Proceedings of the

30th International Workshop on Statistical Modelling

volume 1

July 6 – 10, 2015

Linz, Austria

Herwig Friedl, Helga Wagner

(editors)

Proceedings of the 30th International Workshop on Statistical Modelling,
volume 1,
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

Preface

Dear Participants,

We welcome you to the 30th International Workshop on Statistical Modelling (IWSM) in Linz, Austria. Starting in 1986 in Innsbruck (Austria) with 40 participants, the workshop has been held yearly in various places in Europe, USA, and Australia and has grown to a conference with 150 – 200 participants coming from all over the world. Over these 30 years the workshop has kept its special format with a limited number of distinguished invited talks and only one plenary session throughout the week, thus stimulating discussion and exchange of ideas, encouraging interchange between junior and senior scientists and creating a special, family-like atmosphere.

This year's IWSM, which will bring together about 170 international researchers, follows the well-established traditions of the IWSM. The high standards of the conference and the quality of all presentations is ensured by the scientific committee, who invited renowned experts for keynote presentations and reviewed the contributions for oral and poster presentation. We are glad that Richard Chandler, Andrew Finley, Sylvia Frühwirth-Schnatter, Sonja Greven, and John Hinde have accepted the invitation to give a one hour presentation. The scientific committee also selected 54 contributions for oral presentation and 74 contributions will be presented in two poster sessions. Preceding the workshop, Alan Agresti gave a one day short course on Modeling Ordinal Categorical Data, which was attended by more than 40 participants.

As a tradition, the IWSM encourages PhD students not only to attend the workshop but also to present their work. Three students will receive awards for the best student paper, the best student oral presentation and the best student poster, respectively, at the conference dinner Thursday evening. Furthermore, two student travel grants have been provided by the Statistical Modelling Society.

Finally, we thank all authors who contributed to these proceedings for participating in this workshop and for carefully preparing their talks and posters. The contributions are also provided as a PDF-file from the conference website. We also acknowledge the valuable editorial help of Judith Kloas and Daniela Pauger, who both spent heaps of time reading proofs and ultimately harmonizing the manuscripts.

We wish you a pleasant stay in Linz and hope that you will experience an interesting and stimulating conference.

Herwig Friedl and Helga Wagner
Linz, July 2015

Contents

Part I – Invited Papers

Richard E. Chandler, Chiara Ambrosino: Spatiotemporal modelling for environmental applications: challenges and opportunities 3

Andrew O. Finley, Yuzhen Zhou, Sudipto Banerjee, Bruce D. Cook: Bayesian functional data models for coupling high-dimensional LiDAR and forest variables over large geographic domains 23

Sylvia Frühwirth-Schnatter: Recent advances in sparse Bayesian factor analysis 35

Sonja Greven: A general framework for functional regression 51

John Hinde: All statistical models are wrong, some are usable!.... 67

Part II – Contributed Papers

Abdelkader Ameraoui, Kamal Boukhetala, Jean-François Dupuy: Bayesian estimation of the tail index of a heavy tailed distribution with censoring 89

Francesco Bartolucci, Brian Francis, Silvia Pandolfi, Fulvia Pennoni: Robust maximum likelihood estimation of latent class models 94

Moritz Berger, Gerhard Tutz: An extended adjacent categories model accounting for response styles 100

Marco Bottone, Mauro Bernardi, Lea Petrella: Quantile regression: a Bayesian robust approach 106

Vincent Bremhorst, Michaela Kreyenfeld, Philippe Lambert: Time dependent variables in cure survival models: An application to the timing of a third pregnancy 112

Sarah Brockhaus, Andreas Fuest, Andreas Mayr, Sonja Greven: Functional regression models for location, scale and shape applied to stock returns 117

Carlo G. Camarda, Paul H.C. Eilers: Regression models with digit preference 123

Roberto Colombi, Sabrina Giordano: A mixture model for multidimensional ordinal data 129

Jesus Crespo Cuaresma, Bettina Grün, Paul Hofmarcher, Stefan Humer, Mathias Moser: Understanding economic growth regressions using latent class analysis 135

Wim De Mulder, André Grow, Geert Molenberghs, Geert Verbeke: Application of statistical emulation to an agent-based model: assortative mating and the reversal of gender inequality in education in Belgium 139

