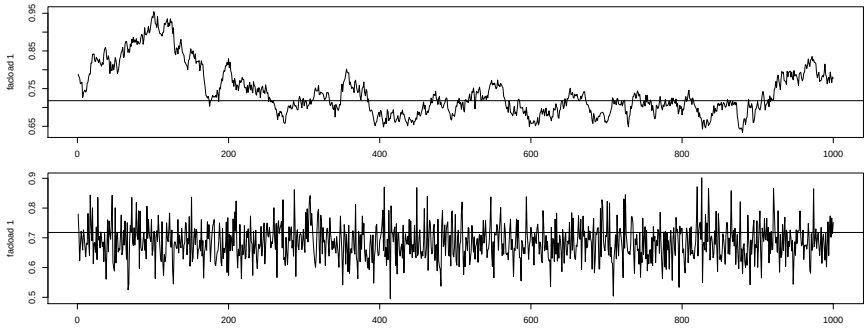


FIGURE 1. Trace plots for 1000 draws, exemplified for the first loading on factor one. Top panel: No interweaving. Bottom panel: ASIS. The draws are thinned; every 100th draw is displayed. Horizontal lines indicate data generating values.

burn-in of 50 000 draws. There seems to be little reason to believe that the sampler has converged at all. The bottom panel displays draws obtained from the sampler using ASIS, exhibiting practically no autocorrelation.

3 Prediction

To illustrate the feasibility of our approach for obtaining draws from the m-dimensional predictive distribution, we consider 20 exchange rates previously analyzed in Kastner et al. (2014). Figure 2 illustrates the bivariate marginals from the one-day-ahead predictive distribution on 2008-05-16 (arbitrarily chosen). Evaluating the predictive density at the observed value gives immediate rise to the predictive likelihood. Analogously to Kastner (forthcoming), this measure of forecasting accuracy can straightforwardly be used for model comparison.

References

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, **18**, 338 – 357.

Chib, S., Nardari, F., and Shephard, N. (2006). Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics*, **134**, 341 – 371.

Griffin, J.E. and Brown, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171 – 188.

Kastner, G. (forthcoming). Dealing with stochastic volatility in time series using the R package stochvol. *Journal of Statistical Software*.

Kastner, G. and Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics and Data Analysis*, **76**, 408 – 423.

Kastner, G., Frühwirth-Schnatter, S., and Lopes, H.F. (2014). Analysis of exchange rates via multivariate Bayesian factor stochastic volatility models. In: *The Contribution of Young Researchers to Bayesian Statistics – Proceedings of BAYSM2013*, Springer Proceedings in Mathematics & Statistics, **63**, Lanzarone, E. and Ieva, F. (Eds.), 181 – 186.

Yu, Y. and, Meng, X.-L. (2011). To center or not to center: that is not the question—An ancillarity-suffiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, **20**, 531 – 570.

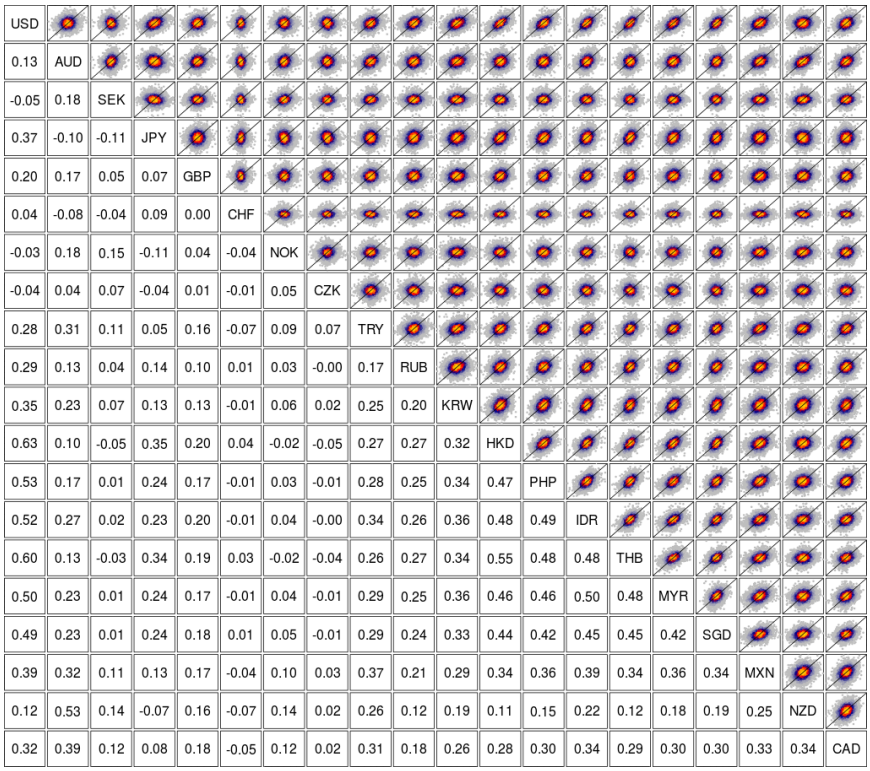


FIGURE 2. Pairwise scatterplots and empirical correlation coefficients of draws from the one-day-ahead predictive distribution for 2008-05-16, obtained from a six factor model.

Model based segmentation of the Czech hospitals according to their longitudinal financial performance in 2007 – 2011

Lenka Komárková[1], Táňa Hajdíková[1], Arnošt Komárek[2]

[1] Faculty of Management, University of Economics in Prague, Czech Republic
[2] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

E-mail for correspondence: komarkol@fm.vse.cz

Abstract: This contribution shows usage of a model-based clustering technique towards classification of Czech hospitals using their longitudinal financial performance (reached economic result) during the period 2007 – 2011.

Keywords: Classification; Generalized linear mixed model; Longitudinal data.

1 Introduction

Evaluation of performance of hospitals is a topic of interest in all countries with a functional system of public health care. Due to significant share of costs on consumed health care, the financial performance and effectiveness are then the most important quantities next to the quality ones. In this paper, it is our aim to analyze a financial performance of hospitals located in the Czech Republic during the period 2007 – 2011. Among other things, this period is interesting due to the fact that it followed a change of a legal form of many hospitals from a non-for-profit type of the health care provider to a standard business company (mostly of the type of the joint-stock company). Such changes were initiated by the regional governments in some parts of the country (Czech Republic is divided into 14 regions which are into some extent autonomous) in belief that increased flexibility of the business company (compared to the non-for-profit organization) will have a positive impact on the financial performance of the hospital expressed by its positive or balanced economic result.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

2 Data and research problems

Data regarding the financial performance of hospitals for the period 2007 – 2011 were obtained from publicly available annual reports and from the public register of companies. A primary (longitudinal) outcome is the economic result of the hospital in a particular year. Data for at least one of the considered years were available from 155 hospitals out of a total of 189 hospitals being registered in the Czech Republic. Our primary goal will be to use longitudinally observed economic result to find groups (segments) of hospitals with its similar development. Secondary, we would like to relate the clusters to some characteristics of the hospitals, in the first place, to their regional affiliation.

3 Model-based clustering

To proceed, we introduce some notation. Let $X_{i,t}$ denote the economic result (in CZK) of hospital i ($i = 1, \ldots, N$, $N = 155$) at year $2007 + t$ ($t = 0, \ldots, T$, $T = 4$). Segmentation of hospitals will be performed by a model-based clustering methodology of Komárek and Komárková (2013). To this end, we specify a mixture model for two outcomes derived from the values of the economic result $X_{i,t}$. The first derived outcome, denoted here as $Y_{i,t}$ takes into account both a numeric value of the economic result and its sign and is defined as $Y_{i,t} = \mathrm{sgn}(X_{i,t}) \log_{10}(1 + X_{i,t})$. The second derived outcome, $Z_{i,t}$, is only an indicator of a profit or balanced economy, i.e., $Z_{i,t} = 1$ if $X_{i,t} \geq 0$ and $Z_{i,t} = 0$ if $X_{i,t} < 0$.

In the mood of a model-based clustering, let $U_i \in \{1, \ldots, K\}$ denote a (hidden) allocation of the ith hospital ($i = 1, \ldots, N$) into one of K segments with $\mathsf{P}(U_i = k) = w_k$, $k = 1, \ldots, K$, where $\boldsymbol{w} = \big(w_1, \ldots, w_K\big)^{\mathrm{T}}$ are unknown proportions of the segments. Initially, K is assumed to be known. Given the ith hospital belongs to segment k, i.e., given $U_i = k$, the following multivariate generalized linear mixed model (MGLMM) is assumed for the derived outcomes $\big\{(Y_{i,t},\, Z_{i,t}) : t \in \mathcal{T}_i\big\}$, where $\mathcal{T}_i \subseteq \{0, \ldots, T\}$:

$$\left.\begin{aligned} Y_{i,t} \,|\, \mathbf{B}_i &\sim \mathcal{N}\big(B_{i,1} + B_{i,2}\, t,\, \sigma^2\big), \\ \mathsf{P}\big(Z_{i,t} = 1 \,\big|\, \mathbf{B}_i\big) = p_{i,t}(\mathbf{B}_i), &\qquad \mathrm{logit}\big\{p_{i,t}(\mathbf{B}_i)\big\} = B_{i,3}, \\ \mathbf{B}_i = \big(B_{i,1},\, B_{i,2},\, B_{i,3}\big)^{\mathrm{T}} &\sim \mathcal{N}_3\big(\boldsymbol{\mu}_k,\, \mathbb{D}_k\big). \end{aligned}\right\} \qquad (1)$$

The kth segment ($k = 1, \ldots, K$) is then characterized by (a) the mean $\boldsymbol{\mu}_k = \mathsf{E}\big(\mathbf{B}_i \,\big|\, U_i = k\big)$ and (b) the covariance matrix $\mathbb{D}_k = \mathsf{var}\big(\mathbf{B}_i \,\big|\, U_i = k\big)$ of the hospital specific intercept ($B_{i,1}$) and slope ($B_{i,2}$) of the trend of the transformed economic result $Y_{i,t}$ and of the hospital specific logit transformed probability ($B_{i,3}$) of making a profit.

A Bayesian approach implemented in the R package mixAK (Komárek and Komárková, 2014) is used to infer on unknown model parameters vector

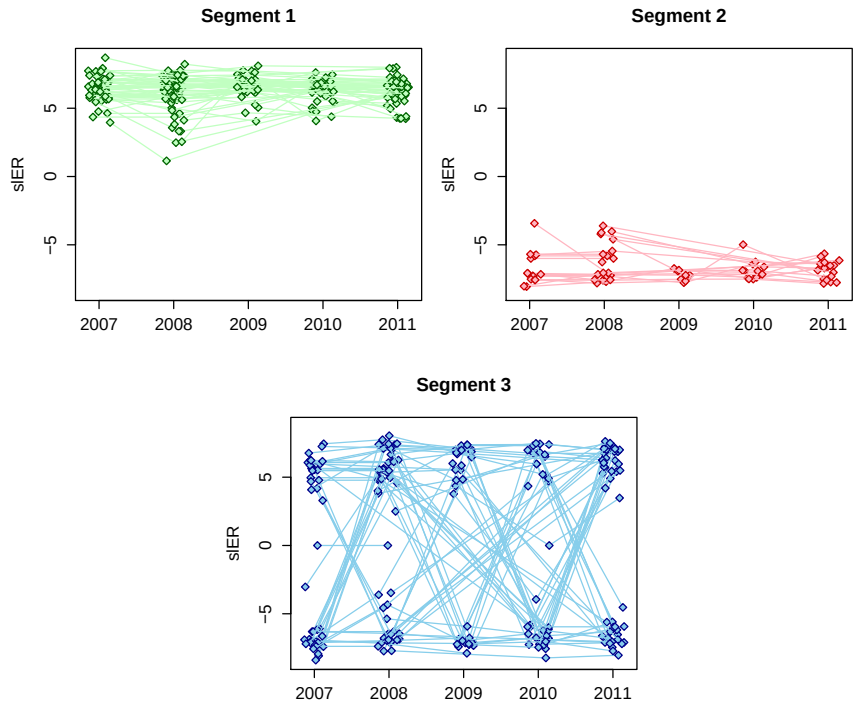


FIGURE 1. Observed longitudinal profiles of the transformed economic result by segments.

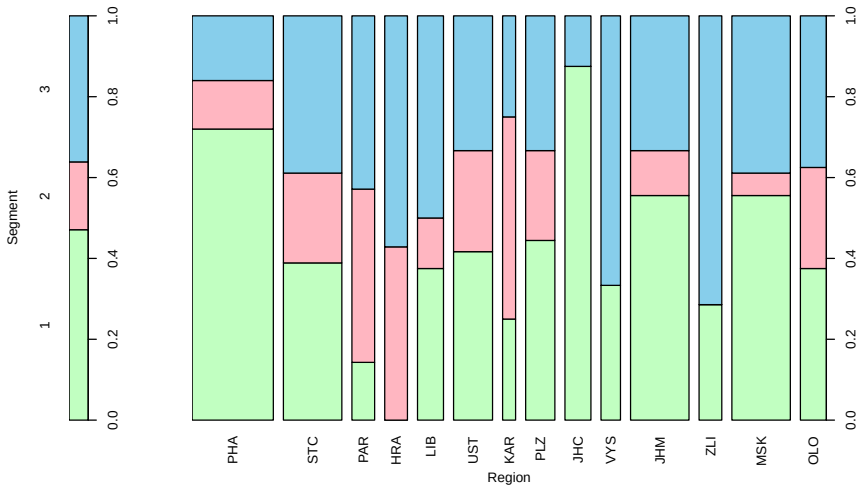


FIGURE 2. Mosaic plots for the representation of the three segments in each of 14 regions.

$\boldsymbol{\theta} = \left(\mathbf{w}^{\mathrm{T}}, \boldsymbol{\mu}_1^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_K^{\mathrm{T}}, \mathsf{vec}^{\mathrm{T}}(\mathbb{D}_1), \ldots, \mathsf{vec}^{\mathrm{T}}(\mathbb{D}_K), \sigma^2\right)^{\mathrm{T}}$. Segmentation of the hospitals is then based on the posterior distributions of the individual component probabilities, i.e., on the posterior distributions of derived parameters $\mathsf{p}_{i,k}(\boldsymbol{\theta}) = \mathsf{P}\Big(U_i = k \Big| \{(Y_{i,t},\, Z_{i,t}) : t \in \mathcal{T}_i\}\Big)$, $i = 1, \ldots, N$, $k = 1, \ldots, K$, expression of which follows from the assumed model (1). Selection of a number of clusters is based on the concept of a penalized expected deviance (PED, Plummer, 2008).

4 Results and discussion

The optimal number of segments according to the PED appeared to be three, i.e., $K = 3$, with the posterior means of their weights (proportions) being 0.437, 0.159, 0.405. The three segments are graphically visualized on Figure 1 showing the observed longitudinal trajectories of the transformed economic result $Y_{i,t}$ by segments. It is seen that segment 1 contains only hospitals that made, in all considered years, a profit. With this respect, segment 2 is opposite to segment 1 in the sense that all hospitals being classified in segment 2 managed with loss in all years 2007 – 2011. The third segment is then mostly composed of hospitals that alternated between loss and profit.

Selected results of additional exploration of the relationship between some characteristics of the hospitals and the three segments are found on Figure 2. It shows representation of the segments in each of the fourteen regions of the Czech Republic. Interestingly, none of the hospitals of the *South Bohemian* region (JHC) that were all changed from a non-profit type of the health care provider into a joint-stock company shortly before 2007, is classified in segment 2 (economic loss in all years 2007 – 2011). This suggests that the expectation of the regional South Bohemian government of an economically positive impact of this change was justifiable.

References

Komárek, A. and Komárková, L. (2013). Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*, **7**, 177 – 200.

Komárek, A. and Komárková, L. (2014). Capabilities of `R` package `mixAK` for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software,* **59**, 1 – 38.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics,* **9**, 523 – 539.

A simulation study assessing the advantages of cumulative link models

Altea Lorenzo-Arribas[1,2], Mark J. Brewer[1], Antony M. Overstall[2]

[1] Biomathematics and Statistics Scotland, UK
[2] University of Glasgow, UK

E-mail for correspondence: altea.lorenzo-arribas@bioss.ac.uk

Abstract: Discussion over the appropriateness of arbitrary categories number assignment (Senn, 2007), threshold structure specification (Christensen, 2015) and the bias introduced by fitting non-ordinal regression models to ordered-response variables (Agresti, 2010) is still ongoing. By means of simulations and analysis of real socio-economic data, we demonstrate some of the benefits of the family of cumulative link models. Precision and user-friendliness are at the heart of this study which aims to provide additional evidence to existing guidelines on model choice procedures for ordinal data (McKinley et al., 2014).

Keywords: Ordinal response data; Simulation; Socio-economics.

1 Ordinal regression models

Power benefits from ordinal regression models are often highlighted (Capuano et al., 2007). The present study outlines additional advantages for ordinal models over simpler alternatives, and identifies problems from inappropriate use of the latter. We focus our study on cumulative logit models as defined by (McCullagh, 1980):

$$\text{logit}(P(Y_i \leq j)) = \log\left\{\sum_{k=1}^{j} \pi_{ik} / \sum_{k=j+1}^{C} \pi_{ik}\right\}, \qquad j = 1, \ldots, C-1,$$

where Y_i is an ordinal response variable with C categories following a multinomial distribution and where parameters $\boldsymbol{\pi}$ are the probabilities of being in each category and $P(Y_i \leq j)$ are the cumulative probabilities.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

These models rely on the concept of thresholds on a continuous latent scale which are cut-points defining the categories of the ordinal variable. Thresholds can be constrained, for example as either symmetric or equidistant. The proportional odds model is defined within this framework as follows:

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k X_{ik}, \qquad j = 1, \ldots, C-1,$$

where Y_i is interpreted in terms of a same covariate effect β for each response category, independently of the thresholds $\alpha_1 < \alpha_2 < \cdots < \alpha_{C-1}$. More generally, we have partial proportional odds models, for example:

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p_0} \beta_k X_{ik} + \sum_{k=p_0+1}^{p} \beta_{jk} X_{ik}, \qquad j = 1, \ldots, C-1,$$

which allow some or all coefficients to vary by category (here β_{jk}).

Simulation study: Figure 1 highlights the consequences of using two different approaches for modelling ordinal responses. We suppose the existence of data on an underlying continuous latent response, and show simulated points plotted against a covariate x. The true relationship is a straight line, drawn here in black. The horizontal, dotted lines show the cut-points used to create an ordinal response (with values shown on the left-hand y-axes) from the latent response; note that subsequently treating this ordinal response as numeric introduces a "warping" of the true scale, indicated by the values on the right-hand y-axes (where midpoints of the classes are shown).

Treating the derived ordinal scale as numeric and using linear regression produces the red lines, which miss the true relationships owing to the bias caused by the (unknown) warping of the real scale. Using a cumulative logit model, the scale is not warped in this way, and back-calculating the fitted lines onto the numeric scale produces fits (shown in blue) which are hard to distinguish from the true relationships. If the thresholds had been equispaced, we would not expect to see this bias; for the symmetric threshold spacing in (b), the bias appears in the estimated slope, whereas for the asymmetric spacing in (a), the bias is visible in the intercept. We suggest the result for (b) is more serious, as this is affecting the size of the estimated effect of the covariate, and may impact any associated inference.

To consider the effect of modelling choice on inference, we have simulated repeated data sets in the manner of those in Figure 1. We study the proportion of occurrences of statistically significant and non-significant results for asymmetric thresholds at both the 5% and 1% significance level for modelling: (i) the original latent response via linear regression; (ii) the ordinal response as if it were continuous via linear regression; and (iii) the ordinal response via a cumulative logit model.

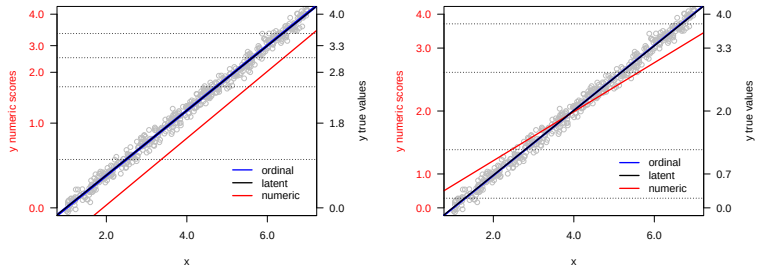


FIGURE 1. Model fitting comparison for simulated thresholds.

Preliminary results from the simulations show that there is a closer agreement between the results of (iii) and (i) than between (ii) and (i). In around one third of the simulated data sets, model (iii) alone gave the same results as for the latent response model (the "correct" model) for 5% significance (slightly fewer for 1%). All three models resulted in the same decisions around 50% of the time for 5% significance and 70% for 1%. Only rarely (3% of the time for both 5% and 1% significance) did the "ordinal as continuous" results alone match the correct model's results.

2 Socio-economics case study

Ordinal responses are common in social and economics research, where variables are often measured in the form of classes or by means of scales (e.g. Likert); examples include measures of wellbeing and opinion surveys. Our case study assesses the effects of the interaction of connectedness to nature and experience. 350 individuals were asked to provide ratings of pleasantness (as a response) for two categories of experience (nature and shopping) given their reported measurement for connectedness to nature (CNS), which is a 3-level factor ranging from 1 = hard to 3 = light. From exploratory analysis, the presence of an interaction between the two independent variables becomes apparent. Table 1 summarises results of the analysis of the data using a linear model.

TABLE 1. Results: linear model.

	Estimate	Std. error	t value	Pr($>\|t\|$)
CNS2	0.2865	0.2299	1.2462	0.2135
CNS3	0.4737	0.2259	2.0969	0.0367
Experience Shopping	−1.7576	0.2411	−7.2889	< 0.001
CNS2:Experience Shopping	−0.5139	0.3258	−1.5776	0.1156
CNS3:Experience Shopping	−1.3188	0.3195	−4.1275	< 0.001

TABLE 2. Results: partial proportional odds model, symmetric thresholds.

	Estimate	Std. error	z value	$\Pr(>\|z\|)$
CNS2	0.7142	0.2710	2.636	0.0084
CNS3	1.1568	0.2842	4.070	< 0.001
CNS2:Experience Shopping	−0.9070	0.3596	−2.523	0.0117
CNS3:Experience Shopping	−1.9097	0.3693	−5.171	< 0.001

We found evidence that the proportional odds assumption was not satisfied. Table 2 shows results of the best fitting model for the data, i.e., a partial proportional odds model with symmetric thresholds, relaxing the proportional odds assumption for the variable experience. Results qualitatively differed in terms of parameter significance, and did not show signs of poor fit during model checking. Using a model appropriate to the response data type has thus produced different inference compared to a simpler model. All simulations and modelling were performed in R (R Core Team, 2015).

Acknowledgments: A. Lorenzo-Arribas and M.J. Brewer were funded by the Rural and Environment Science and Analytical Services Division of the Scottish Government.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. New Jersey: Wiley.

Capuano, A.W., Dawson, J.D., and Gray, G.C. (2007). Maximizing power in seroepidemiological studies through use of the proportional odds model. *Influenza and Other Respiratory Viruses*, **1**, 87 – 93.

Christensen, R.H.B. (2015). Analysis of ordinal data with cumulative link models - estimation with the R-package ordinal. *R vignette*.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B*, **42**, 109 – 142.

McKinley, T.J., Morters, M., and Wood, J.L.N. (2015). Bayesian model choice in cumulative link ordinal regression models. *Bayesian Analysis*, **10**, 1 – 30.

R Core Team (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/.

Senn, S. (2007). Drawbacks to Noninteger Scoring for Ordered Categorical Data. *Biometrics*, **63**, 296 – 299.

Bayesian variable selection in semi-parameteric growth regression

Gertraud Malsiner-Walli[1], Bettina Grün[1], Paul Hofmarcher[2]

[1] Institut für Angewandte Statistik, Johannes Kepler Universität Linz
[2] Institute for Monetary and Fiscal Policy, Department of Economics, WU Wien

E-mail for correspondence: gertraud.malsiner_walli@jku.at

Abstract: Competing theories advancing different determinants for economic long-term growth have been proposed and thus a large set of potential determinants of economic growth may be included in empirical analyses. In addition to the uncertainty about the variables to include in a regression, the functional relationship between the determinants and the dependent variable is also not known and most empirical analyses impose a linear relationship neglecting the possibility of a non-linear relationship.
In this paper we investigate the use of the spike-and-slab prior to combine Bayesian variable selection with penalized spline regression and thus identify robust determinants of economic growth while allowing for a non-linear relationship between the determinants and long-term economic growth.

Keywords: Stochastic search; Spike-and-slab prior; NMIG prior; R package spikeSlabGAM; Non-linear regressors.

1 Introduction

Competing theories to explain long-term economic growth propose different determinants leading to a vast number of potential variables which need to be considered in empirical research. This uncertainty about the determinants to include as explanatories in a regression with long-term economic growth as dependent variable has led to the development of methods which allow for the identification of robust determinants of economic growth. Sala-i-Martin et al. (2004) for example analyze a data set consisting of 88 countries and 67 explanatory variables using Bayesian model averaging (BMA, Hoeting et al., 1999) of classical linear regression models and identify 18 out of the 67 explanatory variables as being significantly and robustly partially correlated with long-term growth.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

The approach by Sala-i-Martin at al. (2004) accounts for the model uncertainty regarding the variables to include. However, the method requires the assumption that the functional relationship between the determinants and long-term economic growth is linear. This assumption might be too restrictive and constitute a misleading oversimplification such that important determinants, which are related to long-term economic growth through a non-linear functional relationship, remain undetected.
There is growing empirical evidence in the growth literature that there are non-linear functional relationships present. Henderson et al. (2011) for instance use non-parameteric estimation methods on only a small subset of potential determinants of long-term economic growth to identify which determinants have a non-linear functional relationship with the dependent variable. However, they are not able to fully take the model uncertainty regarding the vast number of potential determinants into account.
In this paper we try to address both types of model uncertainty within the same modeling framework: uncertainty about the functional form of the influence (linear versus non-linear effects) and uncertainty about the explaining variables (variable uncertainty). We use recently developed methods for semi-parametric regression (Scheipl et al., 2012) to investigate potential non-linear determinants of long-term economic growth which can be identified in a robust way despite model uncertainty. The unknown effects are modeled by flexible penalized spline functions. Stochastic search techniques are used to identify relevant variables along with their functional relationships. The regression analysis is performed with the R package spikeSlabGAM (Scheipl, 2011).

2 Model

In the Bayesian framework, the problem of selecting an appropriate subset of explanatory variables and at the same time determining, whether linear or more flexible functional forms are required to model the relationship between explanatory variables and response, is translated into the issue of estimating marginal posterior probabilities.
The R package spikeSlabGAM implements fully Bayesian variable selection in combination with the determination whether linear or more flexible functional forms are required to model the effects of the respective covariates. A spike-and-slab prior structure is employed for the prior variance of a regression coefficient or a coefficient group resulting in a two-component scale mixture around zero as a prior, with one narrow component (spike) and a flat component covering a wide range of values for the coefficient or coefficient group. Then, for the specific coefficient or coefficient group, the posterior mixture weight for the slab component can be interpreted as the posterior probability of variable inclusion into the model.
In order to model regressor effects of an unknown shape for p potential

covariates, i.e.

$$y = f_1(x_1) + \cdots + f_p(x_p) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

the unknown functions f_j are approximated by a linear combination of B-spline basis functions. A generous number of basis functions is selected and overfitting is avoided by penalizing the basis coefficients in order to induce smooth fitting results (Eilers and Marx, 1996). The penalization is achieved by specifying an improper Gaussian random walk prior of second order on the coefficients, with the variance of the prior acting as the penalization parameter. This choice leads to a model specification where the stochastic search is performed by selecting the linear terms separately from the higher order terms and a distinction between linear effects and additional non-linear effects for each covariate is possible based on the posterior inclusion probabilities of each of these effects.

3 Application

In order to identify robust non-linear predictors of economic growth, the model is fitted to the data set used by Sala-i-Martin et al. (2004). The prior specifications for the spike-and-slab approach are $\tau^2 \sim \mathcal{G}^{-1}(5, 25)$ for the variance of the slab and $\nu_0 = 0.0000025$ for the variance of the spike. The results are summarized in Table 1. If the model is fitted by only allowing linear terms, the most important predictors identified by Sala-i-Martin et al. (2004) in the context of BMA of classical linear regression models can be reproduced using a slightly different model specification regarding the prior structure as well as estimation method. However, if in addition smooth effects are included in the model, three new predictors ("Population Density", "Fraction GDP in Mining", and "Higher Education") appear among the list of the "top ten" regressors.

References

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

Henderson, D.J., Papageorgiou, C., and Parmeter, C.F. (2011). Growth empirics without parameters. *The Economic Journal*, **122**, 125 – 154.

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, **14**, 382 – 401.

Sala-i-Martin, X., Doppelhofer, G., and Miller, R.I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *The American Economic Review*, **94**, 814 – 835.

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, **107**, 1518 – 1532.

Scheipl, F. (2011). spikeSlabGAM: Bayesian variable selection, model choice and regularization for generalized additive mixed models in R. *Journal of Statistical Software*, **43**, 1 – 24.

TABLE 1. Data set by Sala-i-Martin et al. (2004). Posterior inclusion probabilities (PIP) of the top 15 explanatory variables identified by Sala-i-Martin et al. (2004) ("SDM") and the spike-and slab framwork (Scheipl et al., 2012) with only linear terms ("SpSl (lin)") or additional smooth terms ("SpSl (lin+smo)"). "Rk" denotes rank, "Pop." and "Ex."abbreviates population and exchange respectively, and possible terms are factor ("fct"), linear ("lin") and smooth ("smo").

SpSl (lin+smo)			SpSl (lin)		SDM		
Rk	Term	PIP	Rk	PIP	Rk	PIP	Variables
1	fct	0.752	1	0.825	1	0.823	East Asian Dummy
3	lin	0.581	3	0.392	2	0.796	Primary Schooling '60
8	lin	0.331	5	0.334	3	0.774	Investment Price
			6	0.306	4	0.685	log GDP '60
5	lin	0.370	4	0.337	5	0.563	Fraction of Tropical Area
			10	0.153	6	0.428	Coastal Pop. Density '60
6	lin	0.357	2	0.427	7	0.252	Malaria Prevalence '60
11	lin	0.232	8	0.269	8	0.209	Life Expectancy '60
4	fct	0.408	7	0.270	9	0.206	Fraction Confucian
10	fct	0.238	15	0.144	10	0.154	African Dummy
					11	0.149	Latin American Dummy
7	smo	0.349			12	0.124	Fraction GDP in Mining
			14	0.132	13	0.123	Spanish Colony
			9	0.167	14	0.119	Years Open 1950–94
					15	0.114	Fraction Muslim
			11	0.151			Fraction Buddhist
			12	0.149			Gov. Consum. Share '60
			13	0.148			Real Ex. Rate Distortions
2	smo	0.704					Pop. Density '60
9	smo	0.263					Higher Education '60
12	smo	0.223					Tropical Climate Zone
13	smo	0.218					Ethnolinguistic Fractional.
14	smo	0.218					Political Rights
15	smo	0.214					Fraction Pop. in Tropics

Simulated adjustment of the signed scoring rule root statistic

Valentina Mameli[1], Monica Musio[2], Laura Ventura[3]

[1] Dept. of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy
[2] Dept. of Mathematics and Computer Science, University of Cagliari, Italy
[3] Dept. of Statistical Sciences, University of Padova, Italy

E-mail for correspondence: mmusio@unica.it

Abstract: We focus on adjustments of the signed scoring rule root statistic generalizing results for likelihood quantities. In particular, for a scalar parameter of interest, we investigate a bootstrap adjustment of the signed scoring rule root statistic. An example is discussed.

Keywords: Asymptotic expansions; Bootstrap; Higher-order inference; Tsallis scoring rule.

1 Introduction

A *scoring rule* is a loss function $S(x, Q)$ measuring the quality of a quoted probability distribution Q for the random variable X, in the light of the realized outcome x of X. It is *proper* if, for any distribution P for X, the expected score $S(P, Q) := E_{X\sim P}S(X, Q)$ is minimized by quoting $Q = P$. There is a wide variety of proper scoring rules (see, e.g., Dawid and Musio, 2014, and references therein). A prominent example is the *log-score* $S(x, Q) = -\log q(x)$, with $q(\cdot)$ the density (or the probability mass function) of X. Another useful example is the *Tsallis* score, given by

$$S(x, Q) = (\gamma - 1)\int q(y)^{\gamma} d\mu(y) - \gamma q(x)^{\gamma-1}, \qquad \text{with } \gamma > 1, \qquad (1)$$

also called the *density power score* (Basu et al., 1998). Other examples of special proper scoring rules are given in Dawid and Musio (2014). Proper scoring rules, different from the log-score, can be used as an alternative to the full likelihood, when the aim is to increase the robustness or to simplify

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

computations. For instance, Dawid et al. (2014) give sufficient conditions that guarantee the robustness of the estimator based on (1).
Proper scoring rule inference is usually based on the first-order approximations to the distribution of the scoring rule estimator or of the scoring rule ratio test statistic. However, several examples (see e.g. Dawid et al., 2014, Mameli and Ventura, 2015, and reference therein) illustrate the inaccuracy of first-order methods, even in models with a scalar parameter, when the sample size is small or moderate. For more accurate inference refinements can be considered to improve the first-order approximations.
Analytical higher-order asymptotic expansions for proper scoring rules, generalizing results for likelihood quantities but allowing for the failure of the information identity, have been discussed in Mameli and Ventura (2015). However, the calculation of the quantities involved in the analytical adjustments of the signed and signed profile scoring rule root statistic is cumbersome when the dimension of the parameter (or of the nuisance parameter) is large, even for simple models.
Paralleling results for likelihood statistics (see, e.g., Young, 2009), the aim of this paper is to discuss the alternative approach to higher-order adjustments, based on a parametric bootstrap. In particular, focus is on the signed profile scoring rule root statistic.

2 Background on first-order inference

Suppose that we wish to fit a parametric statistical model $F_\theta = F(x;\theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^k$, based on the random sample $(x_1, \ldots, x_n)$.
To estimate θ, we might consider the goodness-of-fit by the total empirical score $S(\theta) = \sum_{i=1}^n S(x_i, F_\theta)$. Asymptotic arguments indicate that $\hat{\theta}_S = \arg\min_\theta S(\theta) \to \theta_0$ as $n \to \infty$, where θ_0 is the true parameter value. Maximum likelihood estimation, as well as composite likelihood estimation, are special cases of score estimation when $S(\theta)$ is the negative log-likelihood (Dawid and Musio, 2014).
Score estimation forms a special case of M-estimation (Huber and Ronchetti, 2009). Indeed, let $s(x,\theta) = \partial S(x, F_\theta)/\partial\theta$. Under suitable regularity conditions on the scoring rule and on the statistical model, θ can be estimated by $\hat{\theta}_S$, the root of the estimating equation $s(\theta) = \sum_{i=1}^n s(x_i,\theta) = 0$, which is an unbiased estimating function. In particular, $\hat{\theta}_S$ is consistent and asymptotically normal with mean θ and variance $V(\theta) = K(\theta)^{-1}J(\theta)K(\theta)^{-1}$, with $J(\theta) = E_\theta(s(\theta)s(\theta)^\top)$ and $K(\theta) = E_\theta(\partial s(\theta)/\partial\theta^\top)$. The form of $V(\theta)$ is due to the failure of the information identity since, in general $K(\theta) \neq J(\theta)$. In view of this, the asymptotic distribution of the scoring rule ratio statistic $W^S(\theta) = 2\{S(\theta) - S(\hat{\theta}_S)\}$ departs from the familiar likelihood result, and involves a linear combination of independent chi-square random variables
When θ is partitioned as $\theta = (\psi, \lambda)$, where ψ is a scalar parameter of interest and λ is a $(k-1)$-dimensional nuisance parameter, the profile scoring

rule ratio statistic for ψ is given by $W_p^S(\psi) = 2(S(\hat{\theta}_{S\psi}) - S(\hat{\theta}_S))$, where $\hat{\theta}_{S\psi} = (\psi, \hat{\lambda}_{S\psi})$ represents the constrained score estimate. The asymptotic distribution of $W_p^S(\psi)$ is a linear combination of independent chi-square random variables (see Dawid et al., 2014). Also the asymptotic distribution of the profile signed scoring rule root statistic

$$r_p^S(\psi) = \text{sgn}(\hat{\psi}_S - \psi)\sqrt{W_p^S(\psi)} \tag{2}$$

departs from the familiar likelihood result. A simple adjustment of (2), which recovers normality, is $r_{p\,adj}^S(\psi) = \mu_1(\psi)^{-1/2} r_p^S(\psi)$, with $\mu_1(\psi) = [G^{\psi\psi}(\hat{\theta}_{S\psi})^{-1} H^{\psi\psi}(\hat{\theta}_{S\psi})]^{-1}$, where $G^{\psi\psi}(\theta)$ and $H^{\psi\psi}(\theta)$ are sub-matrices of the inverses of $G(\theta)$ and $H(\theta)$, respectively.

3 Adjustments of $r_p^S(\psi)$

Let us focus on the profile signed scoring rule root statistic (2) for a scalar parameter of interest. As for likelihood quantities, theory in Mameli and Ventura (2015) shows that a suitable centering and scaling of $r_p^S(\psi)$, i.e.

$$r_{pM}^S(\psi) = \frac{r_p^S(\psi) - m(\psi)}{\sqrt{\mu_1(\psi) + v(\psi)}}, \tag{3}$$

improves the accuracy of the asymptotic normal approximation to the distribution of $r_{p\,adj}^S(\psi)$. In (3) we have that $m(\psi)$ is of order $O(n^{-1/2})$ and $v(\psi)$ is of order $O(n^{-1})$. The analytical expressions of $m(\psi)$ and $v(\psi)$ are derived in Mameli and Ventura (2015) and they involve several expected values of scoring rules derivatives. However, the analytical calculations of $m(\psi)$ and $v(\psi)$ are cumbersome as the dimension of the nuisance parameter is large, even for simple models.
Here we exploit a parametric bootstrap approach in order to compute (3). The idea is, for a fixed ψ, to draw B samples from the distribution $F(x; \hat{\theta}_{S\psi})$ and compute (3) using the bootstrap mean and variance of $r_p^S(\psi)$. In the classical likelihood approach, the parametric bootstrap provides a $O(n^{-3/2})$ order of accuracy, and the resulting approximation is sometimes called constrained pre-pivoting of the signed likelihood root statistic (see DiCiccio et al., 2001).

Example: Let us consider the linear regression model $y = X\beta + \sigma\epsilon$, where X is a $n \times p$ fixed matrix, $\beta \in \mathbb{R}^p$ $(p \geq 1)$, $\sigma = 1$, and ϵ an n-dimensional Gaussian vector. Let $\psi = \beta_2$ be the scalar parameter of interest, and let $\lambda = (\beta_1, \beta_3)$ be the nuisance parameter. We ran a simulation experiment, for $n = 10, 20$ and $\beta = (1, 2, 3)$, in order to assess the accuracy of the parametric bootstrap Tsallis modified profile signed scoring rule root statistic

($r^T_{pMb}(\beta_2)$). Note that the Tsallis score estimator is B-robust since the influence function is bounded (see Dawid et al., 2014 and Mameli and Ventura, 2015). Table 1 gives the results of the study based on 10,000 simulations with $B = 500$ bootstrap replications. We note that the accuracy of the parametric bootstrap depends mainly on the choice of the estimate to use for generating samples. Indeed, parametric bootstrap of the Tsallis profile signed scoring rule root statistic $r^T_{pM}(\psi)$ under the model $F(y;\hat{\theta}_{S\psi})$, provided B is large enough, yields an accurate parametric inference approach, bypassing any analytical computation. On the contrary, results, not shown here, indicated that the same accuracy is not retained when we sample from $F(y;\hat{\theta}_S)$. Note that also in the likelihood framework $\hat{\theta}_{S\psi}$ is the best choice to providing accurate inference (Di Ciccio et al., 2001).

TABLE 1. Empirical coverages of 95% confidence intervals for β_2. Pivots used: parametric bootstrap higher-order signed profile likelihood root (r^*_{pb}) and parametric bootstrap Tsallis modified profile signed scoring rule root (r^T_{pMb}) with $\gamma = 1.2$.

n	r^*_{pb}	r^T_{pMb}
10	0.9498	0.9500
20	0.9536	0.9547

References

Basu, A., Harris, I.R., Hjort, N.L., and Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549 – 559.

Dawid, A.P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, **72**, 169 – 183.

Dawid, A.P., Musio, M., and Ventura, L. (2014). Minimum scoring rule inference. ArXiv:1403.3920.

Di Ciccio, T.J., Martin, M.A., and Stern, S.E. (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. *Canadian Journal of Statistics*, **29**, 67 – 79.

Huber, P.J. and Ronchetti, E.M. (2009). *Robust Statistics*. John Wiley and Sons, Inc.

Mameli, V. and Ventura, L. (2015). Higher-order asymptotics for scoring rules. *Journal of Statistical Planning and Inference*, to appear, doi:10.1016/j.jspi.2015.03.005.

Young, G.A. (2009). Routes to higher-order accuracy in parametric inference. *Australian and New Zealand Journal of Statistics*, **51**, 115 – 126.

A flexible bivariate location-scale finite mixture approach to economic growth

Alessandra Marcelletti[1], Antonello Maruotti[2], Giovanni Trovato[3]

[1] Dipartimento di Economia Diritto ed Istituzioni, Università di Roma "Tor Vergata", Roma, Italy; Dipartimento di Scienze Politiche, LUISS Guido Carli, Roma, Italy
[2] Southampton Statistical Sciences Research Institute, University of Southampton, UK; Dipartimento di Scienze Politiche, Università di Roma Tre, Roma, Italy
[3] Dipartimento di Economia Diritto ed Istituzioni, Università di Roma "Tor Vergata", Roma, Italy

E-mail for correspondence: marcelletti@economia.uniroma2.it

Abstract: We introduce a multivariate multidimensional mixed-effects regression model in a finite mixture framework. We relax the usual unidimensionality assumption on the random effects multivariate distribution. Thus, we introduce a multidimensional multivariate discrete distribution for the random terms, with a possibly different number of support points in each univariate profile, allowing for a full association structure. Our approach is motivated by the analysis of economic growth. Accordingly, we define an extended version of the augmented Solow model. Indeed, we allow all model parameters, and not only the mean, to vary according to a regression model. Moreover, we argue that countries do not follow the same growth process, and that a mixture-based approach can provide a natural framework for the detection of similar growth patterns. Our empirical findings provide evidence of heterogenous behaviors and suggest the need of a flexible approach to properly reflect the heterogeneity in the data. We further test the behavior of the proposed approach via a simulation study, considering several factors such as the number of observed units, times and levels of heterogeneity in the data.

Keywords: Economic growth, Country classifications, Finite mixture model.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

1 Introduction

In modelling panel economic data, it is common to account for the unobserved heterogeneity between sample units, that is, the heterogeneity that cannot be explained by means of observable covariates. Addressing the heterogeneity of analyzed processes is of fundamental importance to the study of economic growth, since sample units (i.e. countries) could be characterized by heterogeneous income performances and has led to a substantial evidence for the existence of variations in growth patterns across countries. Indeed, since Solow's seminal paper (1956), different econometric and statistical approaches are used to look at countries' growth. Dynamic panel data with fixed effect (Caselli et al., 1996), as well as extreme bound analysis (Levine and Renelt, 1992), Bayesian model averaging (Fernandez et al., 2001) or model on varying coefficients are performed to deal with the main empirical challenges in growth theory: unobserved heterogeneity (see e.g. Caselli, 1996), uncertainty (Temple, 2000) and omitted variable bias (Durlauf and Quah, 1999). Recently, data-driven approaches to estimate multiple (heterogeneous) growth processes have been employed within the wide class of mixture models (Alfó et al., 2008), and extensions of the finite mixture approach for panel data are available in the literature (Pittau et al., 2010).
We contribute to the literature about finite mixture models for panel data by extending the approach introduced by Alfó et al. (2008), and providing an empirical formulation of the augmented Solow model based on a multivariate-multidimensional specification, that allows to solve the unobserved heterogeneity issue. We address the heterogeneity issues related to: varying parameters across countries, omitted variables and non-linearities in the production function.

2 Empirical model

We propose an approach to panel growth data based on a flexible bivariate location-scale finite mixture approach allowing for all model parameters to depend on covariates in a regression framework. We relax the common unidimensionality assumption of the random effects distribution, allowing for a general and flexible association structure among the outcomes. Indeed, we develop an endogenous clustering approach lying on a bivariate bidimensional model recovering Bernanke (2002) intuition: country's rate of investment and of human capital and the population growth rate are correlated with the long run growth rate of output per capita. Furthermore, we jointly determine the evolution of income per capita and volatility of growth by expliciting the variance of the growth rate as dependent on explanatory variables. Thus, we introduce two separates equations for the location and scale parameters of the dependent variables, such that

the explanatory variables are also associated to the unpredictability of the variable itself.
Computational complexity is often the price we have to pay to flexibility. However, we show that parameter estimates can be obtained by extending the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for finite mixture to the multidimensional case. Furthermore, we avoid any restriction on the covariance structure of the random effects as assumed e.g. by the so-called one-factor model (Winkelmann, 2000), which is more parsimonious but could be hard to justify in empirical applications. By allowing the number of mixture components to grow with the sample size, the proposed model can be also used as a semiparametric estimator of multivariate mixed effects models, where the distribution of the random effects is estimated by a discrete multivariate random variable with a finite number of support points. This can be seen as a possible solution to computational issues arising with multivariate mixed models.
We illustrate the proposal by a simulation study in order to investigate the empirical behaviour of the proposed approach with respect to several factors, such as the number of observed units and times and the distribution of the random term (with varying number of support points). The simulation study addresses the properties of the maximum likelihood estimator we propose for the multidimensional bivariate random model, and the goodness of classification.