Matthew C. Edwards, Renate Meyer, Nelson Christensen: Signal extraction and power spectral density estimation: A Bayesian semi-parametric approach 145

Paul H.C. Eilers: Separation of signals with tailored penalties ... 151

Lida Fallah, John Hinde: Modelling homogeneous and heterogenous populations under joint progressive type-II censoring 157

Amanda Fernández-Fontelo, Pere Puig, Anna Alba: Generalized INAR models with trend and seasonality for veterinary syndromic surveillance 163

Gianluca Frasso, Philippe Lambert, Wilfried Bonou: A Bayesian model for the Ebola epidemic in Sierra Leone 169

Sylvia Frühwirth-Schnatter, Alexandra Posekany: Merging parallel MCMC output with yin-yang sampling 175

Jutta Gampe, Hein Putter, Paul H.C. Eilers: Hazard modelling for interval-censored data by smoothing within the EM algorithm 181

Antonio Gasparrini, Ben Armstrong, Michael G. Kenward: Penalized distributed lag non-linear models 187

Georg M. Goerg, Yuxue Jin, Nicolas Remy, Jim Koehler: How many people visit YouTube? Imputing missing events in panels with excess zeros 192

Andreas Groll, Trevor Hastie, Gerhard Tutz: Regularization in Cox frailty models 198

Marco Grzegorczyk, Andrej Aderhold, Dirk Husmeier: Network reconstruction with realistic models 204

Gillian Heller, Dominique-Laurent Couturier, Stephane Heritier: Dispersion modelling for Poisson-inverse Gaussian regression 210

Lisa Hermans, Geert Molenberghs, Michael G. Kenward, Wim Van der Elst, Vahid Nassiri, Marc Aerts, Geert Verbeke: Clusters with random size: maximum likelihood versus weighted estimation 215

Benjamin Hofner, Adam Smith: Boosted negative binomial hurdle models for spatiotemporal abundance of sea birds 221

Chiu-Hsieh Hsu, Jeremy Taylor, Chengcheng Hu: Analysis of accelerated failure time data with dependent censoring 227

Maria Iannario, Joseph B. Lang: Testing conditional independence in sets of $I \times J$ tables 232

Stijn Jaspers, Dankmar Böhning, Geert Verbeke, Marc Aerts: Estimation of an MIC distribution using the vertex exchange method 238

Maria Kateri, Alan Agresti: Average effect measures for group comparisons in ordinal response models 244

Nadja Klein, Helmut Herwartz, Thomas Kneib: Geoadditive stochastic frontier analysis 249

Nadja Klein, Thomas Kneib: Scale-dependent priors for variance parameters in distributional regression 255

Bernhard Klingenberg: Multiple comparisons of marginal probabilities following GEE estimation 261

Ioannis Kosmidis, Anyi Zou, Kavya Jagan: Beyond beta regression: modelling percentages and fractions in the presence of boundary observations 267

Philippe Lambert, Michel Denuit: Density estimation from tabulated summary statistics 273

Denis Larocque, Jeffrey S. Simonoff: Model-based mixed effects trees for longitudinal and clustered data 279

Robson J.M. Machado, Ardo van den Hout: Functional forms for transition intensities in a progressive three-state model 285

Brian D. Marx: Varying single-index signal regression 290

Andreas Mayr, Benjamin Hofner, Matthias Schmid: Boosting the discriminatory power of survival models via the C-index and stability selection 296

Damien McParland, Isobel Claire Gormley: Clustering mixed data via latent variable models 302

Rafael A. Moral, John Hinde, Clarice G.B. Demétrio, Edwin M.M. Ortega: Assessing goodness-of-fit for accelerated failure rate models: An insect ecology case-study 308

Izabela R.C. Oliveira, Geert Molenberghs, Clarice G.B. Demétrio, Carlos T.S. Dias, Cláudio L. Souza: Quantifying the genetic contribution to the variability of count traits 314

Jean Peyhardi: New qualitative choice models incorporating individual and choice characteristics 319

Johannes Reichl: Estimating marginal likelihoods from the posterior draws through a geometric identity 324

María Xosé Rodríguez-Álvarez, María Durbán, Dae-Jin Lee, Paul H.C. Eilers: Fast estimation of multidimensional adaptive P-spline models 330

Gunther Schauberger, Gerhard Tutz: BTL-Lasso – A penalty approach to heterogeneity in paired comparison data 336