3 Results

We test the proposal by analysing a sample taken from the Summers-Heston Penn World Tables (PWT) version 8.0 and the World Bank database for years 1975 – 2005 for non-oil countries. In order to avoid the endogeneity problems related to growth model estimation, we consider non-overlapping 5-year period with explanatory variable averaged over the corresponding time period; while the dependent variables are taken 5 periods ahead.
Our empirical findings confirm the augmented Solow model. The intercept term in the GDP level equation for the location scale is left free to vary among clusters, in order to capture the omitted country-specific features, such as, institutional characteristic and technological factors and may contribute to reach (or not) economic convergence. This relies on the idea that accumulation driven growth equation is incomplete (see e.g. Alfó et al., 2008). Different levels of heterogeneity are detected in GDP and GDP growth, respectively. We find that our sample is much more heterogeneous with respect to GDP levels than growth patterns. Although this result sounds obvious, previous empirical results, based on unidimensional specification of the latent structure, were not able to distinguish for different heterogeneity levels (see e.g. Alfó et al., 2008).
Entering into details, six clusters are identified with respect to GDP levels, in which, coherently with the literature, we find the highest value for the

random effect for the component clustering the richest and more industrialized countries, such as USA and UK. Results show the existence of two groups with respect to the growth rate of GDP, representing high-growth and low-growth countries: the first group characterized by a negative and significant effect of the initial level of GDP on the growth pattern, confirming economics theory about convergence; the second group is characterized by the possible existence of multipla equilibria and the lack of convergence. These results suggest the presence of a convergence club.

References

Alfó, M., Trovato, G., and Waldmann, R.J. (2008). Testing for country heterogeneity in growth models using a finite mixture approach. *Journal of Applied Econometrics*, **23**, 487 – 514.

Bernanke, B.S. and Gürkaynak, R.S. (2002). Is growth exogenous? taking Mankiw, Romer, and Weil seriously. In: *NBER Macroeconomics Annual 2001*, **16**. MIT Press, 11 – 72.

Caselli, F., Esquivel G., and Lefort F. (1996). Reopening the convergence debate: a new look at cross-country growth empirics. *Journal of Economic Growth*, **1**, 363 – 389.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1 – 38.

Durlauf, S.N. and Quah, D.T. (1999). The new empirics of economic growth. *Handbook of Macroeconomics*, **1**, 235 – 308.

Fernandez, C., Ley, E., and Steel, M.F. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, **16**, 563 – 576.

Levine, R. and Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American Economic Review*, 942 – 963.

Pittau, M.G., Zelli, R., and Johnson, P.A. (2010). Mixture models, convergence clubs, and polarization. *Review of Income and Wealth*, **56**, 102 – 122.

Solow, R.M. (1956). A contribution to the theory of economic growth. *The quarterly journal of economics*, 65 – 94.

Temple, J. (2000). Growth regressions and what the textbooks don't tell you. *Bulletin of Economic Research*, **52**, 181 – 205.

Winkelmann, R. (2000). Seemingly unrelated negative binomial regression. *Bulletin of Economics and Statistics*, **62**, 553 – 560.

An approach to determine clusters overlap for K-means clustering

Kenan Matawie[1], Arshad Mehar[1], Anthony Maeder[1]

[1] School of Computing, Engineering and Mathematics. University of Western Sydney, Australia

E-mail for correspondence: k.matawie@uws.edu.au

Abstract: Clustering is one of the major and interesting tools for many data analysis in business, science, medical, social network and other sources. Various clustering methods are available and applied in different fields, but unfortunately they are still limited especially when the data modelled are not completely in k disjoint partitions and may belong to multiple clusters. One of the ways to solve such a clustering problem or limitation is to find and calculate the overlapping proportions in order to better understand and model the data structure. Many overlapping approaches are proposed in the literature based on different methods, in this paper we using and extending the recent new approach based on incidental and proportion matrices related to forward and backward movement of the objects at different number of clusters when K-means method/algorithm is applied. The degree of separation and overlap between these clusters is evaluated, discussed and analysed through a simulated dataset.

Keywords: K-means; Clustering algorithm; Clusters overlap; Proportion matrix.

1 Introduction and related work

Often studies utilizing clustering techniques for large datasets have not been able to produce useful results especially for clustering analysis due to lack of prior knowledge. A variety of clustering algorithms are available and established in the literature but for complex and sparse data (especially in health and economics) there are still many challenges and problems remain to be solved for better and effective outcomes especially in relation to cluster overlapping. Using a simple clustering approach may not produce a useful knowledge for large and complex datasets. In this presentation, we will focus on K-means clustering results for adjacent and non-adjacent set of clusters to determine the degree of overlapping between clusters. This

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

will be based on the forward and backward mapping of proportion of common elements to develop combined information of the K-means clustering algorithm results.

In cluster analysis especially for K-means clustering user specified the K number of clusters and can produce as many clusters as N-1 clusters from the dataset D with N observations. The main issue is to find, if exist, the clusters set that is completely disjoined or non-disjoint with degree of overlapping between clusters. Numbers of studies are carried out in cluster analysis, which includes determining the stability of cluster analysis and estimating the optimal values of K in the datasets using K-means clustering. Different types of validation techniques for K-means clustering results have been found in the literature by Dunn (1974), Kaufman and Rousseeuw (1990), and the formulation of these indexes has been briefly described Mehar et al. (2013).

Extraction of hidden patterns and strategic knowledge from large datasets is a growing challenge facing organizations, analyst and researchers. The new MMM approach or technique Mehar et al. (2013) provided information and initial approach for optimal number of clusters that are completely isolated and amount of overlapping when non-disjoint clusters are exist.

2 Clusters overlapping method

In the real world datasets most of the clusters obtained may have inherently some degree of overlapping. A forward and backward approach mapping the adjacent clusters and combining the information was used to develop MMM index described in Mehar et al. (2013) with one distant step $k+1$ from adaptive application of K-means algorithm. This method compares cluster mapping at adjacent consecutive k numbers. In this paper we will further map the proportion of common elements at $k+2, k+3, \ldots, k+9$ including non-adjacent distance steps. This new approach is based on changes in cluster membership as k varies with non-adjacent proportion of mapping common elements, and provides more systematic solution which is more independent of cluster characteristics, and more consistent in its behaviour across a wider range of potential k values that was introduced and discussed in Mehar et al. (2010, 2013). The underlying idea behind this approach is to analyse the movement of elements between clusters to find the common elements for range of $k+1, k+2, \ldots, k+9$ adjacent and non-adjacent distant steps. The approach is described as follows. For a given choice of k = number of clusters, a given choice of clustering technique U, and a given choice of V = set of parameters $v_1, \ldots, v_n$ used to control the clustering technique, we first construct a set of clusters $C_k(U,V) = C_{k,i}$ and $C_{k+1}(U,V) = C_{k+1,i}$ using the same clustering technique, where $i = 1, \ldots, k$. We may write these cluster sets more simply as C_k and C_{k+1}. Now these two (C_k, and C_{k+1}) consecutive groups at k will be used to find the

number of common elements from C_k to C_{k+1} and C_{k+1} to C_k, to create a rectangular mapping of common elements matrix of size $k \times k+1$, where k and $k+1$ correspond to rows and columns of forward matrix $P_{k,k+1}$. We denote the proportion of data elements in common between a particular pair of clusters, say cluster $C_{k,i}$ from C_k and cluster $C_{k+1,j}$ from C_{k+1} by $p(C_{k,i}, C_{k+1,j})$, which can be abbreviated to $p_{k,k+1,i,j}$. Similarly, we can compute $P_{k+1,k}$ to create another rectangular matrix of size $(k+1) \times k$ where $k+1$ correspond to rows and k to columns. In general $P_{k+1,k}$ is not equal to $P_{k,k+1}$ and they have different dimension and size.

To investigate how much movement of objects occurs from C_k to C_{k+1} and from C_{k+1} to C_k we multiply the corresponding proportion matrices to obtain $k \times k$ matrix with the entries showing the joint probabilities of the forward/back movement of the objects between the set of clusters from k to $k+1$ and $k+1$ to k. The results will be called joint information matrix $Q_{k,k}$ and that is $Qk, k = P_{k,k+1}P_{k+1,k}$, where, $P_{k,k+1} = p_{k,k+1,i,j}$ is the $k \times k+1$ forward proportion matrix from k to $k+1$, $P_{k+1,k} = p_{k+1,k,j,i}$ is the $k+1 \times k$ backward proportion matrix from $k+1$ to k, $i = 1, \ldots, k$ and $j = 1, \ldots, k+1$.

Due to the row sum constant of 1 the resultant $Q_{k,k}$ is also known as a row stochastic matrix Johnson (1981). This $Q_{k,k}$ matrix will be used to determine clusters overlap or the degree of separation from one another.

The elements of $[Q_{k,k}]$ matrices represents the proportion of clusters separation or the level of overlap between clusters, and the mean value of the diagonal for each $[Q_{k,k}]$ matrix is computed as $\mu_{(k,k+1)} = \sum_{i=1}^{k} q_{ii}/k$ to represent different $Q_{k,k}$ combined information for different distant steps.

3 Analysis and results

Usually the clustering techniques are heuristic calculations making it impossible to assess which is the best or most informative K clusters. The degree and details of the clusters overlap and the compactness of cluster memberships at any choice of k is essential and important for understanding and explaining the homogeneity and separation of clusters when applying the k-means clustering algorithm for two or multidimensional datasets. In this presentation we will only focus on synthetic dataset with two dimensions, real dataset with multi-dimensions will not be discussed here. Computational experiments start with applying the k-means clustering algorithm for $k = 2$ to 11 to cover and create all possible homogenous groups for each dataset. Using the resultant clusters from K-means we will calculate the overlap $Q_{k,k}$ matrices with adjacent $k+1$ and non-adjacent distant steps $k+2, k+3, \ldots, k+9$. For this purpose we have simulated 2-dimensional dataset (Dataset2D with 700 observations) using normal distribution with different values of parameters mean (μ) and standard deviation (σ); $n = 400$ with $\mu = (40, 30)$, $\sigma = 7$ and $n = 300$ with $\mu = (60, 50)$,

$\sigma = 7$. This data is presented in Figure 1 plot (a) below, and plots (b)-(d) corresponding to clusters when $k = 2, 3$, and 4. Plot (e) shows the line plot for different $Q_{k,k}$ of the combined proportion mean values with coloured lines for different distant steps. It is clear from plot (e) and at $k = 2$ clusters are not completely isolated and the degree of overlap is less than 2% compared to other values of k where the degree of overlapping among them are much higher. This matrix also showed the changes in the overlap proportions associated with different k values. These information are important to better fit and model the structure of the data, it also help and can be used as part of criteria in the construction of the clusters especially when it shows the proportions are monotonically decreasing with k. This new approach will also allow us to handle complex problems in large datasets by detecting the level of overlaps. Efficient exploitation of the new approach in the simulated and variety of real datasets will lead us to solve more challenging problems in this area of analysis and research.

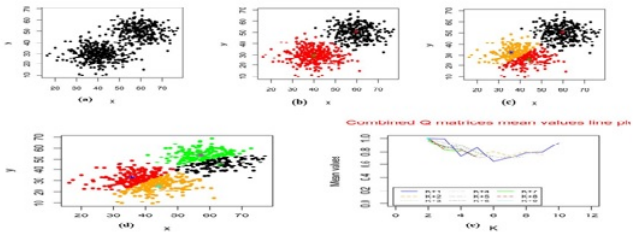


FIGURE 1. Dataset2D clusters at $k = 3$ and $k + 1 = 4$.

References

Dunn, J.C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, **4**, 95 – 104.

Johnson, C.R. (1981). Row stochastic matrices similar to doubly stochastic matrices. *Linear and Multilinear Algebra*, **10**, 113 – 130.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data, an Introduction to Cluster Analysis*. New York: Wiley.

Mehar, A.M., Maeder, A., Matawie, K., and Ginige, A. (2010). *Blended clustering for health data mining in E-health*. Springer, 130 – 137.

Mehar, A.M., Matawie, K., and Maeder, A. (2013). Determining an optimal value of k in k-means clustering. In: *IEEE International Conference on Bioinformatics and Biomedicine*. Shanghai, China, 51 – 55.

Nonparametric estimation of the survival function for ordered multivariate failure time data: a comparative study

Luís Meira-Machado[1], Marta Sestelo[1], Andreia Gonçalves[1]

[1] Centre of Mathematics & Department of Mathematics and Applications, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal

E-mail for correspondence: lmachado@math.uminho.pt

Abstract: In longitudinal studies of disease, patients may experience several events through a follow-up period. In these studies, the sequentially ordered events are often of interest and lead to problems that have received much attention recently. Issues of interest include the estimation of bivariate survival, marginal distributions and the conditional distribution of the second gap time given the first gap time. In this work we consider the estimation for the survival given the first gap time. Different nonparametric approaches will be considered for estimating these quantities, all based on the Kaplan-Meier estimator of the survival function. Real data illustration based on a German breasts cancer study is included.

Keywords: Conditional survival; Gap times; Kaplan-Meier; Nonparametric estimation; Recurrent events.

1 Introduction

In many medical studies individuals can experience several events across a follow-up study. The events of concern can be of the same nature (e.g., cancer patients can experience recurrent disease episodes) or represent different states in the disease process (e.g., alive and disease-free, alive with recurrence and dead). If the events are of the same nature, this is usually referred as recurrent events, whereas if they represent different states they are usually modeled through their intensity functions (Andersen et al., 1993). In this studies several issues are often of interest and lead to problems that have received much attention. Most of the times, one will be interested in describing the distribution of the joint gap times (see e.g., Lin

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

et al., 1999; de Uña-Álvarez and Meira-Machado, 2008; de Uña-Álvarez and Amorim, 2011). In other cases the interest is more focussed in the survival function, such as the estimation of the bivariate survival (Wang and Wells, 1997), the estimation of gap time survival functions or the conditional survival function of the gap times (Wang and Chang, 1999; Schaubel and Cai, 2004). In this work we propose four estimators for the conditional survival function in a three state progressive model. The proposed methods can be easily extended to the k-state progressive model.

2 Nonparametric estimators

Consider n independent and identically distributed pairs of successive failure (gap) times (T_{1i}, T_{2i}), $1 \leq i \leq n$. These pairs of gap times are subject to univariate right-censoring at times C_i which we assume to be independent of (T_{1i}, T_{2i}). Because of this, we only observe $(\widetilde{T}_{1i}, \widetilde{T}_{2i}, \Delta_1, \Delta_2)$ where $\widetilde{T}_{1i} = \min(T_{1i}, C_i)$, $\Delta_{1i} = I(T_{1i} \leq C_i)$, $\widetilde{T}_{2i} = \min(T_{2i}, C_{2i})$, $\Delta_{2i} = I(T_{2i} \leq C_{2i})$ where $C_{2i} = (C_i - T_{1i})I(T_{1i} \leq C_i)$. Let $T = T_1 + T_2$ be the total time and put $\widetilde{T} = \min(T, C)$. Since the censoring time is assumed to be independent of the process, the survival function of the first gap time T_1, say S_1, may be consistently estimated by the Kaplan-Meier estimator based on the $(\widetilde{T}_1, \Delta_1)$. Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on the $(\widetilde{T}_i, \Delta_{2i})$'s. In this work we are interested in the estimation of the conditional survival function $S(y \mid x) = P(T > y \mid T_1 > x)$.

Recently de Uña-Álvarez and Meira-Machado (2008) proposed estimators to empirically estimate the bivariate distribution function for censored gap times. The idea behind estimation is to use the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. Since $S(y \mid x) = P(T > y | T_1 > x) = \frac{P(T>y, T_1>x)}{P(T_1>x)}$, a natural estimator for the conditional survival function is obtained using the same ideas (i.e., Kaplan-Meier weights). The proposed estimator (Kaplan-Meier Weighted Estimator, KMW) is given by $\hat{S}^{\texttt{KMW}}(y \mid x) = \sum_{i=1}^{n} W_i I(\widetilde{T}_{1i} > x, \widetilde{T}_i > y)/\hat{S}(x)$, where $\hat{S}(x)$ is the Kaplan-Meier estimator of survival of T and where W_i are Kaplan-Meier weights attached to $\widetilde{T}_i$ when estimating the marginal distribution of T from $(\widetilde{T}_i, \Delta_i)$'s.

The conditional survival will be hard to estimate in the right tail where censoring effects are stronger. Because of this we consider alternative expressions for the conditional survival $S(y \mid x) = 1 - \frac{P(T_1>x, T\leq y)}{1-P(T_1\leq x)}$. The corresponding estimator (transformed Kaplan-Meier Weighted Estimator, tKMW) can be obtained in a similar way as introduced the KMW estimator.

Another way to introduce a nonparametric estimator for the conditional survival is by considering specific subsamples or portions of data at hand.

For example, given the time point x, to estimate $S(y \mid x) = P(T > y | T_1 > x)$ the analysis can be restricted to the individuals with a first gap time greater than x. Let $n_x = \#\{i : T_{1i} > x\}$ and introduce $\hat{S}^{\texttt{cKMW}}(y \mid x) = 1 - \sum_{i=1}^{n_x} W_i^x I(\widetilde{T}_i \leq y)$, the survival function of T computed from such a subset.

The standard error of the cKMW estimator may be large when the censoring is heavy, particularly with a small sample size. Interestingly, the variance of this estimator may be reduced by presmoothing (de Uña-Álvarez and Amorim 2011). The corresponding presmoothed estimator (Kaplan-Meier presmooth weighted estimator, cKMPW) involves replacing the censoring indicators in the building of the Kaplan-Meier weights, W_i^x, by a smooth fit (e.g. using logistic regression). In the limit case of no presmoothing, the cKMPW estimator reduces to the cKMW estimator.

3 Example of application

To illustrate our methods we will use data from a German Breast cancer study. In this dataset, a total of 686 woman with primary node positive Breast cancer were recruited in the period between 1984 and 1989. From this total 299 developed a recurrence and among these 171 died. For each patient, the two gap times (time to recurrence and time from recurrence to death) and the corresponding indicator status is recorded. Other covariates were also recorded. The covariate recurrence is the only time-dependent covariate, while the other covariates included are fixed. Recurrence can be considered as an intermediate transient state and modeled using a three-state progressive model with states "Alive and disease-free", "Alive with Recurrence" and "Dead". For illustration purposes we show in Figure 1 the plot for $S(y \mid x)$ for all four methods by fixing $T_1 = 1084$ and $T_1 = 1684$. From this plot we can see the behavior of all methods. With the exception of the KMW estimator, all perform similarly. As expected, the cKMPW estimator has less variability.

4 Conclusions

In this paper, the problem of estimating the conditional survival function for ordered multivariate failure time data has been reviewed, and four estimators has been considered. Two new sets of estimators have been proposed. Simulation results, not reported here, reveal that a new proposals perform favorably when compared with the competing methods.

Acknowledgments: The authors acknowledge financial support from FEDER Funds through "Programa Operacional Factores de Competitividade - COMPETE" and by Portuguese Funds through FCT - "Fundação para a Ciência

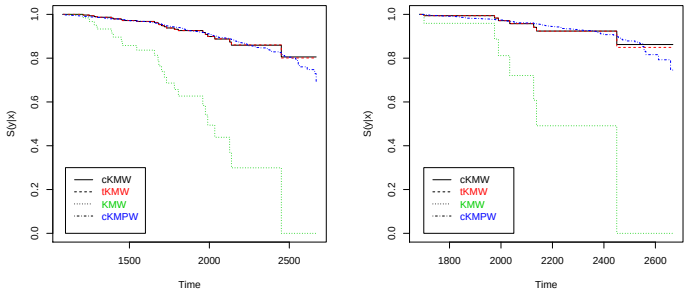


FIGURE 1. Estimated conditional survival for $S(y|x)$, $x = 1084$ (left) and $x = 1684$ right. Breast cancer data.

e a Tecnologia", in the form of grant PEst-OE/MAT/UI0013/2014. Marta Sestelo acknowledges financial support from Grant SFRH/BPD/93928/2013 of Portuguese "Fundação para a Ciência e a Tecnologia" and by FEDER Funds through "Programa Operacional Factores de Competitividade - COMPETE".

References

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer: New York.

de Uña-Álvarez, J. and Meira-Machado, L. (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters*, **78**, 2440 – 2445.

de Uña-Álvarez, J. and Amorim, A.P. (2011). A semiparametric estimator of the bivariate distribution function for censored gap times. *Biometrical Journal*, **53**, 113 – 127.

Lin, D., Sun, W., and Ying, Z. (1999). Nonparametric estimation of the time distributions for serial events with censored data. *Biometrika*, **86**, 59 – 70.

Schaubel, D.E. and Cai, J. (2004). Non-parametric estimation of gap time survival functions for ordered multivariate failure time data. *Statistics in Medicine*, **23**, 1885 – 1900.

Wang, M. and Chang, S. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*, **94**, 146 – 153.

Wang, W. and Wells, M. (1997). Nonparametric estimators of the bivariate survival function under simplified censoring conditions. *Biometrika*, **84**, 863 – 880.

Flexible models in survival analysis: an illustration

Shirin Moghaddam[1], John Hinde[1], Milovan Krnjajić [1]

[1] School of Mathematics, Statistics and Applied Mathematics, NUI Galway, Ireland

E-mail for correspondence: s.moghaddam1@nuigalway.ie

Abstract: This project involves development of flexible statistical regression models with a particular focus on the modelling requirements for types of data that arise in survival analysis. We apply nonparametric Bayesian (NPB) methods, which substantially enhance the flexibility of standard parametric models while providing a full probabilistic framework for inference. Under the NPB paradigm, the unknown distributions of the model are treated as random (infinite-dimensional) parameters necessitating specification of stochastic nonparametric priors, such as Dirichlet or Gaussian proceses, over spaces of distributions. Here we will develop BNP survival models and associated MCMC fitting methods.

Keywords: Time to event data; Survival analysis; Bayesian nonparametric methods; Medical applications.

1 Introduction

A standard way to develop a parametric regression model is to allow parameters to depend on covariates in a pre-specified way. Classical semiparametric methods specify parametrically the regression relationship between the response and covariates, but leave the actual survival distribution unspecified. The disadvantages of these methods stem from inflexible functional forms of parametric models and limited inference of classical semiparametric techniques. In particular, fixed specification of distributional properties for the random error terms in the model, while typically mathematically convenient, may be inadequate for the actual data, which clearly calls for more flexible and robust modelling approaches. We use the nonparametric Bayesian (NPB) methods, which substantially enhance the flexibility of standard parametric models while providing a coherent unified probabilistic framework for inference. We propose NPB regression models for survival

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

data and use Dirichlet process (DP) mixture models as nonparametric priors for the survival distributions.

2 Nonparametric Bayes

A DP mixture model is a mixture with a parametric kernel and a random mixing distribution modeled with a DP prior, see Ferguson (1973, 1974), Antoniak (1974), Escobar and West (1995). The definition of the DP involves a parametric distribution function G_0, the center or base distribution of the process, and a positive scalar precision parameter ν. The larger the value of ν the closer a realization of the process is to G_0. A DP mixture model is given by

$$F(\cdot; G) = \int K(\cdot|\theta)G(d\theta),$$

where $K(\cdot|\theta)$ is the distribution function of the parametric kernel of the mixing and $G \sim \mathrm{DP}(\nu G_0)$.

3 Bayesian nonparametric survival analysis

Bayesian nonparametric methods are very well suited for survival data analysis, enabling flexible modelling for the unknown survival function, cumulative hazard function or hazard function, and providing techniques to handle censoring and truncation. The DP mixture modelling framework allows the incorporation of prior information and provides a rich inferential framework that can be extended to distributions with support on $\mathbb{R}^+$ for application to survival data (Kottas, 2006).

Modelling DP mixtures for the distributions that have support on R usually employs normal kernels, but in the context of survival analysis choosing the kernel becomes more delicate. In this case we usually use Lognormal, Weibull or Gamma kernels (Kottas, 2006). The Weibull distribution seems preferable because it has an increasing hazard and also its survival function is in closed form.

In the case of Weibull distribution where $K_W(t|\alpha, \lambda) = 1 - \exp(-\lambda^{-1}t^{\alpha})$ and $k_W(t|\alpha, \lambda) = \lambda^{-1}\alpha t^{\alpha-1}\exp(-\lambda^{-1}t^{\alpha})$, the base distribution is assumed to be as follows:

$$G_0(\alpha, \lambda|\phi, \gamma) = \mathrm{Uniform}(\alpha|0, \phi) \times \mathrm{IGamma}(\lambda|d, \gamma).$$

Denoting the survival times by t_i, $i = 1, \ldots, n$, following Kottas (2006) the

full model can be written in the hierarchical form:

$$
\begin{aligned}
&t_i|\alpha_i,\lambda_i \overset{ind}{\sim} K_W(t_i|\alpha_i,\lambda_i), \qquad i=1,\ldots,n,\\
&(\alpha_i,\lambda_i)|G \overset{iid}{\sim} G, \qquad i=1,\ldots,\\
&G|\nu,\gamma,\phi \sim \mathrm{DP}(\nu G_0),\\
&\nu,\gamma,\phi \sim \mathrm{Gamma}(\nu|a_\nu,b_\nu)\mathrm{Gamma}(\gamma|a_\gamma,b_\gamma)\mathrm{Pareto}(\phi|a_\phi,b_\phi). \qquad (1)
\end{aligned}
$$

We model the unknown survival distribution with a Weibull Dirichlet process mixture, mixing on both the shape and scale parameters.

4 Simulation based model fitting

Here we present the Gibbs sampler designed to fit model (1). To simplify some of the full conditionals, we assume here that all of the survival times are uncensored. Our initial goal is to generate from (α_i,λ_i), which subsequently allows generation from the predictive distribution t_i.

In Bayesian nonparametric method the discreteness of the random distribution G induces a clustering of (α_i,λ_i). Let n^* be the number of clusters in the vector $((\alpha_1,\lambda_1),\ldots,(\alpha_n,\lambda_n))$ and denoted by (α_j^*,λ_j^*), $j=1,\ldots,n^*$, the distinct (α_i,λ_i). The vector of configuration indicators $s=(s_1,\ldots,s_n)$, defined by $s_i=j$ if and only if $(\alpha_i,\lambda_i)=(\alpha_j^*,\lambda_j^*)$, $i=1,\ldots,n$, determines the clusters. Let n_j be the size of cluster j, i.e., $|\{i:s_i=j\}|$, $j=1,\ldots,n^*$.

Gibbs sampling to draw from $((\alpha_1,\lambda_1),\ldots,(\alpha_n,\lambda_n),\nu,\gamma,\phi|data)$ is based on the following full conditionals:

1. $[(\alpha_i,\lambda_i,s_i)|\{(\alpha_{i'},\lambda_{i'},s_{i'}),i^{'}\neq i\},\nu,\gamma,\phi,data]$, for $i=1,\ldots,n$. Once this step is completed, we have a specific number of clusters (n^*), a specific configuration $(s=(s_1,\ldots,s_n))$, and the associated cluster locations (α_j^*,λ_j^*), $j=1,\ldots,n^*$.

2. $[(\alpha_j^*,\lambda_j^*)|s,n^*,\gamma,\phi,data]$, for $j=1,\ldots,n^*$. This step improves the mixing of the chain by moving the cluster locations.

3. $[\nu|n^*,data],[\phi|\{(\alpha_j^*,\lambda_j^*),\ j=1,\ldots,n^*\},n^*]$ and $[\gamma|\{(\alpha_j^*,\lambda_j^*),\ j=1,\ldots,n^*\},n^*]$. In this step we draw from the full conditionals.

The full conditionals in the first step are obtained by using the Pólya urn representation of a DP. Assume t_i, $i=1,\ldots,n$, corresponds to the uncensored survival times, we could draw from the first step full conditionals based on the following mixed distribution:

$$
\frac{q_0^0 h^0(\alpha_i,\lambda_i|\gamma,\phi,t_i)+\sum_{j=1}^{n^{*-}} n_j^- q_j^0 \delta_{(\alpha_j^*,\lambda_j^*)}(\alpha_i,\lambda_i)}{q_0^0+\sum_{j=1}^{n^{*-}} n_j^- q_j^0},
$$

where $q_j^0 = k_W(t_i|\alpha_j^*, \lambda_j^*)$ and $q_0^0 = \frac{d\nu\gamma^d}{\phi t_i} \int_0^\phi \frac{\alpha t_i^\alpha}{(\gamma + t_i^\alpha)^{d+1}} d\alpha$.

Also $h^0(\alpha_i, \lambda_i|\gamma, \phi, t_i) \propto k_W(t_i|\alpha_i, \lambda_i) g_0(\alpha_i, \lambda_i|\phi, \gamma)$ and g_0 is the density function of G_0. In this step with probability $\frac{q_0^0}{q_0^0 + \sum_{j=1}^{n^{*-}} n_j^- q_j^0}$ we generate a new value from h^0 and otherwise we sample from the previous (α^*, λ^*). In the above formulas "-" denotes all relevant quantities when (α_i, λ_i) is removed from the vector $((\alpha_1, \lambda_1), \ldots, (\alpha_n, \lambda_n))$.

In step 2 for each $j = 1, \ldots, n^*$,

$$[(\alpha_j^*, \lambda_j^*)|s, n^*, \gamma, \phi, data] \propto g_0(\alpha_j^*, \lambda_j^*|\phi, \gamma) \prod_{\{i:s_i=j\}} k_W(t_i|\alpha_j^*, \lambda_j^*).$$

Finally, we go to step 3, where we update ν, the precision parameter of the Dirichlet process, using the augmentation method in Escobar and West (1995). Also the full conditional for ϕ and γ are proportional to $[\phi] \prod_{j=1}^{n^*} \phi^{-1} 1_{(\phi \geq \alpha_j^*)}$ and $[\gamma] \prod_{j=1}^{n^*} \gamma^d \exp(-{\lambda_j^*}^{-1}\gamma)$, where $[\phi]$ and $[\gamma]$ are the prior density in model (1).

5 Future work

We will complete the implementation of this model. Then we will incorporate censoring in the model and, finally, try to use other models instead of the Weibull distribution.

Acknowledgments: Research supported by Irish Research Council (IRC) Government of Ireland Postgraduate Scholarship GOIPG/2013/1314.

References

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *The Annals of Statistics*, **2**, 1152 – 1174.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577 – 588.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209 – 230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615 – 629.

Kottas, A. (2006). Nonparametric Bayesian survival analysis using mixture of Weibull distributions. *Journal of Statistical Planning and Inference*, **136**, 578 – 596.

Spatially adaptive probabilistic temperature forecasting using Markovian EMOS

Annette Möller[1]

[1] Department of Animal Sciences, Biometrics & Bioinformatics Group, Georg-August University of Göttingen, Germany

E-mail for correspondence: annette.moeller@agr.uni-goettingen.de

Abstract: We propose a spatially adaptive extension of the state-of-the-art EMOS postprocessing model. As our method introduces a Markovian dependence structure on the model parameters by employing Gaussian Markov random fields, we call it Markovian EMOS (MEMOS). For fitting the MEMOS model in a Bayesian fashion we utilize the recently developed INLA approach that allows for fast and accurate approximation of the posterior distributions of the parameters. We apply the MEMOS method to 24-h forecasts of surface temperature over Germany for the years 2010 and 2011, using the 50-member ensemble of the European Center for Medium-Range Weather Forecasts (ECMWF). Our proposed method outperforms the raw ensemble and the ensemble postprocessed with standard univariate global and local EMOS.

Keywords: Probabilistic weather forecasting; Ensemble postprocessing; Gaussian Markov random fields, Integrated nested Laplace approximation.

1 Introduction

Ensemble forecast systems aim to reflect and quantify sources of uncertainty in the numerical weather prediction (NWP) model forecasts. However, they tend to be biased and underdispersed (Hamill and Colucci, 1997). Therefore statistical postprocessing methods for ensemble forecasts are required to properly calibrate the resulting distribution.
Univariate postprocessing methods such as Ensemble Model Output Statistics (EMOS), introduced by Gneiting et al. (2005) have been developed successfully for a variety of univariate weather quantities. However, they are designed only for a single weather quantity, for fixed locations and fixed forecast horizons. Therefore, they are not accounting for multivariate dependence structures, as for example inter-variable, spatial and temporal

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

dependencies. To account for spatial dependence structures, we develop a spatially adaptive Bayesian extension of the EMOS model.

2 Markovian EMOS

Let y_s denote the surface temperature at location s, $s = 1, \ldots, n$, where n is the total number of stations that are considered. Further, $x_1, \ldots, x_m$ denote the m ensemble forecasts. Our proposed Markovian EMOS (MEMOS) model then assumes

$$y_s|\eta_s, \sigma^2, x_{1,s}, \ldots, x_{m,s} \sim N(\eta_s, \sigma^2), \quad (1)$$

$$\eta_s = \gamma + a_s + b_s\bar{x}_s. \quad (2)$$

Here, η_s is the linear predictor and σ^2 the variance of the error term ϵ with $\epsilon_s \sim N(0, \sigma^2)$ for all $s = 1, \ldots, n$. Further, γ denotes an overall fixed effect intercept, $\bar{x}_s$ the mean over the m ensemble members at location s, and a_s as well as b_s represent random effects bias correction coefficients. We assume a_s and b_s to be realizations of latent Gaussian fields (GFs) $\mathbf{a}(s)$ and $\mathbf{b}(s)$, respectively.

To fit the MEMOS model in a Bayesian framework and at the same time in an efficient way, we utilize two recently developed approaches. The INLA methodology (Rue et al., 2009) allows to perform fast and accurate approximation of the posterior margins of the parameters. It is combined with the SPDE approach proposed by Lindgren et al. (2011) whose basic building block is the fact that a GF with Matérn covariance function is the stationary solution to a certain linear fractional stochastic partial differential equation (SPDE). By solving the SPDE on a triangulated domain with n vertices, an explicit Gaussian Markov random field (GMRF) representation $x_n(s)$ of the GF $x(s)$ can be obtained via a linear combination of piecewise linear basis functions.

To approximate the posterior margins (especially) of η_s and σ^2 in model (1), the `R-SPDE-INLA` package (see `www.r-inla.org`) is used.

3 Application to ECMWF temperature forecasts

We apply our method to 24-h ahead surface temperature forecasts of the 50-member European Center for Medium-Range Weather Forecasts (ECMWF) ensemble (Buizza, 2006) over Germany. Our data covers the time period February 2, 2010 to April 30, 2011, with a total of $n = 518$ available stations. The verifying observations are provided by the German Weather Service (DWD).

In our case study we compare the predictive performance of the proposed MEMOS model with the raw ensemble and two variants of EMOS, global

and local EMOS. While global EMOS estimates a single set of parameters for all stations, local EMOS estimates a separate set of parameters at each considered station, however, without accounting for spatial dependence structures. Therefore, local EMOS is a natural benchmark for our MEMOS method. For fitting the postprocessing models, a 25 days rolling training period is used.
The predictive performance of the models is measured in terms of several univariate verification scores, namely the continuous ranked probability score (CRPS), the mean absolute error (MAE) and the root mean square error (RMSE) (Wilks, 2011), on the test set containing the dates March 24, 2010 to April 30, 2011. In addition, we investigate the verification rank histogram of the raw ensemble and PIT histograms (Wilks, 2011) of the postprocessing methods to check the calibration of the respective method. Table 1 shows that our proposed MEMOS method clearly outperforms the raw ensemble as well as global and local EMOS in terms of all considered scores. When looking at the rank histograms in Figure 1 the raw ensemble exhibits a strong underdispersion indicated by the U-shape of the histogram. Applying a postprocessing method such as global or local EMOS improves the calibration to a high extent. When inspecting the PIT histogram for MEMOS, a further improvement in calibration is visible. The MEMOS PIT histogram is closest to uniformity in comparison to the PIT histograms of both EMOS versions, while (especially local) EMOS still exhibits a slight underdispersion.

TABLE 1. Predictive performance of the different postprocessing methods and the raw ensemble in terms of univariate verification scores, aggregated over all days and stations in the test set.

	CRPS	MAE	RMSE
Raw ECMWF	2.498	2.807	3.762
Global EMOS	1.792	2.489	3.242
Local EMOS	1.415	1.964	2.551
MEMOS	1.384	1.938	2.508

4 Discussion

We propose a spatially adaptive Bayesian extension of the standard EMOS postprocessing model, that induces a Markovian dependence structure on the model parameters. By this, the predictive performance and calibration can be improved further in comparison to EMOS. In a next step we propose to combine the basic MEMOS model with the ECC method (Schefzik et al., 2013), as ECC is computationally efficient and aims at recovering the multivariate dependency information present in the raw ensemble.

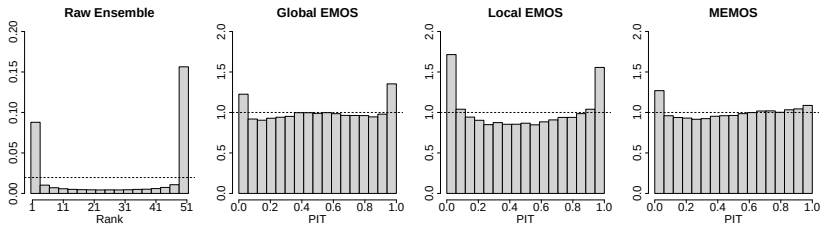


FIGURE 1. Univariate verification rank histogram and PIT histograms over all stations and dates in the test set.

Acknowledgments: The presented work is based on results of the PhD thesis of the author under supervision of Tilmann Gneiting, Thordis Thorarinsdottir and Alex Lenkoski, and was supported by the German Research Foundation (DFG) within the programme "Spatio-/Temporal Graphical Models and Applications in Image Analysis" grant GRK 1653.

References

Buizza, R. (2006). The ECMWF ensemble prediction system. In: Palmer, T.N. (Ed.) *Predictability of Weather and Climate*, Cambridge University Press, 459 – 489.

Gneiting, T., Raftery, A., Westveld, A., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098 – 1118.

Hamill, T.M. and Colucci, S.J. (1997). Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, **125**, 1312 – 1327.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B*, **73**, 423 – 498.

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

Schefzik, R., Thorarinsdottir, T.L., and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, **28**, 616 – 640.

Wilks, D.S. (2011). *Statistical Methods in the Atmospheric Sciences* (3rd ed.). Academic Press.

Estimation of order-restricted mean structure for independent and correlated data using Bayesian variable selection models

Leacky Muchene[1], Ziv Shkedy[1], Tom Jacobs[2]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Hasselt, Belgium,
[2] Janssen pharmaceutical companies of Johnson and Johnson, Beerse, Belgium

E-mail for correspondence: leacky.muchene@uhasselt.be

Abstract: In the pharmaceutical industry, dose-response studies are a central experimental tool for drug development. Within the dose-response framework, order restricted hypotheses are usually of interest when a differential effect of the dose is foreseen. Usually, in determining the correct dose-response profile for the trend or the minimum effective dose, several shapes of the dose-response trend can be fitted to the data and compared with goodness of fit statistics (such as AIC, BIC, etc.). However, estimation and inference ignores the fact that there are different models that can be fitted and estimation and inference is thus based on a selected model. Such a procedure is called a post selection estimation and inference procedure. In order to account for model uncertainty, we propose Bayesian variable selection (BVS) models for order-restricted hypothesis. In BVS, several models are fitted simultaneously and the posterior probability for each model given the observed data computed. Parameter estimates and inference is based on all possible (order-restricted) models and therefore take into account model uncertainty for both estimation and inference. Furthermore, the estimated dose-response curve can be interpreted as a model average of all possible order restricted models that can be fitted to the data. The proposed method is applied to independent and correlated datasets. In particular, the properties of the posterior distribution of the BVS estimates are investigated.

Keywords: Bayesian variable selection models; Dose-response analysis; Hierarchical Bayesian models; Order-restricted models.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

1 Introduction

Statistical modelling of dose-response data usually involves several steps. First, a candidate set of R models $g_1 \dots g_R$ is proposed. For each candidate model g_r, statistical analysis is performed and a comparison between competing models is usually done based on information criteria such as Akaike information criterion and Bayesian information criterion (Lin et al., 2012; Otava et al., 2014). Subsequently, inference for parameters of interest such as the minimum effective dose (MED) are based on the selected model. The shortcoming however is that inference is performed ignoring the uncertainty with regards to the best model amongst candidate models (Cooke, 2009; Claeskens and Hjort, 2008). Several approaches to account for model uncertainty have been proposed (Kato et al., 2006). In particular, within the model averaging approach, estimation and inference is based on the model average of all candidate models. Weights for model averaging are computed using the information criterion for each model. Thus, instead of having a point estimate from the selected 'best' model, a weighted-average of estimates from all models is used (Claeskens and Hjort, 2008). In this paper, we propose a Bayesian variable selection (BVS, O'Hara et .al., 2009) approach to model binary dose-response data while accounting for model uncertainty specifically for monotone increasing dose-response profiles, although the methodology is applicable to non-monotone dose-response profiles as well.

2 Methodology

Consider a binomial outcome Y_i which denotes the number of patients who were pain-free after receiving a dose $i = \{0, 2.5, 5, 10, 20, 50, 100, 200\}$ of a given treatment;

$$\begin{aligned} Y_i &\sim \text{Binomial}(\pi_i, n_i) \\ \text{logit}(\pi_i) &= \beta_0 + \sum_{j=1}^{J} Z_j \delta_j d_i \end{aligned} \tag{1}$$

π_i denotes the probability of being pain-free for patients who received dose i while n_i denotes the total number of patients who received dose i. The parametrization of the dose-response relationship in (1) includes the BVS procedure via an indicator variables Z_j. We specify a hierarchical Bayesian model in which the prior distributions assumed for the model parameters are shown in (2).

$$\begin{aligned} \pi_i &\sim \text{Uniform}(0, 1) \\ \beta_0 &\sim \text{Normal}(0, 0.001) \\ \delta_j &\sim \text{Normal}(0, 0.001) I(0). \end{aligned} \tag{2}$$

The indicator variable Z_j is assumed to be a Bernoulli random variable $Z_j \sim \text{Bernoulli}(\phi_j)$ sampled with a uniform probability $\phi_j \sim \text{Uniform}(0, 1)$.

Note that a monotone increasing dose-response profile is specified by constraining the prior for δ_j to positive values. As shown in Lin et al. (2012), the posterior mean of Z_j denotes the posterior model probability $\bar{Z}_j$ for the model corresponding to that given set of sampled Z_j. Moreover, the posterior mean of ϕ_j denotes the posterior inclusion probability $\bar{\phi}_j$ of the corresponding dose. By using the posterior mean of the model probabilities as the models' weights, a model-averaged estimate of the parameters of interest can be obtained.

3 Application to a binary dose-finding study

Using binomial *migraine* data from the DoseFinding R package. Pinheiro et al. (2013) applied non-linear sigmodal Emax models of different shapes to this data. Since there were seven active doses, there were 128 plausible monotone increasing dose-response profiles resulting from the mean structure specified in (1]. Figure 1 shows the model-averaged fit for the dose-response profile based on the BVS logistic regression model and the sigmoidal Emax model. The BVS model provided a good fit for the data comparable to the sigmoidal Emax model.

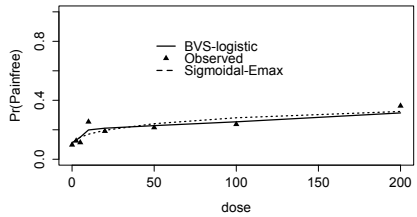


FIGURE 1. Migraine data: Dose-response profile based on Bayesian variable selection as well as sigmoidal Emax model.

The model weights for all the 128 models are shown in Figure 2. The most probable model had a probability of 4%. In addition, only 73 out of the 128 models had a non-zero probability; hence these 73 models are the only ones that contributed to the model-averaged estimates plotted in Figure 2.

4 Conclusion

In this paper, while modelling dose-response profiles, model uncertainty was accounted for through BVS techniques. By introducing appropriate selection indices and specifying appropriately constrained priors, monotone increasing profiles of different shapes were fitted simultaneously. The results

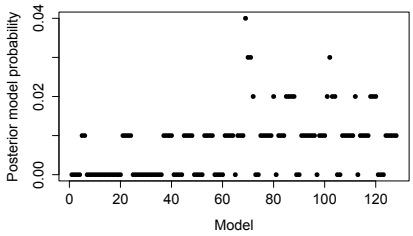


FIGURE 2. Posterior model probability from BVS model fitting. These probabilities are the weights for BMA.

shown in this paper were based on the model averaging approach for which 128 monotone increasing models were simultaneously fitted and averaged. Thus, the simultaneous fitting of the candidate set of models while only requiring the specification of only one model for which all the subsequent models are derived is a desirable property of BVS. While we illustrated the methodology with independent binary data, the concept of Bayesian variable selection is applicable to other types of outcomes as well as for correlated data.

References

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Cooke, R.M. (2009). Quantifying dose-response uncertainty using Bayesian model averaging. *Uncertainty Modeling in Dose Response: Bench Testing Environmental Toxicity*. Wiley, 165 – 79.

Kato, B.S. and Herbert, H. (2006). A Bayesian approach to inequality constrained linear mixed models: estimation and model selection. *Statistical Modelling*, **6**, 231 – 249.

Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijnens, L. (2012). *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R: Order Restricted Analysis of Microarray Data*. Springer.

O'Hara, R. and Sillanpaa, M. (2009). A Review of Bayesian variable selection methods: what, how, and which. *Bayesian Analysis*, **4**, 85 – 118.

Otava, M., Shkedy, Z., Lin, D., Hinrich, W.H.G., Bijnens, L., Talloen, W., and Kasim, A. (2013). Dose-response modeling under simple order restrictions using Bayesian variable selection methods. *Statistics in Biopharmaceutical Research*, **6**, 252 – 262.

Pinheiro, J., Björn, B., Ekkehard, G., and Frank, B. (2013). Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine*, **33**, 1646 – 1661.

The Birnbaum-Saunders generalized-t distribution for positive skewed data

Luiz R. Nakamura[1], Robert A. Rigby[2], Dimitrios M. Stasinopoulos[2], Roseli A. Leandro[1], Cristian Villegas[1]

[1] Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, 13418-900, Piracicaba, São Paulo, Brazil
[2] STORM, London Metropolitan University, London N7 8DB, United Kingdom

E-mail for correspondence: lrnakamura@usp.br

Abstract: In this work we consider a new four parameter distribution, called the Birnbaum-Saunders generalized-t distribution, that includes some other models as special cases, such as the Birnbaum-Saunders-t (BS-t) and Birnbaum-Saunders power exponential (BSPE) distributions. We study its probability density function, provide maximum likelihood estimation and present an application to show the flexibility of this new model, comparing it with Birnbaum-Saunders, BS-t and BSPE distributions.

Keywords: Generalized Birnbaum-Saunders distributions; Generalized t distribution; Heavy-tailed distribution.

1 Introduction

In the last few years different distributions are being created in order to solve problems that demand very flexible models. Díaz-García and Leiva (2005) proposed a generalization of the Birnbaum-Saunders (BS) distribution (Birnbaum and Saunders, 1969) in order to develop really flexible distributions called the Generalized Birnbaum-Saunders (GBS) family of distributions, wherein they present eight new models besides the standard BS. We say that a positive random variable T that follows a GBS distribution, denoted by $T \sim GBS(\alpha, \beta; g)$, can be defined by a transformation from any symmetric model as

$$T = \beta \left[\frac{\alpha X}{2} + \sqrt{\left(\frac{\alpha X}{2}\right)^2 + 1} \right]^2, \qquad x \in \mathbb{R} \tag{1}$$

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

where $\alpha > 0$ represents the shape parameter, β is the scale parameter and also the median of the distribution and $X \sim S(\mu, \sigma; g)$ with location parameter $\mu = 0$ and scale parameter $\sigma = 1$. As $\alpha \to 0$, the BS distribution becomes symmetrical around β, whereas when α grows the distribution becomes positively asymmetric. The probability density function (pdf) corresponding to (1) is given by

$$f_T(t|\alpha, \beta; g) = c\frac{t^{-\frac{3}{2}}(t+\beta)}{2\alpha\beta^{\frac{1}{2}}} g\left(\frac{1}{\alpha^2}\left[\frac{t}{\beta} + \frac{\beta}{t} - 2\right]\right), \qquad t > 0,$$

where $g(\cdot)$ is the kernel of the density of X and c is the normalizing constant such that $f_X(x)$ is a proper density.
In this present work, we will propose a new model, that belongs to the class of GBS distributions, called Birnbaum-Saunders generalized-t (BSGT, for short) distribution, based on the BS transformation (1) from the generalized t (GT) distribution for X. Its usefulness is illustrated through a real data set application regarding female patients who died from lung diseases in the United Kingdom (UK).

2 The BSGT distribution

The GT distribution, $GT(\mu, \sigma, \nu, \tau)$, was proposed by McDonald and Newey (1988) and its pdf is given by

$$f_X(x|\mu, \sigma, \nu, \tau) = \frac{\tau}{2\sigma\nu^{\frac{1}{\tau}} B\left(\frac{1}{\tau}, \nu\right)\left(1 + \frac{|x-\mu|^\tau}{\nu\sigma^\tau}\right)^{\nu+\frac{1}{\tau}}}, \qquad x \in \mathbb{R}, \qquad (2)$$

where $\mu \in \mathbb{R}$ is a location parameter, $\sigma > 0$ is the scale parameter, $\nu > 0$ and $\tau > 0$ control the shape of the density and $B(a, b) = \int_0^1 w^{a-1}(1-w)^{b-1}dw$ is the beta function. Small values of ν, the parameter that can be associated with the student-t distribution will result in heavier tails. Small values of τ, the parameter that can be associated with the power exponential (PE) distribution, will result in thicker tails and a sharper peak, while high values of τ will result in lighter tails and a flatter peak.
Applying the Birnbaum-Saunders transformation (1) to $X \sim GT(0, 1, \nu, \tau)$ gives the Birnbaum-Saunders generalized-t distribution, denoted by $T \sim BSGT(\alpha, \beta, \nu, \tau)$, with pdf given by

$$f_T(t|\alpha, \beta, \nu, \tau) = \frac{\tau t^{-\frac{3}{2}}(t+\beta)}{4\alpha\beta^{\frac{1}{2}}\nu^{\frac{1}{\tau}} B\left(\frac{1}{\tau}, \nu\right)}\left(1 + \frac{1}{\nu\alpha^\tau}\left|\frac{t}{\beta} + \frac{\beta}{t} - 2\right|^{\frac{\tau}{2}}\right)^{-\left(\nu+\frac{1}{\tau}\right)},$$

where $\alpha, \beta > 0$ are defined in (1); and $\nu, \tau > 0$ are the parameters related to the tails of the distribution and as in (2) small values of both ν and τ

result in heavier tails. Similarly, larger values of ν or τ will produce lighter tails and as in (2) low or high values of τ will produce sharper or flatter peaks, respectively.

We provide estimation of the parameters of BSGT distribution through the maximum likelihood method. Let T_i, $i = 1, \ldots, n$, be a random variable following a BSGT distribution with parameter vector $\theta = (\alpha, \beta, \nu, \tau)^\top$. The total log-likelihood function for θ is given by

$$\begin{aligned} l(\theta) &= n\log(\tau) - 2n\log(2) - n\log(\alpha) - \frac{n}{2}\log(\beta) - n\log\left[B\left(\frac{1}{\tau}, \nu\right)\right] \\ &\quad -\frac{n}{\tau}\log(\nu) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) - \left(\nu + \frac{1}{\tau}\right)\sum_{i=1}^{n}\log\left(1 + \frac{|\omega_i|^{\frac{\tau}{2}}}{\nu\alpha^\tau}\right) \\ &\quad +\sum_{i=1}^{n}\log(t_i + \beta) \end{aligned} \tag{3}$$

where $\omega_i = t_i/\beta + \beta/t_i - 2$.

The elements of the score vector of θ are obtained easily by taking the derivatives of (3) with respect of each of its parameters. The maximum likelihood estimate (MLE) $\hat{\theta}$ of θ is obtained solving the equations $U_\alpha(\theta) = 0$, $U_\beta(\theta) = 0$, $U_\nu(\theta) = 0$ and $U_\tau(\theta) = 0$. This could be achieved by a numerical maximization algorithm using the R functions optim or gamlss.

3 Application

The data refer to the monthly female deaths from bronchitis, emphysema and asthma in the UK in years 1974 – 1979 and is presented on Diggle (1990) and is also available on R software. Table 1 displays the MLEs and standard errors (SE) of the BS, BS-t, BSPE and BSGT parameters and their corresponding AIC and BIC values.

We can clearly see that the BSGT distribution outperformed all its special-cases, since it produced the smallest values of AIC and BIC (928.514 and 937.620, respectively). The high SE value from parameter τ could be related to its high correlation with parameter ν. Plots of the fitted distributions are displayed in Figure 1.

4 Conclusion

We presented a new four parameter distribution, called the BSGT distribution and show its flexibility through a real data set application related to deaths caused by lung diseases. Due its flexibility this distribution could be applied to different fields of application.

TABLE 1. MLEs and SE (in parentheses) of the model parameters for the lung diseases data and their corresponding model AIC and BIC values.

Model	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\nu}$	$\hat{\tau}$	AIC	BIC
BSGT	0.454	555.573	0.337	21.200	928.514	937.620
	(0.039)	(13.184)	(0.450)	(5.506)		
BSPE	0.519	561.577	–	4.659	932.227	939.057
	(0.035)	(17.172)	–	(1.129)		
BS-t	0.304	535.683	100*	–	942.961	949.791
	(0.026)	(19.149)	(fixed)	–		
BS	0.306	535.928	–	–	940.587	945.141
	(0.025)	(19.037)	–	–		

*BS-t did not converge and so parameter ν was fixed resulting in the BS distribution

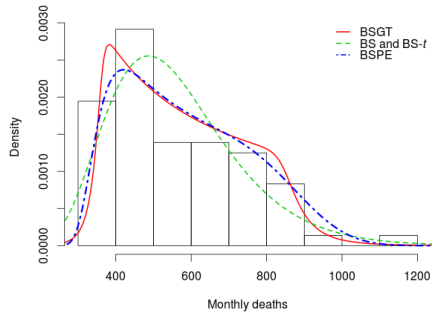


FIGURE 1. Comparison of the fitted distributions to the lung diseases data set.

Acknowledgments: The first author gratefully acknowledge grant from CAPES (Brazil) under the process number 99999.009857/2014-01.

References

Birnbaum, Z.W. and Saunders, S.C. (1969). A new family of life distributions. *Journal of Applied Probability*, **6**, 319 – 327.

Díaz-García, J.A. and Leiva, V. (2005). A new family of life distributions based of the elliptically contoured distributions. *Journal of Statistical Planning and Inference*, **128**, 445 – 457.

Diggle, P.J. (1990). *Time Series: A Biostatistical Introduction*. Oxford: OUP.

McDonald, J.B. and Newey, W.K. (1988). Partially adaptive estimation of regression models via the generalized t distribution. *Econometric Theory*, **4**, 428 – 457.

Inference in nonlinear differential equations

Mu Niu[1], Maurizio Filippone[2], Dirk Husmeier[1], Simon Rogers[2]

[1] School of Mathematics and Statistics, University of Glasgow, Glasgow UK
[2] School of Computing Science, University of Glasgow, Glasgow UK

E-mail for correspondence: mu.niu@glasgow.ac.uk

Abstract: Parameter inference in mechanistic models of coupled differential equations is a challenging problem. We propose a new method using kernel ridge regression in Reproducing Kernel Hilbert Spaces (RKHS). A three-step gradient matching algorithm is developed and applied to a realistic biochemical model.

Keywords: RKHS; Nonlinear ordinary differential equation; Gradient matching.

1 Introduction

Many processes in science and engineering can be described by dynamical systems models based on nonlinear ordinary differential equations (NODEs). However, the parameters of the NODEs are often unknown and not directly measurable. Direct inference requires a computationally expensive numerical integration of the NODEs every time the parameters are adapted. For that reason, approximate methods based on gradient matching have recently gained much attention; see e.g. Dondelinger et al. (2013) and Heinonen et al. (2014). The purpose of the present article is to try a new variant of this approach based on reproducing kernel Hilbert spaces (RKHS). We consider systems governed by first order multivariate NODEs:

$$\dot{x} = \frac{dx}{dt} = f\big(x(t), \theta\big), \tag{1}$$

with initial value $x(t_1) = x_1$, where $x(t)$ is an r dimensional vector of state variables ($x(t) \in \mathbb{R}^r$). We observe the states of the system at n time points $(y_1, ..., y_n)$ and assume that the observations consist of the states corrupted by Normal additive iid noise $y_i = x(t_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ iid. In this work, a framework of penalized regression for vector-valued functions in RKHS is used to make predictions of the system state. A three step gradient matching approach is developed to learn the NODEs parameters:

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

1. Learn a smooth function g over the observations using an RKHS approximation with observation $(y_1, \ldots, y_n)$ at time points $(t_1, \ldots, t_n)$.
2. Given g, learn the NODEs parameters θ by gradient matching to minimize the difference between $\dot{g}(t)$ and $f(g(t), \theta)$.
3. Re-estimate g using the optimised θ and observed trajectory y. Given re-estimated g, optimise θ as in step 2.

2 Methodology

For each state variable x_s indexed by $s = 1, \ldots, r$, a positive definite kernel $k_s : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is used to define the Hilbert space $\mathcal{G}_s$ of the smoother function g_s. The squared exponential kernel (Gaussian kernel) is used in this study with lengthscale parameter l: $k(t_k, t_i) = \exp\big(-l^{-2}(t_k - t_i)^2\big)$. In step one, kernel ridge regression is used to learn g_s, the smoother of the sth state with the following loss function:

$$\mathcal{L}_1(l_s) = \sum_{i=1}^{n} (g_s(t_i) - y_{s_i})^2 + ||g_s||^2 \tag{2}$$

$$g_s(t_i) = \sum_{k=1}^{n} b_{s_k} k_s(t_k, t_i) \tag{3}$$

$$\bar{b}_s = (K_s + \lambda_s I)^{-1} \overline{y}_s, \qquad ||g_s||^2 = \lambda_s \bar{b}_s^T K_s \bar{b}_s, \tag{4}$$

where $k_s(\cdot, t_i)$ is the ith basis function and K_s is the Gram kernel matrix. $\overline{y}_s$ is the vector of observations for the sth state, $||g_s||$ is the norm of $\mathcal{G}_s$ and the regularisation factor λ_s is estimated using leave-one-out cross validation. I is the $n \times n$ identity matrix. Different states are assumed to have different lengthscales, l_s, and these are estimated independently through a gradient-based quasi-Newton optimisation routine. With the optimised l_s, the state estimation $g_s(t_j)$ of $x_s(t_j)$ is made on a uniform grid of m time points indexed by $j = 1, \ldots, m$. The gradient of the smoother is:

$$\dot{g}_s(t_j) = \sum_{i=1}^{n} b_{s_i} \frac{dk(t_i, t_j)}{dt_i}. \tag{5}$$

In step 2, the NODE parameters are estimated by gradient matching for all states with the loss function:

$$\mathcal{L}_2(\theta) = \sum_{s=1}^{r} \sum_{j=1}^{m} \Big(\dot{g}_s(t_j) - f\big(g_s(t_j), \theta\big)\Big)^2, \tag{6}$$

where $f(g_s(t_j), \theta)$ is the target gradient generated by feeding the output of the sth smoother, $g_s(t)$, into the NODEs system. θ is optimised using a gradient-based quasi-Newton routine.

Finally, in step 3, we use the estimated θ from step 2 to construct the NODE system. The lengthscale vector $\bar{l}$ for all state variables is then re-estimated using the full loss function.

$$\mathcal{L}_3(\bar{l}) = \sum_{s=1}^{r} \Big(\sum_{i=1}^{n} (g_s(t_i) - y_{s_i})^2 + ||g_s||^2 + \sum_{j=1}^{m} (\dot{g}_s(t_j) - f\big(g_s(t_j), \theta\big)\big)^2 \Big). \quad (7)$$

The updated smoother $\tilde{g}(t)$ is calculated using $\bar{l}$. The refined θ can then be optimised iteratively by repeating the second step.

3 Application: calcium model

The calcium model (Peifer and Timmer, 2007) represents the oscillations of calcium signaling in eukaryotic cells via a dynamic system with states corresponding to the concentrations of free calcium in cytoplasm C_{ac} and endoplasmic reticulum C_{ar}, as well as active G_α. and phospholipase-C, P_c.

$$\begin{aligned}
\frac{dG_\alpha}{dt} &= k_1 + k_2 G_\alpha - k_3 P_c R_1(G_\alpha) - k_4 C_{ac} R_2(G_\alpha) \\
\frac{dP_c}{dt} &= k_5 G_\alpha - k_6 R_3(P_c) \\
\frac{dC_{ac}}{dt} &= k_7 P_c C_{ac} R_4(C_{ar}) + k_8 P_c + k_9 G_\alpha - k_{10} R_5(C_{ac}) - k_{11} R_6(C_{ac}) \\
\frac{dC_{ar}}{dt} &= -k_7 P_c C_{ac} R_4(C_{ar}) + k_{11} R_6(C_{ac}), \qquad (8)
\end{aligned}$$

where $R_i(x) = \frac{x}{x+km_i}$. We followed Oates et al. (2014) and fixed the km_i parameters, leaving $k_{1:11}$ to be inferred.

Three scenarios were tested: the noise-free case, and adding iid Gaussian noise with signal-to-noise ratio $SNR = 50db$ and $SNR = 10db$ to the numerical solution of the model. We sampled $n = 100$ ($t_1 = 0, t_{100} = 20$) regularly spaced observations, and made state predictions on $m = 200$ grid points. For $10db$ and $50db$ noise, 100 independent data sets were generated. The mean NODEs parameter estimates with error bars are shown in Figure 1. We see that when no noise is added, our estimates closely agree with the true values. The parameter estimates at $SNR = 50db$ noise are similar to the noise-free case when the parameters are smaller than 10, but drift away from the true value for parameters larger than 10. For the highest noise level, $SNR = 10db$, the NODEs parameter estimates become worse, although we still have 4 reliable estimates out of 11 parameters. For the $SNR = 50db$ case, the length scale parameters $l_{1:4}$ for the 4 states at the optimum after step 2 are $[0.06, 0.18, 0.01, 0.05]$. After step 3, these change to [0.06,0.81,0.22,0.79], suggesting that a regularising effect from the NODEs exists.

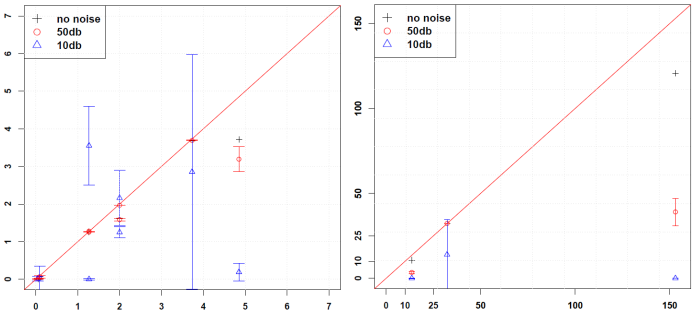


FIGURE 1. Scatter plot of true versus estimated NODEs parameters in three noise scenarios. No noise (black crosses), $SNR = 50db$ (red circles) and $SNR = 10db$ (blue triangles) noise. Left and right plots show parameters with true values < 10 and > 10 respectively.

4 Conclusion

We have described an RKHS based gradient matching approach for parameter inference in a system of NODEs. In low noise scenarios we obtain good parameter estimates for a realistic biochemical application, and we have quantified the deterioration resulting from increased noise levels. We have demonstrated that in the proposed three-step algorithm, the length scale parameters of the kernels are regularised by the gradient matching step. A potential explanation for the poor performance in the high-noise regime is related to the choice of kernel. In our future work, we will explore the effect of using non-stationary kernels (like the MLP kernel) and less smooth kernels (like the Matérn class kernels) as alternatives to the squared exponential kernel.

References

Dondelinger, F., Fillipone, M., Rogers, S., and Husmeier, D. (2013). ODE parameter inference using adaptive gradient matching with Gaussian processes. *JMLR W&CP*, **31**, 216 – 228.

Heinonen, M. and d'Alch-Buc, F. (2014). Learning nonparametric differential equations with operator-valued kernels and gradient matching. *arXiv:1411.5172*.

Oates, C.J., Dondelinger, F., Bayani, N., Korkola, J., Gray, J.W., and Mukherjee, S. (2014). Causal network inference using biochemical kinetics. *Bioinformatics*, **30**, i468 – i474.

Peifer, M. and Timmer, J. (2007). Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting. *Systems Biology, IET*, **1.2.**, 78 – 88.

Emulation of ODEs with Gaussian processes

Umberto Noè[1], Maurizio Filippone[2], Dirk Husmeier[1]

[1] School of Mathematics and Statistics, University of Glasgow, UK
[2] School of Computing Science, University of Glasgow, UK

E-mail for correspondence: u.noe.1@research.gla.ac.uk

Abstract: Inference in nonlinear ordinary differential equations (ODEs) is challenging due to the high computational complexity of the numerical integration. In the present paper, we explore an emulation-based approach for approximating the likelihood, based on Gaussian processes, with the objective to reduce the number of numerical integration steps. We assess the viability of the scheme on the Lotka-Volterra model of predator-prey interactions.

Keywords: Gaussian processes; Kernel density estimation; Likelihood; ODEs.

1 Introduction

Ordinary Differential Equations (ODEs) arise in the study of several areas of science and engineering. However, carrying out parameter inference is challenging for a number of reasons. First, the likelihood can be highly multi-modal. Second, direct numerical integration of ODEs for several settings of the parameters can be prohibitively expensive to be feasible. Motivated by encouraging results reported in Wilkinson (2014), we explore the feasibility of emulation based on Gaussian processes (GPs) for accelerated inference in ODEs such that an explicit numerical solution is only required for a comparatively small set of parameters. We report an experimental evaluation of the proposed emulation approach on a two-parameter Lotka-Volterra (LV) model where we can gain insights into the potential and the limitations of the proposed method.

2 GP emulation for ODEs

A general continuous time dynamical system described by the interaction of S state variables can be modelled by a functional equation of the form

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t); \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Omega[\boldsymbol{\theta}] \subset \mathbb{R}^M$, where the S states at time t are $\boldsymbol{x}(t) = [x_1(t), \ldots, x_S(t)]^{\mathrm{T}}$ and the vector valued function describing their evolution over time is $\boldsymbol{f} \equiv [f_1, \ldots, f_S]^{\mathrm{T}}$. We define $\boldsymbol{X} = [\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_T)]^{\mathrm{T}} \in \mathbb{R}^{T \times S}$ to be the numerical solution of the ODE for a set of times $\boldsymbol{t} = [t_1, \ldots, t_T]^{\mathrm{T}}$ and we assume that observations $\boldsymbol{Y}$ are a noisy realisation of $\boldsymbol{X}$. Any optimisation or inference scheme for ODE parameters would entail repeatedly solving ODEs for different configurations of the parameters. Instead, we propose an emulation approach based on GPs as follows. We consider N parameter configurations $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_n\}_{n=1}^N$. For each of the N configurations of the parameter $\boldsymbol{\theta}$ we compute the numerical solution $\boldsymbol{X}(\boldsymbol{\theta}_n)$ of the ODE. Then we compare these numerical solutions with the noisy signal $\boldsymbol{Y}$ using the Residual Sum of Squares (RSS) score $\mathsf{rss}(\boldsymbol{\theta}_n) = \|\mathsf{vec}[\boldsymbol{Y}] - \mathsf{vec}[\boldsymbol{X}(\boldsymbol{\theta}_n)]\|_2^2$ obtaining the vector $\tilde{\boldsymbol{r}} = [\mathsf{rss}(\boldsymbol{\theta}_1), \ldots, \mathsf{rss}(\boldsymbol{\theta}_N)]^{\mathrm{T}}$. We then fit a GP to the training dataset $\mathcal{D} = \{(\boldsymbol{\theta}_n, r_n)\}_{n=1}^N = \{\boldsymbol{\Theta}, \boldsymbol{r}\}$, where the training vector $\boldsymbol{r}$ is the normalised $\tilde{\boldsymbol{r}}$ vector: the latter minus its mean and divided by its standard deviation, $\boldsymbol{r} = \beta^{-1}(\tilde{\boldsymbol{r}} - \alpha \mathbf{1})$, where $\alpha = \tilde{\boldsymbol{r}}^{\mathrm{T}}\mathbf{1}/N$ and $\beta = (N-1)^{-1/2}\|\tilde{\boldsymbol{r}} - \alpha\mathbf{1}\|_2$. In this way, we can infer the optimal ODE parameters relying on the GP emulation. The proposed hierarchical non-parametric Bayesian model makes use of a Squared Exponential kernel whose hyperparameters are $\boldsymbol{\psi} = [\tau^2, l_1, \ldots, l_M]^{\mathrm{T}}$. The regression model $r_n = z(\boldsymbol{\theta}_n) + \epsilon_n$ postulates our training outputs as observations from a latent function $z(\cdot)$ which is given a Gaussian process prior, corrupted by additive i.i.d. $\mathsf{N}(0, \upsilon^2)$ noise:

$$\begin{aligned} \boldsymbol{r} \mid \boldsymbol{z}, \upsilon^2 &\sim \mathsf{N}_N(\boldsymbol{z}, \upsilon^2 \boldsymbol{I}) \\ z(\cdot) \mid \boldsymbol{\psi} &\sim \mathsf{GP}(m(\cdot), \kappa_{\boldsymbol{\psi}}(\cdot, \cdot)) \\ \boldsymbol{\psi}, \upsilon^2 &\sim P(\tau^2)P(l_1)\ldots P(l_M)P(\upsilon^2), \end{aligned} \tag{1}$$

where $\boldsymbol{z} = [z(\boldsymbol{\theta}_1), \ldots, z(\boldsymbol{\theta}_N)]^{\mathrm{T}}$, $\boldsymbol{r} = [r_1, \ldots, r_N]^{\mathrm{T}}$ and as a consequence of normalisation we assume that $m(\boldsymbol{\theta}) \equiv 0$ for all $\boldsymbol{\theta} \in \Omega[\boldsymbol{\theta}]$. The GP formulation yields predictions for the normalised RSS score $r(\cdot)$ corresponding to any ODE parameters $\boldsymbol{\theta} \in \Omega[\boldsymbol{\theta}]$ using standard properties of GPs. In the case of observations $\boldsymbol{Y}$ assumed to be distributed as a Gaussian centred at the solution of the ODE with variance σ^2, we can interpret our approach as emulating the negative logarithm of a power of the likelihood of the model: $\ell(\boldsymbol{\theta}) = \log P(\mathcal{D} \mid \boldsymbol{\theta}) = \text{const} - \frac{1}{2\sigma^2}\mathsf{rss}(\boldsymbol{\theta}) = \text{const} - \frac{\beta}{2\sigma^2} r(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Omega[\boldsymbol{\theta}]$.

3 Experimental evaluation

We test the performance of our method on a non-standard variant of the Lotka-Volterra system, introduced in Mingari Scarpello and Ritelli (2003):

$$\dot{x}_1(t) = a\,[1 - \exp\{x_2(t)\}] \qquad \dot{x}_2(t) = -c\,[1 - \exp\{x_1(t)\}]\,, \tag{2}$$

where the components of $\boldsymbol{x} \equiv [x_1, x_2]^{\mathrm{T}}$ represent the populations of 'log' preys and 'log' predators respectively. We obtained a numerical solution

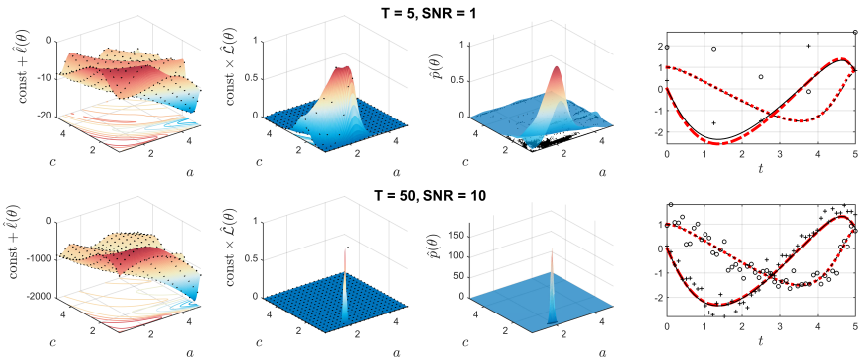


FIGURE 1. The emulated log likelihood (left); the emulated likelihood (centre left); the kernel density estimate of the optima from 5000 different datasets (centre right); the ODE solution using $\boldsymbol{\theta}_{KDE} = \arg\max_{\boldsymbol{\theta}} \hat{p}(\boldsymbol{\theta})$ as parameter shown in red and the true signal shown in black (right) for two different T and SNR settings. See text for details.

$\boldsymbol{X}$ at times $\boldsymbol{t}$ in a given scenario where $\boldsymbol{\theta} = \{a, c\} = \{2, 1\}$ and we added i.i.d. Normal noise with signal-to-noise ratio SNR to obtain $\boldsymbol{Y}$. The priors over GP hyper-parameters used in our model are $P(\tau^2) = \mathsf{Ga}(\tau^2 \mid 1, \text{rate} = 10)$; $P(\log(l_1)) \propto 1$, $P(\log(l_2)) \propto 1$ and v^2 was fixed to 10^{-5}. We then obtained a grid of $G = 20$ values for the a, c parameters around their true values $\{2, 1\}$. Figure 1 shows the emulated log-likelihood, $\text{const} + \hat{\ell}(\boldsymbol{\theta}) = -\beta/(2\sigma^2) m^{\star}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) - \max[-\beta/(2\sigma^2) m^{\star}_{\boldsymbol{\zeta}}(\boldsymbol{\theta})]$, on the left and the emulated likelihood, $\text{const} \times \hat{\mathcal{L}}(\boldsymbol{\theta}) = \exp\{\text{const} + \hat{\ell}(\boldsymbol{\theta})\}$, on the centre left; both transformations of the GP posterior mean $m^{\star}_{\boldsymbol{\zeta}}(\cdot)$ with hyperparameters $\boldsymbol{\zeta} = \{\tau^2, l_1, l_2, v^2\}$ optimised by taking the Maximum A Posteriori (MAP).

Given the GP posterior mean we estimated the ODE parameters by the maximum of the emulated likelihood, using a trust-region-reflective algorithm. We optimised $m^{\star}_{\boldsymbol{\zeta}}(\cdot)$ in 5000 different datasets, each with 50 starting points for $\boldsymbol{\theta}$ in the interval $[0.75, 5]^2$. The chosen design included the first three smallest values in $\boldsymbol{r}$ and we added 47 randomly initialised parameter configurations. The simulation results show that the likelihood is characterised by some local optima in the case of higher uncertainty ($T = 5$ and $\text{SNR} = 1$) while in the case $T = 50$ and $\text{SNR} = 10$ we find a distribution of points scattered around the true configuration $(2, 1)$. In order to more clearly understand the distribution of the optima we fitted a multivariate kernel density estimator (KDE), shown in Figure 1 (centre right). The argument that maximises the density estimate in the first scenario is $\boldsymbol{\theta}_{KDE} = [2.18, 0.96]^{\mathrm{T}}$ and $\boldsymbol{\theta}_{KDE} = [1.98, 1.01]^{\mathrm{T}}$ in the last. This allows us to make a comparison in the parameter space with the true configuration $\boldsymbol{\theta} = [2, 1]^{\mathrm{T}}$. In order to compare the estimates with the true parameter in

the function space instead, in Figure 1 (right) we show the solution of the ODE for each $\boldsymbol{\theta}_{KDE}$ (in red) and we can compare it with the true signal (in black). The symbols + and ∘ represent the noisy observations on preys and predators respectively. We can see that the maximum of the distribution of the optimal parameters, obtained by using a KDE applied to a sample from 5000 independent data instantiations, is an approximately unbiased estimator of the true ODE parameter vector.

4 Conclusions and future work

In this paper we investigated an emulation approach based on GPs to optimise ODE parameters. The emulation entails fitting a GP to a normalised version of the RSS evaluated on a set of configurations of parameter where ODEs are explicitly solved.
Working with a GP-based emulator of the normalised RSS has strong advantages over direct numerical integration of the ODEs. The hyperpriors on the GP hyperparameters have an easy interpretation and the computational time was reduced. To solve the same problem by dealing with the true RSS using an interior-point algorithm we would need 12 hours and a half, while in only 5 hours we solved half a million optimisation tasks which involved fitting 10000 different GPs.
Also, the results show that the parameter configuration that has the highest estimated density of the optimised parameters is a very good estimate of the true one. In our future work, we will apply the proposed method to more complex ODEs with higher-dimensional parameter vectors, and we will investigate the use of adaptive design strategies.

Acknowledgments: UN is grateful to Biometrika Trust for his research studentship. MF and DH are partially funded by the EPSRC (EP/L020319/1).

References

Mingari Scarpello, G. and Ritelli, D. (2003). A new method for the explicit integration of Lotka-Volterra equations. *Divulgaciones Matemáticas*, **11**, 1 – 17.

Wilkinson, R. (2014). Accelerating ABC methods using Gaussian processes. *JMLR-WCP (AISTATS)*, **33**, 1015 – 1023.

Within lake clustering of high resolution satellite retrievals - a functional data and clustering approach

Ruth O'Donnell[1], Claire Miller[1], Marian Scott[1]

[1] University of Glasgow, UK

E-mail for correspondence: ruth.haggarty@glasgow.ac.uk

Abstract: Satellite retrievals for lakes provide high resolution spatiotemporal images. This paper proposes a within lake dimensionality reduction approach using functional data analysis to enable computationally efficient clustering of temporal patterns within lakes.

Keywords: Remote sensing; Dimensionality reduction; Functional data; Clustering.

1 Introduction

Earth Observation instruments such as MERIS (Medium-Spectral Resolution, Imaging Spectrometer) and AATSR (Advanced Along-Track Scanning Radiometer) from the European Space Agency's (ESA's) Envisat satellite platform have been commonly used for ocean color and sea surface temperature retrievals, respectively. Recent developments have enabled these instruments to now be applied to lakes to investigate lake water quality and lake surface water temperature (LSWT). These expansive spatiotemporal data sets simultaneously enable global assessment of environmental changes and present new statistical challenges.
GloboLakes (www.globolakes.ac.uk) is a 5-year Natural Environment Research Council consortium project involving 6 UK research groups. One of the aims of Globolakes is to investigate temporal coherence (similarities in major fluctuations in a set of time series) of water quality for 1000 lakes using a 20-year archive of satellite based spatial images (e.g. bi-monthly at 1° resolution). Determinands of interest include temperature, chlorophyll and coloured dissolved organic matter.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

The aim of this paper is to present a dimensionality reduction and functional data approach for the clustering of remotely sensed spatiotemporal data from within lakes. The approach proposed is applied to LSWT data and provides a means of dealing with long periods of ice cover in some lakes.

1.1 Data

The ESA funded ARC-Lake project (MacCallum and Merchant, 2012) has employed the use of the AATSR instrument in order to derive observations of LSWT data for a large number of lakes across the globe. The data considered in this paper are spatially and temporally complete reconstructions of the ARC-Lake LSWT products for one lake, Lake Superior, from the ARC-Lake version 3 data-set (see www.geos.ed.ac.uk/arclake/data for details). The data-set for the lake is comprised of bi-monthly spatial images (each with 4094 pixels) for the 18-year period from 1995 to 2012. The time series includes periods of zeros which indicate time points with ice cover.

2 Methods

We propose initially to reduce the dimensionality of the individual lake spatiotemporal images. For each pixel within each lake, a functional data analysis (FDA) approach has been taken where each time series is represented as,

$$y_i(t) = G_i(t) + \epsilon_i(t) \qquad i = 1, \ldots, n,\ t = 1, \ldots, T,$$

where G_i is a smooth curve and ϵ_i is a normally distributed independent random error term. The curve G_i is a spline function of degree d which can be expressed as a linear combination of B-splines, written in the following functional form for the spline

$$G_i(t) = \sum_{l=1}^{K+d-1} \beta_{i,l} B_l(t),$$

where $\beta_i = (\beta_{i,l}, \ldots, \beta_{i,K+d-1})^{\mathrm{T}}$ is a vector of real-valued coefficients, $B = (B_1(t), \ldots, B_{k+d-1}(t))^{\mathrm{T}}$ are the B-spline basis functions and K is the number of knots. To accommodate periods of ice cover we propose using an over-saturated basis to represent the curve, removing basis functions at periods of ice cover and then applying a smoothing parameter to fit the curve. Therefore, the β_i vector is estimated by least squares with a penalty:

$$(B^{\mathrm{T}}B + \lambda D)^{-1} B^{\mathrm{T}} y,$$

where λ is the smoothing parameter and D is a penalty matrix based on the integral of the squared second derivative of G. The curve G_i is then approximated by $\hat{G}_i(t)$.

Functional principal components analysis (FPCA) is then applied to the smooth curves, $\hat{G}$, to reduce dimensionality, and hence identify the dominant modes of variation in the data set. This provides a very computationally efficient way of exploring any underlying structure in the data, producing functional component scores:

$$f_{ik} = \int \xi_k(t) G_i(t) dt, \qquad i = 1, \ldots, N,\ k = 1, \ldots, K,\ t = 1, \ldots, T,$$

where ξ is eigenfunction k. Finally, a variety of clustering approaches have been applied (k-means, hierarchical and model-based (Fraley and Raftery, 1998)) to the functional component scores (formed as $f_{ik} w_j$). These have been adjusted to identify coherent regions within each lake, where w_j is a weight to account for the proportion of variability each functional component explains. Spatial correlation is accounted for via weights within the clustering procedure.

3 Results

Curves were fitted to the time series data for each pixel in Lake Superior using a cubic b-spline basis of 150 equally spaced functions. The smoothing parameter was set so that, after removal of basis functions in areas where there were zeros, there was one degree of freedom per 3 month season (66 in total). An example of a fitted curve for a single pixel is shown in the left panel of Figure 1 where points represent the data and the solid line represents the fitted smooth curve. As can be seen, there is good agreement between the fitted curve and the observations and the curve has captured the periods of ice cover well.

After applying FPCA and clustering, the statistically optimal number of clusters for describing the underlying variability in the pixel curves can be determined using approaches such as the L-curve or Gap statistic (Tibshirani et al., 2001). Four clusters was found to be statistically optimal in terms of describing variability in Lake Superior. The clusters are displayed on a map of Lake Superior in Figure 1 (right). There were clear distinctions between the mean functions corresponding to these groups with the key discrepancy being the amplitude of the seasonal patterns each year. Group 1 pixel curves had the greatest amplitude (cooler in winter, warmer in summer) whilst Group 4 had the smallest.

4 Conclusions

The use of weighted functional PCs based on penalised over-saturated B-splines enables the data dimensions to be reduced substantially, while appropriately accounting for sequences of zeros in the time series. While methods have been developed for LSWT, we are currently applying and extending them for water quality measures such as chlorophyll.

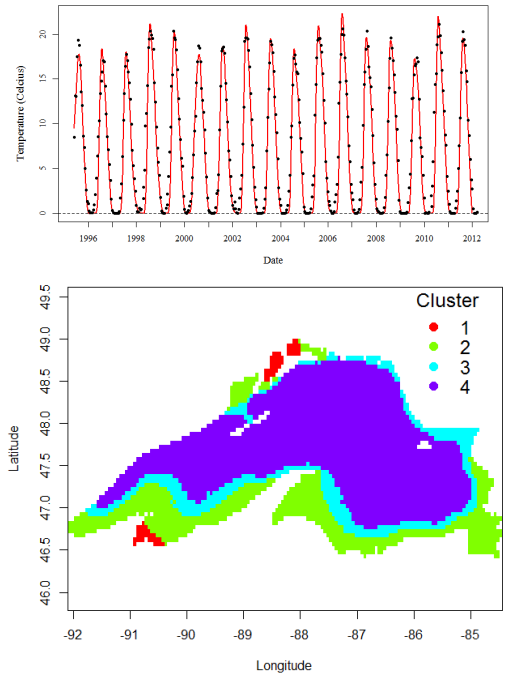


FIGURE 1. Left: Time series for a randomly selected pixel (points) and fitted spline function (solid line). Right: Map of Lake Superior, Canada, showing spatial distribution of estimated clusters.

Acknowledgments: O'Donnell, Scott and Miller were partly funded for this work through the NERC GloboLakes project (NE/J022810/1). The authors gratefully acknowledge the ARC lake project for access to the data.

References

Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, **41**, 578 – 588.

MacCallum, S.N. and Merchant, C.J. (2012). Surface water temperature observations of large lakes by optimal estimation. *Canadian Journal of Remote Sensing*, **38**, 25 – 45.

Tibshirani, R, Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B*, **63**, 411 – 423.

Bayesian model averaging for copula-based estimation of upper tail dependence in loss distributions

Adrian O'Hagan[1]

[1] School of Mathematical Sciences, University College Dublin, Ireland

E-mail for correspondence: adrian.ohagan@ucd.ie

Abstract: An important duty for actuarial and financial modelers is to assess the dependence between loss distributions from related lines of business. Copulas have become a prevalent tool for analyzing the characteristics of such random variables, including tail dependence. A range of parametric copulas is available for such tasks, but it is not always clear which copula provides optimal fit to a given data set. Bayesian model averaging is introduced as a suitable tool for combining information from multiple candidate copula models for loss distributions and tested on both real and simulated data sets.

Keywords: Bayesian model averaging; Copulas; Extreme loss distributions; Financial modeling; Upper tail dependence coefficient.

1 Data

Two data sets are used to assess the performance of Bayesian model averaging for copula-based estimation of upper tail dependence.

1) Bivariate-t distribution using real loss data – 300 bivariate-t data points are simulated from a t copula based on an underlying dependence structure sourced from *real* insurance loss data. Hence the t copula should provide the best fit and give the most accurate estimate of the upper tail dependence coefficient λ_u, where the true λ_u can be calculated in advance and implies weak upper tail dependence.

2) Bivariate gamma and beta distribution using simulated loss data - 300 bivariate data points from gamma and beta distributions respectively are simulated, based on an underlying dependence structure sourced from *simulated* insurance loss data. The t copula should again provide the best fit

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

and give the most accurate estimate of λ_u, where the true λ_u is again known in advance and implies strong upper tail dependence.

2 Methods

2.1 Copulas

Sklar's Theorem (Sklar, 1959) describes the dependence between two or more random variables $X_1, \ldots, X_d$. It states that the joint cumulative distribution function (CDF) of the random variables, $H(x_1, \ldots, x_d)$, can be expressed as a copula function C of the marginal CDFs, $F_1(x_1), \ldots, F_d(x_d)$ and is unique if all marginal CDFs are continuous:

$$\begin{aligned} H(x_1, \ldots, x_d) &= P(X_1 \leq x_1, \ldots, X_d \leq x_d) \\ H(x_1, x_2, \ldots x_d) &= C(F_1(x_1), \ldots, F_d(x_d)). \end{aligned}$$

2.2 The upper tail dependence coefficient

Upper tail dependence between two random variables is the phenomenon whereby knowledge of the realisation of a large (tail) value for one random variable increases the probability of a large value being realised for the other random variable. The concept extends to the general case of multiple random variables. The upper tail dependence coefficient (Fischer and Klein, 2007), λ_u, is a numerical value that captures this information. For the bivariate case with marginal random variables X and Y, the upper tail dependence coefficient λ_u is defined as:

$$\lambda_u = \lim_{\eta \to -1} \left[P(Y > F_Y^{-1}(\eta) \,|\, X > F_X^{-1}(\eta)) \right] = \lim_{\eta \to -1} \left[\frac{(1 - 2\eta + C(\eta, \eta))}{(1 - \eta)} \right].$$

The copulas selected for consideration under the Bayesian model averaging framework here are the t copula, the Gumbel copula and the Joe copula, all of which are gaining increasing traction in the insurance industry (Demarta and McNeil, 2005). However the method can easily be extended to a wider range of copulas. The t copula has been envisaged as a potential widespread successor to the Gaussian copula, given its incorporation of tail dependence (namely symmetric upper and lower tail dependence). Its upper tail dependence coefficient is calculated according to the formula:

$$\lambda_{u,t} = 2P\left(t_{\eta+1} < \frac{-\sqrt{\eta + 1}\sqrt{1 - \rho}}{\sqrt{1 + \rho}} \right),$$

where η is the copula degrees of freedom and ρ is the value of the correlation coefficient between the two marginal distributions. This "known" value of λ_u for the t copula will be exploited in assessing the performance of the BMA method for the motivating data sets. The Gumbel and Joe copulas also incorporate upper tail dependence, with zero lower tail dependence. All three are freely available for application using the **R** package **copula**.

2.3 Bayesian model averaging

Bayesian model averaging (BMA) is a statistically robust, widely used means of reliably combining information from multiple candidate models for a given data set, which has been shown to improved predictive performance in a range of applications (Hoeting et al., 1999). BIC_j is the Bayesian information criterion (BIC) value associated with the j^{th} copula, $j = 1, \ldots, J$, and is calculated as $BIC_j = -2 \log L_j + k_j \log n$, where the number of parameters associated with the j^{th} copula is k_j, n is the number of observations in the data set and L_j is the likelihood value associated with the j^{th} copula. The BMA-based weighted estimate of the upper tail dependence coefficient is then calculated as $\lambda_{u,BMA} = \sum_{j'=1}^{J} W_{C_j} \lambda_{u,C_j}$, where the weight associated with the j^{th} copula, W_{C_j} is given by:

$$W_{C_j} = \frac{\exp(-0.5BIC_j)}{\sum_{j'=1}^{J} \exp(-0.5BIC'_j)}.$$

3 Results

3.1 Bivariate t results

The true value of λ_u corresponding to the underlying t copula can be calculated in advance as 0.232 (weak upper tail dependence). The BIC for the "best" (correct) model is significantly greater than that for the other models (see Table 1): $\exp^{-0.5BIC_j} \approx 0$ for all copulas aside from the t copula. Hence in this case the BMA weighting approach simply, but valuably, identifies the correct model as giving the best estimate of upper tail dependence. The BMA weighted estimate $\lambda_{u,BMA}$ is equal to $\lambda_{u,t}$, which is very close to the actual value for λ_u.

TABLE 1. Results for upper tail dependence coefficient and BIC for the t, Gumbel and Joe copulas applied to Bivariate t data simulated from a t copula with weak upper tail dependence.

Copula	Upper tail dependence coefficient	BIC
t	0.238	-1205.35
Gumbel	0.471	-146.74
Joe	0.620	-157.58

3.2 Bivariate gamma and beta results

The true value of λ_u corresponding to the underlying t copula can be calculated in advance as 0.785 (strong upper tail dependence). The BICs

for all three candidate copulas are relatively close (see Table 2), with the t copula yielding the optimum BIC value, as expected given the data were in fact simulated from a t copula (notably we would not have knowledge of this fact in real world applications). The BMA-based weights are calculated as $W_{C_t} = 0.884$, $W_{C_{Gumbel}} = 0.104$ and $W_{C_{Joe}} = 0.012$, with significantly stronger weight being given to the t copula. This results in an overall BMA-weighted estimate of upper tail dependence of $\lambda_{u,BMA} = 0.779$ versus the true value of 0.785.