Micha Schneider, Gerhard Tutz: Variable selection in mixture models for ordinal response with an uncertainty component 342

Marianna Siino, Salvatore Fasola, Vito M.R. Muggeo: Penalized logistic regression for small or sparse data: interval estimators revisited 348

Jose Luiz P. da Silva, Enrico A. Colosimo, Fabio N. Demarqui: Doubly robust-based GEE for the analysis of longitudinal ordinal missing data 354

Bernhard Spangl, Sascha Desmettre, Peter Ruckdeschel: Statistical models for dynamics in extreme value processes 360

Reto Stauffer, Nikolaus Umlauf, Jakob W. Messner, Georg J. Mayr, Achim Zeileis: Spatio-temporal censored model of precipitation climatology 366

Federico Torretta, Vito M.R. Muggeo, Paul H.C. Eilers: P-spline quantile regression with a mixed model algorithm 372

Laura Vana, Bettina Grün, Kurt Hornik: Modeling creditworthiness using a generalized linear mixed-effects approach 377

Elisabeth Waldmann, David Taylor-Robinson: Joint modelling of quantile regression and survival time of lung function decline in cystic fibrosis patients 381

Matthew Weldon, Andrew Titman: Estimation of two-sided choice models: an application to public school choice 387

Ariane Würbach, Sabine Zinn: Estimation of a general heaping model via random-walk Metropolis algorithms 392

Index 398

Part I – Invited Papers

Spatiotemporal modelling for environmental applications: challenges and opportunities

Richard E. Chandler[1], Chiara Ambrosino[1]

[1] University College London, UK

E-mail for correspondence: r.chandler@ucl.ac.uk

Abstract: Many developments in spatial and spatiotemporal modelling have been motivated by problems that are related in some way to environmental systems, together with their potential impacts on society and human health. Although a variety of approaches have been developed for specific applications, most modern statistical models for spatiotemporal data are based at some level on Gaussian processes, either as direct representations of the quantities of interest or as latent processes. The (latent) Gaussian formulation is enormously powerful and rather seductive, and has become almost the default approach for any application requiring a statistical model for spatiotemporal data. However, there are many environmental applications for which either the available data or expertise will not support its implementation; and there are many others for it does not provide a satisfactory framework within which to answer the substantive questions of interest. In this contribution, we give a flavour of some of these challenges based on our own experience with hydrological and climatological applications, and we provide some suggestions as to how they might be overcome. In so doing, we hope to highlight the many opportunities for the statistical modelling community to develop new tools for the study of spatiotemporal systems.

Keywords: Composite likelihood; Dependence; Gaussian process; Multivariate process; Precipitation.

1 Introduction

The routine modelling and analysis of large space-time datasets is one of the big success stories for statistics over the last two decades, with many application areas including the environmental sciences, epidemiology and medical imaging. A broad survey of the area is given by Finkenstädt et al. (2007). Approaches include stochastic models for specific classes of problem, such as the modelling of disease outbreaks (Diggle, 2007) and tumour

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

progression (Vedel Jensen et al., 2007). In modern statistical practice however, by far the most common approach to the analysis of space-time data is based on hierarchical models using Gaussian processes (see, for example, Banerjee et al., 2004, Chapter 8; Cressie and Wikle, 2011).

In its simplest form, the hierarchical Gaussian process approach can be summarised as follows: if $Y(\mathbf{s},t)$ denotes the value of some variable of interest at spatial location $\mathbf{s}$ and time t, then the field $\{Y(\mathbf{s},t) : \mathbf{s} \in \mathcal{S}, t \in \mathcal{T}\}$ is modelled as $Y(\mathbf{s},t) = \mu(\mathbf{s},t) + \varepsilon(\mathbf{s},t)$ where $\mu(\mathbf{s},t)$ is the mean function and $\varepsilon(\mathbf{s},t)$ is a zero-mean Gaussian process. The mean surface $\mu(\mathbf{s},t)$ is itself modelled as a space-time Gaussian process, possibly with its own mean function depending on covariates, and with some covariance structure controlling the smoothness of the surface. In this formulation, the space-time covariance function of $\{Y(\mathbf{s},t)\}$ is a sum of two components: one contribution from the mean surface $\mu(\mathbf{s},t)$, and the other from the 'residual' $\varepsilon(\mathbf{s},t)$. Without imposing strong constraints on the form or parameters of these components, it is usually hard to identify them unambiguously from observations. The simplest such constraint is to assume that the residuals at any finite collection of locations are uncorrelated, at least in space. More realistically, prior or hyperprior specifications may be used in a Bayesian framework to specify slow decay in the correlation function(s) for $\mu(\mathbf{s},t)$ (reflecting a prior belief that this surface will vary relatively slowly in both space and time) and more rapid decay for $\varepsilon(\mathbf{s},t)$.