TABLE 2. Results for upper tail dependence coefficient and BIC for the t, Gumbel and Joe copulas applied to Bivariate gamma and beta data simulated from a t copula with strong upper tail dependence.

Copula	Upper tail dependence coefficient	BIC
t	0.781	-1439.96
Gumbel	0.764	-1435.68
Joe	0.759	-1431.34

4 Conclusions

Bayesian model averaging is a computationally efficient, statistically robust method of identifying when a copula model for tail dependence is significantly better than other candidates; or to blend information from multiple copula models for tail dependence when more than one copula is plausible.

References

Demarta, S. and McNeil, A. (2005). The t copula and related copulas. *International Statistical Review*, **73**, 111 – 129.

Fischer, M. and Klein, I. (2007). Some results on weak and strong tail dependence coefficients for means of copulas. *Friedrich-Alexander-University Erlangen-Nuremburg*, **78**.

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382 – 417.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Inst. Statist. Univ. Paris*, **8**, 229 – 231.

Generalized linear mixed models applied to overdispersed proportion data in a fungal occurrence study

Thiago P. Oliveira[1], Rafael A. Moral[1], John Hinde[2], Clarice G.B. Demétrio[1], Silvio S. Zocchi[1], Ana B.R. Zanardo[1], Italo Delalibera Jr.[1]

[1] Universidade de São Paulo, Piracicaba, Brazil
[2] National University of Ireland Galway, Galway, Ireland

E-mail for correspondence: `thiago.paula.oliveira@gmail.com`

Abstract: When analysing proportion data, a useful framework is that of generalized linear models. Random effects may be included in the linear predictor for different reasons, e.g., to incorporate correlation between observations taken within the same subject or to model overdispersion. In this work, we use binomial mixed models to model the occurrence of entomopathogenic fungi in five different Brazilian biomes in the dry and humid seasons of 2012. We add an observation-level random effect to incorporate overdispersion and test for the significance of the interaction effect between biome and season.

Keywords: Overdispersion; Biological control; `lme4` package; Logistic-normal model.

1 Introduction

In Brazil there are several ecosystems and living organisms that are associated to the climatic conditions, soil, water and other factors. A range of microrganisms live in the soil, amongst them the fungi, which may contribute to arthropod regulation. There are several products made with fungi used to control insects in different crops, e.g., *Beauveria bassiana*, *Isaria fumosorosea* and *Metarhizium anisopliae*, see Faria and Wraight (2007). In this context, it is important to understand the occurrence of these entomopathogenic fungi in different Brazilian biomes to plan preservation strategies and test their potential in controlling different pests.

---

2 Material and methods

2.1 Case-study

To study the occurrence of entomopathogenic fungi, soil samples were collected in areas that represented the native vegetation of the Brazilian biomes Amazon, Caatinga, Atlantic Forest, Cerrado and the Pampas in the humid and dry seasons of 2012. Sampling was made in four different farms per biome and in each farm, six points were sampled, totaling 120 observations per season. Then, 10 larvae of *Galleria mellonella* and 10 larvae of *Tenebrio molitor* were exposed to each soil sample and after three weeks, the number of insects dead due to fungal infection was observed. This is a hierarchical design and observations within the same farm are correlated so this must be taken into account in the modelling process.

2.2 Statistical models

The data in this experiment consist of proportions so a reasonable assumption is that the number of insects dead due to fungal infection $Y_{ijk} \sim \text{Bin}(m_{ijk}, \pi_{ijk})$. An initial step was to fit a binomial generalized linear mixed model with logit link and the linear predictor

$$\text{logit}(\pi_{ijk}) = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \sigma_Z z_{ij}, \tag{1}$$

where μ is the intercept, α_i is the effect of the ith biome, γ_j is the effect of the jth season, $(\alpha\gamma)_{ij}$ is the interaction between the ith biome and the jth season, σ_Z^2 is the variance of the random effect associated with the farms and Z_{ij} are standard normal random variables.
To model overdispersion, a random effect at the observation level was included in the linear predictor, so that model (1) became

$$\text{logit}(\pi_{ijk}) = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \sigma_Z z_{ij} + \sigma_W w_{ijk}, \tag{2}$$

where σ_W^2 is the variance of the random effect at the observation level and W_{ijk} are standard normal random variables. This model is also called a logistic-normal model, see Demétrio et al. (2014).
To test for the significance of fixed effects, we used likelihood-ratio tests for nested models. Goodness-of-fit was assessed via half-normal plots with simulation envelopes (Moral et al., 2014) using R (R Core Team, 2014).

3 Results and discussion

We verified that for both species model (1) did not fit well to the data, see Figures 1(a) and (c). So we studied the inclusion of an observation-level random effect (model (2)) to incorporate overdispersion, resulting in a better model fit, see Figures 1(b) and (d). The interaction between biome

and season was significant at 5% significance level for both species, see Table 1 for estimated parameters. It was found that a larger proportion of entomopathogenic fungi infected *G. mellonella* in the dry season when compared to the humid, however the opposite result was obtained for *T. molitor* in the Amazon, see Figure 2. Moreover, there is a clear difference between the proportion of infected insects for humid and dry seasons in the Caatinga biome for *G. mellonella*, which was not found for *T. molitor*. Apparently, the patterns are the same for both species in the biomes Cerrado, Atlantic Forest and Pampas. The data is zero-inflated for some biome×season combinations and this is subject of ongoing work.

TABLE 1. Parameter estimates for model (2) fitted to both species' data.

	Species	
Parameter	*G. mellonella*	*T. molitor*
Intercept (μ)	−0.28 (0.30)	−3.88 (0.49)
Caatinga (α_2)	1.24 (0.43)	1.09 (0.62)
Cerrado (α_3)	−1.89 (0.46)	1.46 (0.61)
Atlantic Forest (α_4)	−1.66 (0.46)	−0.15 (0.69)
Pampas (α_5)	−1.92 (0.46)	4.92 (0.61)
Humid season (γ_2)	−3.34 (0.55)	1.78 (0.60)
Caatinga × humid season ($\alpha\gamma_{22}$)	3.74 (1.21)	−1.42 (0.82)
Cerrado × humid season ($\alpha\gamma_{32}$)	5.96 (0.73)	0.61 (0.78)
Atlantic Forest × humid season ($\alpha\gamma_{42}$)	1.75 (0.77)	−2.53 (0.99)
Pampas × humid season ($\alpha\gamma_{52}$)	3.81 (0.73)	−1.99 (0.77)
σ_Z	0.0143	0.0898
σ_W	1.2766	1.4082

Acknowledgments: Special Thanks to CNPq and FAPESP.

References

Demétrio, C.G.B., Hinde, J., and Moral, R.A. (2014). Models for overdispersed data in entomology. In: Ferreira, C.P. and Godoy, W.A.C. (Eds.) *Ecological Modelling Applied to Entomology*, Springer, 219 – 259.

Faria, M.R. and Wraight, S.P. (2007). Mycoinsecticides and mycoacaricides: A comprehensive list with worldwide coverage and international classification of formulation types. *Biological Control*, **43**, 237 – 256.

Moral, R.A., Hinde, J., and Demétrio, C.G.B. (2014). hnp: Half-normal plots with simulation envelopes. R package version 1.0. URL http://CRAN.R-project.org/package=hnp.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

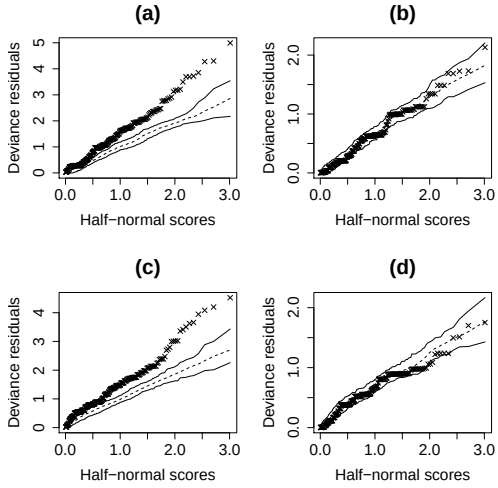


FIGURE 1. Half-normal plots with simulation envelopes for (a) model (1) and (b) model (2) for *G. mellonella* and (c) model (1) and (d) model (2) for *T. molitor*.

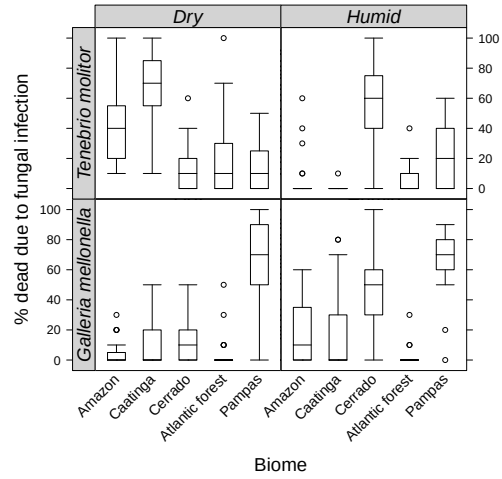


FIGURE 2. Box-plots for each biome×season combination for both species.

A general class of bivariate regression models for mixed discrete and continuous responses

Willian Luís Oliveira[1], Carlos Alberto Ribeiro Diniz[1], María Durbán[2]

[1] Department of Statistics, Federal University of São Carlos, Brazil
[2] Department of Statistics, Carlos III University of Madrid, Spain

E-mail for correspondence: willian26oliveira@gmail.com

Abstract: Bivariate regression models for mixed discrete and continuous responses whose joint distributions are constructed by the conditional approach (probability density functions, pdf, as the product of a marginal pdf and a conditional pdf) have been found in the literature (Fitzmaurice and Laird, 1995; Yang et al., 2007). However, the lack of more flexible and general models is noticed. In this paper a wide general class of models for mixed responses is proposed. It is assumed that the distribution of the discrete response and the conditional distribution of the continuous response given the discrete variable belong to one- or two-parameter exponential family of distributions. Furthermore, the marginal means are related to the covariates by link functions using linear and/or nonlinear predictors and a dependency structure between the responses is inserted into the model via the conditional mean. Classic estimation method, diagnostic analysis and influence techniques are presented as well as a simulation study considering a Bernoulli-exponential model, a particular case of the proposed class. Finally, the proposed model is used in a real data set involving the total cost of care for each patient during hospitalization, the use or not of the intensive treatment unit and the age of the patient.

Keywords: Bivariate model; Discrete and continuous responses; Conditional approach.

1 Introduction

In many areas of science, such as economics, medicine and psychology, are common situations which are simultaneously observed two responses associated with the same individual. In these cases the intrinsic relationship between the two variables should be considered in the analysis. In this

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

context, bivariate models, which allocate a dependent structure between the two responses, should be taken into account.
Bivariate distributions constructed by the conditional approach has been used in the literature. Catalano and Ryan (1992) considered a bivariate model assuming normal distribution for the continuous response and a Bernoulli distribution for the discrete response given the continuous response. Laird and Fitzmaurice (1995) proposed a bivariate model in which it is assumed that the discrete response variable follows the Bernoulli distribution and the continuous response variable follows a normal distribution given the Bernoulli response. Yang et al. (2007) developed a regression model for mixed Poisson and continuous responses while George et al. (2007) formulated a fully parametric regression model for clusters of bivariate data with binary and continuous components in which data from the same cluster are assumed to be exchangeable. However the lack of more flexible and general models is noticed, especially the models where the conditional distribution is not normal. In this article we propose a wide general class of models for mixed responses whose joint distributions are constructed using the conditional approach.

2 Bivariate models for mixed responses

It is assumed that the distribution of the discrete response (Y_i) and the conditional distribution of the continuous response given the discrete variable $(X_i|Y_i = y_i)$ belong to one- or two-parameter exponential family of distributions. Further, covariates are available to predict Y_i and X_i and are related to the marginal means by link functions using linear and/or nonlinear predictors.
The dependence between the response variables is determined in the model by the conditional mean $\lambda_i = E(X_i|Y_i = y_i)$. This mean is related to μ_{iX} (marginal mean of X_i), to Y_i (discrete response), to μ_{iY} (marginal mean of Y_i) and to γ (parameter included in the model, which may be the correlation coefficient or other measure of association between the response variables), by a linear or nonlinear function.
Maximum likelihood method is considered where iterative methods such as Newton-Raphson and Fisher score are required to solve the system of the resulting likelihood equations. The asymptotic covariance matrix of the maximum likelihood estimates of the parameters is obtained by inversion of the expected or observed information matrix. A diagnostic analysis and local and global influence measures are presented for the proposed model.
Introduced by Fitzmaurice and Laird (1995), Bernoulli-normal bivariate regression model is a particular case of the proposed class of models. In this particular case we assume that $Y_i \sim \text{Bernoulli}(\mu_{iY})$ and $X_i|Y_i = y_i \sim \text{N}(\lambda_i, \sigma^2)$, $i = 1, \ldots, n$ with $\lambda_i = \mu_{iX} + \gamma(y_i - \mu_{iY})$ and σ^2 unknown. In addition, p covariates are available for predicting Y_i and X_i, considering

both linear predictors. In this case, the likelihood equations can be solved iteratively using Fisher and reweighted least squares methods.
Introduced by Yang et al. (2007), Poisson-normal bivariate regression model is another particular case of the proposed class. This model is analogous to the Bernoulli-normal model except that now $Y_i \sim \text{Poisson}(\mu_{iY})$.
The Bernoulli-exponential model is also a particular case of the proposed class. Consider $Y_i \sim \text{Bernoulli}(\mu_{iY})$ and $X_i|Y_i = y_i \sim \text{Exponential}(\frac{1}{\lambda_i})$ with $\lambda_i = (1 - y_i + \mu_{iY})\mu_{iX}$. Logit and logarithmic link functions are considered to relate μ_{iY} and μ_{iX} to the available covariates. Linear predictors are adopted. The likelihood equations can be solved iteratively using Newton-Raphson algorithm. For this model, a simulation study is carried out, and the model is used in a real data set.

3 Simulation

A simulation study is carried out in different predetermined scenarios in order to analyze the behavior of maximum likelihood estimates of the Bernoulli-exponential model parameters with respect to the bias, the square root of the mean square error ($\sqrt{\text{MSE}}$), the standard deviation of the estimates (Sd), the average of the asymptotic standard errors (ASE_M) and the coverage probability of the asymptotic confidence intervals. The behavior of different residuals is also analyzed via simulation studies. The results are satisfactory in all treated scenarios. In fact, the bias values are quite small and generally decreases with increasing sample size which also occur with the values of $\sqrt{\text{MSE}}$, Sd and ASE_M. Regarding the coverage probability, the results are good overall, with estimated coverage probability close to the nominal coverage probability.

4 Real data set

A real data set containing information related to admissions of patients provided by a managed care plans is analysed by using the Bernoulli-exponential model. Composed by information from 300 admissions, the total cost of care for each patient during hospitalization and the use or not of the intensive treatment unit are adopted as continuous and discrete response variables, respectively. The age of the patient is categorized into less than 25 in group 1, 26-44 years in group 2, 45-64 years in group 3, 65-83 years in group 4 and above 84 in group 5. Thus, there are 4 dummy variables (z_{i1}, z_{i2}, z_{i3} and z_{i4}) which are considered as covariates, where $z_{ij} = 0$ or 1, for $j = 1, \ldots, 4$ and $i = 1, \ldots, 300$. Table 1 presents the maximum likelihood estimates of the model parameters and theirs estimated standard errors. A residual and influence analysis are performed. The results are satisfactory, consistent with those found in the simulation studies.

TABLE 1. Maximum likelihood estimates and theirs standard errors.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimates	−1.0662	2.2342	1.4229	1.1681	−0.1119
(Standard errors)	(0.3153)	(0.4240)	(0.3690)	(0.3721)	(0.4941)
	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$	$\hat{\delta}_4$
Estimates	7.1431	0.4331	1.1015	1.3170	1.3412
(Standard errors)	(0.2162)	(0.2590)	(0.2478)	(0.2514)	(0.3260)

5 Conclusion

The proposed general class of bivariate regression models is more flexible since it encompasses a variety of models. As a particular case, it is considered the Bernoulli-exponential model. Using a simulation study in some predetermined scenarios we can observe the good properties of the model parameter estimates as well as the behavior of the diagnostic techniques.

Acknowledgments: Special thanks to Brazilian organization *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).*

References

Catalano, P.J. and Ryan, L.M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, **87**, 651 – 658.

Fitzmaurice, G.M. and Laird, N.M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, **90**, 845 – 852.

George, E.O., Armstrong, D., Catalano, P.J., and Srivastava, D.K. (2007). Regression models for analyzing clustered binary and continuous outcomes under an assumption of exchangeability. *Journal of Statistical Planning and Inference*, **137**, 3462 – 3474.

Yang, Y., Kang, J., Mao, K., and Zhang, J. (2007). Regression models for mixed Poisson and continuous longitudinal data. *Statistics in Medicine*, **26**, 3782 – 3800.

Functional regression model of pentadal rainfall of Valle del Cauca (Colombia) in the period (1993 – 2011), integrating data from meteorological stations and satellite

Johann Ospina-Galindez[1], Mercedes Andrade-Bejarano[1], Ramón Giraldo-Henao[2]

[1] Universidad del Valle, Colombia
[2] Universidad Nacional, Colombia

E-mail for correspondence: johann.ospina@correounivalle.edu.co

Abstract: This paper presents a model of functional regression between weather data from meteorological stations and satellite data. The model was fitted by integrating spatial correlation structure in the errors. The study aims to integrate rainfall and ground data using remote sensing, in order to make predictions in unobserved sites. For the study, 92 meteorological stations from Valle del Cauca (Colombia) and Tropical Research Measurement Mission (TRMM) satellite data for the period 1993 – 2011 were used. The results showed the existence of spatial correlation due to the proximity among the meteorological stations. Finally, the satellite information had a significant contribution in the fitted model throughout the analysis period.

Keywords: Functional regression; Spatial correlation; Rainfall; Satellite data; Meteorological station.

1 Case study

Rainfall is measured through the meteorological stations. These provide specific measurements, but insufficient in some regions. On the other hand, there is a weather satellite, it can measure rainfall covering the whole area of study, however, contains errors associated equipment limitations and weather variability (Funk and Verdin, 2003). Since there are two types of data, it is important to study methodologies to integrate the data for the accuracy of the meteorological stations and satellite coverage, in order to

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

predict the rainfall in unobserved sites to generate information for farmers, research centers, health institutions, etc. With the advancement of information systems has emerged the need to develop statistical methods for analyzing large data sets (Finley et al., 2009). As alternative, the analysis of functions (Noguerales et al., 2010) arises. The statistical study of functions is framed in an area called Functional Data Analysis (FDA) (Noguerales et al., 2010). Modeling pentadal rainfall through the FDA, can represent an ongoing process. However, the model of functional regression assumes no correlation structure between the functional data (Febrero, 2008; Silverman and Ramsay, 2005), generating a statistical problem, since the data to be modeled may have a spatial correlation due to the closeness among the meteorological stations. In this research we propose to work with concurrent functional regression model incorporating the structure of spatial correlation in the errors.

2 Material and methods

Data from 92 meteorological stations from Valle del Cauca (Colombia) was used. In a first step the data was obtained from historical series of total daily rainfall, which came from the meteorological stations. The satellite rainfall data used was TRMM (Tropical Rainfall Measurement Mission) with a resolution of 5x5 km. For this research a model of functional regression with correlated errors is fitted. The model is fitted by using Generalised Least Square method and the covariance structure of the errors is modelled by using geostatistics for functional data (Giraldo, 2009). The fitted model has the form:

$$\mathbf{Y}(t) = \mathbf{X}^*(t)\mathbf{b} + \epsilon(t), \tag{1}$$

where $\mathbf{Y}(t)$ are functional data from meteorological stations, $\mathbf{X}^*(t)$ are functional data from satellite, $\mathbf{b}$ are coefficients associated with the model parameters, and $\epsilon(t)$ are functional errors. Using generalized least squares, we have to find $\mathbf{b}$ that minimize:

$$MCG(\mathbf{b}) = \int \left[\mathbf{Y}(t) - \mathbf{X}^*(t)\mathbf{b}\right]^{\mathrm{T}} \Omega^{-1} \left[\mathbf{Y}(t) - \mathbf{X}^*(t)\mathbf{b}\right] dt,$$

where the matrix Ω is the matrix of variances and covariances of the residuals $\epsilon(t)$ of the model estimated from functional geostatistics. Finally the parameters from generalized least squares are:

$$\hat{\beta}(t) = \Theta\hat{\mathbf{b}}.$$

3 Results

We displayed only the results of 2011. The mean and standard deviation curve, had higher values when the rainfall comes from meteorological stations compared with satellite data (Figure 1). The presence of spatial autocorrelation is evident. Trace-variogram bin shows a lower semivariance

close to zero values on the axis of the semivariance. It indicates that meteorological stations located at a shorter distance have similar rainfall values, while at greater distances (30 – 50 km), is found higher semivariance values, indicating that meteorological stations far away among them have less similar rainfall values (Figure 2 left). The model fitted with spatially correlated errors, showed a significant contribution of functional curves satellite in the period of analysis, this is denoted, because the confidence bands (95%) of the functional coefficient were above zero in the most pentadales times (Figure 2 right).

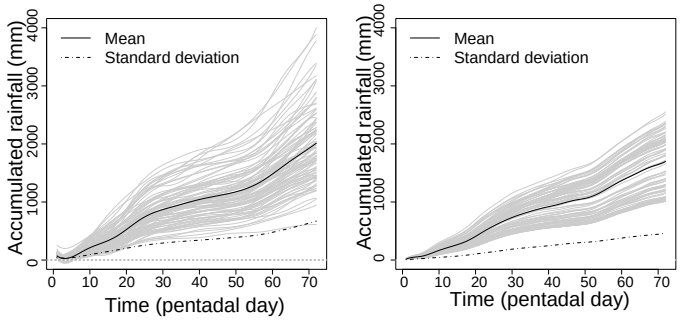


FIGURE 1. Left: functional data from meteorological stations. Right: functional data from satellite, period 2011.

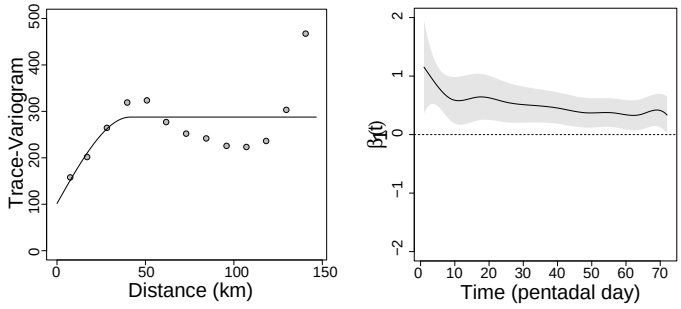


FIGURE 2. Left: trace-variogram bin of functional residual. Right: functional regression coefficient associated of the satellite data, period 2011.

4 Conclusions

The methodology enables a robust statistical analysis with high volume of data and includes in the functional linear regression model, the modeling

of the spatial correlation in the errors due to the closeness among meteorological stations. The inclusion of satellite data in the model was important throughout the period analysed, providing an additional tool in estimating rainfall in unobserved sites at the Valle del Cauca, Colombia. These results are important in emerging areas where ground monitoring networks are scarce, since the measurement via satellite allows real-time estimates of rainfall.

References

Febrero, M. (2008). A present overview on functional data analysis. *Boletín de Estadística e Investigación Operativa, BEIO*, **24**, 6 – 12.

Finley, A.O., Sang, H., Banerjee, S., and Gelfand, A.E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, **53**, 2873 – 2873.

Funk , C. and Verdin, J. (2003). Comparing satellite rainfall estimates and reanalysis rainfall fields with station data for western Kenya. *In Proc. Int. workshop on crop monitoring for food security in Africa, European Joint Research Centre UN Food and Agriculture Organization, Nairobi, Kenya*, 28 – 30.

Giraldo, R. (2009). *Geostatistical Analysis of Functional Data*. PhD thesis: Universitat Politécnica de Catalunya.

Noguerales, J.L.T., González, A.C., and Fernández, C.S.C. (2010). *Análisis de datos funcionales, clasificación y selección de variables*. Master thesis: Universidad Autónoma de Madrid.

Silverman, B. and Ramsay, J. (2005). *Functional Data Analysis*. Springer.

A comparison of different priors for sparse Bayesian modelling in regression models with ordinal predictors

Daniela Pauger[1], Helga Wagner[1]

[1] Johannes Kepler University Linz, Austria

E-mail for correspondence: daniela.pauger@jku.at

Abstract: Sparse modelling is an important issue particularly in regression models with ordinal predictors. One way to reduce the dimension of the model is to fuse categories with the same effect on the response. We compare the performance of two different prior specifications that encourage sparsity by effect fusion, i.e. spike and slab prior and normal-gamma prior, to an uninformative prior using a simulation study.

Keywords: Sparse modelling; Ordinal predictors; Bayesian; Prior.

1 Introduction

In regression models the collected covariates often are measured on an ordinal scale. The usual strategy of using dummy variables for modelling the effect of one level with respect to the reference category can easily lead to high-dimensional models. A sparser model can be achieved by removing covariates with no effect on the response or by fusing levels with the same effect.
In this work we compare two different prior specifications which encourage sparsity by effect fusion of ordinal predictors to an uninformative prior in regression models.

2 Model specification

Let y denote the normal response in a standard linear regression model with $j = 1, \ldots, p$ ordinal covariates c_j. We assume that the jth covariate has $K_j + 1$ categories $0, \ldots, K_j$ where 0 defines the reference category. To

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

achieve sparsity in modelling the effect of c_j we follow Gertheiss and Tutz (2009) and use split-coded regressors X_{jk}, i.e.

$$X_{jk} = \begin{cases} 1 & \text{for } c_j \geq k \\ 0 & \text{otherwise} \end{cases} \qquad k = 1, \dots K_j$$

and specify the linear regression model as

$$y = \mu + \sum_{j=1}^{p} \sum_{k=1}^{K_j} X_{jk}\theta_{jk} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Here, the regression coefficient θ_{jk} can be interpreted as the effect difference of level k with respect to the previous level $k-1$. Both levels of covariate c_j have a different effect on the response if $\theta_{jk} \neq 0$ and they have the same effect, if $\theta_{jk} = 0$.

3 Priors and inference

We consider two different types of priors for sparse modelling that encourage shrinkage of small regression effects to zero: The first, the spike and slab prior, is a finite mixture prior of two components and the second is the Normal-Gamma prior.

3.1 Spike and slab prior

A spike and slab prior (George and McCulloch, 1993) for the regression coefficient θ_{jk} can be specified hierarchically as

$$\begin{aligned} p(\theta_{jk}|\delta_{jk}, \tau_j^2) &\sim \delta_{jk}\mathcal{N}(0, \tau_j^2\sigma^2) + (1-\delta_{jk})\mathcal{N}(0, r\tau_j^2\sigma^2) \\ p(\delta_{jk} = 1) &= w_j \\ p(w_j) &\sim \mathcal{B}(a_0, b_0) \\ p(\tau_j^2) &\sim \mathcal{G}^{-1}(s_0, S_0), \end{aligned}$$

where r is a small value and δ_{jk} is an indicator for the slab component. w_j corresponds to the mixture weight for variable c_j and τ_j^2 is the variance parameter. Note that both, the spike and the slab component, are specified as a scale mixture of Normal distributions with an Inverse Gamma mixing distribution, which results in a scaled t-distribution. Sparsity is accomplished by the spike component: If $\delta_{jk} = 0$, θ_{jk}, is assigned to the spike component and hence shrunk aggressively to zero, whereas it experiences only little shrinkage when assigned to the slab component.

3.2 Normal-Gamma prior

The Normal-Gamma prior, introduced by Brown and Griffin (2010), is also a scale mixture of Normal distributions, with a Gamma mixing distribution. We consider two versions of this prior: The first is specified hierarchically as

$$p(\theta_{jk}|\psi_{jk}) \sim \mathcal{N}(0, \psi_{jk}\sigma^2)$$
$$p(\psi_{jk}|\lambda, \gamma) \sim \mathcal{G}(\lambda, 1/(2\gamma^2)),$$

whereas in the second version we put a covariate specific hyperprior on the scale parameter γ_j of the Gamma distribution

$$p(\gamma_j) \sim \mathcal{G}^{-1}(g_j, G_j).$$

Also the Normal-Gamma prior shrinks small effects severely to zero, however in contrast to the spike and slab prior, effects are not explicitly classified as zero or non-zero.

4 Simulations

4.1 Simulation setup

We compare the performance of the various prior settings in terms of accurate coefficient estimation in a simulation study, where we consider three ordinal covariates, each with 11 categories. 100 observations are generated from the model $y = 0.5 + \boldsymbol{X}\boldsymbol{\beta} + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$ with dummy coded regressors. Regression effects are $\beta_1 = (0, 0, 0, 0, 2, 2, 2, 5, 5, 5)$, $\beta_2 = (2, 2, 4, 4, 5, 5, 5, 7, 7, 7)$, whereas no level of the third covariate has an effect, i.e. β_3 is a vector of zeros.
We specify the spike and slab prior with $s_0 = 5, S_0 = 25$ and $a_0 = b_0 = 1$. Following Brown and Griffin (2010) we set $\lambda = \gamma = 1$ for the Normal-Gamma prior with fixed hyperparameters and choose $g_j = 2$ and $G_j = 1$ for the more flexible variant. For comparison we estimate also the full model where the regression coefficients are assigned a flat Normal prior with mean zero and variance 10,000.
Bayesian inference is accomplished by sampling from the posterior distribution using MCMC methods. Performance of the various prior specifications is compared by the mean squared errors (MSE) of the estimated coefficients. Figure 1 shows boxplots of the results for 100 simulation runs.

4.2 Results

Generally, priors encouraging sparsity outperform the uninformative prior and the spike and slab prior outperforms both variants of Normal-Gamma

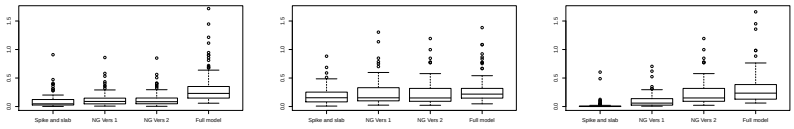


FIGURE 1. MSE of coefficient estimation.

prior for all three covariates. The most pronounced difference in the MSE is observed for the third covariate, where all level effects are equal to zero. For this covariate the normal-gamma prior with fixed parameters performs better than the version with hyperprior, whereas there is little difference for the other two covariates.
With respect to computational effort the prior specifications that encourage a sparser model are comparable, the uninformative prior is considerably faster.

5 Conclusion

We compared the performance of various prior specifications that encourage sparsity by effect fusion, i.e. spike and slab prior and two variants of Normal-Gamma prior, with each other as well as to an uninformative prior. For all covariates the priors for effect fusion perform better than the uninformative prior and spike and slab outperforms both variants of Normal-Gamma prior. An improvement by covariate specific hyperpriors on the parameters of the Normal-Gamma prior could not be observed.

Acknowledgments: We acknowledge gratefully financial support by the Austrian Science Fund FWF, project number P25850 'Sparse Bayesian modelling for categorical predictors'.

References

George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881 – 889.

Gertheiss, J. and Tutz, G. (2009). Penalized regression with ordinal predictors. *International Statistical Review*, **77**, 345 – 365.

Griffin, J.E. and Brown, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171 – 188.

Partially linear models with autoregressive symmetric errors

Gilberto A. Paula[1], Carlos Eduardo M. Relvas[1]

[1] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

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

Abstract: In this paper we discuss estimation and model selection procedures for partially linear models with first-order autoregressive symmetric errors. Estimation is performed by maximum penalized likelihood and by using natural cubic splines. A reweighed iterative process based on the back-fitting algorithm is derived for the parameter estimation and the inference is based on the penalized Fisher information matrix. Some model selection procedures are derived and a real data set is analyzed under the proposed models.

Keywords: Correlated data; Model selection; Sensitivity; Student-t models.

1 Introduction

Partially linear models have been investigated under independent errors (see, e.g., Ibacache-Pulgar et al., 2013, and the references therein), but few has been investigated under correlated data. The aim of this paper is to derive an estimation procedure and some model selection methods for partially linear models with first-order autoregressive (AR(1)) symmetric errors. The idea is to improve the goodness-of-fit of linear models with AR(1) errors by taking into account the relationship between time and response in the systematic component, and since this kind of form is in general nonlinear, nonparametric forms seem to be more flexible than parametric ones. From an appropriate penalized log-likelihood function a back-fitting algorithm is derived for the parameter estimation. Some model selection procedures are also derived and a real data set is analyzed under the proposed models.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

2 The model

Partially linear models with first-order autoregressive symmetric errors assume the following relationship between the response and the explanatory variable values:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f(t_i) + \epsilon_i, \tag{1}$$

where $\epsilon_i = \rho\epsilon_{i-1} + e_i, \quad -1 < \rho < 1$, for $i = 1, \ldots, n$, y_i denotes the ith response value, $\mathbf{x}_i$ is a $p \times 1$ vector of explanatory variable values from the ith experimental unit, $\boldsymbol{\beta}$ is the $p \times 1$ fixed parameter vector, ρ is the autoregressive coefficient, $f(t_i)$ is an arbitrary univariate smooth function of the time that contributes nonparametrically on the response y_i and e_i is a random error assumed to follow a symmetric distribution with zero mean and dispersion parameter ϕ, namely $e_i \overset{\text{iid}}{\sim} \mathrm{S}(0, \phi)$. According to Green and Silverman (1994) we apply the following penalized log-likelihood function $Ł_p(\boldsymbol{\theta}, \alpha) = Ł(\boldsymbol{\theta}) - \frac{\alpha}{2}\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f}$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \mathbf{f}^{\mathrm{T}}, \phi, \rho)^{\mathrm{T}}$, $Ł(\boldsymbol{\theta})$ denotes the regular conditional log-likelihood function, $\alpha > 0$ is the smoothing parameter, $\mathbf{K}$ is a $q \times q$ non-negative definite smoothing matrix and $\mathbf{f} = (f(t_1^0), \ldots, f(t_q^0))^{\mathrm{T}}$ with t_j^0, for $j = 1, \ldots, q$, denoting the distinct and ordered values of t_i, for $i = 1, \ldots, n$, namely $a \leq t_1^0 < t_2^0 < \cdots < t_q^0 \leq b$.

3 Parameter estimation

Applying the Fisher scoring method, the $(u+1)$th step of the iterative process for obtaining the maximum penalized likelihood estimates of $\boldsymbol{\beta}$ and $\mathbf{f}$, by fixing ρ and ϕ, may be expressed as

$$\begin{pmatrix} \boldsymbol{\beta}^{(u+1)} \\ \mathbf{f}^{(u+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_\beta \left[\mathbf{r}_\beta^{(u)} + \left\{\mathbf{I}_n - \frac{\mathbf{D}^{(u)}(\mathbf{v})}{4d_g}\right\}\mathbf{A}\boldsymbol{\mu}^{(u)}\right] \\ \mathbf{S}_f \left[\mathbf{r}_f^{(u)} + \left\{\mathbf{I}_n - \frac{\mathbf{D}^{(u)}(\mathbf{v})}{4d_g}\right\}\mathbf{A}\boldsymbol{\mu}^{(u)}\right] \end{pmatrix}, \tag{2}$$

for $u = 0, 1, \ldots$, where $\mathbf{A} = \mathbf{A}(\rho)$ is an $n \times n$ matrix appropriately defined for AR(1) errors, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{N}\mathbf{f}$, $\mathbf{X}$ is an $n \times p$ matrix with rows $\mathbf{x}_i^{\mathrm{T}}$, $\mathbf{D}(\mathbf{v}) = \mathrm{diag}\{v_1, \ldots, v_n\}$ with $v_i > 0$ named weights, $i = 1, \ldots, n$, $\mathbf{N}$ is the $n \times q$ incidence matrix, $\mathbf{I}_n$ is the identity matrix of order n, d_g are known quantities derived for each symmetric distribution,

$$\mathbf{S}_\beta = \{(\mathbf{A}\mathbf{X})^{\mathrm{T}}(\mathbf{A}\mathbf{X})\}^{-1}(\mathbf{A}\mathbf{X})^{\mathrm{T}} \text{ and } \mathbf{S}_f = \left\{(\mathbf{A}\mathbf{N})^{\mathrm{T}}(\mathbf{A}\mathbf{N}) + \frac{\phi\alpha\mathbf{K}}{4d_g}\right\}^{-1}(\mathbf{A}\mathbf{N})^{\mathrm{T}}$$

are the smoothing matrices, and

$$\mathbf{r}_\beta^{(u)} = \mathbf{A}\left\{\frac{\mathbf{D}^{(u)}(\mathbf{v})\mathbf{y}}{4d_g} - \mathbf{N}\mathbf{f}^{(u)}\right\} \quad \text{and} \quad \mathbf{r}_f^{(u)} = \mathbf{A}\left\{\frac{\mathbf{D}^{(u)}(\mathbf{v})\mathbf{y}}{4d_g} - \mathbf{X}\boldsymbol{\beta}^{(u)}\right\}$$

are named partial residuals with $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. In particular, for $\mathbf{A} = \mathbf{I}_n$ we recover the independent case. The iterative process (2) should be alternated with the iterative process

$$\phi^{(s+1)} = \phi^{(s)} + \{\mathrm{U}_p^{\phi}/\mathrm{I}_p^{\phi\phi}\}^{(s)} \text{ and } \rho^{(s+1)} = \rho^{(s)} + \{\mathrm{U}_p^{\rho}/\mathrm{I}_p^{\rho\rho}\}^{(s)}, \qquad (3)$$

for $s = 0, 1, \ldots$, where U_p^{ϕ} and U_p^{ρ} are the penalized score functions whereas $\mathrm{I}_p^{\phi\phi}$ and $\mathrm{I}_p^{\rho\rho}$ are the respective penalized Fisher information of ϕ and ρ, respectively. Expression for the effective degrees of freedom $\mathrm{df}(\alpha)$ and the generalized cross-validation score $\mathrm{GCV}(\alpha)$ were obtained. The model selection may be performed by minimizing either $\mathrm{GCV}(\alpha)$ or $\mathrm{AIC}(\alpha) = -2\mathrm{L}_p(\hat{\boldsymbol{\theta}}, \alpha) + 2\{2 + p + \mathrm{df}(\alpha)\}$. The approximate variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is derived from the inverse of the penalized Fisher information matrix for $\hat{\boldsymbol{\theta}}$. So, we have $\widehat{\mathrm{Var}}_{\mathrm{approx}}(\hat{\boldsymbol{\theta}}) = \mathbf{I}_p^{\theta\theta^{-1}}|_{\hat{\theta}}$, where $\mathbf{I}_p^{\theta\theta}$ denotes the penalized Fisher information matrix.

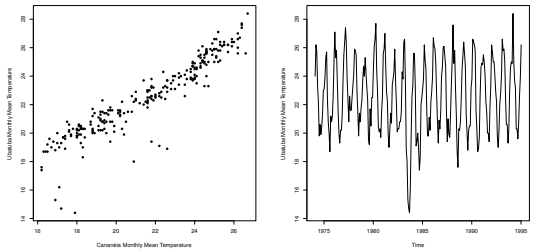


FIGURE 1. Dispersion graph between the monthly mean temperatures of Ubatuba and Cananéia (left) and Ubatuba monthly mean temperature from January 1974 to December 1994 (right).

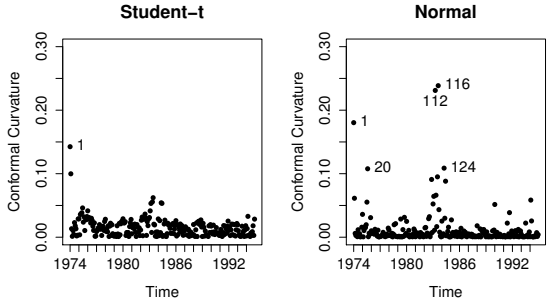


FIGURE 2. Sensitivity graph for the estimate $\hat{\mathbf{f}}$ under the case-weight perturbation scheme. Pointed out observations such that $B_i > 0.10$.

4 Application

As illustration we consider the analysis of the relationship between the monthly mean temperatures of two beach cities, Ubatuba and Cananéia, in São Paulo state, Brazil, from January 1974 to December 1994 (see Fig. 1). We suppose the model $y_i = \beta_0 + \beta_1 x_i + f(t_i) + \epsilon_i$, for $i = 1, \ldots, 252$, where y_i and x_i denote, respectively, the mean temperature of Ubatuba and Cananéia at the ith time t_i (ith month), whereas $f(t_i)$ is an arbitrary univariate smooth function, $\epsilon_i = \rho\epsilon_{i-1} + e_i$ with $e_i \overset{\text{iid}}{\sim} \text{S}(0, \phi)$.
Comparing normal and Student-t error models we found that the Student-t model with $\nu = 3$ degrees of freedom presents the lowest AIC(α) and accommodated better the outlying observations observed in Figure 1. In addition, the parameter estimates from the Student-t model appear less sensitive than the remaining error models under various perturbation schemes, as illustrated in Figure 2 for $\hat{\mathbf{f}}$ under the case-weight perturbation scheme. For the selected model we found $\hat{\beta}_0 = 5.190(0.318)$, $\hat{\beta}_1 = 0.808(0.014)$, $\hat{\phi} = 0.167(0.010)$ and $\hat{\rho} = 0.342(0.042)$. Using the model selection methods AIC(α) and GCV(α) jointly with df(α) we obtain $\alpha = 2000$ and df(α) $= 20.57$. In Figure 3 we use the data from January 1974 to December 1990 for predicting the 48 months ahead. We may notice a very good agreement between the observed and estimated values.

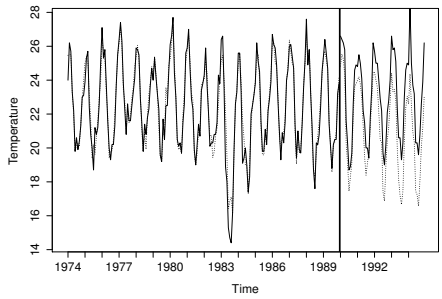


FIGURE 3. Observed (black line), estimated and predicted (dotted line) monthly mean temperature of Ubatuba from the Student-t model.

Acknowledgments: Special Thanks to CNPq and FAPESP, Brazil.

References

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.

Ibacache-Pulgar, G., Paula, G.A., and Cysneiros, F.J.A. (2013). Semiparametric additive models under symmetric distributions. *TEST*, **22**, 103 – 121.

D_s-optimality for discriminating between copula models: a first example

Elisa Perrone[1], Werner Müller[1]

[1] Department of Applied Statistics, JKU, Linz, Austria

E-mail for correspondence: elisa.perrone@jku.at

Abstract: Copulas are a very flexible tool to highlight structural properties of the design for a wide range of dependence structures. A natural question for these models is whether design techniques might be used in order to discriminate between different dependencies. In this work we introduce the D_s-optimality criterion for special models constructed as convex combination of copulas, with the aim of finding the design that best discriminates the two copulas involved.

Keywords: Copulas; D_s-optimality; Equivalence theorem, D_s-efficiency.

1 Introduction

In many areas of applied statistics, copula functions are largely employed as a flexible tool to describe the behaviour of the dependence between random variables. However, the use of such a function in design theory is a new part of this field. A first step was made by Denman et al. (2011), while a more complete framework for copula models was described in our previous work (Perrone and Müller, 2014), where a Kiefer-Wolfowitz type equivalence theorem for D-optimality was also provided.
The tools reported in Perrone and Müller (2014) allow one to find the D-optimal design for a particular copula model. However, a natural question regards the choice of the copula, an issue that can be seen in the design framework as a problem of discrimination between different models.
In this work, we focus on this new aspect and we provide a way to use D_s-optimality for discriminating between two different copulas. In fact, by considering the meaningful interpretation of the copula parameters, we construct a flexible model where one of the copula parameter is a link parameter between two copula families and, then, we use the well-known D_s-criterion to study the impact of the link parameter on the design ob-

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

tained. Then, in order to check how good the D-optimal design fares for discrimination, we compare this design with the D_s-optimal one.

2 The general framework

We shall consider a vector $\mathbf{x}^{\mathrm{T}} = (x_1, \ldots, x_r) \in \mathcal{X}$ of control variables, where $\mathcal{X} \subset \mathbb{R}^r$ is a compact set. The results of the observations and of the expectations in a regression experiment are the vectors $\mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}), y_2(\mathbf{x}))$,

$$\mathbf{E}[\mathbf{Y}(x)] = \mathbf{E}[(Y_1, Y_2)] = \eta(\mathbf{x}, \boldsymbol{\beta}) = (\eta_1(\mathbf{x}, \boldsymbol{\beta}), \eta_2(\mathbf{x}, \boldsymbol{\beta})),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ is a certain unknown parameter vector to be estimated and η_i $(i = 1, 2)$ are known functions. Let us call $F_{Y_i}(y_i(\mathbf{x}, \boldsymbol{\beta}))$ the margins of each Y_i for all $i = 1, 2$ and $f_{\mathbf{Y}}(\mathbf{y}(\mathbf{x}, \boldsymbol{\beta}), \boldsymbol{\alpha})$ the joint probability density function of the random vector $\mathbf{Y}$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ is the unknown (copula) parameter vector.
According to Sklar's theorem (see Nelsen, 2006), let us assume that the dependence between Y_1 and Y_2 is modeled by a copula function

$$C_\alpha(F_{Y_1}(y_1(\mathbf{x}, \boldsymbol{\beta})), F_{Y_2}(y_2(\mathbf{x}, \boldsymbol{\beta}))).$$

The Fisher Information Matrix for a single observation is a $(k+2) \times (k+2)$ matrix whose elements are

$$\mathbf{E}\left(-\frac{\partial^2}{\partial \gamma_i \partial \gamma_j} \log \left[\frac{\partial^2}{\partial y_1 \partial y_2} C_\alpha(F_{Y_1}(y_1(\mathbf{x}, \boldsymbol{\beta})), F_{Y_2}(y_2(\mathbf{x}, \boldsymbol{\beta}))) \right] \right), \qquad (1)$$

where $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_{k+2}\} = \{\beta_1, \ldots, \beta_k, \alpha_1, \alpha_2\}$. The aim of design theory is to quantify the amount of information on both sets of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively, from the regression experiment embodied in the Fisher Information Matrix. For a concrete experiment with N independent observations at $n \leq N$ support points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the corresponding information matrix then is

$$M(\xi, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{N} \sum_{i=1}^{n} w_i m(\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}), \sum_{i=1}^{n} w_i = 1, \xi = \left\{ \begin{array}{ccc} \mathbf{x}_1 & \ldots & \mathbf{x}_n \\ w_1 & \ldots & w_n \end{array} \right\}.$$

The approximate design theory is concerned with finding $\xi^*(\gamma)$ such that it maximizes some scalar function $\phi(M(\xi, \gamma))$, the so-called design criterion.