When studying non-Gaussian fields, modelling options are more limited, mainly due to a lack of tractable models for joint distributions with non-Gaussian margins. Thus, if the context dictates that $\{Y(\mathbf{s},t)\}$ cannot be modelled adequately as a Gaussian field, many modern applications work with hierarchical models in which $Y(\mathbf{s},t)$ has mean $\mu(\mathbf{s},t) = g^{-1}[\eta(\mathbf{s},t)]$, say, where $g(\cdot)$ is a monotonic link function chosen so that $\{\eta(\mathbf{s},t)\}$ can plausibly be modelled as a Gaussian field. The $\{Y(\mathbf{s},t)\}$ are then often taken as conditionally independent given the $\{\eta(\mathbf{s},t)\}$.

This modelling approach has many appealing features, not least of which is that there are few alternatives that are both tractable and generally applicable; and it has acquired an almost canonical status in the toolkit for statistical analysis of space-time data. Unfortunately however, its appropriateness is sometimes questionable. The most serious of its deficiencies relates to the assumption, often made in non-Gaussian settings, that the responses are conditionally independent given the mean field $\{\mu(\mathbf{s},t) = g^{-1}[\eta(\mathbf{s},t)]\}$. This assumption is often unrealistic: consequences of its violation include underestimation of uncertainty (the problem is analogous to ignoring serial dependence when fitting regression models to longitudinal or time series data) and, potentially, biased estimation of spatially or temporally aggregated quantities of interest. Further issues include a relative scarcity of space-time covariance models that are both realistic and tractable; and, more pragmatically, the need for a high level of statistical literacy to use the methodology appropriately.

Of course, although the hierarchical Gaussian approach dominates modern space-time modelling practice, there are many other active lines of research. An example is the growing body of literature on spatial and space-time extremes: for an excellent review of this area, see Davison et al. (2012).
Against this background, our view is that despite many successes, the analysis of spatial and space-time data still presents many exciting challenges. We aim here to highlight some of them, exemplified by a case study involving the modelling of daily weather sequences for hydrological applications: this case study is introduced in Section 2, highlighting the major issues that must be considered. Section 3 discusses some of the options that are currently available to address these issues, and suggests areas where more work is needed. Section 4 concludes.

2 Case study: daily weather modelling

There is widespread recognition of the need for national and international strategies to cope with the effects of a changing climate; see, for example, European Commission (2013). Strategies are usually informed by models of environmental, engineered and social system response to climate projections that are derived from comprehensive models of the global climate system. However, despite recent advances, for many purposes the spatial resolution of most climate model outputs is too coarse. Therefore, climate projections are routinely 'downscaled' to a finer resolution. Many downscaling techniques are available; see Maraun et al. (2010) for a review. Here we focus on statistical models for the space-time fields of interest, conditioned on large-scale information that can be extracted from climate models.
We consider a specific case study that arose from a collaborative project involving climate scientists, hydrologists and hydrogeologists. The aim was to study the effects of climate change upon the water cycle, and to examine the implications for applications including flood risk assessment and water resource management. The catchment of the River Thames in England was selected for detailed study. The approach taken was to identify relevant large-scale atmospheric 'drivers' of weather within the study area (Lavers et al., 2012); to build statistical models for simulating space-time weather sequences, conditioned on climate model projections of these drivers, throughout the Thames catchment; and to use these simulations as input to land surface and hydrological models to obtain projections of system response. The land surface model used was the Joint UK Land Environment Simulator (JULES: see `www.jchmr.org/jules/`), which takes several meteorological variables as input, at an hourly time scale and at a spatial resolution of $1 \times 1\text{km}^2$. Seven key variables were identified, from which the remaining JULES inputs could be derived: precipitation, air pressure, air temperature, wind speed, downward short-wave radiation, wet bulb temperature and cloud cover. In the first instance, for testing purposes