2.1 D_s-optimality

In our previous work we consider as design criterion the *D-optimality*. In this work we focus on the D_s-criterion, i.e. $\phi_s(M) = \log\det(M_{11} - M_{12}M_{22}^{-1}M_{12}^{\mathrm{T}})$, if M is nonsingular and where

$$M = \left(\begin{array}{cc} M_{11} & M_{12} \\ M_{12}^{\mathrm{T}} & M_{22} \end{array} \right),$$

with M_{11} is the $(s \times s)$ minor related to the estimated parameters. Starting from the Kiefer-Wolfowitz type equivalence theorem proved in Perrone and Müller (2014), a generalization of such type is also possible for the D_s-optimality, as showed by the following Theorem, proved elsewhere.

Theorem 1 *For a localized parameter vector $(\bar{\gamma})$, the following properties are equivalent:*

- *ξ^* is D_s-optimal;*
- *let us call $A = (I_s\ 0)$, then $\forall x \in \mathcal{X}$*

$$tr[M(\xi^*, \bar{\gamma})^{-1} A (A^{\mathrm{T}} M(\xi^*, \bar{\gamma})^{-1} A)^{-1} A^{\mathrm{T}} M(\xi^*, \bar{\gamma})^{-1} m(x, \bar{\gamma})] \leq s;$$

- *over all $\xi \in \Xi$, ξ^* minimize*

$$\max_{x \in \mathcal{X}} tr[M(\xi^*, \bar{\gamma})^{-1} A (A^{\mathrm{T}} M(\xi^*, \bar{\gamma})^{-1} A)^{-1} A^{\mathrm{T}} M(\xi^*, \bar{\gamma})^{-1} m(x, \bar{\gamma})].$$

For the comparison of designs we define D_s-*Efficiency* of the design ξ with respect to the design ξ^* as the ratio

$$\left(\frac{|M_{11}(\xi, \bar{\gamma}) - M_{12}(\xi, \bar{\gamma}) M_{22}^{-1}(\xi, \bar{\gamma}) M_{12}^{\mathrm{T}}(\xi, \bar{\gamma})|}{|M_{11}(\xi^*, \bar{\gamma}) - M_{12}(\xi^*, \bar{\gamma}) M_{22}^{-1}(\xi^*, \bar{\gamma}) M_{12}^{\mathrm{T}}(\xi^*, \bar{\gamma})|} \right)^{1/s}.$$

3 Example

We analyze an example with possible application in clinical trials. We consider a bivariate binary response (Y_{i1}, Y_{i2}), $i = 1, \ldots, n$ with four possible outcomes $\{(0,0), (0,1), (1,0), (1,1)\}$ where 1 usually represents a success and 0 a failure (of e.g. a drug treatment where Y_1 and Y_2 might be efficacy and toxicity). For a single observation denote the joint probabilities of Y_1 and Y_2 by $p_{y_1,y_2} = \Pr(Y_1 = y_1, Y_2 = y_2)$ for $(y_1, y_2) = (0,1)$. Now, define

$$\begin{array}{rclrcl} p_{11} & = & C_\alpha(\pi_1, \pi_2), & p_{10} & = & \pi_1 - p_{11}, \\ p_{01} & = & \pi_2 - p_{11}, & p_{00} & = & 1 - \pi_1 - \pi_2 + p_{11}. \end{array} \tag{2}$$

Let us now allow the strength of the dependence itself be dependent upon the regressors x. As in our context only positive associations make sense we consider in the following the τ modelled by a logistic:

$$\tau(x, \alpha_1) = \frac{e^{\alpha_1 x - c}}{1 + e^{\alpha_1 x - c}},$$

where c is a constant chosen such that τ takes values in $[\epsilon, 1]$ for $\alpha_1 \in [0,1]$; for our computations we chose $\epsilon = 0.05$. Then, using the relationship between the Kendall's τ and the copula parameter, we model p_{11} by a

convex combination of the Clayton and the Gumbel copulas by linking them at the same τ values, so we end up with

$$C(\pi_1, \pi_2; \alpha_1, \alpha_2) = \alpha_2 C_1(\pi_1, \pi_2; 2e^{\alpha_1 x - c}) + (1 - \alpha_2) C_2(\pi_1, \pi_2; 1 + e^{\alpha_1 x - c}).$$

In this model, the impact of the dependence structure and the association level is reflected by two different parameters, as the α_1 parameter is only related to the measure of association Kendall's τ, while the α_2 parameter is strictly related to the structure of the dependence. Therefore, applying the D_s-criterion on α_2, we find a design for discriminating, in this specific model, between the copula Clayton and Gumbel. We compare the design obtained for different τ intervals and localized values for α_2 with the D-optimal design obtained for the same localized values. Analysing the rather high losses in D_s-efficiency reported in Table 1, it shows that the D-criterion alone is not sufficient when we require information about the structure of the model.

TABLE 1. Losses in D_s-efficiency in percent.

$\bar{\alpha}_2$	$\tau \in [0.05, 0.3]$	$\tau \in [0.05, 0.95]$	$\tau \in [0.05, 0.995]$
0.1	38.66	45.01	43.34
0.5	53.54	39.61	45.52
0.9	38.13	33.39	42.93

References

Denman, N.G., McGree, J.M., Eccleston, J.A., and Duffull, S.B. (2011). Design of experiments for bivariate binary responses modelled by copula functions. *Computational Statistics and Data Analysis*, **55**, 1509 – 1520.

Heise, M.A. and Myers, R.H. (1996). Optimal design for bivariate logistic regression. *Biometrics*, **52**, 613 – 624.

Nelsen, R.B. (2006). *An Introduction to Copulas (2nd ed.)*. New York: Springer-Verlag.

Perrone, E. and Müller, W.G. (2014). Optimal designs for copula models. *Submitted*.

A comparative review of generalizations of the Gumbel extreme value distribution

Eliane C. Pinheiro[1], Silvia L.P. Ferrari[1]

[1] Department of Statistics, University of São Paulo, Brazil

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

Abstract: The generalized extreme value distribution and its particular case, the Gumbel extreme value distribution, are widely applied for extreme value analysis. Our goal is to extensively collect in the present literature the distributions that contain the Gumbel distribution embedded in them and to identify those that have flexible skewness and kurtosis, are heavy-tailed and could be competitive with the generalized extreme value. The generalizations of the Gumbel distribution are described and compared using an application to a wind speed data set and Monte Carlo simulations. We show that some distributions suffer from over-parameterization and coincide with other generalized Gumbel distributions with a smaller number of parameters, i.e., are non-identifiable. Our study suggests that the generalized extreme value distribution and a mixture of two extreme value distributions should be considered in practical applications.

Keywords: Generalized extreme value distribution; Gumbel distribution; Heavy-tailed distribution; Non-identifiable model.

1 Introduction

Extreme value data usually exhibit excess kurtosis and/or heavy right tails. This is particularly common in environmental data, e.g., maximum water level, maximum wind speed, spatial and temporal variability of turbulence, daily maximum ozone measurement, and largest lichen measurements. The generalized extreme value distribution (GEV) is fairly well-accepted as a standard working model. Despite of such well-established theory, extreme-value distributions are not always preferred in studies of empirical data which do not contemplate the conditions to use extreme value theory results. Sometimes, the fit for finite samples is poor. To surpass these issues other generalizations of the Gumbel distribution were proposed. The Gumbel distribution is also used to model extreme values. However, its skewness

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

and kurtosis coefficients are constant, and its right tail is light. Generalizations of the Gumbel distribution with flexible skewness and kurtosis coefficients could provide a better fit for extreme value data.
We present a comparative review of distributions that contain the Gumbel distribution as a special or limiting case. Some distributions suffer from overparameterization and coincide with other generalized Gumbel distributions with smaller number of parameters. We limit our study to the identifiable family of distributions only. We compare their coefficients of skewness and kurtosis (which are invariant under location-scale transformations) that are primarily controlled by the extra parameters. We highlight those that can achieve high values of skewness and kurtosis with a heavy right tail. To study the tail behavior of the distributions, we also employ the regular variation theory (de Haan, 1970) and a criterion proposed by Rigby et al. (2014).

2 The Gumbel distribution and its generalizations

The Gumbel distribution is one of the three possible limiting laws of the standardized maximum of independent and identically distributed random variables (Gnedenko, 1943). This distribution is also known as the extreme value ($\text{EV}(\mu, \sigma)$) or type I extreme value distribution.
Table 1 gives distributions that contain the Gumbel distribution as a special or limiting case; the nonidentifiable distributions are marked with an asterisk (*). The coefficients of skewness and kurtosis of the Gumbel distribution are constant ($\gamma_{1,\text{EV}} = 1.14$ and $\gamma_{2,\text{EV}} = 5.4$, respectively), i.e., parameter independent. The other distributions have a range of values of skewness and kurtosis coefficients which are primarily controlled by the extra parameters. The GEV distribution is the only one that allows unlimited skewness and kurtosis.
From the regular variation theory (de Haan, 1970), the generalized extreme value distribution $\text{GEV}(\mu, \sigma, \alpha)$ has a heavy right tail with tail index α when $\alpha > 0$ (Fréchet family). It has a non-heavy right tail when $\alpha = 0$ (Gumbel family). The other identifiable distributions addressed in this paper are all non-heavy right tailed distributions. Hence, among the identifiable distributions addressed in this work, the GEV distribution is the only one with a heavier right tail with respect to the Gumbel distribution under the tail index approach.
To distinguish among the generalized distributions we used the Rigby et al. (2014) criterion. The criterion splits the distributions in three types, I, II and III, in decreasing order of heaviness of the tails. The right tail of the GEV distribution with $\alpha > 0$ is of type I. If $\alpha > 1$, the GEV distribution has a heavier right tail than the Cauchy distribution. If $\alpha > 1/2$, the right tail heaviness of the GEV distribution is greater than that of the Student-t distribution with two degrees of freedom, which is uncommon in real data.

TABLE 1. Generalizations of the Gumbel distribution.

Distribution	Proposed by
Generalized extreme value GEV(μ, σ, α)	Jenkinson (1955)
Type IV generalized logistic GLIV(μ, σ, α)	Prentice (1975)
Two-component extreme value TCEV$(\mu, \sigma, \mu_1, \sigma_1, \alpha)$	Rossi et al. (1984)
Three parameter exponential-gamma EGa(μ, σ, α)	Ojo (2001)
Exponentiated Gumbel EGu(μ, σ, α)	Nadarajah (2006)
Transmuted extreme value TEV(μ, σ, α)	Aryal & Tsokos (2009)
Generalized three-parameter Gumbel GGu3(μ, σ, α)	–
Generalized type I extreme value GGu$(\mu, \sigma, \alpha, \beta)$*	Dubey (1969)
Exponential-gamma ExpGamma$(\mu, \sigma, \alpha, \beta)$*	Adeyemi, Ojo (2003)
Beta Gumbel BG$(\mu, \sigma, \alpha, \beta)$*	Nadarajah, Kotz (2004)
Kummer beta generalized Gumbel KBGGu$(\mu, \sigma, \alpha, \beta, \gamma)$*	Pescim et al. (2012)
Kumaraswamy Gumbel KumGum$(\mu, \sigma, \alpha, \beta)$*	Cordeiro et al. (2012)
Exponentiated generalized Gumbel EGGu$(\mu, \sigma, \alpha, \beta)$*	Cordeiro et al. (2013)

The Gumbel distribution and all of the other generalizations are of type II. The EGu, EGa, and GLIV distributions have heavier right tail than the Gumbel distribution when $\alpha < 1$. The TEV, GGu3 and TCEV distributions have lighter right tail than the GEV, EGu, EGa, and GLIV distributions. The TEV distribution when $\alpha < 0$ and the TCEV$(0, 1, 10, 5, \alpha)$ distribution have heavier right tail than the Gumbel distribution. The right tail of the GGu3 distribution is lighter than that of the Gumbel distribution.

3 Simulation and application

A simulation study was carried out to compare the distributions and an application to a maximum monthly wind speed in Florida, was done and will be given in the poster presentation. Our simulation and application results revealed that the GEV distribution is more flexible in fitting data with a heavy right tail than the other generalizations of the Gumbel distribution. The TCEV distribution can also be a good choice.

Acknowledgments: We thank FAPESP, CNPq, and CAPES – Brazil.

References

Adeyemi, S. and Ojo, M.O. (2003). A generalization of the Gumbel distribution. *Kragujevac Journal of Mathematics*, **25**, 19 – 29.

Aryal, G.R. and Tsokos, C.P. (2009). On transmuted extreme value distribution with application. *Nonlinear Analysis: Theory, Methods & Applications*, **71**, 1401 – 1407.

Cordeiro, G.M., Nadarajah, S., and Ortega, E.M.M. (2012). The Kumaraswamy Gumbel distribution. *Statistical Methods & Applications*, **21**, 139 – 168.

Cordeiro, G.M., Ortega, E.M.M., and da Cunha, D.C.C. (2013). The exponentiated generalized class of distributions. *Journal of Data Science*, **11**, 1 – 27.

de Haan, L. (1970). *On Regular Variation and Its Application to the Weak Convergence of Sample Extremes*. Amsterdam: Mathematics Centre Tracts 32.

Dubey, S.D. (1969). A new derivation of the logistic distribution. *Naval Research Logistics Quarterly*, **16**, 37 – 40.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, **44**, 423 – 453. Translated and reprinted in (1992): *Breakthroughs in Statistics I*, (Kotz, S. and Johnson, N.L., eds.) 195 – 225., Springer-Verlag.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, **81**, 158 – 171.

Luceño, A. (2005). Fitting the generalized Pareto distribution to data using maximum goodness of fit estimators. *Computational Statistics & Data Analysis*, **51**, 904 – 917.

Nadarajah, S. and Kotz, S. (2004). The beta Gumbel distribution. *Mathematical Problems in Engineering*, **4**, 323 – 332.

Nadarajah, S. (2006). The exponentiated Gumbel distribution with climate application. *Environmetrics*, **17**, 13 – 23.

Ojo, M.O. (2001). Some relationships between the generalized Gumbel and other distributions. *Kragujevac Journal of Mathematics*, **23**, 101 – 106.

Pescim, R.R., Cordeiro, G.M., Demétrio, C.G.B., Ortega, E.M.M., and Nadarajah, S. (2012). The new class of Kummer beta generalized distributions. *SORT-Statistics and Operations Research Transactions*, **36**, 153 – 180.

Prentice, R.L. (1975). Discrimination among some parametric models. *Biometrika*, **62**, 607 – 614.

Rigby, R.A., Stasinopoulous, D.M., Heller, G., and Voudouris, V. (2013). *The Distribution Toolbox of GAMLSS*. www.gamlss.org

Rossi, F., Fiorentino, M., and Versace, P. (1984). Two-component extreme value distribution for flood frequency analysis. *Water Resources Research*, **20**, 847 – 856.

Constrained hypotheses testing via quasi U-statistics and its application to undergraduate performance assessment

Hildete P. Pinheiro[1], Pranab K. Sen[2], Aluísio Pinheiro[1], Samara F. Kiihl[1]

[1] Department of Statistics, University of Campinas, Brazil
[2] Department of Biostatistics, University of North Carolina at Chapel Hill, USA

E-mail for correspondence: hildete@ime.unicamp.br

Abstract: We propose new methodologies to assess Undergraduate performance dissimilarities. Emphasis is given to the sector of High School education from which the College student comes – private or public. Due to the complex structure of Undergraduate courses, the overall performance of a student is not based only upon its GPA (grade point average). The sample consists of all undergraduate students entering Unicamp at years 2000 – 2005 as follows. For each student a vector is formed by his/her grades in each course taken: multiple scores are considered whenever fail/pass grades happen. These vectors are then used in pairwise comparisons of common courses grades between all individuals who entered college in the same year taking into account the entrance rank. These forms a generalized U-statistic based on the classical signed rank kernel. We apply the decomposability of quasi U-statistics to define average distance measures within and between groups. Test statistics for homogeneity among groups are developed and asymptotic normality under mild conditions can be proved for the test statistic under the null hypothesis.

Keywords: Diversity measures; Jackknife; Nonparametric methods; U-statistics.

1 Introduction

We propose new methodologies to evaluate the differences in performance of students from entrance to graduation in Undergraduate School. Some work has been done addressing this problem using other methodologies (Pedrosa et al., 2007). Here, we will consider at each year the average of the entrance exam score (EES) in Undergraduate School and the grades

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

in all the courses taken by each individual. U-statistics theory and all the pairwise comparisons among individuals will be used.
Average distance measures within and between groups for each year is defined. Following Pinheiro et al. (2009, 2011), a test statistic for homogeneity tests among groups can be developed and asymptotic normality under mild conditions can be proved for the test statistic under the null hypothesis. An application with real data will be illustrated comparing performances of the students of the University of Campinas (Unicamp) from different groups, for example, those came from Public High Schools (PrS) compared with those who came from Private High Schools (PuS). The data set consists of all the students who were admitted at Unicamp from 2000 to 2005. Unicamp is a public institution, located in the State of São Paulo and one of the top research universities in Brazil.

2 Notation and U-statistics

Let $\mathbf{Z}_{ai} = (Z_{ai1}, \ldots, Z_{aik_i})^{\mathrm{T}}$ be the vector of grades for individual i who entered at year a, $l = 1, \ldots, k_i$ be the index indicating the course taken by individual i. Also, let $\bar{Z}_{0i}$ and $\bar{Z}_{0j}$ be the average of the EES for individuals i and j, respectively. Note that even though Z's can be thought as continuous random variables, we actually observe discrete grades, since they are rounded by one decimal point. So, let Y_{ail} be the discrete grades for individual i. For instance, $y_{ail} \in \{0.0, 0.1, 0.2, \ldots, 10.0\}$, i.e., $y_{ail} = 0.0$, if $y_{ail} \in [0.0, 0.05)$, $y_{ail} = 0.1$, if $y_{ail} \in [0.05, 0.15)$, $\ldots$, $y_{ail} = 9.9$, if $y_{ail} \in [9.85, 9.95)$, $y_{ail} = 10.0$, if $y_{ail} \in [9.95, 10.0]$. Analogously, we can define $\bar{Y}_{0i}$ and $\bar{Y}_{0j}$ as the discrete versions of the EES $\bar{Z}_{0i}$ and $\bar{Z}_{0j}$ for individuals i and j, respectively. Also, let $K_{ij} = k_i \cap k_j$ be the total number of courses taken in common by individuals i and j, K be the total number of courses offered by the University and let $I_l(i,j) = I(i \text{ took course } l) \times I(j \text{ took course } l)$, where $I(A) = 1$, if A is true and 0 otherwise. Define

$$\begin{aligned} \phi(\mathbf{Y}_i, \mathbf{Y}_j) &= \phi(Y_{ail}, Y_{ajl}, \bar{Y}_{0i}, \bar{Y}_{0j}) = I(Y_{ail} > Y_{ajl}) I(\bar{Y}_{0i} < \bar{Y}_{0j}) \\ &+ I(Y_{ail} < Y_{ajl}) I(\bar{Y}_{0i} > \bar{Y}_{0j}) - I(Y_{ail} > Y_{ajl}) I(\bar{Y}_{0i} > \bar{Y}_{0j}) \\ &- I(Y_{ail} < Y_{ajl}) I(\bar{Y}_{0i} < \bar{Y}_{0j}). \end{aligned} \tag{1}$$

If we want to give weights according to the type of courses taken by the students, say "required" (the number of required courses is K_{ij1}) and "elective" (the number of elective courses is K_{ij2}) courses, a weighted version

of incomplete U-statistics can be written as

$$\bar{W}^{(1)}_{n,a,gg} = \sum_{i<j:(i,j)\in n^{(1)}_{ag}} \phi^{*}_{(1)}(\mathbf{Y}^{g}_{i}, \mathbf{Y}^{g}_{j})/n^{(1)}_{ag}, \quad (2)$$

$$\bar{W}^{(2)}_{n,a,gg} = \sum_{i<j:(i,j)\in n^{(2)}_{ag}} \phi^{*}_{(2)}(\mathbf{Y}^{g}_{i}, \mathbf{Y}^{g}_{j})/n^{(2)}_{ag}, \quad (3)$$

$$\bar{W}^{(3)}_{n,a,gg} = \sum_{i<j:(i,j)\in n^{(3)}_{ag}} \phi^{*}_{(3)}(\mathbf{Y}^{g}_{i}, \mathbf{Y}^{g}_{j})/n^{(3)}_{ag}, \quad (4)$$

where $n^{(1)}_{ag} = \#(i,j) : K_{ij1} > 0, K_{ij2} > 0$; $n^{(2)}_{ag} = \#(i,j) : K_{ij1} > 0, K_{ij2} = 0$ and $n^{(3)}_{ag} = \#(i,j) : K_{ij1} = 0, K_{ij2} > 0$, with $\phi^{*}_{(1)}(\mathbf{Y}_i, \mathbf{Y}_j) = \left[\sum_{t=1}^{2} \frac{1}{K_{ijt}} \sum_{l_t=1}^{K_{ijt}} w_t I(l \in l_t)\phi(Y_{ial_t}, Y_{jal_t}, \bar{Y}_{0i}, \bar{Y}_{0j})\right] I(K_{ij1} > 0, K_{ij2} > 0)$; $\phi^{*}_{(2)}(\mathbf{Y}_i, \mathbf{Y}_j) = \left[\frac{1}{K_{ij1}} \sum_{l_1=1}^{K_{ij1}} w_1 I(l \in l_1)\phi(Y_{ial_1}, Y_{jal_1}, \bar{Y}_{0i}, \bar{Y}_{0j})\right] \times I(K_{ij1} > 0, K_{ij2} = 0)$ and $\phi^{*}_{(3)}(\mathbf{Y}_i, \mathbf{Y}_j) = \left[\frac{1}{K_{ij2}} \sum_{l_2=1}^{K_{ij2}} w_2 I(l \in l_2)\phi(Y_{ial_t}, Y_{jal_t}, \bar{Y}_{0i}, \bar{Y}_{0j})\right] I(K_{ij1} = 0, K_{ij2} > 0)$. l_1 stands for the required courses and l_2 for the elective ones and w_1 and w_2 are the respective weights for required and elective courses.

Then, for each year a, in each set (i.e., set (1): $\{K_{ij1} > 0, K_{ij2} > 0\}$, set (2): $\{K_{ij1} > 0, K_{ij2} = 0\}$ and set (3): $\{K_{ij1} = 0, K_{ij2} > 0\}$), we may define overall measures of divergence within and between groups as $\bar{W}_{n,a,gg} = \bar{W}^{(1)}_{n,a,gg} + \bar{W}^{(2)}_{n,a,gg} + \bar{W}^{(3)}_{n,a,gg}$ and $\bar{W}_{n,a,gg'} = \bar{W}^{(1)}_{n,a,gg'} + \bar{W}^{(2)}_{n,a,gg'} + \bar{W}^{(3)}_{n,a,gg'}$. Using the theory for incomplete U-statistics, we know that $\bar{W}^{(d)}_{n,a,gg}$ and $\bar{W}_{nagg'}$ are asymptotically normal distributed. Following Pinheiro et al. (2009, 2011), $B_{nagg'} = 2\bar{W}_{nagg'} - \bar{W}_{nagg} - \bar{W}_{nag'g'}$ is a quasi U-statistics and also follows a normal distribution.

Under the null hypothesis, we expect the distributions for the groups to be the same, i.e., $H_0 : 2\alpha_{gg'} - \alpha_{gg} - \alpha_{g'g'} = 0$ for all $g \neq g'$, where α_{gg} is a measure of divergence within group g and $\alpha_{gg'}$ is a measure of divergence between groups g and g', with individuals i and j from group g and from groups g and g', respectively, given that they took K_{ij} courses in common. Under some general conditions (normality, for instance) this is a one sided test and some hypothesis tests for the interaction effect can be done using the union intersection principle (UIP) discussed in Sen and Silvapulle (2005) in order to maximize the power of the tests.

3 Description of the data set

The dataset is composed by 12168 (57.3% male and 43.7% female) students which have enrolled at Unicamp at years 2000 to 2005 in Bachelor's degree courses of the areas of Arts (A), Health Science (HS), Engineering and

Exact Sciences (EngES) and Social Sciences (SS). The academic situation of these students were classified as following: Active (students who were still enrolled in the University and had not graduated yet - 0.9%), Graduates (students who have already graduated - 77.1%), and Others (the ones who drop out from the University - 22.0%). The students were, in their majority, between 16 and 23 years old (94.3%) from all Brazilian regions and enrolled in 45 different courses from the areas of HS (19.8%), EngES (55.7%), SS (18.5%) and A (6%). About 70% of students who enrolled between 2000 and 2005 come from PrS.
The data shows that the performance at the EES of students who studied in PrS seems to be better than for those coming from PuS, but once they get into the University and we look at their GPA scores, the situation seems to get reversed or at least, on average, they have equal GPA scores.
As we know that students from different Areas and different years of entrance are more likely to take different courses, we will separate the analysis of students' performance by Area and by Year of Entrance.
The main interest here is to test the following hypotheses: H_{01}: *There is no difference in performance between female and male*; H_{02}: *There is no difference in performance between students coming from Public and Private High Schools*; H_{03}: *There is no interaction between sex and type of High School*.

Acknowledgments: H.P. Pinheiro, S.F. Kiihl and A. Pinheiro's research was partially supported by CNPq and FAPESP, Brazil.

References

Pinheiro, A., Sen, P.K., and Pinheiro, H.P. (2009). Decomposability of high-dimensional diversity measures: Quasi U-statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis*, **100**, 1645 – 1656.

Pinheiro, A., Sen, P.K., and Pinheiro, H.P. (2011). A class of asymptotically normal degenerate quasi U-statistics. *Annals of the Institute of Mathematical Statistics*, **63**, 1165 – 1182.

Pedrosa, R.H.L., Dachs, J.N.W., Maia, R.P., Andrade, C.Y., and Carvalho, B.S. (2007). Academic performance, student's background and affirmative action at a Brazilian research University. *Higher Education Management and Policy*. **19**.

Sen, P.K. and Silvapulle, M.J. (2005). *Constrained Statistical Inference: Inequality, order, and shape restrictions*. Wiley Series in Probability and Statistics. Wiley-Interscience.

Structural equation models for dealing with spatial confounding

Hauke Rennies[1], Thomas Kneib[1]

[1] Georg-August University of Goettingen, Germany

E-mail for correspondence: hrennie@gwdg.de

Abstract: In regression analyses of spatially structured data, it is a common practice to introduce spatially correlated random effects into the model to reduce or even avoid bias in the estimation of other covariate effects. If besides the response also the covariates are spatially correlated, the spatial random effects may confound the effect of the covariates or vice versa. In these cases the model fails at identifying the true covariate effect due to multicollinearity problems. For highly collinear continuous covariates path analysis and structural equation modeling techniques proved to be helpful to disentangle direct covariate effects from effects arising from correlation with other variables. This work discusses the applicability of these techniques in regression setups where spatial and covariate effects coincide at least partly and classical geoadditive models fail to separate these effects.

Keywords: Spatial confounding; Path analysis; Effect separation.

1 Introduction

Spatial confounding in regression analysis arises when both the response and the covariate are spatially structured on the same or a similar spatial scale. The resulting estimated effects are most likely biased and interpretation in terms of causal relationships is no longer possible. Aim of current research has therefore become the development of techniques to avoid confounding bias. Hodges and Reich (2010), for instance, encountered this problem while analyzing the relationship between the stomach cancer incidence ratio and the socio economic status on level of municipalities in Slovenia. Paciorek (2010) describes the influence of the spatial scales on which response and covariate vary on the degree of confounding bias. Hughes and Haran (2013) provide methods to alleviate spatial confounding

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

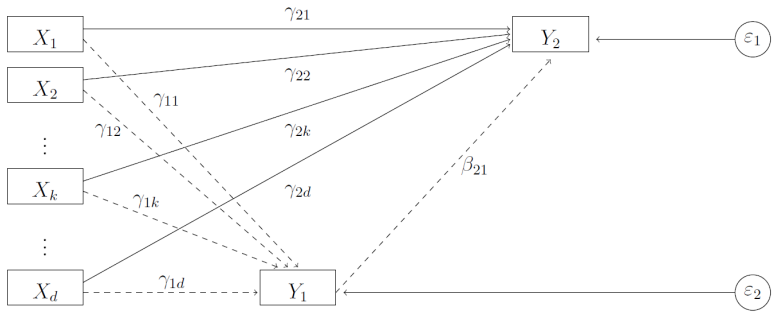


FIGURE 1. Path diagram of spatial confounding when space is measured discretely. Each region $k \in \{1, \ldots, d\}$ has an influence on the response Y_2 directly and through the covariate effect β_{21}.

by a reparametrization of geoadditive models. The common factor of the mentioned papers is that they provide Bayesian methodology and assign all occurring spatial variation to the covariate and hence don't allow for additional spatial structure in the response.
In our contribution, we interpret the role of space conceptually different in the sense that space is allowed to have an effect on the covariate and the response simultaneously. Figure 1 illustrates this idea for discretely measured space (regions $1, \ldots, d$). The total spatial information in the response Y_2 is composed of the direct spatial effect (solid arrows) and the spatial effect in the covariate Y_1 transported to the response through the covariate effect β_{21} (dashed arrows). We establish a frequentist estimation strategy which is based on multiple regression equations unified in the framework of Structural Equations Models (SEM). SEM techniques provide us with the methodology to estimate all occurring effects shown in Figure 1 in one single step. Estimation is based on the joint likelihood of Y_1 and Y_2 conditioned on the location of the observations $(X_1, \ldots, X_2)$. We investigate under which circumstances classical geoadditive approaches fail to separate the effects and illustrate the applicability of SEM techniques in this context.

2 Structural equation models with observed variables

Using standard SEM notation, we denote by $\mathbf{y}_i = (y_{1i}, y_{2i})^{\mathrm{T}}$ the i-th observation of the endogenous variables, where y_{2i} is the response variable and y_{1i} corresponds to the covariate of interest. Space is considered as an

exogenous variable. Figure 1 can be translated into two structural equations

$$y_{1i} = \sum_{k=1}^{d} x_{k_i}\gamma_{1k} + \varepsilon_{1i} \tag{1}$$

$$y_{2i} = \sum_{k=1}^{d} x_{k_i}\gamma_{2k} + y_{1_i}\beta_{21} + \varepsilon_{2i}, \tag{2}$$

where we assume $\varepsilon_{ji} \overset{iid}{\sim} \mathcal{N}(0, \sigma_j^2)$, $i = 1, \ldots, n$, for $j = 1, 2$ and independence between ε_1 and ε_2. The variables x_{k_i} are dummies indicating the location of observation i. Combining Equations (1) and (2) yields the bivariate model equation

$$\mathbf{y}_i = \mathbf{B}\mathbf{y}_i + \mathbf{\Gamma}\mathbf{x}_i + \boldsymbol{\varepsilon}_i,$$

where the two spatial effects are summarized in $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^{\mathrm{T}}$ and $\mathbf{B} = \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix}$ contains the covariate effect. The estimation is based on the resulting likelihood arising from the normality assumptions.

3 Results

To illustrate the applicability of the introduced method a simulation study was performed, similar to that of Hughes and Haran (2013). Data was generated according to

$$\begin{aligned} y_{1i} &= \gamma_{1i} + \varepsilon_{1i} \\ y_{2i} &= \beta_{21} y_{1i} + \gamma_{2i} + \varepsilon_{21} \\ \varepsilon_{1i} &\sim \mathcal{N}(0, \sigma_1^2) \\ \varepsilon_{21} &\sim \mathcal{N}(0, \sigma_2^2) \end{aligned}$$

with a fixed sample size of $n = 100$ and variance combinations $(\sigma_1^2, \sigma_2^2) = (0.2, 1), (1, 1)$ and $(1, 0.2)$. Here, γ_{j_i} denotes the spatial components for covariate and response and was generated on an artificial map consisting of 49 regions. Figure 2 shows the estimated covariate effect for increasing variability σ_1^2 of the covariate beyond the spatial scale. The results for a classical geoadditive approach (red boxplots) are compared to the results from decomposing the effects using path analysis methodology (black boxplots). Geoadditive regression models are incapable of avoiding bias if the covariate does not vary (or only few) beyond the spatial scale. This result is in line with Paciorek (2010). Our methodology reduces the bias in this case, but shows increased uncertainty. For higher variability in the covariate both models lead to unbiased estimates.

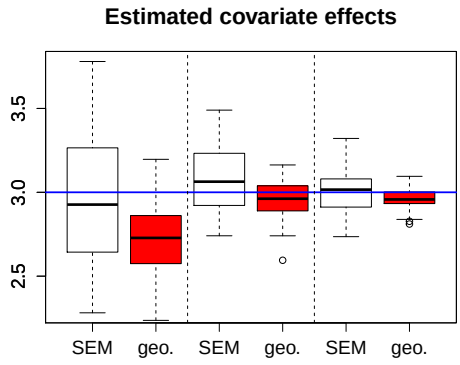


FIGURE 2. Simulation results for geoadditive (red) and SEM (black) estimation for increasing variability of the covariate beyond the spatial scale (from left to right). The true covariate effect is $\beta_{21} = 3$ (blue horizontal line).

4 Discussion

In the present study we incorporate discrete spatial information into the framework of observed variable SEM and provide a methodology to reduce the bias in the estimation that results from spatial confounding. In case of a strong spatial structure within the covariate estimation with the SEM approach is still unbiased but shows increased uncertainty. In contrast to existing techniques to alleviate confounding bias our method allows for spatial variation in both, the covariate and the response.

References

Hodges, J.S. and Reich, B.J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, **64**, 325 – 334.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B*, **75**, 139 – 159.

Paciorek, C.J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, **25**, 107 – 125.

Bayesian approach for modelling bivariate survival data through the PVF copula

Jose S. Romeo[1], Renate Meyer[2]

[1] Department of Mathematics, University of Santiago, Chile
[2] Department of Statistics, University of Auckland, New Zealand

E-mail for correspondence: jose.romeo@usach.cl

Abstract: In this work we study the Archimedean copula model generated by the Laplace transform (LT) of the Power Variance Function (PVF) frailty distribution to model the dependence structure of multivariate lifetime data. The PVF copula family includes, among others, the Clayton, Gumbel (Positive Stable) and Inverse Gaussian copulas as special or limiting cases. Dependence properties of the copula models are described. From a Bayesian framework, parameters of the marginal distributions and the PVF copula are simultaneously estimated using piecewise exponential distributions. We illustrate the usefulness of the methodology using data from the Australian NH&MRC Twin registry.

Keywords: Multivariate survival analysis; Archimedean Copulas; Dependence.

1 Introduction

The independence assumption no longer holds for multivariate survival data in situations when each individual can experience the serial recurrence of the same type of event such as recurrent asthma attacks, or different parallel events such as the onset of retinopathy in left and right eye, nor does it hold for clustered failure times such as failure times of twins or patients in the same hospital. In these situations, the association between the time-to-events within an experimental unit or a cluster needs to be taken into account. The two most popular approaches for analysing multivariate survival data are copula models and frailty models. In a frailty model, the survival times are considered conditionally independent, given a latent subject-specific random effect, taking implicitly the correlation structure between failure times into account. However, an explicit modelling of the dependence is possible via a copula model, that models the marginal distributions and the association structure separately. In this work we study

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

the Archimedean copula model generated by the LT of the PVF frailty distribution. It includes, among others, three common choices of copulas as special or limiting cases: the gamma (Clayton), positive stable (Gumbel) and inverse Gaussian distributions. We provide a comprehensive study of dependence properties of the PVF copula family. Using a Bayesian approach to the joint modelling of bivariate survival data using the PVF copula which allows a one-stage estimation of marginal hazards and copula simultaneously, we illustrate the methodology on data sets from the Australian Twin Study (Duffy et al., 1990) with the main objective to compare monozygotic (MZ) and dizygotic (DZ) twins as to the strength of correlation of the times to appendectomy between the twins.

2 The PVF copula model

A random variable W has a three-parameter PVF(α, δ, θ) distribution if the probability density function is given by

$$f(w|\alpha,\delta,\theta) = -\frac{1}{\pi w}\exp(-\theta w+\delta\theta^{\alpha}/\alpha)\sum_{k=1}^{\infty}\frac{\Gamma(k\alpha+1)}{\Gamma(k+1)}\left(\frac{-w^{-\alpha}\delta}{\alpha}\right)^{k}\sin(\alpha k\pi)$$

for $0 < \alpha < 1$, $\delta > 0$, and $\theta \geq 0$. The LT takes the form $\mathrm{L}_W(s) = \exp\left\{-\frac{\delta}{\alpha}\left[(\theta+s)^{\alpha}-\theta^{\alpha}\right]\right\}$. If $\theta > 0$ all positive moments exist, the expected value and variance are given by $\mathrm{E}[W] = \delta\theta^{\alpha-1}$ and $\mathrm{Var}[W] = \delta(1-\alpha)\theta^{\alpha-2}$, respectively. As shown by Oakes (1989), any multiplicative frailty model for multivariate survival data has an Archimedean copula representation. In the bivariate case, let (T_1, T_2) be continuous random variables with marginal survival functions given by (S_1, S_2) and corresponding copula C. Then for non-negative (T_1, T_2) that are conditionally independent given a random variable W, i.e. $S(t_1, t_2|w) = S_1(t_1|w)S_2(t_2|w)$, the joint survival function can be expressed as $S(t_1, t_2) = C(S_1(t_1), S_2(t_2)) = \varphi^{-1}\left[\varphi(S_1(t_1))\varphi(S_2(t_2))\right]$, with generator $\varphi^{-1}(\cdot) = \mathrm{L}_W(\cdot)$ equal to the inverse LT of W. By considering the PVF distribution as the frailty distribution, we obtain the Archimedean PVF copula. Under the reparametrization $\delta = \eta^{1-\alpha}$ and $\theta = \eta$ with $\alpha \in (0,1)$ and $\eta > 0$, the bivariate PVF copula function is

$$C_{\alpha,\eta}(u_1,u_2) = \exp\left\{-\frac{1}{\alpha}\left[\eta^{1-\alpha}\left(g(u_1)^{\frac{1}{\alpha}}+g(u_2)^{\frac{1}{\alpha}}-\eta\right)^{\alpha}-\eta\right]\right\},$$

where $g(s) = \eta^{\alpha} - \alpha\eta^{\alpha-1}\log s$ and $u_j = S_j(t_j)$, with $j = 1, 2$. For $\alpha \to 0$ we obtain the Clayton copula, if $\eta = 0$ the Gumbel copula and the inverse Gaussian copula can be obtained by considering $\alpha = 0.5$. For the PVF copula Kendall's tau does not have a closed-form expression but can be computed numerically and is given by $\tau = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt$, where $0 < \tau < 1$. The survival times (T_1, T_2) are independent if either $\alpha \to 1$ for all

$\eta \in R^+$ or when $\eta \to \infty$ for all $\alpha \in (0,1)$. The comonotonicity copula, is obtained when both parameters go to zero. The PVF copula has no tail dependence. The Clayton family has lower-tail dependence, $LTD_C = 2^{-1/\eta}$, but does not have upper-tail dependence. The Gumbel family has no lower-tail dependence but has upper-tail dependence, $UTD_C = 2 - 2^{\alpha}$. The inverse Gaussian has no tail dependence. For the PVF copula the cross-ratio function can be expressed by $\chi(v) = 1 - \frac{1-\alpha}{\alpha \log v - \eta}$. Note that $\chi(v) = \chi(S(t_1, t_2))$ is decreasing in both t_1 and t_2. As $\alpha \to 0$, $\chi(v) = \eta^{-1} + 1$, which is constant over (t_1, t_2) (Clayton copula). If $\eta = 0$, $\chi(v) = 1 - \frac{1-\alpha}{\alpha \log v}$ (Gumbel copula), and when $\alpha = 0.5$, $\chi(v) = 1 - (\log v - 2\eta)^{-1}$ (inverse Gaussian copula), which are both decreasing functions in t_1 and t_2.

3 Illustration

We apply the PVF copula to analyze data on the time to appendectomy for adult twins in the Australian NH&MRC Twin Registry given in Duffy et al. (1990). This study was conducted to investigate whether the strength of dependence within twin pairs as to the risk of the onset of various diseases, including acute appendicitis, is different for monozygotic (MZ) and dizygotic (DZ) twins. We present the results of the analysis of female and male twin pairs who had an appendectomy. These comprised of 1798 MZ (1231 female and 567 male pairs) and 1098 DZ twins (748 female and 350 male pairs). We fit three copula models (PVF, Clayton, Gumbel) with piecewise exponential (PWE) marginal distributions implementing the joint one-stage estimation procedure as in Romeo et al. (2006). Assuming prior distributions for the parameters of the marginal distributions and independent prior distributions for the copula parameters, we use the Gibbs sampler for posterior computations which is implemented using the `WinBUGS` software (Lunn et al., 2000). Instead of the conventional approach of analysing MZ and DZ separately, we allow the association parameter to depend on covariates (Meyer and Romeo, 2015), i.e., including the type of zygosity as a dichotomous covariate as well as the sex of the twins. Specifically, we consider the inclusion of the covariates through the logit link in the dependence parameter α for PVF and Gumbel models and through the log link for the η parameter in the Clayton model. The PVF- and Gumbel copula models, see Table 1, provide a better fit than Clayton model. As the differences in DIC and LPML values between PVF and Gumbel copula models are negligible and the posterior estimate of η (0.007 ± 0.014) in the PVF model is close to zero, in Table 2 we summarize the posterior distributions of the parameters for the Gumbel model but as functions of the Kendall's τ coefficient. As expected, the dependence between MZ pairs was stronger than between DZ pairs suggesting a genetic component to the disease. Also, the dependence between female pairs is stronger than male twins.

TABLE 1. Model selection criteria for copula models, Australian twin data.

Model	AIC	BIC	DIC	p_D	LPML
PVF-PWE	15122.71	15254.07	15100.77	22.06	-7551.83
Clayton-PWE	15147.71	15273.10	15127.03	21.33	-7564.39
Gumbel-PWE	15122.61	15248.00	15101.74	21.13	-7551.68

TABLE 2. Posterior Kendall's τ, Gumbel copula model, Australian twin data.

Parameter	Mean	Median	SD	HPD
$\tau_{MZ-Female}$	0.229	0.230	0.024	0.183, 0.276
$\tau_{MZ-Male}$	0.143	0.143	0.031	0.079, 0.197
$\tau_{DZ-Female}$	0.141	0.141	0.025	0.094, 0.193
$\tau_{DZ-Male}$	0.085	0.083	0.026	0.040, 0.137

4 Conclusion and discussion

In this work we presented a Bayesian analysis of the Archimedean copula model induced by the PVF frailty distribution. The methodology was applied to the Australian twin data on appendectomy. Parameters of the marginal distributions and the PVF copula were simultaneously estimated. Allowing the association parameter to depend on covariates, lets the estimation of different Kendall's τ parameters for MZ and DZ twins by sex. In general, such modelling of the copula parameter as a function of other covariates offers considerable flexibility.

References

Duffy, D.L., Martin, N.G., and Mathews, J.D. (1990). Appendectomy in Australian twins. *American Journal of Human Genetics*, **47**, 590 – 592.

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

Meyer, R. and Romeo, J.S. (2015). Bayesian semi-parametric analysis of recurrent failure time data using copulas. *Submitted*.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, **84**, 487 – 493.

Romeo, J.S., Tanaka, N.I., and Pedroso de Lima, A.C. (2006). Bivariate survival modeling: A Bayesian approach based on copulas. *Lifetime Data Analysis*, **12**, 205 – 222.

Posterior sensitivity of variance priors in Bayesian structured additive regression

Helene Roth[1], Stefan Lang[1]

[1] Department of Statistics, University of Innsbruck, Austria

E-mail for correspondence: helene.roth@uibk.ac.at

Abstract: The inverse Gamma distribution is a standard prior for variances of nonlinear and random effects in Bayesian structured additive and related regression models. In this talk we provide an in-depth investigation of the sensitivity of estimation results on the choice of the hyperparameters of the inverse Gamma distribution. We focus on the estimated effects, model fit as well as model selection. We additionally discuss the sensitivity on the Bayesian credible intervals (or more generally on posterior quantiles) of nonlinear curves and random effects. Our analysis is based on both theoretical results and on extensive simulation experiments. Our findings are illustrated with an application on Austrian real estate prices.

Keywords: Bayesian hierarchical models; Inverse gamma distribution; Posterior sensitivity; Mixed models; P-splines.

1 Structured additive regression models

Suppose the observations $(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, where y_i is a continuous response variable, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq})^{\mathrm{T}}$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ are vectors of covariates. For the variables in $\boldsymbol{z}$ possibly nonlinear effects are assumed whereas the variables in $\boldsymbol{x}$ are modeled in the usual linear way. The components of $\boldsymbol{z}$ are not necessarily continuous covariates. A component may also indicate a time scale, a cluster- or a spatial index (e.g. municipality, district or county) a particular observation pertains to. We assume an additive decomposition of the effects of z_{ij} (and x_{ij}) and obtain the model

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma} + \varepsilon_i. \tag{1}$$

Here, $f_1, \ldots, f_q$ are nonlinear functions of the covariates $\boldsymbol{z}_i$ and $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}$ is the usual linear part of the model. The errors ε_i are assumed to be mutually independent Gaussian with mean 0 and variance σ^2, i.e. $\varepsilon_i \sim N(0, \sigma^2)$.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

The nonlinear effects in (1) are modeled by a basis functions approach, i.e. a particular function f of covariate z is approximated by a linear combination of basis or indicator functions

$$f(z) = \sum_{k=1}^{K} \beta_k B_k(z). \tag{2}$$

The B_k's are known basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^{\mathrm{T}}$ is a vector of unknown regression coefficients to be estimated.
Effect modeling and priors depend on the covariate or term type. In a Bayesian framework a standard smoothness prior is a (possibly improper) Gaussian prior of the form

$$p(\boldsymbol{\beta}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{rk(\boldsymbol{K})/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\beta}\right) \cdot I(\boldsymbol{A}\boldsymbol{\beta} = \mathbf{0}), \tag{3}$$

where $I(\cdot)$ is the indicator function. The key components of the prior are the penalty matrix $\boldsymbol{K}$, the variance parameter τ^2 and the constraint $\boldsymbol{A}\boldsymbol{\beta} = \mathbf{0}$. Usually the penalty matrix is rank deficient, i.e. $rk(\mathbf{K}) < K$, resulting in a partially improper prior.
The amount of smoothness is governed by the variance parameter τ^2. A conjugate inverse Gamma prior is employed for τ^2, i.e. $\tau^2 \sim \Gamma^{-1}(a, b)$. The two hyperparameters a and b of the density $p\left(\tau^2\right) \propto \left(\tau^2\right)^{-a-1} \exp\left(-\frac{b}{\tau^2}\right)$ determine the shape of the distribution. In cases of limited information for τ^2 the specific choice of the hyperparameters may affect posterior inference. Fahrmeir and Kneib (2009) elaborate a variety of special cases based on the shape and scale parameter and also show necessary or sufficient conditions for the propriety of the posterior:

- $a = -1, b = 0$ result in an improper flat prior $\tau^2 \sim const.$
- $a = -1/2, b = 0$ corresponds to a flat prior for the standard deviation $\sqrt{\tau^2} \sim const.$
- $a = b = \epsilon$ with a small $\epsilon > 0$ serve as a proper approximation to Jeffrey's prior with $\epsilon = 0$ which is uniform on the log scale.
- $a = 5, b = 25$ are popular for spike and slab priors.

2 Theoretical results

Our results are illustrated with the following one component random effects model

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\gamma} &\sim N\left(\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}\right) \\ \boldsymbol{\gamma}|\tau^2 &\sim N\left(0, \tau^2\boldsymbol{I}\right) \\ \tau^2 &\sim \Gamma^{-1}\left(a, b\right). \end{aligned}$$

The posterior mode can be computed iteratively using similar methodology as in usual linear mixed models. The estimator for the variance parameter is then of the form

$$\hat{\tau}^2 = \frac{\frac{\boldsymbol{\gamma}^T\boldsymbol{\gamma}}{2} + b}{\frac{q}{2} + a + 1} \tag{4}$$

with q being the number of effects $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)$. The estimator is based on the average effect error with a (resp. b) reducing (increasing) the estimated variance.

3 Simulation

Three different simulation designs are used to characterize the relationship of the hyperparameter and the model estimation:

- A *linear* model is set up with a second order random walk P-spline to investigate the automatic model selection quality.
- A sinus in the *smooth* model emulates the necessity of functional flexibility.
- Varied prior shapes may influence the size and significance of group effects in a *random* intercept model.

We simulated models with different sample sizes and error variances. Hyperparameter combinations of $a = 0.001, 1, 5, 25$ and $b = 0.001, 1, 5, 25$ together with $a = -1, b = 0$ and $a = -1/2, b = 0$ are tested.
Evaluated are the overall performance via the deviance information criterion (DIC), the quality of the estimated functions with the mean squared error (MSE) together with tabulated median posterior distributions to capture the variability of the posterior distribution.
The posterior distribution of the parameter variance is shifted towards (resp. away from) zero with increased a (resp. b). (4) already suggests that behavior from the ML estimate. Furthermore the variance of the estimated parameter variance decreases (increases) with increased shape parameter a (scale parameter b). The uninformative hyperparameter combinations produce overestimated results by tendency, however some other combinations are clearly worse.
Hyperparameter changes only affect the corresponding flexible term. There we observe some small effect on the mean and median estimate. The credibility intervals are influenced and sensitivities can change. Intuitively hyperparameter combinations resulting in larger averaged median variance estimate also produce larger credibility intervals. Improved data quality (increased sample size or decreased error variation) reduce this effect.
Even though one can observe an increase (decrease) in the median DIC with an increase of the scale parameter b (shape parameter a) in most of

the cases these changes are not substantial based on Spiegelhalter, et al. (2002). For that reason the uninformative hyperparameter combinations do not perform substantially better, but it also means that one can simply remain with them. More details on our results will be given in the talk.

4 Application

The application is based on house prices from 2770 owner-occupied single-family houses in Austria from 1997 to 2008 collected by the UniCredit Bank Austria AG to assess credit risk. Details will be given in the conference talk.

Acknowledgments: This work was supported by funds of the Österreichische Nationalbank (Österreichische Nationalbank, Anniversary Fund, project number: 15309). This work was supported by the Austrian Ministry of Science BMWF as part of the UniInfrastrukturprogramm of the Focal Point Scientific Computing at the University of Innsbruck.

References

Fahrmeir, L. and Kneib, T. (2009). Propriety of posteriors in structured additive regression models: Theory and empirical evidence. *Journal of Statistical Planning and Inference*, **139**, 843 – 859.

Spiegelhalter, D., Best, N., Carin, B., Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **4**, 583 – 639.

Analysis of Brazil's presidential election via Bayesian spatial quantile regression

Bruno Santos[1], Heleno Bolfarine[1]

[1] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

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

Abstract: We show an extension of Bayesian quantile regression models when the response variable is reported as a proportion and spatial correlation is present. We are specially interested in the data of the last presidential election in Brazil, which was decided by 2% of the valid votes, approximately. We use quantile regression models to show how some sociodemographic variables are associated with different quantiles of the distribution of votes in this close election.

Keywords: Bayesian spatial quantile regression; Asymmetric Laplace predictive process; Brazil's presidential election data.

1 Introduction

Quantile regression has become a effective method when one considers that the conditional distribution of the response variable has more to offer than just the conditional mean. This is usually the case when the data presents heteroskedasticity, for which quantile regression can estimate regression parameters for different quantiles. A Bayesian version with a convenient Gibbs sampler of this model was proposed by Kozumi and Kobayashi (2011). Following, Santos and Bolfarine (2014) extended this model to allow the response variable to vary only between zero and one, and also with the possibility of having a mass of points at zero or one.
Moreover, we are interested in Brazil's last presidential election. Brazil reelected its current president, Dilma Rousseff, in 2014 and the election ended with a very small difference between the final two candidates. She won this election by around 2% of the valid votes and the distribution of votes had, to some extent, some spatial clusters, as one can see in Figure 1. Given this result, we decided to study this data in the light of a Bayesian spatial quantile regression model. This paper is organized as follows. We make a

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

quick review of Bayesian spatial quantile regression models in Section 2. We address Brazil's election problem in Section 3, highlighting the main contributions we aim with our work.

2 Bayesian spatial quantile regression model

When the response variable is defined as a proportion, i.e., being in the interval (0,1), Santos and Bolfarine (2014) showed that a Bayesian quantile regression is suitable to explain its conditional quantiles. The authors show that it is possible to consider the posterior inference consistent in this case. Although there is no need to transform the response variable to estimate the Bayesian quantile regression for this type of data, it is important to avoid considering densities with values outside the proper range (0,1) in the likelihood, thus they suggest working with $h(Y)$ instead of Y, where $h : (0, 1) \to \mathbb{R}$.

In order to take into account the spatial correlation, Reich et al. (2011) and Lum and Gelfand (2012) proposed Bayesian spatial quantile regression models. The former paper considers spatially varying regression coefficients, which are defined as a weighted sum of Bernstein basis polynomials, while the latter uses a representation of the asymmetric Laplace distribution as location-scale mixture of normal and exponential distributions.

Following Lum and Gelfand (2012), a Bayesian spatial quantile regression model through an asymmetric Laplace process (ALP) can be defined as

$$Y(s) = x^{\mathrm{T}}(s)\beta(\tau) + \epsilon_\tau(s), \qquad \epsilon_\tau(s) = \psi\sqrt{\sigma\xi(s)}Z(s) + \theta\xi(s)$$
$$Z(s) \sim GP(0, \rho_Z(s, s'; \lambda),$$

where $\theta = (1 - 2\tau)/(\tau(1 - \tau))$ and $\psi^2 = 2/(\tau(1 - \tau))$; $\sigma > 0$ is a scale parameter; $Z(s)$ is a Gaussian process with zero mean and valid covariance function between two locations, for which we take a exponential correlation function with parameter λ; $\beta(\tau)$ is the quantile regression coefficient for the τth quantile. For this alternative representation of the asymmetric Laplace distribution to be correct, we assume $Z(s)$ and $\xi(s)$ to be independent. Due to computational difficulties, we consider the iid assumption for $\xi(s)$, i.e., $\xi(s) \overset{iid}{\sim} \text{Exp}(\sigma)$. Marginally, we still continue with the errors $\epsilon_\tau(s)$ being distributed according to a asymmetric Laplace distribution with location parameter equal to zero, scale parameter σ and skewness parameter τ.

Due to the size of the dataset of interest, Brazil's presidential election data with its more than 5000 cities, the ALP is too computational demanding. Considering high dimensional cases like this, Lum and Gelfand (2012) proposed an asymmetric Laplace predictive process (ALPP), which is related to Gaussian predictive process models proposed by Banerjee et al. (2008). We make use of this method when we analyze our data in the next section. To complete our Bayesian specification we need to elicit our priors, but in the interest of conciseness we leave it for a extended version of this

manuscript. We make use of similar posterior distributions of Lum and Gelfand (2012).

3 Brazil's presidential election data

The last presidential election in Brazil, which happened in 2014, will be remembered as one of the closest decisions, as the incumbent president won it by 2%, approximately, with more than 100 million people voting. As a result of a such close call, one could be interested in understanding what factors could be associated with the win for one candidate or with the loss for the other. A similar exercise was done by Shikida et al. (2009), where the authors looked at the data for the reelection of president Lula in 2006. However, their article relies only on the conditional mean of the distribution of votes given other variables. Our work aims to help this discussion by considering a more complete picture of this distribution, through the use of a Bayesian spatial quantile regression model. We extend the results of Santos and Bolfarine (2014) to allow the response variable to have a spatial correlation structure.

The data illustrated in Figure 1 shows the proportion of votes candidate Dilma Rousseff received in the second round of voting, where there were only two candidates still competing to be the next president of Brazil. The data is accessible at `http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais` and available for every city in the country. In order to understand what could have affected this election, we selected for the explanatory variables: dummies for the five regions of Brazil, leaving the Southeast region as the reference; human development index (HDI); income per capita (INCPC); growth of income per capita between 2000 and 2010 (G_INCPC); ratio of Gini index between 2000 and 2010; proportion of people older than 18 years old, who are economically active in the population of the city (EAP); proportion of families who receive help from a Conditional Cash Transfer program called *Bolsa Família* (P_BF); average value that each family receives from *Bolsa Família* (Val_BF). The data of the explanatory variables is available at `http://www.atlasbrasil.org.br/2013/en/download/`.

The main result of our work is to show how quantile regression is advantageous when dealing with this type of data, while adjusting for spatial correlation. We are able to infer that for several variables the parameter coefficient is not constant over different quantiles, which could not be verified by looking just at the conditional mean. For example, although the coefficient is always positive for P_BF, the magnitude of this effect is higher for smaller quantiles. In a interesting manner, the estimates for the South Region are positive for $\tau < 0.5$ and negative for $\tau > 0.8$. Also, the estimates for HDI are only significant and positive for $\tau < 0.4$. Furthermore, the estimates for EAP are significant and negative for $\tau > 0.7$.

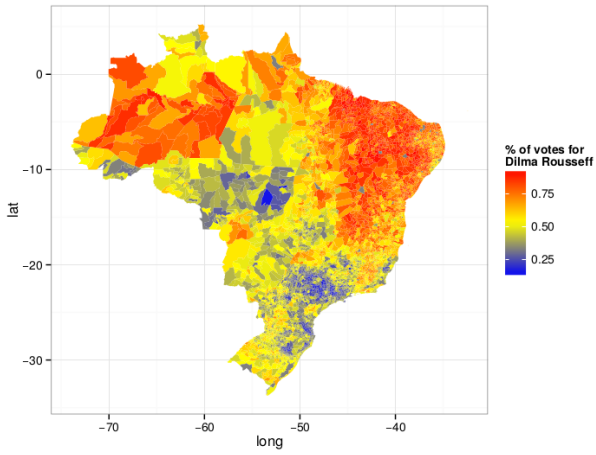


FIGURE 1. Distribution of votes for reelected president Dilma Rousseff, in percentage of total votes for each city of Brazil.

Acknowledgments: The authors thank the financial support by FAPESP, through grants 2012/20267-9 and 2013/04419-6.

References

Banerjee, S., Gelfand, A.E., Finley, A.O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, **70**, 825 – 848.

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, **81**, 1565 – 1578.

Lum, K. and Gelfand, A. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, **7**, 235 – 258.

Reich, B.J., Fuentes, M., and Dunson, D.B. (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association*, **106**, 6 – 20.

Santos, B. and Bolfarine, H. (2014). Bayesian analysis for zero-or-one inflated proportion data using quantile regression. *Journal of Statistical Computation and Simulation*, DOI: 10.1080/00949655.2014.986733.

Shikida, C.D., Monasterio, L.M., de Araujo Jr., A.F., Carraro, A., and Damé, O.M. (2009). "It is the economy, companheiro!": an empirical analysis of Lula's re-election based on municipal data. *Economics Bulletin*, **29**, 976 – 991.

Modelling wind direction with an application

Lukas M. Schäfer[1], Bruce J. Worton[1]

[1] School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, Edinburgh, UK

E-mail for correspondence: Bruce.Worton@ed.ac.uk

Abstract: In this paper we investigate various approaches to model directional data. The motivating example relates to a wind direction dataset collected at the University of Edinburgh. Of particular interest is the need for a flexible model to determine appropriate summaries of the extensive data and thus to gain a better understanding of features of weather patterns. A simulation study was used to evaluate the properties of a numerical estimation procedure.

Keywords: Directional data analysis; Likelihood inference; Weather patterns.

1 Introduction

We consider methods for modelling direction data. In particular, we investigate modelling extensive data collected by the School of Geosciences at the University of Edinburgh. Datasets for the weather station are available at http://www.geos.ed.ac.uk/abs/Weathercam/station/data.html over several years. In this paper we consider a subset of the wind direction data in our analyses; the dataset relates to wind directions with wind speeds greater than 5ms^{-1} and recorded at 12 noon in the year of 2013.

2 The WFGSN model

We investigate the Wrapped Flexible Generalized Skew Normal (WFGSN) model proposed by Hernández-Sánchez and Scarpa (2012) for analysis of the wind direction data. The WFGSN density may be expressed as

$$f(\theta)=\frac{2}{\omega}\sum_{r=-\infty}^{\infty}\phi\left(\frac{\theta+2\pi r-\xi}{\omega}\right)\Phi\left\{\alpha\left(\frac{\theta+2\pi r-\xi}{\omega}\right)+\beta\left(\frac{\theta+2\pi r-\xi}{\omega}\right)^3\right\}$$

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

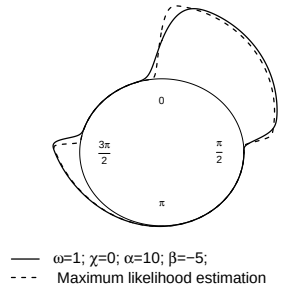


FIGURE 1. Comparison of densities: true and maximum likelihood estimated using an artificial dataset with $n = 50$.

for $0 \leq \theta \leq 2\pi$, $\xi \in \mathbb{R}$, $\omega > 0$, $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ and WFGSN is a bimodal and asymmetrical distribution. In our modelling we also considered a more general version of the WFGSN distribution of degree $2k-1$ with $k = 1, 2, \ldots$ defined by

$$f(\theta; \xi, \omega, \boldsymbol{\kappa}) = \frac{2}{\omega} \sum_{r=-\infty}^{\infty} \phi\left(\frac{\theta + 2\pi r - \xi}{\omega}\right) \Phi\left\{\sum_{i=1}^{k} \kappa_i \left(\frac{\theta + 2\pi r - \xi}{\omega}\right)^{2i-1}\right\},$$

for $0 \leq \theta \leq 2\pi$, $\xi \in \mathbb{R}$, $\omega > 0$ and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_k)^{\mathrm{T}} \in \mathbb{R}^k$.

3 Simulation study

We investigate the ML estimation of the parameters from the WFGSN model. The log likelihood function, based on observations $\theta_1, \ldots, \theta_n$, for the first version of WFGSN is given by

$$l(\xi, \omega, \alpha, \beta) = n \log(2) - n \log(\omega) +$$
$$\sum_{i=1}^{n} \log\left[\sum_{r=-\infty}^{\infty} \phi\left(\frac{\theta_i + 2\pi r - \xi}{\omega}\right) \Phi\left\{\alpha\left(\frac{\theta_i + 2\pi r - \xi}{\omega}\right) + \beta\left(\frac{\theta_i + 2\pi r - \xi}{\omega}\right)^3\right\}\right].$$

We used the function `nlminb` in R to maximize the log likelihood, and this seemed to work well for both the simulation study and the application to real data which is given in the following section. Figure 1 gives a graphical comparison of a true model and a fitted model to an artificial dataset with sample size $n = 50$. This provides an illustration that ML estimation can work well. Table 1 gives the results of a simulation study, which is part of a larger study. In some cases the ML estimators were biased but generally they seemed satisfactory.

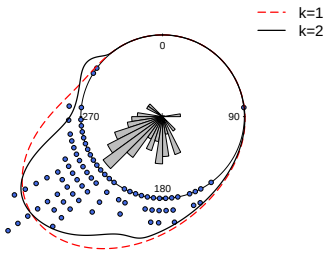


FIGURE 2. Rose diagram with densities of fitted WFGSN distributions, $k = 1, 2$.

4 Application to wind direction dataset

We now consider application of the WFGSN model to the wind direction dataset, to investigate the nature of the weather patterns. Figure 2 shows densities of fitted models for WFGSN distributions, $k = 1, 2$. For comparison, Figure 3 gives a fitted mixture of von Mises distributions (Mardia and Jupp, 2000; Jammalamadaka and SenGupta, 2001).

TABLE 1. Simulation study results.

		Parameters			
		ξ	ω	α	β
	True value	0.000	1.0000	10.0000	−5.0000
$n = 100$	Mean	−0.0055	1.0025	17.0381	−8.5205
	Bias	−0.0055	0.0025	7.0381	−3.5205
	SD	0.0247	0.0723	11.0265	5.1575
	RMSE	0.0253	0.0723	13.0813	6.2445
	CI Lower	−0.0104	0.9881	14.8502	−9.5438
	CI Upper	−0.0006	1.0168	19.2260	−7.4971
$n = 500$	Mean	−0.0022	0.9987	10.6971	−5.3579
	Bias	−0.0022	−0.0013	0.6971	−0.3579
	SD	0.0108	0.0313	1.6519	0.9740
	RMSE	0.0110	0.0313	1.7930	1.0377
	CI Lower	−0.0044	0.9925	10.3694	−5.5512
	CI Upper	−0.0001	1.0049	11.0249	−5.1647

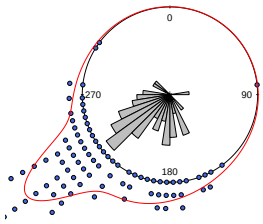


FIGURE 3. Rose diagram with density of fitted mixture of von Mises distributions.

5 Discussion

We have seen that the WFGSN can produce fits which are flexible. Our initial modelling involved using finite mixture models, and we had some issues with possibly poor fits, but the WFGSN approach seemed to provide an attractive alternative model which suited the data well. The simulation study illustrated that the WFGSN model seemed to have some nice features with regard to the estimation of the parameters.

Acknowledgments: We are particularly grateful to the School of Geosciences at the University of Edinburgh for making the wind direction dataset available.

References

Hernández-Sánchez, E. and Scarpa, B. (2012). A wrapped flexible generalized skew-normal model for a bimodal circular distribution of wind directions. *Chilean Journal of Statistics*, **3**, 129 – 141.

Jammalamadaka, S.R. and SenGupta, A. (2001). *Topics in Circular Statistics*. Singapore: World Scientific.

Mardia, K.V. and Jupp, P.E. (2000). *Directional Statistics*. New York: Wiley.

Quantifying LD decay by quantile regression – a case study

Sabine K. Schnabel[1], Federico Torretta[2], Matthias Westhues[3]

[1] Biometris, Wageningen University and Research Centre, The Netherlands
[2] Università di Palermo, Italy
[3] Universität Hohenheim, Germany

E-mail for correspondence: sabine.schnabel@wur.nl

Abstract: Through recent developments in genotyping (e.g. of plant populations) more information on genetic markers become available. In order to perform powerful genome-wide association studies (GWAS) it is important to analyse linkage disequilibrium (LD) through pairwise comparisons of genetic markers. Large numbers of these markers pose new problems in terms of analysis and visualization. In a case study for Maize we explore and quantify LD decay using monotone quantile regression.

Keywords: Quantile regression; Smoothing; Monotonicity; Linkage; LD decay.

1 Introduction and motivation

Genome-wide association studies have emerged as a great tool for the localization of QTLs (quantitative trait loci) in plant and animal breeding programs. However, a crucial requirement to powerful GWAS is the investigation of the genetic relatedness (kinship matrix) (Astle and Balding, 2009). For an appropriate kinship matrix, insight into LD between genetic markers is necessary. This matrix is best based on a set of independent markers (Listgarten et al., 2012). To find such a set of suitable markers (e.g. single nucleotide polymorphism – SNP) we need to explore LD decay over the whole genome. LD is commonly measured in terms of the squared Pearson correlation coefficient R^2 between pairs of genetic markers (Hill and Robertson, 1968). As an example, we are using data from chromosome 1 of a Maize population consisting of 123 European Dent inbred lines and 114 European Flint inbred lines (Fischer et al., 2008). Figure 1(A) shows an LD decay plot for this part of the genome.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

2 Analysis and application

Chromosome 1 of the above described Maize genome has almost 5000 markers, resulting in more than 12 million pairwise comparisons between two markers on the same chromosome. Visualization of these large data sets is challenging. We used a scatterplot smoother (Eilers and Goeman, 2004) for depicting global LD decay on chromosome 1 (length $\approx$ 300 Mbp (Mega base pairs)) in Figure 1(A). As mentioned above, the most common measure for LD decay is the squared Pearson correlation R^2. In order to improve the quality of the fit as well as the visualization, we advocate to use $\sqrt{|R|}$ instead. Further investigation concerning this transformation will be reported elsewhere. In our case, it is of interest to examine what happens to LD decay on a smaller scale. Therefore we investigate local LD decay in subsequent overlapping sliding windows of 2.5 Mbp width and fit a set of quantile curves to each of these sections of the whole plot. We use non-parametric quantile regression with a monotonicity constraint, $\mu_\tau = s_\tau(d)$, where μ_τ is the quantile function at percentile τ, d is the SNP distance between pairs of markers and $s_\tau(\cdot)$ is a smooth and unknown function (Muggeo et al., 2013; Bollaerts et al., 2006). We choose $P-$splines for a smooth functional form. By imposing $b_k < b_{k-1}$, with b_k as the coefficient of the kth spline, a monotone decreasing curve is ensured (that is in line with the underlying biological assumptions). This analysis can be done for any quantile of interest. In Figure 1(B) quantile curves for $\tau = 0.25, 0.5, 0.75$ are plotted as well as a threshold in terms of LD decay (on the initial scale of R^2=0.1, therefore here at $\sqrt[4]{0.1} = 0.56$). For the exploration of local LD decay we might be interested in the distance $\Delta_{0.5}$ from which onwards the LD decay is falling beneath the threshold. This can be interpreted as the average distance within this window from which onwards two marker loci are considered to be independent of each other. Such two markers could, for example, be selected for the computation of a kinship matrix. In the example, local LD decay in terms of median distance at threshold $\sqrt[4]{0.1}$ is 249666 bp. On Chromosome 1 of this population we have about 1000 sliding windows of 2.5 Mbp width with an average of 2000 points falling into one window. The distances Δ_τ for all sliding windows are collected and plotted in Figure 1(C) at the respective center of the window. We collect these data for τ values that are of interest. While these data points are an interesting result as such to quantify local LD decay, it is hard to judge by eye the relationship that is plotted in Figure 1(C). Therefore we used $P-$splines to fit a smooth curve to these results as displayed by the curve in Figure 1(D). For our example, we observe a bi-modal form of the relationship. The black vertical line in the graph indicates the so-called centromere. We have reason to believe that the left mode is around the centromere. However, this bi-modal phenomenon could not be observed for all of the 10 chromosomes in this data set.

3 Conclusion and discussion

This is a case study of how to explore and quantify local LD decay patterns in Maize. We are using quantile regression with monotonicity constraints for a first summary of the LD decay. On top of that we are applying $P-$splines to smooth the median local LD decay. These curves are easier to interpret and to inspect for the collaborating biologists.
While the presented steps are a good tool to quantify local LD decay, they have also been instrumental in identifying problems with the underlying genotypic data that have previously been overlooked. In this sense they can serve as a diagnostic tool. On the one hand we discovered sliding windows with low sample sizes which suggests undercoverage in certain distances in LD decay. While fitting the smooth curves in Figure 1 (D), we observed a noticeable clustering in terms of correlation values in some of the subsets of the data. This phenomenon was unknown to date in this data set and has lead to adjustments in subsequent data analyses.
More results from this case study will be reported elsewhere.

Acknowledgments: This case study was performed while the second and third author were visiting at Biometris at Wageningen University and Research Centre in winter 2014/2015. We are indebted to the group of Prof. Dr. Ruedi Fries, from Technische Universität München, for the SNP genotyping of the parental lines, which was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr SynbreedSynergistic plant and animal breeding (FKZ:0315528d).

References

Astle, W. and Balding, D.J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, **24**, 451 – 471.

Bollaerts, K., Eilers, P.H.C., and Aerts, M. (2006). Quantile regression with monotonicity restrictions using P-splines and the L1-norm. *Statistical Modelling*, **6**, 189 – 207.

Eilers, P.H.C. and Goeman, J.J. (2004). Enhancing scatterplots with smoothed density. *Bioinformatics*, **20**, 623 – 628.

Fischer, S., Möhring, J., Schön, C.C., Piepho, H.-P., Klein, D., Schipprack, W., Utz, H.F., Melchinger, A.E., and Reif, J.C. (2008). Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim. *Plant Breeding*, **127**, 446 – 451.

Hill, W.G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38**, 226 – 231.

Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature Methods*, **9**, 525 – 526.

Muggeo, V. , 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.

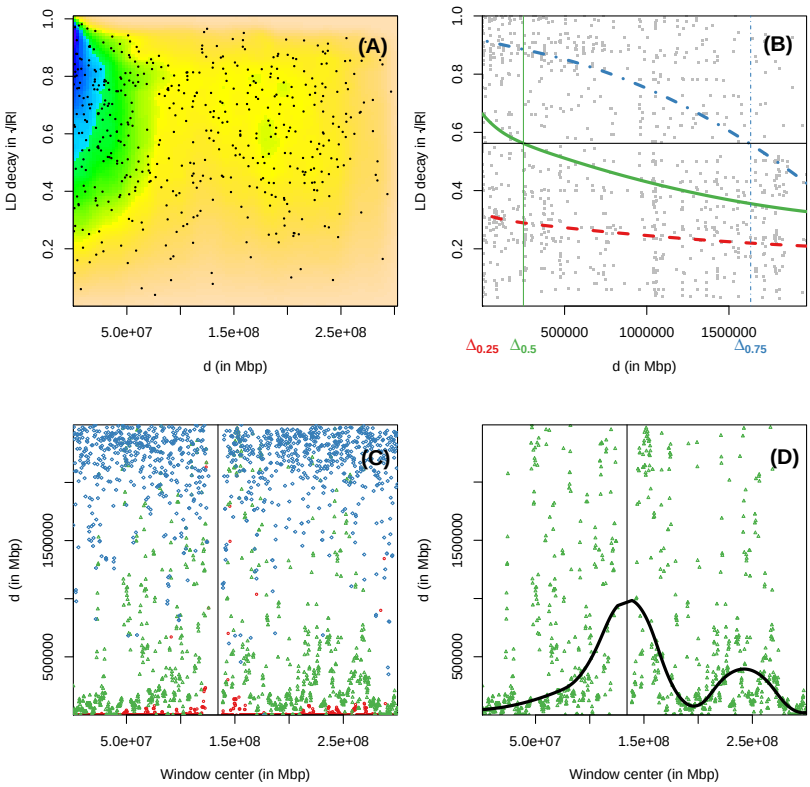


FIGURE 1. (A) Plot of LD decay for Chromosome 1 using `scattersmooth`. (B) Sample of data with threshold R^2=0.1($\sqrt{|R|} = 0.56$) and $\Delta_{0.25,0.5,0.75}$ shown. (C) Collection of $\Delta_{0.25,0.5,0.75}$ in Red/Green/Blue for Chromosome 1 at the center of the sliding window with indication of the centromere. (D) Smooth fit to $\Delta_{0.5}$.

Spatial variation of drivers of agricultural abandonment with spatially boosted models

Max Schneider[1,2], Gilles Blanchard[1], Christian Levers[2], Tobias Kümmerle[2]

[1] Institute of Mathematics, University of Potsdam, Germany
[2] Institute of Geography, Humboldt University of Berlin, Germany

E-mail for correspondence: maxsch03@uni-potsdam.edu

Abstract: Agricultural abandonment (AA) is a significant land use process in the European Union (EU) and modeling its driving factors has great scientific and policy interest. Past studies of drivers of AA in Europe have been limited by their restricted geographic regions and their use of traditional statistical methods which fail to consider the spatial variation in both predictors and AA itself. In this study, we implement a modeling framework based on boosted classification with spatially-varying terms, choosing the squared loss function and P-splines for the base learner, for their preferred statistical properties. By comparing models containing both constant and spatially-varying coefficients, we observe telling spatial trends in the relationship between the drivers and AA.

Keywords: Boosted regression; Spatial modeling; Agricultural abandonment.

1 Motivation and background

The abandonment of agricultural land is a key process of land use and management over the recent history of the European Union. Here, the extent of agricultural land has been declining for numerous reasons, e.g., increasing yields on productive lands, conservation policies or urban pull factors (Cramer et al., 2008). This shift in land use has both negative and positive effects on, e.g., local biodiversity, occurrence of fire and cultural changes in rural areas (Benayas et al., 2007; Baumann et al., 2011). There is thus pressing scientific and policy interest in better understanding of the process of agricultural abandonment (AA), which stems from improved models of its leading drivers.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

A number of studies have previously modeled AA; however, two key limitations remain. Available studies focus on spatially limited regions in Europe, counterintuitive to the many EU-wide regulations and programs related to AA (Baumann et al., 2011; Gellrich et al., 2007). In addition, the majority use basic forms of logistic regression to model the binary response variable (agriculture abandoned or not?) with a set of regressors driving the response. This traditional approach ignores the spatial trend that inevitably exists both within AA and the corresponding explanatory variables.
In this study, we identify drivers of agricultural abandonment in a wall-to-wall fashion for the first 27 member states of the EU (the EU27). Not only do we cover a broader spatial region (and thus, a far larger dataset) than previous work, but we also explicitly modelize spatial characteristics of the hypothesized drivers and AA itself. For this aim, we build a series of models using a modified form of boosted classification as proposed by Hothorn et al. (2011). This results not only in variable selection for the top drivers of AA but also model selection between a set of candidate models which isolate the importance of the spatial variation of these drivers. Comparing across models, we find noticeable differences in the behavior of top drivers over the spatial domain.

2 Spatial boosting as a classification framework

In classical boosted classification (a form of boosted regression), a set of p regressor variables $x = (x_1, \ldots, x_p)$ is related to a binary response variable, y_i $(i = 1, \ldots, n)$, $y_i \in \{0, 1\}$, using a regression technique referred to as a base learner $h(x, \hat{\theta})$. The parameters for the base learner, $\hat{\theta}$, are chosen such that they minimize a loss function, $L(y, f)$, where f is the model predicting y. We compute the negative gradient of this loss function, evaluated with f as the best-fit regression model given the empirical dataset. This is a vector of values and is denoted $\hat{F}$. The base learner is then used to construct a model to fit x onto the negative gradient vector, minimizing the model's residual sum of squares. This process is repeated iteratively over subsequent loss negative gradients and in each step, a weighted version of the parameter values are added to the previous iteration's parameter values. Algorithm 1 describes the steps for the specific procedure we implement.
The loss function used in our case is the squared loss $L(y, f) = (y - f)^2/2$, which means that its negative gradient is simply the residual, $U_i = y_i - \hat{F}_m(x)$. This method is called L2Boost and has been proven to improve the mean squared error (MSE) over a standard linear learner, the regression tool commonly found in the literature (Bühlmann and Yu, 2003). It was further shown that smoothing splines will produce a minimax-optimal MSE when taken as the base learners in L2Boost; we thus use a penalized version (called the P-spline) as the base learner. A constant smoothing parameter is used for all regressors and is drawn from the literature (Hothorn et al., 2011).

Another critical parameter in the boosting procedure is the stopping iteration m_{stop}, as allowing for too many iterations results not only in overfitting and also increased computational cost. Theoretical study of early stopping iterations has discovered that an m_{stop} obtained using cross validation on the data will converge in probability to the minimax-optimal iteration (Caponnetto and Yao, 2010). We thus choose a cross-validated m_{stop} to terminate model fitting.

Algorithm,

Step 1. Given data $(y_i, x_{ij}), i = 1, \ldots, n, j = 1, \ldots, p$, fit an (initial) P-spline $\hat{F}_0(x) = h(x; \hat{\theta}_{Y,x})$. Set $m = 0$.

Step 2. Calculate residuals $U_i = y_i - \hat{F}_m(x)$. Fit a P-spline to the current iteration's residuals. The current iteration's fit is denoted $\hat{f}_{m+1}(\bullet)$. Then, update $\hat{F}_{m+1}(\bullet) = \hat{F}_m(\bullet) + \hat{w}_{m+1}\hat{f}_{m+1}(\bullet)$, where $\hat{w}_{m+1}$ is the weight given to that $(m+1)$-th iteration's fit.

Step 3 (iteration). Iteration index $m = m+1$ and repeat step 2. Continue until reaching the stopping iteration, m_{stop}.

In order to understand the effect of the regressors' spatial variation on AA, we build several classification models, each consisting of various model components. Following earlier uses of boosting in georegression (Kneib et al., 2009), we apply different forms of P-splines. One set of model components consist of the P-spline base-learner applied to each individual regressor, $f_{const}(x_j) = h_j(x_j)$; these do not consider any spatial information and thus have constant coefficients over space. Another set of terms in the model is $f_{vary}(x_j, s) = \beta(s)x_j$, or the regressor with a spatially varying coefficient $\beta(s)$, which we generate using a bivariate tensor product P-spline, and where s is the location latitude-longitude. The final model component is f_s, a tensor product P-spline over the AA values for the entire region. Using the R package mboost, we construct six boosted models: three models only using one of the above three model components, two models with two of the model components and finally, the full model with all three components. These models are compared by their empirical risk, among other performance metrics; better scores of the models containing $f_{vary}(x_j, s)$ over models containing $f_{const}(x_j)$ indicates that the spatial variation of top predictors is pertinent in process understanding of AA.

3 Data and results

A massive dataset was collected for both AA and its potential drivers for the EU27, containing 2.4 million pixels of agricultural land on a resolution of 1 km^2. The AA data stems from twelve single year classifications (from 2001 to 2012) of fallow and active agricultural land, derived from satellite images. Data for 15 climate, landscape and socioeconomic variables hypothesized to influence AA were obtained from government and academic data sources. A stratified random sample was drawn from the data to speed computation of

analysis of the models; this was stratified to retain the proportion of pixels in the five identified landscape zones of Europe (Mücher et al., 2011). Evaluation of the six models showed that models with spatially-varying terms performed better than those with constant terms. The best model based on both performance metrics and model parsimony was that with only a $f_{vary}(x_j, s)$ component and no $f_s(s)$ or $f_{const}(x_j)$ terms. This model, stopped at a cross-validated iteration, consisted of six drivers: Land Under Agricultural Use (in hectares); High Nature Value Farmland (in percent of pixel); Elevation (in meters); Aridity Index (a ratio of precipitation to potential evapotranspiration); Distance from Nearest Forest (in kilometers); and Population-Density-Weighted GDP. Marked spatial trends across the EU27 in the relationship between the drivers and AA were also observed.

References

Baumann, M., Kuemmerle, T., Elbakidze, M., Ozdogan, M., Radeloff, V., Keuler, N., Prishchepov, A., Kruhlov, I., and Hostert, P. (2011). Patterns and drivers of post-socialist farmland abandonment in Western Ukraine. *Land Use Policy*, **28**, 552 – 562.

Benayas, J., Martins, A., Nicolau, J., and Schulz, J. (2007). Abandonment of agricultural land: an overview of drivers and consequences. *CAB reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, **57**, 1 – 14.

Bühlmann, P. and Yu, B. (2003). Boosting with the L2 Loss: Regression and Classification. *Journal of the American Statistical Association*, **98**, 324 – 338.

Caponnetto, A. and Yao, Y. (2010). Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, **8**, 161 – 183.

Cramer, V., Hobbs, R., and Standish, R. (2008). What's new about old fields? Land abandonment and ecosystem assembly. *Trends in Ecology & Evolution*, **23**, 104 – 112.

Gellrich, M., Baur, P., Koch, B., and Zimmermann, N. (2007). Agricultural land abandonment and natural forest re-growth in the Swiss mountains: a spatially explicit economic analysis. *Agriculture, Ecosystems & Environment*, **118**, 93 – 108.

Hothorn, T., Müller, J., Schröder, B., Kneib, T., and Brandl, R. (2011). Decomposing environmental, spatial, and spatiotemporal components of species distributions. *Ecological Monographs*, **81**, 329 – 347.

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

Mücher, C., Klijn, J., Wascher, D., and Schaminée, J. (2011). A new European landscape classification (LANMAP): A transparent, flexible and user-oriented methodology to distinguish landscapes. *Ecological Indicators*, **10**, 87 – 103.

Self-modeling ordinal model with time invariant covariates – an application to prostate cancer

Aliakbar Mastani Shirazi[1], Kalyan Das[2], Aluísio Pinheiro[1]

[1] Department of Statistics, IMECC, University of Campinas, Brazil
[2] Department of Statistics, University of Calcutta, India

E-mail for correspondence: apinheiro.unicamp@gmail.com

Abstract: The severity of genito-urinary toxicity is assessed for prostate cancer patients who are given different doses of radiation. The ordinal responses, patient's severity of side effects, are recorded longitudinally along with the cancer stage, and the differences among the patients due to time-invariant covariates. To build up a suitable framework for the data analysis, a self-modeling ordinal longitudinal model is employed where the conditional cumulative probabilities for a category of an outcome has a relation with shape-invariant model on the framework of a linear mixed model. The population time curve is modeled with a penalized regression spline. A simulation study is performed via Newton-Raphson (NR) and EM algorithms. The method is applied to a real data set.

Keywords: Longitudinal data; Self-Modeling; Nonparametric; Penalized spline.

1 Introduction

Often, the response of interest is measured in a series of ordered categories. We consider the model proposed by McCullagh (1980). The estimation of the response curve shape is done under the self-modeling approach (Lawton et al., 1972). Each individual's curve is some simple transformation of the common shape curve, leading to the so-called shape invariant (SI) model, a special case of the self-modeling regression method (Altman 2004),

$$Y_{ij} = \alpha_{0i} + \exp(\alpha_{1i})\mu_0(\beta_{0i} + \exp(\beta_{1i})t_{ij}) + \epsilon_{ij}, \tag{1}$$

where Y_{ij} is the observed response on subject i at time t_{ij}. Here α_{0i}, α_{1i}, β_{0i} and β_{1i} are unknown parameters which may be functions of observed covariates, ϵ_{ij} is an unobserved error which may be correlated within subject,

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

and μ_0 is a shape function which is common to all subjects. Inferences based on the SI model are very similar to the use of parametric nonlinear mixed models, in the sense that parameters embody all the information about covariate effects. Inference for the parameters conditional on the fitted shape is straightforward, as the conditional model is an ordinary nonlinear mixed effect model. We consider the SI model defined for the conditional cumulative probabilities for a category of an outcome. Therefore, in our model there is no direct relation between the observed response and parameters. A smooth unknown regression function is estimated by assuming a functional parametric shape constructed via a high dimensional basis function. The dimension of the basis is chosen to achieve the desired flexibility, while the basis coefficients are penalized to ensure smoothness of the resulting functional estimates (Ruppert et al, 2003). These often perform well if the number of the intra-individual measurements is not small and the variability of random effects is not large, but when some of the individuals have sparse data or the variability of the random effects is large there are considerable errors in approximating the likelihood. Exact methods such as Monte Carlo EM algorithm have been used, in which the E step is approximated using simulated samples from the exact conditional distribution of the random effects given the observed data (Wei and Tanner, 1990).
On the application we focus on the question of whether the dose level of radiation affects the severity of genito-urinary (bladder) toxicity as a side effect. In particular, we investigate the interaction between the dose effect and follow-up time. A simulation study not shown due to space constraints was performed. The result provide further evidence of the proposed method's advantages.

2 The model

The conditional cumulative probabilities for the L categories of the outcome y_{ij} as $P_{ijl} = \Pr(y_{ij} \leq l | \cdot \upsilon_i, x_{ij}, z_{ij}) = \sum_{k=1}^{l} p_{ijk}$, where p_{ijk} represents the conditional probability of response in category k. The logistic GLMM for the conditional cumulative probabilities is given by $\log[P_{ijl}/(1 - P_{ijl})] = \eta_{ijl}$, $(l = 1, \ldots, L-1)$, where the $\eta_{ijl} = \tau_l - \omega_{ij}$ and $\omega_{ij} = x_{ij}^{\mathrm{T}}\boldsymbol{\beta} + z_{ij}^{\mathrm{T}}\upsilon_i$.
Given the $L-1$ strictly increasing model thresholds τ_l (i.e., $\tau_1 < \cdots < \tau_{L-1}$). x_{ij} is the $(p+1) \times 1$ covariate vector (including the intercept), and z_{ij} is design vector for the r random effects, both vectors being for the jth timepoint nested within subject i. Also, $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of unknown fixed regression parameters. Let $\upsilon = T\theta$, where $TT^{\mathrm{T}} = \Sigma_\upsilon$ is the Cholesky factorization of random-effect variance covariance matrix Σ_υ.
The self-modeling regression (SEMOR) Model is expressed as

$$y_{ij} = \pi_i\{\mu_0[\kappa_i(t_{ij})]\} + e_{ij},$$

where y_{ij} is the response for curve i, $i = 1, \ldots, N$, measured at n_i times, t_{ij}. $\pi_i(x)$ is a monotone inverse link transforming the regression function

and $\kappa_i(x)$ is a monotone transformation of the time axis. μ_0 is a shape function that is common to all the curves, and e_{ij} are errors. We focus on nonparametric modeling of μ_0 and parametric modeling of $\pi_i(x)$ and $\kappa_i(x)$ with known correlation structure for e_{ij}. We give special attention to shape invariant model and apply the SI model (SIM), so $\omega_{ij} = \alpha_{0i} + \exp(\alpha_{1i})\mu_0(t^*_{ij})$, where $t^*_{ij} = \beta_{0i} + \exp(\beta_{1i})t_{ij}$. Therefore, we have $\eta_{ijl} = \log[P_{ijl}/(1 - P_{ijl})] = \tau_l - [\alpha_{0i} + \exp(\alpha_{1i})\mu_0(t^*_{ij})]$.

If one has physical or theoretical justification to pre-specify $\mu_0(t^*_{ij})$ parametrically, this is just a special case of nonlinear regression. The semiparametric SEMOR model allows flexible modeling by estimating $\mu_0(t^*_{ij})$ non-parametrically. We consider $\theta_i = (\alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i})^{\mathrm{T}}$, so that $\theta_{\mathbf{i}} = \mathbf{X}_i\phi + \mathbf{Z}_i\psi_{\mathbf{i}} + \varepsilon_{\mathbf{i}}$, where $\mathbf{Z}_i$ is the design (or covariate) matrix for the random effect vector ψ_i. If μ_0 is a known parametric function, and if we assume that ψ_i, ε_i and $\mathbf{e}_{ij}$ are normally distributed, we have a parametric nonlinear mixed model and the model can be fitted using maximum likelihood.

We fit μ_0 using the penalized spline model. The use of the spline method with penalty chosen by generalized maximum likelihood is equivalent to fitting the model $\mu_0(t^*_{ij}) = \mathbf{U}\boldsymbol{\gamma} + \mathbf{V}\boldsymbol{\zeta}$, where $\mu_0(t^*_{ij})$ is the vector of means at the transformed times, $\mathbf{U}$ is a design matrix for a cubic polynomial in t^*_{ij}, $\mathbf{V}$ is a design matrix for cubics in t^*_{ij} which are left-truncated at the knot, $\boldsymbol{\gamma}$ is a vector of unknown parameters and $\boldsymbol{\zeta}$ is normally distributed with zero mean and covariance matrix $\boldsymbol{\Sigma}_\zeta$.

3 Likelihood estimation

The conditional log likelihood for the observed responses can be written as

$$\ell(\mathbf{Y}_i|\boldsymbol{\theta}) = \prod_{j=1}^{n_i}\prod_{l=1}^{L}(P_{ijl} - P_{ijl-})^{I_{y_{ij}}(l)},$$

where $\boldsymbol{\theta} = (\boldsymbol{\phi}^{\mathrm{T}}, \sigma^2_\psi, \sigma^2_\varepsilon, \boldsymbol{\tau})^{\mathrm{T}}$ and $I_{y_{ij}}(l) = 1$ if $y_{ij} = l$ and zero otherwise. Then the marginal likelihood function is

$$L(\boldsymbol{\theta}, \mathbf{Y}) = \prod_{i=1}^{N}\int\prod_{j=1}^{n_i}\prod_{l=1}^{L}(P_{ijl} - P_{ijl-1})^{I_{y_{ij}}(l)} f(\boldsymbol{\theta}_i|\boldsymbol{\psi}_i) f(\boldsymbol{\psi}_i|\sigma^2_{\psi_i}) d\boldsymbol{\theta}_i d\boldsymbol{\psi}_i.$$

Since $\boldsymbol{\psi}_i$ and $\boldsymbol{\varepsilon}_i$ have $N_q(\mathbf{0}, \boldsymbol{\Sigma}(\sigma^2_\psi))$ and $N_4(\mathbf{0}, \sigma^2_\varepsilon I_4)$, respectively, we write the joint likelihood of conditional cumulative probabilities of ordered outcomes and the random effects $\boldsymbol{\theta}_i$ and $\boldsymbol{\psi}_i$ as

$$\begin{aligned} l_{ic} &= \sum_{j=1}^{n_i}\sum_{l=1}^{L} I_{y_{ij}}(l)\log(P_{ijl} - P_{ijl-1}) + \log\left(\boldsymbol{\Phi}(\boldsymbol{\theta}_i|\mathbf{X}_i\boldsymbol{\phi} + \mathbf{Z}_i, \boldsymbol{\psi}_i, \sigma^2_\varepsilon I_4)\right) \\ &\quad + \log\left(\boldsymbol{\Phi}(\boldsymbol{\psi}_i|\mathbf{0}, \sigma^2_{\psi_i})\right). \end{aligned}$$

Our proposal's first two steps predict μ_0 and step 3 provides estimates for the parameters for the cummulative logit model through MCNR and MCEM.

4 Application to prostate cancer data

The study aims to assess the effects of the different doses of radiation on the patients. $N = 243$ patients of different ages (A) in two Chicago hospitals ($H = 0, 1$) were treated with randomly prescribed dose levels of radiation ($D = 1, 2, 3$ for weak, medium, strong). Three stages of prostate cancer ($S = 1, 2, 3$ for minor, medium, severe) are recorded and the patients were followed up over 6 years where the measurement time points differ among the patients. Physicians assessed the severity of genito-urinary (bladder) toxicity, which is a side effect of radiation therapy ($gu = 0$ for no symptoms, $gu = 1$ for pain/local bleeding but no required intervention, $gu = 2$ for bleeding lesion and minor intervention, and $gu = 3$ for serious lesion requiring hospitalization). The analysis is based on our proposed model with the severity of toxicity as a response variable.
The deviances for the models D, D and A, D, A and H, or D, A, H and S are given by 2210.1, 2203.0, 2195.8 and 2195.6, respectively. Table 1 shows the estimates of intercepts and coefficients for the latter model. The model can be written as $\log[P_{ij1}/(1-P_{ij1})] = 3.9847 - [\alpha_{0i} + \exp(\alpha_{1i}) + \mu_0(t^*_{ij})]$, $\log[P_{ij2}/(1-P_{ij2})] = 5.7957 - [\alpha_{0i} + \exp(\alpha_{1i}) + \mu_0(t^*_{ij})]$ and $\log[P_{ij3}/(1-P_{ij3})] = 7.0340 - [\alpha_{0i} + \exp(\alpha_{1i}) + \mu_0(t^*_{ij})]$, where $\boldsymbol{\theta}_i = (\alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i})^{\mathrm{T}} = -0.1682 D_i - 0.0221 A_i - 0.3421 H_i + 0.0312 S_i$.

TABLE 1. Estimates - Model *Dose*, *Age*, *Hosp*, and *Stage*.

Par.	Est. (Std.Error)	p-value	Var.	Est. (Std. Error)	p-value
τ_1	3.9847 (0.6714)	0.0000	*Dose*	−0.1682 (0.0847)	0.0520
τ_2	5.7957 (0.6820)	0.0000	*Age*	−0.0221 (0.0085)	0.0065
τ_3	7.0340 (0.7147)	0.0000	*Hosp*	−0.3421 (0.1468)	0.0120
			Stage	0.0312 (0.1147)	0.6814

References

Altman, S. and Villarreal J. (2004). Self-modelling regression for longitudinal data with time-invariant covariates. *Canadian Journal of Statistics*, **32**, 251 – 268.

Lawton, W.H., Sylvestre, E.A., and Maggio, M.S. (1972). Self modeling non-linear regression, *Technometrics*, **14**, 513–532.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Societa, Series B*, **42**, 109 – 142.

Statistical analysis of spatial distribution in populations of microspecies of Alchemilla L.

Georgy Yu. Sofronov[1], Nikolay V. Glotov[2], Olga V. Zhukova[2]

[1] Department of Statistics, Macquarie University, Australia
[2] Department of Biology, Mari State University, Russia

E-mail for correspondence: georgy.sofronov@mq.edu.au

Abstract: In this paper, we consider *Alchemilla vulgaris* L. (or common lady's mantle), which is an herbaceous perennial plant. It is known that within this species it is possible to distinguish microspecies, that is, fairly homogeneous groups having minor morphological differences. We study spatial distributions of the microspecies found in various localities as well as possible interaction between different microspecies.

Keywords: Spatial analysis; Join-count statistics; Plant populations; Microspecies; *Alchemilla vulgaris* L.

1 Introduction

Study of agamic complexes in such genera as *Alchemilla*, *Crepis*, *Citrus*, *Hieracium*, *Poa*, *Potentilla*, *Rubus*, *Taraxacum* etc. is of particular interest in plant biology; see Grant (1981). The complexes can be defined as groups of angiosperm (or flowering) plants that are characterised by apomictic reproduction (or apomixes), which is a form of seed reproduction without fertilization. In this paper, we study *Alchemilla* plants growing within Eastern Europe, which is considered as agamo-sexual complex *Alchemilla vulgaris* L.s.l. (Glazunova (1977)). A number of agamospecies are described within the agamo-sexual complex of *Alchemilla* plants. Such rather homogeneous groups with minor morphological differences are called microspecies. From a taxonomic point of view, it is quite difficult to classify microspecies of agamo-sexual complex *Alchemilla vulgaris*. The taxonomy is usually possible for plants in generative period, which are collected in June, during the first flowering. The microspecies may vary in nature and degree of fluffiness of radical leaves, flowers, generative stems as well as in size and form of

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

lamina, lobes and their teeth. There are about a thousand of microspecies in the genus (e.g., see Fröhner, 1995), in Europe there are more than 300 microspecies, 38 are in Central Russia (Tikhomirov et al., 1995), and 31 are in the Republic of Mari El (the study region); see Abramov (2008).

2 Data and methodology

The data were collected from 28 localities (habitats), which are characterised by different environmental conditions: bottomland meadow, dry meadow, fallow land, edge of mixed and coniferous forest. The plants were considered on square sites with the area of 1 m^2. The surveyed area in the localities varies from 2 to 169 m^2. *Alchemilla* plants of generative period were excavated and put in a herbarium. The diagnostics of microspecies was carried out by a set of qualitative morphological traits of the plants preserved in the herbarium. The number of microspecies within the same locality varies from 1 to 14, while 28 different microspecies were identified. In this paper, we consider locality M1, which has the largest number of sites. Within this locality, we study spatial distribution of each microspecies and co-occurrence of different microspecies.

3 Results and discussion

In each of N sites (quadrants) we record whether the particular microspecies has or has not been observed. The i-th site can be coded as either $x_i = 1$ (B, black) or $x_i = 0$ (W, white). As a result, we have a mosaic map of black and white sites (see Figure 1). Following the chess terminology (e.g., see Upton and Fingleton, 1985), we consider two widely used definitions of contiguity on a lattice: rook's (touching edges) and queen's (either touching edges and touching corners). In order to determine whether neighbouring sites are more likely to be the same colour or different colours, we can count the numbers of BB, BW or WW joins (where, for example, BB denotes a join between two black sites) and compare these numbers with the corresponding expected numbers of joins under the null hypothesis of no spatial autocorrelation among the sites.

Let $\mathbf{W} = \{w_{ij}\}$ be a spatial proximity matrix of size $N \times N$ (where N is the total number of sites) in which $w_{ij} = 1$ if i-th and j-th sites are joined, and $w_{ij} = 0$ otherwise. Then the join-count statistics are given by

$$BB = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}x_i x_j, \qquad BW = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - x_j)^2,$$

$$WW = \text{the total number of joins} - (BB + BW).$$

The expected values and the variances of the join-count statistics under non-free sampling (or sampling without replacement) are given in Cliff and Ord (1981).

Using the `joincount.multi` function in R package `spdep` Bivand (2014), we can find the values of the join-count statistics. Table 1 shows the values of the join-count statistics and the z-values for a particular microspecies. In this example, the observed value of BB differs significantly from its expected value implying clustering of B sites in the locality (positive autocorrelation); see Figure 1. The analysis shows that all microspecies identified in locality M1 are spatially autocorrelated but the measures (based on the normalised join-count statistics) of the spatial autocorrelation considerably vary for different microspecies.

TABLE 1. The observed and expected values of the join-count statistics for microspecies *A. tubulosa* Juz. in locality M1.

	Rook			Queen		
	Observed	Expected	z-value	Observed	Expected	z-value
BB	8	2.9890	3.1982	10	5.7481	1.9325
BW	49	56.7912	−2.2225	107	109.2139	−0.3521
WW	255	252.2198	1.1653	483	485.0380	−0.3644

In order to test whether there exists an interaction between two microspecies, we use a modified chi-squared test of independence in 2×2 contingency tables; see Cerioli (1997). The results of the analysis suggest that there is significant interaction between several pairs of microspecies.

References

Abramov, N.V. (2008). *Flora of the Republic of Mari El: handbook*. Yoshkar-Ola: Mari State University Press.

Bivand, R. (2014). spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.5-77.

Cerioli, A. (1997). Modified tests of independence in 2×2 tables with spatial data. *Biometrics*, **53**, 619 – 628.

Cliff, A.D. and Ord, J.K. (1981). *Spatial processes: models & applications*. London: Pion.

Fröhner, S. (1995). Alchemilla L. In: *Illustrierte Flora von Mitteleuropa, Bd. 4*, Berlin-Wien: Blackwell Wissenschafts-Verlag, Hegi, G. (Ed.), 13 – 242.

Glazunova, K.P. (1977). On a possibility of applying the theory of the agamo-sexual complex to systematic of angiosperms (on example of the genus Alchemilla L.). *Byull. Mosk. Obshch. Ispyt. Prir., Otd. Biol.*, **82**, 129 – 139.

Grant, V. (1981). *Plant Speciation.* (2nd ed.). New York: Columbia University Press.

Tikhomirov, V.N., Notov, A.A., Petukhova, L.V., and Glazunova, K.P. (1995). Genus Alchemilla. In: *Biological flora of the Moscow Region. Volume 10*, Moscow, Pavlov, V.N. and Tikhomirov, V.N. (Eds.), pp. 83 – 118.

Upton, G. and Fingleton, B. (1985). *Spatial Data Analysis by Example. Volume 1: Point Pattern and Quantitative Data.* John Wiley & Sons Ltd.

101	102	103	104	105	106	107	108	109	110	201	211	221
111	112	113	114	115	116	117	118	119	120	202	212	222
121	122	123	124	125	126	127	128	129	130	203	213	223
131	132	133	134	135	136	137	138	139	140	204	214	224
141	142	143	144	145	146	147	148	149	150	205	215	225
151	152	153	154	155	156	157	158	159	160	206	216	226
161	162	163	164	165	166	167	168	169	170	207	217	227
171	172	173	174	175	176	177	178	179	180	208	218	228
181	182	183	184	185	186	187	188	189	190	209	219	229
191	192	193	194	195	196	197	198	199	200	210	220	230
231	232	233	234	235	236	237	238	239	240	261	262	263
241	242	243	244	245	246	247	248	249	250	264	265	266
251	252	253	254	255	256	257	258	259	260	267	268	269

FIGURE 1. Layout of sites in locality M1. The 169 sites are numbered from 101 to 269. If a site has microspecies *A. tubulosa* Juz. then it is colour coded black, otherwise it is white.

Use of genetic relationship matrices in the prediction of breeding values and their accuracy assessment

Katia Stefanova[1], Wallace Cowling[1], Arthur Gilmour [2]

[1] The UWA Institute of Agriculture, University of Western Australia, Australia
[2] Cargo Vale, Cargo, Australia

E-mail for correspondence: katia.stefanova@uwa.edu.au

Abstract: The paper presents a statistical model to obtain the predicted breeding values with accuracy of selection in the context of plant breeding trials in two cycles of selection by using genetic relationship matrices.

Keywords: Linear mixed model; Predicted breeding values; BLUPs; Multi-environment trial; Accuracy of selection.

1 Introduction

Analyses of single trials and multi-environment trials (METs) assume IID normal genetic effects, although for METs genetic variances and correlations may differ between trials. The parental covariances arising from common ancestors are ignored. In reality the test lines in single trials or METs are genetically related and using this information can improve the accuracy of prediction of breeding values. Relatedness may be determined from the pedigree or from genetic markers. This allows the individual genetic effects to be partitioned into *additive* and *non-additive* genetic effects (Oakey et al., 2007). The additive effect reflects the ability of a line as a parent and the non-additive effect is associated with dominance and residual effect.
The paper presents a statistical model incorporating genetic relationship matrix for the analysis of two cycles of selection in a plant breeding trial. Based on the model, predicted breeding values (PBV) and the accuracy of selection are obtained. The approach is illustrated with data from a self-pollinating crop (*Pisum Sativum*) for resistance to ascochyta blight (*Didymella pinodes*) complex (Cowling et al., 2015) A new breeding approach, utilizing F1-recurrent selection and the animal breeding model, is

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

used to accelerate response to selection in this self-pollinating crop, including phenotypic and relationship records from self progeny. Using PBV, narrow-sense heritability and the accuracy of selection allows comparison of breeding strategies.

2 Motivating example

Traditionally in a plant breeding trial, selfed progenies of F1s are intercrossed, not the F1s themselves. The innovation in this study is that using relationships allows evaluation and hence selection of the F1 plants through their progeny. Each self is identified as a genotype with explicitly stated identical parents inside the pedigree tree.

2.1 Phenotypic data

A total of 1139 and 1077 early generation progeny plants were tested in cycles 1 and 2, respectively. Plots were single plants spaced $1m$ apart in rows and columns, with 40 columns by 30 rows in cycle 1 grown in 2010, and 20 columns by 70 rows in cycle 2 grown in 2013. We consider the two cycles as two experiments (trials). Replication and concurrency of the genotypes across experiments were provided by pure lines including founder lines, Australian cultivars and Chinese landraces.

2.2 Experimental design

Trials were laid out as rectangular arrays of rows and columns using a partially replicated (*p-rep*) design (Cullis et al., 2006). The design was generated using package DiGGer (Coombes, 2009).

3 Statistical model

Denote the vector of plot yields $\mathbf{y} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_t^{\mathrm{T}})^{\mathrm{T}}$, t is the number of trials. The model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_a + \mathbf{Z}_g\mathbf{u}_{\bar{a}} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e},$$

where $\mathbf{u}_g = \mathbf{u}_a + \mathbf{u}_{\bar{a}}$ and $\mathbf{u}_g, \mathbf{u}_a, \mathbf{u}_{\bar{a}}$ are respectively the vectors of genetic, additive genetic and non-additive genetic effects, $\boldsymbol{\tau}$ is the vector of fixed effects with design matrix $\mathbf{X}$, $\mathbf{u}_p$ is the vector of random non-genetic effects with design matrix $\mathbf{Z}_p$ and $\mathbf{e}$ is the vector of plot errors combined across trials.

The random effects are assumed to follow Gaussian distribution

$$\begin{pmatrix} \mathbf{u}_a \\ \mathbf{u}_{\bar{a}} \\ \mathbf{u}_p \\ \mathbf{e} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{G}_a & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\bar{a}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R} \end{pmatrix}\right]$$

and

$$\text{var}(\mathbf{y}) = \mathbf{Z}_g\mathbf{G}_a\mathbf{Z}_g^{\mathrm{T}} + \mathbf{Z}_g\mathbf{G}_{\bar{a}}\mathbf{Z}_g^{\mathrm{T}} + \mathbf{Z}_p\mathbf{G}_p\mathbf{Z}_p^{\mathrm{T}} + \mathbf{R}.$$

The variance matrix for non-genetic random effects $\mathbf{G}_p$ is usually a diagonal matrix of scaled identity matrices. The variance matrix for the plot error effects is assumed block diagonal with $\mathbf{R} = \text{diag}(\mathbf{R}_j)$, where $\mathbf{R}_j$ is the plot error matrix for the jth trial. The data are spatially modelled and respectively ordered as rows within columns, reflecting the field layout $\mathbf{R}_j = \sigma_j^2\mathbf{\Sigma}_{c_j} \otimes \mathbf{\Sigma}_{r_j}$, where σ_j^2 is a scale parameter and Σ_{c_j} and Σ_{r_j} are the correlation matrices for column and row dimensions of the trial j, corresponding to autoregressive processes of order one. The variance matrix for the additive genetic effects $\mathbf{G_a}$ is assumed to have the separable form $\mathbf{G}_a = \mathbf{G}_{e_a} \otimes \mathbf{G}_{v_a}$, where matrix $\mathbf{G}_{e_a}$ is the matrix of additive genetic variances/covariances between environments, matrix $\mathbf{G}_{v_a}$ is the matrix of additive genetic variances/covariances between genotypes. We set $\mathbf{G}_{v_a} = \mathbf{A}$ where the known numerator relationship matrix $\mathbf{A} = \{a_{ij}\}$ is given by $a_{ii} = 1 + F_i$ and $a_{ij} = 2f_{ij}$. F_i is the inbreeding coefficient of genotype i and f_{ij} is the coefficient of parentage between genotypes i and j. In a similar manner we assume that the non-additive effects may be presented as a two-way structure of genotype by environments effects, respectively with variance $\mathbf{G}_{\bar{a}} = \mathbf{G}_{e_{\bar{a}}} \otimes \mathbf{G}_{v_{\bar{a}}}$. We assume independence between the non-additive genetic components and set $\mathbf{G}_{v_{\bar{a}}} = \mathbf{I}_m$. The matrix $\mathbf{A}^{-1}$ and the models are computed using ASReml-R (Butler et al., 2009).

4 PBV, accuracy and heritability

PBV is the obtained best linear unbiased predictor (BLUP) for each genotype. More precisely, it is the empirical BLUP (E-BLUP), since the unknown variance parameters have been replaced by their REML estimates in the mixed model equation.

The accuracy r of PBVs is the correlation between the true and predicted breeding values, and is sometimes reported as *reliability*, the squared correlation (r^2) (Mrode, 2005). The accuracy r of the predicted breeding value for the ith genotype at an individual environment was calculated following Gilmour et al. (2009) for the animal model

$$r_i = \sqrt{1 - \frac{s_i^2}{(1 + f_i)\sigma_a^2}},$$

where s_i^2 is the prediction error variance for the ith genotype, σ_a^2 is the additive genetic variance and $1 + f_i$ is the diagonal element of the relationship matrix $\mathbf{A}$. Narrow-sense heritability h^2 was calculated as $h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_{\bar{a}}^2 + \sigma_e^2}$.

5 Results

The fitted model identified significant linear row and column effects for both trials, the local variation ($AR1 \times AR1$) was not that strong. The variance/covariance structure fitted for the genotype by environment interaction effect was general correlation with heterogeneous variances, equivalent to the unstructured in the case of two environments. The latter produced separate estimates of the trials variances and the genetic correlation between the trials which appeared quite high, 0.82. The inclusion of the pedigree in the model showed significantly ($p < 0.001$) better fit. The average accuracy of predicted breeding values in the combined analysis without pedigrees was 0.562 (genotypes in cycle 1) and 0.696 (genotypes in cycle 2), and increased to 0.894 (cycle 1) and 0.807 (cycle 2) when pedigree information was included.

Acknowledgments: The financial support of the Grains Research and Development Corporation of Australia is gratefully acknowledged.

References

Butler, D.G., Cullis, B.R., Gilmour, A.R., and Gogel, B.J. (2009). ASReml-R reference manual. Technical Report, *Queensland Department of Primary Industries* URL http://www.vsni.co.uk/software/asreml/.

Coombes, N. (2009). *DiGGer design search tool in R*. http://www.austatgen.org/software.

Cowling, W.A., Stefanova, K.T., Beeck, C.P., Nelson, M.N., Hargreaves, B.L.W., Sass, O., Gilmour, A.R., and Siddique, K.H.M. (2015). Using the animal model to accelerate response to selection in a selfpollinating crop. *G3: Genes, Genomes, Genetics* (accepted).

Cullis, B.R., Smith, A.B., and Coombes, N.E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 381 – 393.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., and Thompson, R. (2009). *ASReml User Guide Release 3.0.* VSN International Ltd, UK. http://vsni.de/downloads/asreml/release3/UserGuide.pdf

Mrode, R.A. (2005). *Linear Models for the Prediction of Animal Breeding Values*, (2nd ed.). CABI Publishing, UK.

Oakey, H., Verbyla, A.P., Cullis, B.R., Wei, X., and Pitchford, W.S. (2007). Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theoretical and Applied Genetics*, **114**, 1319 – 1332.

A modified binomial likelihood model for zero and n-inflated count data

James Sweeney[1]

[1] School of Business, University College Dublin, Ireland

E-mail for correspondence: james.sweeney@ucd.ie

Abstract: A statistical inconsistency of a zero-inflated binomial likelihood model for count data is identified. This issue occurs when the response, y, is both zero and n-inflated, and results in statistically inconsistent and erroneous parameter inferences being drawn from the data. The zero-modified binomial likelihood is amended to address this issue of *n-inflation*, resulting in a fully symmetric binomial likelihood model for both zero and n-inflated counts. We present a simple regression example from the ecological literature which details the practical application of the new likelihood model.

Keywords: Zero-inflation; n-inflation; Binomial distribution.

1 The zero and n-modified binomial distribution

A frequent feature of data sets involving a count response is their tendency to contain many zeroes. Data sets which contain a higher proportion of zeroes over that expected by the standard statistical families can be modelled through the use of *zero-modified* distributions. However, the use of such distributions, in the context of data which are subject to sum constraints, will result in erroneous and inconsistent parameter inferences. We propose an extension to the existing zero-inflated binomial likelihood model of Hall (2000) which simultaneously addresses possible *n-inflation* in the observed data, specifically,

$$y_i \sim \begin{cases} 0 & \text{with probability } \dfrac{(1-\pi_{1i})\pi_{2i}}{\pi_{1i}+\pi_{2i}-\pi_{1i}\pi_{2i}} \\ n_i & \text{with probability } \dfrac{\pi_{1i}(1-\pi_{2i})}{\pi_{1i}+\pi_{2i}-\pi_{1i}\pi_{2i}} \\ \text{Binomial}(n_i, p_i) & \text{with probability } \dfrac{\pi_{1i}\pi_{2i}}{\pi_{1i}+\pi_{2i}-\pi_{1i}\pi_{2i}} \end{cases} \quad (1)$$

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

Here $(1-\pi_{1i})$ and $(1-\pi_{2i})$ refer to the probability of zero-inflation for y_i and $z_i = n_i - y_i$ respectively. We explicitly assume the constraint that the n_i must be non-zero - this implies a probability correction factor $\pi_{1i}+\pi_{2i}-\pi_{1i}\pi_{2i}$. If there is no n-inflation present in the data the model collapses to that of Hall (2000).

2 Application

We utilise an example from the ecological literature; the data set consists of 61 pollen counts, obtained from lake sediment (Huntley et al., 1993), which we separate into the categories of either warmer or cooler climate-preferring types. A measure of local climate, $GDD5$ (Growing degree days above $5C$), is available for each site. Let $\mathbf{Y} = \{y_1, \ldots, y_{61}\}$, represent the pollen counts of the "cooler" type and $\mathbf{Z} = \{z_1, \ldots, z_{61}\}$ the counts of the "warmer" type. The sum constraint at each location is naturally the sum of the respective pollen types, i.e. $\mathbf{N} = \{n_1, \ldots, n_{61}\} = \mathbf{Y}+\mathbf{Z}$. We assume a simple model for the cooler pollen-$GDD5$ interaction, namely that the proportion of pollen (y_i/n_i) observed at a given site is a logistic-linear function of the $GDD5$ measurement (c_i) at that location, $\text{logit}(p_i) = \beta_0 + \beta_1 c_i = \boldsymbol{\beta}\mathbf{C}_i$.
In the following let $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2\}$

$$\begin{aligned} y_i &\sim \text{zero/n-inflated Binomial}(n_i, p_i, \boldsymbol{\alpha}) \\ \text{logit}(p_i) &= \boldsymbol{\beta}\mathbf{C}_i \\ \pi_{1i} &= (p_i)^{\alpha_1} \\ \pi_{2i} &= (1-p_i)^{\alpha_2} \end{aligned} \tag{2}$$

The log-likelihood for the model in (2) is, up to a constant,

$$\begin{aligned} & L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{Y}, \mathbf{N}) \\ \propto & \sum_{y_i=0} \log\left(\frac{1}{(1+e^{\boldsymbol{\beta}\mathbf{C}_i})^{\alpha_2}} - \frac{e^{\alpha_1\boldsymbol{\beta}\mathbf{C}_i}}{(1+e^{\boldsymbol{\beta}\mathbf{C}_i})^{\alpha_1+\alpha_2}} + \frac{e^{\alpha_1\boldsymbol{\beta}\mathbf{C}_i}}{(1+e^{\boldsymbol{\beta}\mathbf{C}_i})^{\alpha_1+\alpha_2+n_i}}\right) \\ + & \sum_{y_i=n_i} \log\left(\frac{e^{\alpha_1\boldsymbol{\beta}\mathbf{C}_i}}{(1+e^{\boldsymbol{\beta}\mathbf{C}_i})^{\alpha_1}} - \frac{e^{\alpha_1\boldsymbol{\beta}\mathbf{C}_i}}{(1+e^{\boldsymbol{\beta}\mathbf{C}_i})^{\alpha_1+\alpha_2}} + \frac{e^{(\alpha_1+n_i)\boldsymbol{\beta}\mathbf{C}_i}}{(1+e^{\boldsymbol{\beta}\mathbf{C}_i})^{\alpha_1+\alpha_2+n_i}}\right) \\ + & \sum_{y_i\neq 0, n_i} (\alpha_1+y_i)\boldsymbol{\beta}\mathbf{C}_i - (\alpha_1+\alpha_2+n_i)\log\left(1+e^{\boldsymbol{\beta}\mathbf{C}_i}\right) \\ - & \sum_{y_i} \log\left(\frac{e^{\alpha_1\boldsymbol{\beta}\mathbf{C}_i}}{(1+e^{\boldsymbol{\beta}\mathbf{C}_i})^{\alpha_1}} + \frac{1}{(1+e^{\boldsymbol{\beta}\mathbf{C}_i})^{\alpha_2}} - \frac{e^{\alpha_1\boldsymbol{\beta}\mathbf{C}_i}}{(1+e^{\boldsymbol{\beta}\mathbf{C}_i})^{\alpha_1+\alpha_2}}\right) \end{aligned} \tag{3}$$

As the log-likelihood of the model in (3) is simple to maximise due to the small number of model parameters $(\alpha_1, \alpha_2, \beta_0, \beta_1)$, we proceed to do so

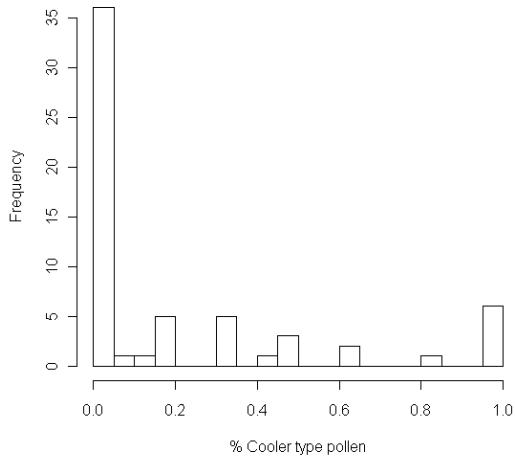


FIGURE 1. Histogram of the "cooler" pollen counts (scaled by sum totals). We note that the counts exhibit clear signs of both zero and n-inflation as evidenced by the excess of observations at 0 and 1.

using the Newton Raphson method. The maximum likelihood estimates are presented in Table 1, as well as the estimates produced for the zero-modified model of Hall (2000) where both $\mathbf{Y}$ and $\mathbf{Z}$ are separately modelled as the response.

TABLE 1. Maximum likelihood estimates for model parameters $\pm$ standard errors obtained from the inverse observed information matrix.

Model	zero & n-inflated	zero-inflated (cooler)	zero-inflated (warmer)
$\log(\alpha_1)$	0.113 ± 0.179	0.701 ± 0.186	–
$\log(\alpha_2)$	-0.701 ± 0.474	–	-0.544 ± 0.453
β_0	105.45 ± 0.07	79.67 ± 0.055	210.99 ± 0.091
β_1	-0.015 ± 0.00002	-0.011 ± 0.000016	-0.030 ± 0.000019
AIC	245.82	764.57	737.54

3 Results

The AIC for the zero & n-inflated model is substantially lower than for the competing zero-inflated model. Notably, the AIC for a binomial model fit is 3876, indicating that the incorporation of at least one zero-inflation aspect to the model does improve fit immeasurably. As the high $GDD5$

site locations should favour the warmer pollen types, the zeroes observed are more likely to be errant zeroes - this is reflected in the AIC for the zero-inflated fit of this model being superior to that of the cooler pollen equivalent. It is apparent that the n_i's of the **Z** counts are clearly as a result of zero-inflation in the **Z** counts but this feature is not captured by the zero-inflated model for the cooler pollen counts. Based on the output of the various model fits presented in table, both the cooler and warmer pollen types exhibit significant degrees of zero inflation.

4 Summary

In this article we have presented a modified binomial likelihood model which addresses zero-inflation of both response variables simultaneously, resulting in consistent parameter inferences regardless of response variable choice. The superiority, in terms of consistency of inference and model fit, is clearly displayed in the ecological regression example presented herein.

References

Hall, D.B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56**, 1030 – 1039.

Huntley, B. (1993). The use of climate response surfaces to reconstruct paleoclimate from quartenary pollen and plant macrofossil data. *Philosophical Transactions of the Royal Statistical Society*, **341**, 215 – 244.

Conditional weight gain using an external reference

Fraser Tough[1], Charlotte M. Wright[1], John H. McColl[1]

[1] University of Glasgow, Scotland

E-mail for correspondence: f.tough.1@research.gla.ac.uk

Abstract: Conditional weight gain is a statistical term used to describe how a change in weight between two time points in one child, differs from that which would be expected given their initial weight. However, when comparing children that derive from different backgrounds, the assumptions for conditional weight gain break down as their distributions differ. This paper provides an adapted version of conditional weight gain, allowing children from different datasets to be compared directly relative to a common reference.

Keywords: Z scores; Fractional polynomials; Conditional weight gain.

1 Introduction

The concept of Z scores is essential to understanding conditional weight gain. A Z score provides an estimate of where measurement X lies relative to its population mean, μ_X, in terms standard deviations (SD) σ_X. Let

$$Z = \frac{X - \mu_X}{\sigma_X}. \tag{1}$$

Since $Z \sim N(\mu_Z, \sigma_Z^2)$, Z can be converted to a centile, providing an estimate of the proportion of the population below or above X.
If the estimates of the mean and SD are from the same population from which X derives from then we can assume $Z \sim N(0, 1)$. If the mean and SD do not derive from the same population as X, then $\mu_Z \neq 0$ and $\sigma_Z \neq 1$. Conditional weight gain uses these properties of Z scores to provide an estimate of how one child's change in weight compares with the rest of the child's population, given their initial weight. If an external reference is used (a reference distribution other than the child's parent population in

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

which the mean and standard deviation is calculated), μ_Z and σ_Z need to be estimated so that the correct centile can be acquired.
To compute conditional weight gain, we firstly obtain Z scores for X_1 and X_2. We then regress Z_2 on Z_1. The resulting linear regression model allows us to determine what we expect Z_2 should be based on Z_1. We then create a Z score for $X = Z_2 - E(Z_2|Z_1) = Z_2 - (\alpha + \beta Z_1)$

$$Z = \frac{X - \mu_X}{\sigma_X} = \frac{Z_2 - \mu_{Z_2} - r\frac{\sigma_{Z_1}}{\sigma_{Z_2}}(Z_1 - \mu_{Z_1})}{\sigma_{Z_2}\sqrt{(1 - r^2)}}$$

The reason for creating a Z score for the difference between Z_2 and its predicted value based on Z_1 is to take regression to the mean into the account. Regression to the mean is a statistical phenomenon that states that if groups of individuals are weighed once then later again, their weight centile will tend to be on average closer to the population mean. Conditional weight gain accounts for this expected movement. When an internal reference is used (i.e. the child's parent population), then $\mu_{Z_1} = 0, \mu_{Z_2} = 0, \sigma_{Z_1} = 1$ and $\sigma_{Z_2} = 1$ and the formula reduces to

$$Z = \frac{Z_2 - rZ_1}{\sqrt{(1 - r^2)}}.$$

If an external reference is used then we cannot assume $Z \sim N(0, 1)$ and need to estimate $\mu_{Z_1}, \mu_{Z_2}, \sigma_{Z_1}$ and σ_{Z_2}.
In both cases, the correlation coefficient, r, must be calculated.

2 Data and aims

Three developing world datasets were made available in this study; South African, Malawian and Pakistani. The aim is to make the datasets comparable relative to a common reference. Only one dataset is summarised within this paper, the South African Vertical Transmission Study (VTS). The VTS was conducted between 2001 and 2006. Anthropometry measurements were taken at 6, 10, 14, 18, 22, and 26 weeks then at months 7, 8, 9, 12, 15, 18, 21 and 24 from 2938 children.

3 Modelling correlation

Correlation coefficients, r, can be calculated between groups of measurements at the scheduled dates available. However to calculate correlations between intermediate values, the correlation surfaces were modelled so that values could be interpolated at any point. Generalised Additive Models (GAM), Fractional Polynomials (FP) (Royston and Altman, 1994) and the Argyle model (Argyle, 2002) were all fitted and assessed using the Akaike

Information Criterion (AIC) and mean square error (MSE). The Argyle model models correlation surfaces on the log scale. When applying GAM and FP models, the Fisher transformation was used. The reason for using the Fisher transformation is because as the true population parameter $|\rho|$ gets closer to 1, the variance of r decreases. The Fisher transformation is an approximate variance stabilising transformation, meaning the variance of the transformed data is approximately constant for all values of the population correlation coefficient. As the models are on two different scales, AIC's cannot be compared directly. However, by multiplying the likelihood L by the Jacobian J the two are comparable. For $f(y) = \log(y)$, $J = \prod \frac{1}{y_i}$ and for $f(y) = \phi(y) = \frac{1}{2}(\log(1+r)/(1-r))$, $J = \prod \frac{1}{(1-r^2)}$ where AIC $= 2k - 2(\log(L) + \log(J))$. Overall, fractional polynomial models fitted best in terms of AIC and MSE, and were adopted to model the correlation surfaces. The covariates used were the difference, Δt, and means (μ_t) between t_1 and t_2.

Fractional polynomials allow more flexibility than normal polynomials by allowing more flexible power selection. A backward elimination algorithm is run to determine which powers to use. The VTS correlation model can be seen below which can be used to interpolate values from any point on the correlational surface.

$$\begin{aligned} E(\phi(r_{t_i}, r_{t_j})) &= \beta_0 + \beta_1 \left(\frac{\Delta t}{10}\right)^{1/2} + \beta_2 \left(\frac{\Delta t \mu_t}{100}\right)^2 \\ &\quad + \beta_3 \left(\frac{\Delta t \mu_t}{100}\right) \log\left(\frac{\Delta t \mu_t}{100}\right) + \beta_4 \left(\frac{\mu_t}{10}\right)^{1/2} + \beta_5 \left(\frac{\mu_t}{10}\right). \end{aligned} \tag{2}$$

4 Using an external reference

The World Health Organisation (WHO) reference is an international standard which uses composite longitudinal growth measurements from children from a number of different backgrounds. These backgrounds include both developing and developed world countries from all over the world. The standard describes how children should grow, not how they do grow, and should therefore be used as a reference to determine where children sit relative to an ideal set of children. Since each dataset is structurally different, their mean values and standard deviations will differ, relative to the WHO reference. By modelling the means and SD's for the VTS infants, values can be interpolated from any point in time to obtain μ_{Z_1}, μ_{Z_2}, σ_{Z_1} and σ_{Z_2}.

GAM's for Location, Scale and Shape (GAMLSS), a semi-parametric modelling procedure which allows users to model complex statistical distributions, were used to model the means and standard deviations (Rigby and Stasinopoulos, 2001). An example of which can be seen in Figure 1.

The plots shown in Figure 1 show mean differences and ratios of SD's between the WHO reference and the VTS study for males. It can be seen

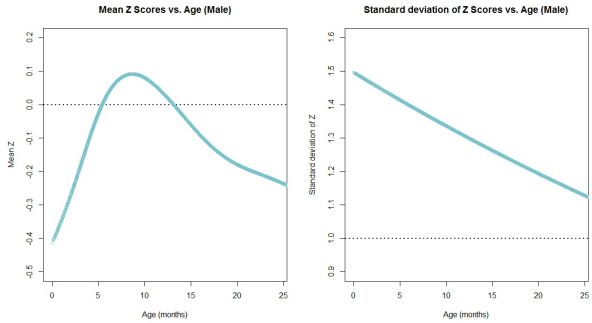


FIGURE 1. Mean & standard deviation of Z scores vs. Age (months) (WHO reference).

that the average Z score sits around -0.4 at age 0, increases above the WHO standard at 6 months and dips back down through 0 at 15 months.
By interpolating from the two plots, $\mu_{Z_1}, \mu_{Z_2}, \sigma_{Z_1}$ and σ_{Z_2} can be estimated at ages t_1 and t_2 and the correlation can be estimated using equation (2). Then, by using equation (1), the conditional weight gain can be calculated.

5 Summary

The adapted methodology provides a framework in which practitioners can compare children's weight gain Z scores from different backgrounds directly with one another, using a common reference. The reference can easily be changed and the resulting means and standard deviations of the Z scores can be modelled with GAMLSS.

Acknowledgments: The authors thank Dr. Ruth Bland for permission to use the VTS data within this paper.

References

Argyle, J. (2002). *Statistical analysis of child growth data.* Retrieved from Durham E-Theses Online: http//etheses.dur.ac.uk/4113/.

Rigby, R.A. and Stasinopoulos, D.M. (2001). The GAMLSS project: a flexible approach to statistical modelling. *New trends in statistical modelling: Proceedings of the 16th IWSM*, 249 – 256.

Royston, P. and Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling *Journal of the Royal Statistical Society, Series C*, **43**, 429 – 467.

Prior elicitation for estimation in a mixed effect logistic model

Diego Tovar-Rios[1], Rafael Tovar-Cuevas[1], Mercedes Andrade-Bejarano[1]

[1] Universidad del Valle, Colombia

E-mail for correspondence: diego.tovar@correounivalle.edu.co

Abstract: In this work we proposed a procedure to elicit prior distributions for the parameters of a generalized linear mixed model, in situations where there not exist specialist on the matter or the data analyst does not have information from historical sources. Our proposed method is supported by heuristics and Bayes empirical methods.

Keywords: Elicitation procedures; Empirical Bayes methods; MCMC.

1 Introduction

The generalized linear mixed models (GLMM) combine the generalized linear models with random effects with normal distribution within the lineal predictor, which can be applied to a wide variety of problems and are particularly useful in longitudinal studies. However, their flexibility is limited by the need to carry out multidimensional numerical integrations, which are computationally complex (Fong et al., 2010).
The hierarchical model formulation where the outcome is modeled conditionally on random effects, which are then modeled in an additional step, makes the Bayesian approximation an attractive methodology but this requires the specification of prior distributions for the model's parameters. In those cases where information about the problem is available, Chen et al. (2003) propose informative prior distributions based on historical data; while Fong et al. (2010) show that it is possible to specify normal prior distributions where the parameters can be obtained by specifying two quantiles of the dependent variable with their probabilities.
However, when no historical data, information or a specialist in the topic of interest is available, it is possible to obtain information for the model's

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

parameters directly from the data (Carlin and Louis, 2000). Another option is to assume that there is sufficient information in the data so that the parameters can be well estimated using a non-informative prior distributions. The main goal of this study was to consider a situation in which little information is available and it is not possible to identify an specialist on the topic. This proposal begins with an analysis of part of the data in order to obtain information that will make it possible to elicit prior distributions for the model's parameters.

2 Case study

We used a database with a total of 266 records, corresponding to individuals diagnosed with differentiated thyroid cancer (DTC), who underwent an ablation procedure. As part of the follow-up to the treatment, the levels of thyrotropin, a thyroid stimulating hormone (TSH), were assessed at three different times after the surgery (3, 9 and 18 mo). According to the management protocols of the American Thyroid Association (2012), the TSH levels in a controlled patient should no be greater than 1 mIU/mL. It was observed that about 90% of the individuals had controlled TSH levels as of the ninth month (Table 1).

TABLE 1. Frequencies of the TSH measures.

TSH	Month 3		Month 9		Month 18	
> 1	83	31.2%	24	9.0%	23	8.7%
≤ 1	183	68.8%	242	91.0%	243	91.4%

3 Statistical modelling

Let a GLMM be for binary data with a logistic link, where θ_i is the probability that the TSH measurement is less than or equal to 1 mIU/mL in the ith individual, then

$$\text{logit}(\theta_i) = \beta_0 + \beta_1 * age + \beta_2 * size + \beta_3 * time + b_i,$$

where β_j, $j = 0, 1, 2, 3$, are the regression coefficients, referred to as fixed effects; and b_i is the random effect, which represents the subject-specific random variation, so that $b_i \sim N(0, \sigma^2)$ (Molenberghs and Verbeke, 2005). In our literature review, we established three covariates related with the persistence/recurrence of DTC, to study their relationship with the TSH after the treatment. It was also possible to make groups of individuals in according with their age after we observed the behavior of the variable in

the sample ($age \leq 38$ years old). For the size of the tumor, we used the cut point reported in the literature ($size \leq 2cm$); and for the time observation, we used the variable in its continuous form ($time$).
In order to run the estimation procedure, it is necessary to specify prior distributions for the fixed effects β_j and the variance component σ^2. Starting initially with a simple model for a "typical" individual (i.e. one with $b_i = 0$) and taking the variable age, the model

$$\text{logit}(\theta_i) = \beta_0 + \beta_1 * age + b_i$$

was adjusted with a subsample ($n = 50$) from the total database; and the probabilities of success $\hat{\theta}_i$ associated with the typical individuals were estimated within the age range. According to Tovar-Cuevas (2012), it is possible to obtain the hyperparameters of a Beta(a, b) associated with the success probabilities, dividing the parametric space in k excluding intervals, so that the sum of its ranges is equal to the range of the complete parametric space and using Chebyshev's inequality, the values of the parameters a and b can be obtained. Once the distributions for the probabilities are established, it is possible to get the distributions of the fixed effects using the inverse transformation method and generate its values using MCMC.
The previous random subsample were resampled m times and simple models were adjusted for each variable; and given the symmetry of the empirical distribution of the m adjusted estimations, a goodness-of-fit normality test were carried out. In those cases where it is not possible to reject the hypothesis of normality, it was assumed that the parameter of interest could have a normal distribution whose hyperparameters μ and σ^2 correspond to the average and the variance of the m observations. We used $m = 1000$ to obtain our estimate values. For those cases where the hypothesis of normality was rejected, an asymmetric distribution was sought, such as the Gamma(a, b); and in order to obtain the hyperparameters, the values of the sample moments were matched with those of the distribution.

4 Results and conclusions

An initial model were adjusted using the adaptive Gaussian quadrature as the estimation method (Molenberghs and Verbeke, 2005). The results (see Table 2) were compared with the Bayesian estimations. For fixed effects, normal distributions with a large variances were used as non informative prior, in this case the posterior distributions had the highest standard error. For informative priors, the empirical distributions of the fixed effects for the variables age and tumor $size$ showed a symmetric behavior with adjustments made to normality. The variable $time$ showed a positive asymmetric behavior and a Gamma$(275.6, 0.001)$ distribution was used. We observed less variability in the estimates in this cases. The variables $size$

and *time* underwent important changes within the model. For the variance component, an Inv-Gamma distribution with large precision was used and it shows a high between-subject variability with a large standard error compared with the maximum likelihood estimation. Using our propose procedure, the posterior estimation is closer to maximum likelihood estimation but with smaller standard error.

TABLE 2. Results of the estimation process.

Parameter	β_0	β_1	β_2	β_3
	Approximation of the integral			
Est.	1.27	0.62	-0.24	0.02
St. Dev.	0.28	0.23	0.23	0.02
Interval	(0.72, 1.80)	(0.16, 1.07)	(-0.69, 0.20)	(-0.02, 0.05)
	Noninformative priors			
Est.	3.70	1.92	-0.78	0.03
St. Dev.	0.96	1.02	1.99	0.02
Interval	(1.78, 5.59)	(-0.05, 3.95)	(-2.64, 1.29)	(-0.02, 0.08)
	Informative priors			
Est.	1.05	0.04	0.04	0.14
St. Dev.	0.046	0.01	0.05	0.01
Interval	(0.97, 1.14)	(0.03, 0.05)	(-0.05, 0.14)	(0.10, 0.12)

References

American Thyroid Association (2012). El folleto de cáncer de tiroides. In: `http://www.thyroid.org/cancer-de-tiroides/`.

Carlin, B.P. and Louis, T.A. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC.

Chen, M.H., Ibrahim, J.G., Shao, Q.M., and Weiss, R.E. (2003). Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference*, **111**, 57 – 76.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, **11(3)**, 397 – 412.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics, Springer.

Tovar-Cuevas, J.R. (2012). Eliciting beta prior distributions for binomial sampling. *Rev. Bras. Biom*, **30**, 159 – 172.

A conceptional Lego toolbox for Bayesian distributional regression models

Nikolaus Umlauf[1], Reto Stauffer[1], Jakob W. Messner[1], Georg J. Mayr[2], Achim Zeileis[1]

[1] Department of Statistics, University of Innsbruck, Austria
[2] Institute of Meteorology and Geophysics, University of Innsbruck, Austria

E-mail for correspondence: Nikolaus.Umlauf@uibk.ac.at

Abstract: Bayesian analysis provides a convenient setting for the estimation of complex generalized additive regression models (GAM). Because of the very general structure of the additive predictor in GAMs, we propose an unified modeling architecture that can deal with a wide range of types of model terms and can benefit from different algorithms in order to estimate Bayesian distributional regression models.

Keywords: Additive models; gamlss; Distributional regression; MCMC.

1 Introduction

Bayesian estimation based on Markov chain Monte Carlo (MCMC) simulation is particularly attractive since it provides valid inference that does not rely on asymptotic properties and allows extensions such as variable selection or multilevel models. Existing estimation engines already provide infrastructures for a number of regression problems exceeding univariate responses, e.g., for multinomial, multivariate normal or mixed discrete-continuous distributed variables. In addition, most of the engines support random effect estimation that can be utilized for setting up complex models with additive predictors (see, e.g., Fahrmeir et al., 2013).
In order to ease the usage of already existing implementations and code, as well as to facilitate the development of new algorithms and extensions, we present an unified and entirely modular architecture for models with additive predictors which does not restrict to any type of regression problem. The approach follows the model class of generalized additive model

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

for location, scale and shape (gamlss, Rigby and Stasinopoulos, 2005) but is more flexible and is sometimes referred to as distributional regression.

2 Model structure

The models discussed assume conditional independence of the response variable $y_1, \ldots, y_n$ given covariates. Within distributional regression, all parameters of the response distribution can be modeled by explanatory variables such that

$$\mathbf{y} \sim \mathcal{D}\big(h_1(\boldsymbol{\theta}_1) = \boldsymbol{\eta}_1, \quad h_2(\boldsymbol{\theta}_2) = \boldsymbol{\eta}_2, \ldots, \quad h_K(\boldsymbol{\theta}_K) = \boldsymbol{\eta}_K\big), \tag{1}$$

where $\mathcal{D}$ denotes any distribution available for the response variable and $\boldsymbol{\theta}_k$ are parameters that are linked to an additive predictor using known monotonic link functions $h_k(\cdot)$. The k-th additive predictor is given by

$$\eta = f_1(\mathbf{x}) + \cdots + f_p(\mathbf{x}), \tag{2}$$

where $\mathbf{x}$ represents a generic vector of all linear and nonlinear modeled covariates. The functions f_j are possibly smooth functions encompassing various types of effects, e.g., linear and nonlinear effects of continuous covariates, two-dimensional surfaces, spatially correlated effects, varying coefficients, random intercepts and random slopes, etc. Using a basis function approach, the vector of function evaluations can be written in matrix notation $\mathbf{f}_j = \mathbf{X}_j\boldsymbol{\beta}_j$ and can also be represented as a mixed model with $\mathbf{f}_j = \tilde{\mathbf{X}}_j\tilde{\boldsymbol{\gamma}}_j + \mathbf{U}_j\tilde{\boldsymbol{\beta}}_j$, where $\tilde{\boldsymbol{\gamma}}_j$ represents the fixed effects parameters and $\tilde{\boldsymbol{\beta}}_j \sim N(\mathbf{0}, \tau_j^2\mathbf{I})$ independent and i.i.d. random effects (see, e.g., Fahrmeir et al., 2013).

3 A conceptional Lego toolbox

For Bayesian inference, prior distributions need to be assigned to the regression coefficients. A general setup is obtained by using normal priors for $\boldsymbol{\beta}_j$ of the form

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{K}_j\boldsymbol{\beta}_j\right), \tag{3}$$

where $\mathbf{K}_j$ is the so called penalty matrix that depends on the functional type chosen for f_j. The variance parameter τ_j^2 is equivalent to the inverse smoothing parameter in a frequentist approach and controls the trade off between flexibility and smoothness. A common choice of prior for the variance parameter is a weakly informative inverse Gamma hyperprior.

The main building block of all estimation engines is the logarithm of the posterior given by

$$\log p(\boldsymbol{\vartheta}|\mathbf{y}) = \ell(\boldsymbol{\vartheta}|\mathbf{y}) + \sum_{k=1}^{K}\sum_{j=1}^{p_k}\left\{\log p(\boldsymbol{\beta}_{jk}|\tau_{jk}^2) + \log p(\tau_{jk}^2)\right\}, \qquad (4)$$

with log-likelihood $\ell(\cdot)$ and priors $p(\cdot)$, e.g., given by (3), where $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\tau}_1^2, \ldots, \boldsymbol{\tau}_K^2)^{\mathrm{T}}$. From a frequentist perspective (4) can be viewed as a penalized log-likelihood.
Moreover, gradient based algorithms require the evaluation of the first derivative or score vector as well as the second derivatives, e.g., when applying a Newton-Raphson type algorithm, or MCMC sampling using IWLS proposals (see, e.g., Fahrmeir et al., 2013). Because these quantities can be nicely decomposed using the chain rule and model terms are represented by an unified approach, algorithms for distributional regression models can be build by combining the following "Lego-bricks":

- The log-likelihood function $\ell(\boldsymbol{\vartheta}|\mathbf{y})$.
- The first order derivatives $\partial\ell(\boldsymbol{\vartheta}|\mathbf{y})/\partial\boldsymbol{\theta}_k$, $\partial\boldsymbol{\theta}_k/\partial\boldsymbol{\eta}_k$ and $\partial\boldsymbol{\eta}_k/\partial\boldsymbol{\vartheta}_k$.
- Second order derivatives $\partial^2\ell(\boldsymbol{\vartheta}|\mathbf{y})/\partial\boldsymbol{\eta}_k\partial\boldsymbol{\eta}_k^{\mathrm{T}}$ (and expectations).
- Derivatives for priors, e.g., $\log p(\boldsymbol{\beta}_{jk}|\tau_{jk}^2)$ and $\log p(\tau_{jk}^2)$.

Hence, a modular system can in principle be used to implement various estimation algorithms (also using existing software). A simple generic algorithm for distributional regression models is outlined by the following pseudo code:

```
while(eps > ε & i < maxit) {
    for(k in 1:K) {
        for(j in 1:p) {
            Compute η_j^[k] = η^[k] − f_j^[k].
            Obtain new (β_j^[k], τ²_j^[k])^T = u_j^[k](y, η_j^[k], X_j^[k], β_j^[k], τ²_j^[k], family, k).
            Update η^[k].
        }
    }
    Compute new eps
}
```

The algorithm does not distinguish between optimization or sampling, because the functions $\mathtt{u}_\mathtt{j}^{[\mathtt{k}]}(\cdot)$ could either return proposals from a MCMC sampler or updates from an optimization algorithm. Moreover, it is possible to use different update functions for model terms within predictors, e.g., IWLS proposals combined with slice sampling or Hamiltonian Monte Carlo. An implementation of the modular infrastructure is provided in the R package **bamlss** (available at `https://R-forge.R-project.org` at the time of writing).

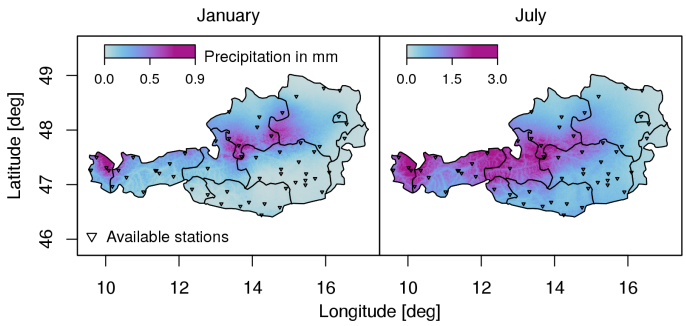


FIGURE 1. Predicted average precipitation for January 10th and July 10th. Animation available at `http://eeecon.uibk.ac.at/~umlauf/data/austria.gif`.

4 Example

As an illustration, we analyze precipitation data taken from the HOMSTART project conducted at the Zentralanstalt für Meteorologie und Geodynamik (ZAMG, see also Umlauf et al., 2012). The aim is to estimate a good climatology which can be used for subsequent meteorological models. Since precipitation data is skewed and exhibits high density at zero observations, we estimate a censored normal additive regression model with latent Gaussian variable $\mathbf{y}^\star$ and observed response $\mathbf{y}$, the square root of daily precipitation observations. The model is given by

$$\mathbf{y}^\star \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2), \quad \boldsymbol{\mu} = \boldsymbol{\eta}_\mu, \quad \log(\boldsymbol{\sigma}) = \boldsymbol{\eta}_\sigma, \quad \mathbf{y} = \max(\mathbf{0}, \mathbf{y}^\star).$$

For both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, we use the following additive predictor:

$$\boldsymbol{\eta} = \beta_0 + f_1(\texttt{day}, \texttt{lon}, \texttt{lat}) + f_2(\texttt{lon}, \texttt{lat}) + f_3(\texttt{day}) + f_4(\texttt{alt}),$$

where function f_1 is a spatially varying seasonal effect, f_2 a spatially correlated effect, f_3 the seasonal and f_4 the altitude effect. The resulting climatology for two particular days of the year are shown in Figure 1.

References

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression Models, Methods and Applications*. Berlin: Springer-Verlag.

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

Umlauf, N., Mayr, G., Messner, J., and Zeileis, A. (2012). Why does it always rain on me? A spatio-temporal analysis of precipitation in Austria. *Austrian Journal of Statistics*, **41**, 81 – 92.

Geostatistical analysis using K-splines in the geoadditive model

Yannick Vandendijck[1], Christel Faes[1], Niel Hens[12]

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

[2] Centre for Health Economic Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

E-mail for correspondence: yannick.vandendijck@uhasselt.be

Abstract: In geostatistics, both kriging and smoothing splines are commonly used to predict a quantity of interest. The geoadditive model proposed by Kammann and Wand (2003) represents a fusion of kriging and penalized spline additive models. The fact that the underlying spatial covariance structure is poorly estimated using geoadditive models is a drawback. We describe K-splines, an extension of geoadditive models such that estimation of the underlying spatial process parameters and predictions of the spatial map are performed with the same accuracy and precision as in kriging.

Keywords: Covariogram; Kriging; Mixed model; Penalized spline.

1 Introduction

The objective of geostatistics is to produce a map of a variable of interest on a specified domain based on observations which are measured with or without noise. Consider the geostatistical model $y(\mathbf{s}_i) = z(\mathbf{s}_i) + \varepsilon_i$, $i = 1, \ldots, n$, where the $y(\mathbf{s}_i)$ are observed data values from the underlying true values $z(\mathbf{s}_i)$. These data values are noise-corrupted by white-noise error terms ε_i. The spatial locations $\mathbf{s}_i$ belong to a specified continuous domain $D \subset \mathbb{R}^d$. The idea of geostatistics is to use the data $y(\mathbf{s}_i)$ to make predictions of $z(\mathbf{s}_0)$ where $\mathbf{s}_0 \in D$. Both kriging and spline methods can be used to handle geostatistical problems. In kriging, the values $z(\mathbf{s}_i)$ are assumed to be the realisations of an autocorrelated random process (Cressie, 1993). Smoothing splines assume that the $z(\mathbf{s}_i)$ are the values of a smooth non-parametric function (see e.g., Hutchinson and Gessler, 1994).

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

In kriging, linearity of the covariate effects is usually assumed. Kammann and Wand (2003) merged kriging and additive models to allow for non-linear relationships between covariates and the response variable in geostatistics. Their so-called geoadditive model consists of a penalized spline additive model with a geostatistical extension. The geoadditive model can be expressed as a linear mixed model which allows for estimation and inference using standard methodology. The drawback of their model is the biased estimation of the underlying spatial process.
Vandendijck et al. (2015) introduced the concept of kriging-splines, abbreviated by K-splines, which extends geoadditive models such that the estimation of the underlying spatial process and prediction of the map of interest is performed with similar accuracy and precision as in kriging. By showing a theoretical connection between kriging and K-splines, it is presented how the spatial covariance structure (covariogram) implied by K-splines is derived. K-splines are also embedded within the linear mixed model framework and the estimation uses a two-step likelihood procedure.

2 K-splines

For simplicity, suppose the data are $(y_i, \mathbf{s}_i, a_i, b_i)$, $1 \leq i \leq n$, where y_i is the value of the ith response, a_i and b_i are the values of two predictor variables a and b, and $\mathbf{s}_i$ represents the geographical location. Suppose the predictor a enters the model linearly and that the predictor b enters the model non-linearly. The geoadditive model is

$$y_i = \beta_0 + \beta_a a_i + f(b_i) + S(\mathbf{s}_i) + \varepsilon_i, \qquad 1 \leq i \leq n, \tag{1}$$

where f is a smooth function of b and S is the geographical component of the model. Both f and S are modelled using penalized spline functions. We use the thin plate spline family to construct f. The spatial component S is modelled through a set of radial basis functions and is of the form $S(\mathbf{s}) = \sum_{k=1}^{K_s} u_i^s g_\phi(\mathbf{s} - \boldsymbol{\kappa}_k^s)$, where g_ϕ can be any of the proper covariance or generalized covariance functions used in kriging. The subscript in g_ϕ is used to denote a possible dependence on a spatial decay parameter ϕ. An overview of the most important covariance functions g_ϕ that can be used are given in Table 1. The vector $(u_1^s, \ldots, u_{K_s}^s)$ contains the K_s unknown knot coefficients that are penalized to overcome overfitting. The K_s knots $\boldsymbol{\kappa}_1^s, \ldots, \boldsymbol{\kappa}_{K_s}^s$ are a representative subset of $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$ used for the construction of the basis functions.
Kammann and Wand (2003) propose to choose ϕ via the simple rule $\hat{\phi} = \max_{1 \leq i,j \leq n} \|\mathbf{s}_i - \mathbf{s}_j\|$. We propose K-splines, a new estimation approach for geoadditive models that allows for accurate estimation of the parameter ϕ. This enables one to estimate the underlying spatial process accurately and precisely. A two-stage iterative estimation approach is proposed to estimate

TABLE 1. Some important and often used covariance functions g_ϕ (where ϕ is positive for each function).

	$g_\phi(\mathbf{x}) =$
Exponential	$\exp\left(-\frac{\|\mathbf{x}\|}{\phi}\right)$
Gaussian	$\exp\left(-\frac{\|\mathbf{x}\|^2}{\phi^2}\right)$
Spherical	$\left(1-\frac{3}{2}\frac{\|\mathbf{x}\|}{\phi}+\frac{1}{2}\frac{\|\mathbf{x}\|^3}{\phi^3}\right)$
Matérn ($\nu=\frac{3}{2}$)	$\exp\left(-\frac{\|\mathbf{x}\|}{\phi}\right)\left(1+\frac{\|\mathbf{x}\|}{\phi}\right)$
Circular	$1-\frac{2}{\pi}\left(\vartheta\sqrt{1-\vartheta^2}+\arcsin\vartheta\right)$, with $\vartheta=\min\left(\frac{\|\mathbf{x}\|}{\phi},1\right)$

the parameters. At the first stage, the linear mixed model representation of (1) is estimated fixing ϕ in g_ϕ at its current value, and in the second stage the parameter ϕ is optimized fixing the linear mixed model parameters. For mode details on the estimation procedure and inference using K-splines, we refer to Vandendijck et al. (2015).

3 Simulation study

We consider as spatial domain the unit square. Data at a spatial location $\mathbf{s}=(s_x,s_y)$ on this square is simulated using the model

$$y_{\mathbf{s}} = S(\mathbf{s}) - 0.5x_{1\mathbf{s}} + \sin(2\pi x_{2\mathbf{s}}) + \varepsilon_{\mathbf{s}}, \tag{2}$$

where $\varepsilon_{\mathbf{s}} \sim \mathcal{N}(0,\sigma^2_\varepsilon = 0.10)$, $x_{1\mathbf{s}} \sim \text{Unif}[0-1]$, $x_{2\mathbf{s}} \sim \text{Unif}[0-1]$ and $S(\mathbf{s})$ is a zero-mean Gaussian Random Field (GRF) (Gelfand et al., 2010) with a Gaussian covariogram without nugget, namely $K(\mathbf{h}) = c_s \exp\left(-\frac{\|\mathbf{h}\|^2}{\tau^2}\right)$. We consider $c_s = 0.50$ and $\tau = 0.15$. We obtain 250 simulated realizations from (2). From each realization we draw a random sample of size $n = 500$. For each simulated dataset, the covariogram parameters (c_s,τ) and the measurement error parameter σ^2_ε were estimated using seven different methods: (1) Direct maximum likelihood (D-ML) parameter estimation for GRFs; (2) Direct restricted maximum likelihood (D-REML) parameter estimation for GRFs; (3) Weighted least squares (WLS) estimation of the empirical semivariogram; (4) Maximum likelihood estimation as described in Kammann and Wand (2003) (KW-ML); (5) Restricted maximum likelihood estimation as described in Kammann and Wand (2003) (KW-REML); (6) Maximum likelihood estimation using K-splines (KS-ML); and (7) Restricted maximum likelihood estimation using K-splines (KS-REML). In addition, the prediction performance at five locations on the considered spatial domain was evaluated. D-ML, D-REML and WLS are kriging approaches in which the covariates enter the mean function linearly. For KW-ML, KW-

REML, KS-ML and KS-REML, we use model (1) where 150 knots are used to model the spatial component S.
Results are displayed in Table 2. It is observed that K-splines perform better for the estimation of the covariogram parameters c_s and τ. Because the covariates enter the mean function linearly in D-ML, D-REML and WLS, the measurement error parameter σ_ε^2 is not well estimated. The estimated covariogram parameters for KW-ML and KW-REML are seriously biased. This can be expected since the approach of Kammann and Wand (2003) does not attempt to estimate these parameters well. In terms of prediction, we see that K-splines perform the best.

TABLE 2. MSE results over 250 simulations for the covariogram parameters and predictions with corresponding 95% confidence intervals coverage.

	c_s	τ	σ_ε^2	c_s/τ	pred.	cov.	cov.[a]
D-ML	1.67	0.02	3.76	58.97	23.65	57.2	
D-REML	1.82	0.02	3.83	62.24	23.64	57.2	
WLS	2.03	0.06	3.88	75.33	23.88	57.6	
KW-ML	$> 10^3$	148.52	3.71	$> 10^3$	14.61	63.4	
KW-REML	$> 10^3$	148.52	3.73	$> 10^3$	14.61	64.4	
K-ML	1.51	0.01	0.01	49.11	2.39	94.8	95.0
K-REML	1.55	0.01	0.01	50.06	2.40	94.8	95.3

a: based on a bootstrap procedure (see Vandendijck et al., 2015)

4 Conclusion

K-splines offer a framework wherein the covariogram parameters in a geoadditive model are estimated accurately and precisely. From simulation studies we can conclude that predictions benefit from this.

References

Cressie, N.A.C. (1993). *Statistics for spatial data*. New York: Wiley.

Gelfand, A.E., Diggle, P.J., Fuentes, M., and Guttorp, P. (Eds.) (2010). *Handbook of Spatial Statistics*. Boca Raton: Chapman & Hall/CRC.

Hutchinson, M.F. and Gessler, P.E. (1994). Splines - more than just a smooth interpolator. *Geoderma*, **62**, 45 – 67.

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. *Journal of the Royal Statistical Society, Series C*, **52**, 1 – 18.

Vandendijck, Y., Faes, C., and Hens, N. (2015). K-splines: A new method towards geostatistical analysis. *Technical Report, Hasselt University, 02/02/2015*.

Do contacts over distance follow a power-law distribution? Estimation of the social contact distance kernel

Kim Van Kerckhove[1], Christel Faes[1], Philippe Beutels[23], Niel Hens[12]

[1] I-Biostat, Hasselt University, Belgium
[2] CHERMID, VAXINFECTIO University of Antwerp, Belgium
[3] School of Public Health and Community Medicine, The University of New South Wales, Australia

E-mail for correspondence: kim.vankerckhove@uhasselt.be

Abstract: Do contacts over distance show a multimodal form, with a peak of contacts close to home and a second peak further away from home, or is a power-law form sufficient? By using data from our social contact study, we were able to test this hypothesis. We exploited various distributions for the contacts at a certain distance, e.g. Poisson, Negative Binomial, ..., and incorporated random effects to account for the clustering of contacts within participants. Various forms of the underlying distribution were tested by integrating their information into the observed categories. The preliminary results support a Weibull form for the distribution of contacts over distance, however subtle differences are present when differentiating by the participant's age and by week or weekend days.

Keywords: Contacts; Distance kernel; Power-law; Random effects.

1 Introduction

Cooper (2006) noticed the lack of appropriately incorporating spatial information about contacts in epidemic models. He reported a study by Riley and Ferguson (2006) in which journey-to-work data were used as a proxy for the spatial distribution of potential contacts, a rough approximation to the truth according to Cooper (2006). He furthermore referred to unpublished work on contacts over distance to indicate the possible biases in using commuting data. Read et al. (2014) noted the same gap and presented data on contacts over distance in the setting of southern China. Their results

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

consisted of comparing the contacts over distance by rural and urban areas, describing the differences by age, but nonetheless no estimation of the distance kernel itself. Differences in the number of contacts over distance by age were shown, whereby the age group between 20 and 29 years of age was most mobile. The estimation of the distance kernel could lead to more realistic models and hence would possibly fill in the gap noted by previous authors. The current study aims at estimating the underlying distribution of contacts over distance, hence the distance kernel, while accounting for clustering and allowing differentiations based on characteristics of the participants.

2 Methodology

Using the time use data and the contact data from our social contact survey (described in detail by Willem et al., 2012) a probabilistic match between contacts and their distances from home was obtained. In the time use part, the participants indicated the place and at which distance they spent most of their time for certain time blocks (e.g. 5-8 h, 8-9 h, 9-10 h, ...). The distance (from home) was divided into 4 categories: 0-1 km, 2-9 km, 10-74 km and more than 75 km. The contact diary part was similar to previous contact surveys (e.g. Mossong et al., 2008), hence information on the location of the contact was available. A contact was defined as either a conversation with a person at less than 3 meters distance or a person they touched (skin-to-skin contact), hence contacts by phone or internet were not recorded. Repeated contacts were collected only once by combining the information. To combine the two parts, contacts were separated by location, if multiple locations were reported. The information from the time use data was summarized to have consecutive time blocks that either differed in the location the participant spent most of their time or by the distance at which the participant spent their time. Contacts at specific locations were further separated to have each of the possible options present from the time use data (by time and distance). As contacts might occur in a short period of time and the time spent at this location might not be of substantial duration, it occurs that a contact cannot be linked to a distance nor time. This is a drawback of the study design, necessitating the omission of these data. Since the probabilistic matching involves uncertainty, weights are given to each link. We refer to a combination of possible links as a distributed contact, and the weights sum to one for each distributed contact.
The information present for a possible link i can be described by $\mathbf{D_i} = (D_{1i}, D_{2i}, D_{3i}, D_{4i})$ with D_{1i} equal to 1 if the link occurs at distance 0-2 km and 0 otherwise. Each element of this stochastic vector follows a Poisson (Negative Binomial) distribution. Taking the weights for each link

into account, the following log-likelihood will be used:

$$\mathrm{ll}(\boldsymbol{\beta}|\mathbf{d}) = \sum_{i=1}^{p}\sum_{j=1}^{4}\left[-\lambda_{ji} + y_{ij}\log(\lambda_{ji}) - \log(y_{ij}!)\right].$$

Information concerning the participant (age, information recorded on week or weekend day) can be incorporated in the model via the parameters and the link functions. Let η_{1i}, η_{2i}, η_{3i}, and η_{4i} denote the linear predictors, which include the information of the participant through $\boldsymbol{\beta}$. These β-estimates can differ for each linear predictor. The parameters can be obtained as follows (with $n_i = \sum_{j=1}^{k} w_{ij}$).

$$\begin{aligned} \lambda_{1i} &= n_i \exp(\eta_{1i}), & \lambda_{2i} &= n_i \exp(\eta_{2i}) \\ \lambda_{3i} &= n_i \exp(\eta_{3i}), & \lambda_{4i} &= n_i \exp(\eta_{4i}). \end{aligned}$$

However, this approach does not take the clustering of contacts within participants into account. Random effects were added for each linear predictor. Furthermore, one can assume an underlying distribution for the contacts over distance, say $f(u, \boldsymbol{\theta})$. The linear predictors include the expression of the underlying distribution as for example $\eta_2 = \int_2^{10} f(u, \boldsymbol{\theta})\mathrm{d}u$. Information concerning the participant are incorporated in the parameters $\boldsymbol{\theta}$ of the underlying distribution. Various distributions including Powerlaw and Weibull are considered for $f(u, \boldsymbol{\theta})$.
Additionally, various models for the total number of contacts were compared to obtain the best fit. As such we are able to combine these two elements in the estimation of the joint density of the total number of contacts and the number of contacts over distance. This density allows us to test the assumption of independence between the total number of contacts and contacts over distance.

3 Preliminary results

A total of 41,327 links were recorded by 1527 participants, after removing the links with missing values for distance. Taking the weighted sum of the distributed contacts shows that most of the contacts (10,119.17) were reported at distance 2-9 km from home, followed by contacts close to home (8374.84) and contacts 10-74 km from home (7567.59). Substantially fewer contacts were reported at a distance of more than 75 km from home (837.51).
In a first attempt, we model the parameters allowing different estimates for various age classes and week or weekend days, but ignoring the clustering within participants. The age classes considered are based on the schooling system: $[0, 3)$, $[3, 6)$, $[6, 12)$, $[12, 18)$, $[18, 25)$, $[25, 45)$, $[45, 65)$ and $[65, 100)$. The saturated model, which allows a different estimate for each age group and week/weekend, has the lowest AIC (-65150) when age and week were

included together with an interaction term. Using an underlying Weibull model lead to the best fit from the various underlying distributions (AIC=-64081), however the saturated model outperforms this model. A further improvement in the Weibull model is to allow the two parameters to vary by age and week.
The estimates for the probabilities ($E(\frac{Y_1}{n_i})$) based on the saturated model showed a discrepancy over age: children showed a large proportion (around 55%) of contacts at home, whereas adults or young adults showed a smaller proportion of contacts close to home (around 25-35%), but an increase in the proportion of contacts at distance 10-74 km from home (around 29-40% for young adults and adults compared to 9-14% for children). Furthermore, during weekends generally a larger proportion of the contacts were reported closer to home (mostly around 30-59%).

4 Discussion

The preliminary results indicate a Weibull underlying distribution, however an extension of this model should give more evidence but clear influences of the participant's age and the day of collection are present. These subtle differences should be incorporated in future models such as agent based models or meta-population models in infectious diseases. Due to the study design not all contacts were linked to a possible distance and hence information was lost.

References

Cooper, B. (2006). Poxy models and rash decisions. *Proceedings of the National Academy of Sciences of the US*, **103**, 12221-12222.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, **5(3)**, e74.

Read, J.M., Lessler, J., Riley, S., Wang, S., Tan, L.J., Kwok, K.O., Guan, Y., Jiang, C.Q., and Cummings, D.A.T. (2014). Social mixing patterns in rural and urban areas of southern China. *Proceedings of the Royal Society B*, **281**, 20140268.

Riley, S. and Ferguson, N.M. (2006). Smallpox transmission and control: Spatial dynamics in Great Britain. *Proceedings of the National Academy of Sciences of the US*, **103**, 12637-12642.

Willem, L., Van Kerckhove, K., Chao, D.L., Hens, N., and Beutels, P. (2012). A nice day for an infection? Weather conditions and social contact patterns relevant to influenza transmission. *PLoS ONE*, **7**, e48695.

A simple and intuitive test for number-inflation or number-deflation

Paul Wilson[1], Jochen Einbeck[2]

[1] School of Mathematics and Computer Science/Statistical Cybermetrics Research Group, University of Wolverhampton, WV1 1LY, United Kingdom
[2] Department of Mathematical Sciences, Durham University, DH1 3LE, United Kingdom

E-mail for correspondence: pauljwilson@wlv.ac.uk

Abstract: We present a test of zero-modification which checks if the number of zeros is consistent with the hypothesized count distribution. This test is easily extended to test for inflation or deflation of *any non-negative values, and, by performing multiple tests of inflation/deflation of the counts present in observed data relative to* any given model, it is possible to assess the suitability of that model. Such multiple testing may be represented diagrammatically. The test for number-inflation/deflation is informally called the "Christmas Eve Test" as the original idea occurred to the main author on December 24th, 2014, and the diagrammatic method the "Durham Diagram" as it was developed during preparation for a talk at Durham University.

1 Problem and methodology

We are given random draws Y_i, $i = 1, \ldots, n$ from some count distribution, which is hypothesized to possess a specific parametric density function $f(y_i, \Theta_i)$. For instance, f may be the Poisson density, and Θ_i may correspond to a linear predictor $z_i^{\mathrm{T}}\beta$, with z_i a vector of covariates and β an unspecified parameter vector. We are interested in testing whether this distributional assumption is correct, or, in other words, whether the observed data are consistent with this specification.
We will consider initially the particular question of whether the observed number of zero's is consistent with this assumption (but generalize this idea to other values later on). Therefore, let $E(Y_i) = \mu_i$ and $p_i = P(Y_i = 0)$. Hence if X_i is a random variable which takes the value 1 if $Y_i = 0$ and 0 otherwise then X_i is a Bernoulli random variable with parameter p_i.

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

If $\mu_1 = \cdots = \mu_n = \mu$, i.e. μ does not depend on covariates, and hence all the p_i's are equal also, then the distribution of the number of zeros among $Y_1, \ldots Y_n$ is the sum of n independent Bernoulli random variables with parameter p, and hence is the binomial distribution $Bin(n, p)$, and thus has mean np and variance $np(1-p)$. Of more interest is when μ_i *does* depend on covariates, and hence the p_i's are not all equal. The sum, S_X, of n independent Bernoulli random variables $X_1, \ldots, X_n$ with parameters $p_1, \ldots, p_n$ respectively is known as a *Poisson–Binomial* distribution (Chen and Liu, 1997):

$$P(S_X = k) = \left\{ \prod_{i=1}^{n} (1 - p_i) \right\} \sum_{i_1 < \cdots < i_k} w_{i1} \cdots w_{ik}, \tag{1}$$

where $w_i = \frac{p_i}{1-p_i}$, $i = 1, \ldots, n$, and the summation is over all possible combinations of distinct $i_1, \ldots, i_k$ from $\{1, \ldots, n\}$. The R package *poibin* (Hong, 2013) implements both exact and approximate methods for computing the cdf of the Poisson-Binomial distribution.

2 The Christmas Eve test

We wish to determine whether data is zero-inflated or zero-deflated relative to a conditional count distribution (for instance, Poisson). The procedure for this is as follows:

- (i) Fit the model according to the hypothesized count distribution;
- (ii) For each Y_i, estimate $P(Y_i = 0) = p_i$;
- (iii) Use *poibin* to determine a (say) 90% confidence interval.

If the observed number of zeros in the data exceeds the upper limit of the confidence interval then we have evidence of zero-inflation. If it is less than the lower limit we have evidence of zero–deflation. (Alternatively *poibin* will return a p value.)

2.1 Example

The $n = 100$ observations in the following table were simulated from a ZIP distribution with zero-inflation parameter 0.2 and a Poisson mean uniformly distributed on $[0.5, 1.5]$.

Y	0	1	2	3	4	5	6	7
Count	39	18	17	16	7	0	2	1

Representing the vector of Poisson means by `Z`, the three steps outlined in the previous subsection are implemented through the following R-code:

```
(i)    mod <- glm(Y ~ Z, family = poisson)
(ii)   mfv <- dpois(0, mod$fitted.values)
(iii)  qpoibin(c(0.05,0.95), pp = mfv)
```

Step (iii) returns the 90% confidence interval $[19, 35]$ for the expected number of zeros, hence as the observed number of zeros is greater than the upper limit of the confidence interval we may reject the null hypothesis that the observed number of zeros is consistent with a Poisson distribution.

2.2 Parameter estimation, power and type-one error rates

As visible from step (i) of the above example, the procedure requires estimation of the means μ_i under the hypothesis that the count distribution is correctly specified. However, these estimates may be poor if this hypothesis is wrong, rendering the distribution (1) incorrect too.
Indeed, we found in further investigation that the estimation of the Poisson parameter will be generally biased (but reasonably precise) if the data are in fact zero–inflated. However, by estimating the mean parameter from the truncated, positive data only, the estimates of the mean parameter became unbiased, but imprecise.
Simulations show that a combination of the two approaches is successful: excellent power and type-one error rates are achieved when the Poisson parameter is estimated as a $2:1$ weighted mean of the the two estimators. The type-one error rates and powers obtained when the Christmas Eve test is used as a test of zero-inflation for 100 observed data are shown in Figure 1. The corresponding rates for a score test and a likelihood ratio test are also illustrated. As is apparent the Christmas Eve test is the most powerful. Its type one error rate is comparable to that of the score test, but behaves better than that of the likelihood ratio test.

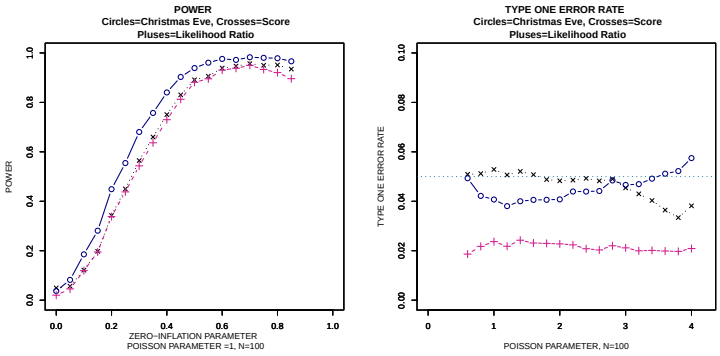


FIGURE 1. Power and Type One Error rates.

3 Extending the test to positive values

The code of Section 2.1 may easily be modified to obtain confidence intervals for other values. For example a 90% confidence interval for the number of 1's under the Poisson model may be estimated to be $[20, 34]$, indicating

that Y is "one-deflated" relative to the fitted Poisson model. Similarly it may be shown that $[23, 41]$, $[14, 30,]$, $[6, 18]$, $[1, 9]$, $[0, 5]$, $[0, 3]$ and $[0, 1]$ are 90% confidence intervals for the number of 2's, 3's, 4's, 5's, 6's and 7's under the Poisson model. This may be illustrated diagrammatically by a "Durham Diagram". In the left-hand diagram of Figure 2 the dotted lines represent the upper and lower limits of the confidence intervals for the counts under the Poisson model, and the dashed line the observed values. If the data is consistent with the reference model the dashed line should in general stay within the confidence bands, and departures from within the confidence bands indicate possible unsuitability of the reference model, and hence the left-hand diagram of Figure 2 indicates that a Poisson model is not suitable. The right-hand diagram of Figure 2 is constructed taking a zero-inflated Poisson distribution as the reference model; here we see that none of the observed counts exceed or fall short of the confidence intervals, indicating that a zero-inflated model may be suitable for the data.

4 Conclusion

The Christmas Eve Test is a highly intuitive test that when used to test zero-inflation has superior power to score and likelihood ratio tests, and an excellent type-one error rate. Whilst the Christmas Eve Test, including its extension to values other than zero, was originally developed with respect to zero-inflation it may be used to assess the suitability of any model for observed data.

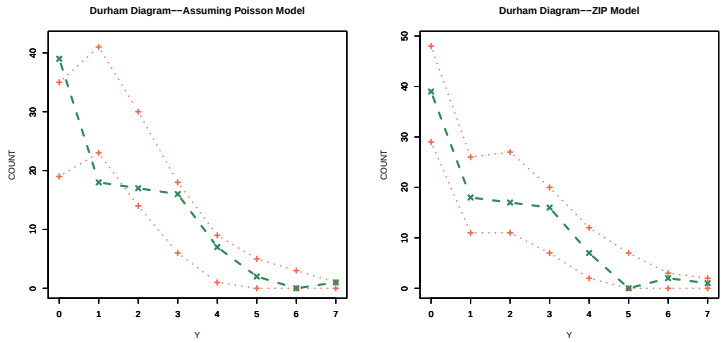


FIGURE 2. Durham Diagrams: Assuming Poisson and Zero-inflated Poisson.

References

Chen, S.X. and Liu, J.S. (1997). Statistical applications of the Poisson-Binomial and conditional Bernoulli distributions. *Statistica Sinica*, **7**, 875 – 892.

Hong, Y. (2013). poibin: The Poisson Binomial distribution. url = http://CRAN.R-project.org/package=poibin.

Index

Aerts, Marc, 59
Agogo, George, 3
Ainsbury, Liz, 123
Alencar, Airlane Pereira, 95
Alvares, Danilo, 7
Alvarez-Iglesias, Alberto, 135
Anan, Orasa, 11
Andrade-Bejarano, Mercedes, 211, 283
Araveeporn, Autcha, 15
Armero, Carmen, 7
Arteaga, Lina, 119

Böhning, Dankmar, 11
Böhnstedt, Marie, 51
Bühner, Markus, 55
Bacci, Silvia, 19
Bakka, Haakon, 23
Banditvilai, Somsri, 27
Bar, Edan, 31
Bar, Haim, 31
Bartolucci, Francesco, 19
Benková, Eva, 35
Beutels, Philippe, 295
Bigirumurame, Theophile, 39
Bitto, Angela, 43
Blanchard, Gilles, 259
Blanco, C., 63
Blas, Betsabé G., 47, 67
Bolfarine, Heleno, 247
Bollmann, Stella, 55
Boshuizen, Hendriek, 3
Brewer, Mark J., 147
Bruyndonckx, Robin, 59

Cakmak, S., 63
Castro, Cristina, 119
Catry, Boudewijn, 59
Cavalcante, Mileno, 67
Chiogna, Monica, 83
Coenen, Samuel, 59
Corduas, Marcella, 71
Cowling, Wallace, 271
Crouse, D., 63
Cysneiros, Audrey H.M.A., 79

Das, Kalyan, 263
Dayaratna, Kevin D., 75
De Bastiani, Fernanda, 79
Delalibera Jr., Italo, 203
Demétrio, Clarice G.B., 203
Diniz, Carlos Alberto Ribeiro, 207
Djordjilović, Vera, 83
Drovandi, Christopher C., 111
Durbán, María, 207

Einbeck, Jochen, 299
Enea, Marco, 127
Espendiller, Michael, 87

Faes, Christel, 291, 295
Ferrari, Silvia L.P., 227
Fiaccone, Rosemeire L., 91
Filippone, Maurizio, 187, 191
Fonseca, Eder Lucio da, 95
Forte, Anabel, 7
Frühwirth-Schnatter, Sylvia, 43, 139

Galea, Manuel, 79
Galipienso, Luis, 7
Gampe, Jutta, 51
Gilmour, Arthur, 271
Giraldo-Henao, Ramón, 211

Glotov, Nikolay V., 267
Gonçalves, Andreia, 167
Gopal Pillay, Khuneswari, 99
Grün, Bettina, 151
Guédon, Yann, 103
Guo, Jingyi, 107

Hölzl, Andreas, 55
Hainy, Markus, 111
Hajdíková, Táňa, 143
Harman, Radoslav, 35
Hebbern, C., 63
Heene, Moritz, 55
Henderson, Robin, 91
Hens, Niel, 59, 291, 295
Hermann, Philipp, 115
Hernández, Freddy, 119
Higueras, Manuel, 123
Hinde, John, 171, 203
Hofmarcher, Paul, 151
Hossain, Abu, 127
Husmeier, Dirk, 187, 191

Iannario, Maria, 131

Jacobs, Tom, 179
Jalali, Amirhossein, 135

Küchenhoff, Helmut, 55
Kümmerle, Tobias, 259
Kasim, Adetayo, 39
Kastner, Gregor, 139
Kateri, Maria, 87
Kedem, Benjamin, 75
Kiihl, Samara F., 231
Kneib, Thomas, 235
Komárek, Arnošt, 143
Komárková, Lenka, 143
Krnjajić, Milovan, 171

Lang, Stefan, 243
Latour, Katrien, 59
Leandro, Roseli A., 183
Levers, Christian, 259
Lopes, Hedibert Freitas, 139
Lorenzo-Arribas, Altea, 147

Möller, Annette, 175
Müller, Werner, 223
Maeder, Anthony, 163
Mahakeeta, Saowapa, 27
Malsiner-Walli, Gertraud, 151
Mameli, Valentina, 155
Marcelletti, Alessandra, 159
Maruotti, Antonello, 11, 159
Matawie, Kenan, 163
Mayr, Georg J., 287
McColl, John H., 99, 279
Mehar, Arshad, 163
Meira-Machado, Luís, 167
Messner, Jakob W., 287
Meyer, Renate, 239
Miller, Claire, 195
Moghaddam, Shirin, 171
Moral, Rafael A., 203
Morettin, Pedro Alberto, 95
Moriña, David, 123
Muchene, Leacky, 179
Musio, Monica, 155

Nakamura, Luiz R., 183
Newell, John, 135
Niu, Mu, 187
Noè, Umberto, 191

O'Donnell, Ruth, 195
O'Hagan, Adrian, 199
Oliveira, Thiago P., 203

Oliveira, Willian Luís, 207
Ospina-Galindez, Johann, 211
Overstall, Antony M., 147

Pandolfi, Silvia, 19
Pauger, Daniela, 215
Paula, Gilberto A., 219
Perrone, Elisa, 223
Perualila-Tan, Nolen, 39
Piccolo, Domenico, 131
Pinheiro, Aluísio, 231, 263
Pinheiro, Eliane C., 227
Pinheiro, Hildete P., 231
Puig, Pedro, 123

Relvas, Carlos Eduardo M., 219
Rennies, Hauke, 235
Riebler, Andrea, 107
Rigby, Robert A., 127, 183
Rogers, Simon, 187
Romeo, Jose S., 239
Roth, Helene, 243
Rothkamm, Kai, 123
Rubio, Luis, 7
Rue, Håvard, 107

Santos, Bruno, 247
Schäfer, Lukas M., 251
Schnabel, Sabine K., 255
Schneider, Max, 259
Scott, Marian, 195
Sen, Pranab K., 231
Sestelo, Marta, 167
Shirazi, Aliakbar Mastani, 263
Shkedy, Ziv, 39, 179
Sofronov, Georgy Yu., 267
Stasinopoulos, Dimitrios M., 127, 183
Stauffer, Reto, 287
Stefanova, Katia, 271
Stehlík, Milan, 115
Sweeney, James, 275

Torres-Taborda, Mabel, 119
Torretta, Federico, 255
Tough, Fraser, 279
Tovar-Cuevas, Rafael, 283
Tovar-Rios, Diego, 283
Trijsburg, Laura, 3
Trovato, Giovanni, 159

Umlauf, Nikolaus, 287
Uribe-Opazo, Miguel Angel, 79

van der Voet, Hilko, 3
van Eeuwijk, Fred A., 3
Van Kerckhove, Kim, 295
van't Veer, Pieter, 3
Vandendijck, Yannick, 291
Vanos, J., 63
Ventura, Laura, 155
Vicent, Antonio, 7
Villegas, Cristian, 183

Wagner, Helga, 215
Westhues, Matthias, 255
Wilson, Paul, 299
Worton, Bruce J., 251
Wright, Charlotte M., 279

Zanardo, Ana B.R., 203
Zeileis, Achim, 287
Zhukova, Olga V., 267
Zocchi, Silvio S., 203

30th IWSM 2015 Sponsors

We are very grateful to the following organisations for sponsoring the 30th IWSM 2015.

- Statistical Modelling Society
- Toyota Motor Corporation
- Leonard N. Stern School of Business, New York University
- CRC Press
- Springer
- Johannes Kepler University Linz, Austria
- Land Oberösterreich
- Stadt Linz
- Pöttinger