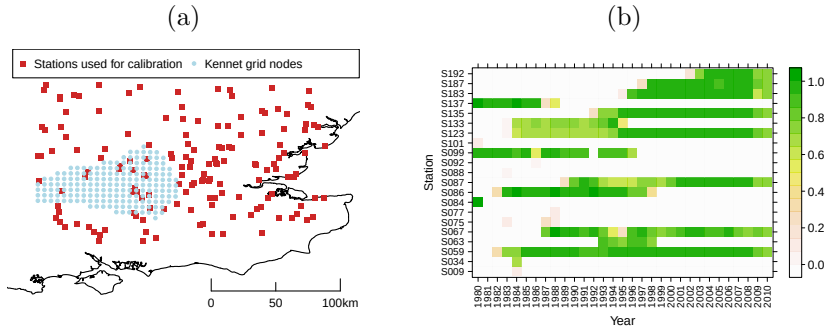

FIGURE 1. (a) Map of south-east England, showing the locations for which data are nominally available along with the locations at which inputs are required for JULES model simulations of the Kennet sub-catchment (b) availability of daily pressure data for Thames catchment, 1980 – 2010. The colour scale indicates the proportions of days each year for which data are available from each station.

it was agreed to focus on a well-studied subcatchment of the Thames (the Kennet) and to provide daily weather sequences at $5 \times 5\text{km}^2$ resolution. Figure 1(a) shows a map of the study region, including the 186 locations in the Kennet subcatchment for which JULES inputs are to be provided. Also shown are the locations of 157 weather stations for which daily data on some or all of the required variables are nominally available, from the British Atmospheric Data Centre (`badc.nerc.ac.uk/home/`), for the period 1980–2010. In reality however, no station records all of the variables: for example, temperature is reported at 99 of the stations, precipitation at 53 and radiation at just 11. Moreover, for stations that do provide data, many records are short or incomplete. This is illustrated in Figure 1(b) which shows that of the 21 stations reporting pressure, none have complete records and many returned hardly any data at all. Precipitation and temperature tend to be recorded more consistently; but it is nonetheless not unusual to encounter such patchy data in applications.

This application poses many challenges for a modeller tasked with simulating daily weather sequences. In the first instance, the variables have very different marginal distributions: temperature and pressure can plausibly be modelled using Gaussian fields but daily wind speed distributions are typically skewed; cloud cover is expressed as a proportion rounded to the nearest multiple of 0.125; and daily precipitation has a mixed distribution with an atom of probability at zero and a highly skewed continuous component on $\mathbb{R}^+$. Next, there are complex inter-variable relationships that are potentially dependent on large-scale atmospheric conditions and other factors. For example, precipitation cannot occur without clouds, which limit both incoming and outgoing radiation and are therefore associated with warmer temperatures during winter but cooler temperatures in summer.

This suggests that precipitation and temperature will tend be positively associated in winter and negatively associated in summer; but despite this, hot summer days with high humidity are prone to thunderstorms that produce intense precipitation. Such relationships are important because several critical land surface and hydrological phenomena, in particular evapotranspiration, depend nonlinearly upon combinations of the variables. If inter-variable relationships are not simulated adequately therefore, the variability of these phenomena will be misrepresented in projections of system response (because the variance of a function of several variables depends on the covariance between them); this in turn could lead to under- or overestimation of risk from extreme events such as floods or droughts.

A further requirement is that the simulations must incorporate realistic space-time dependencies, which may also vary with large-scale conditions and other factors. In northern Europe for example, temporal autocorrelation in precipitation tends to be weaker in summer than in winter, because precipitation is often associated with short-lived convective events in summer but with large-scale weather systems in winter. Again, the dependence can be an important determinant of system response. For example, the response of a catchment to a precipitation event will depend on soil wetness, which itself depends on the preceding weather: if the soil is relatively dry then the precipitation can be absorbed and, potentially, infiltrate to the groundwater. However, if the soil is already saturated then even a moderate amount of precipitation can lead to flooding. Spatial dependence is similarly important: increased flow in a single small tributary will usually have little impact on the flooding potential of a major river, but increases from many tributaries simultaneously can have very different consequences. Note that this type of spatial dependence is, in general, not synonymous with 'systematic regional variation' which, in the context of the discussion in Section 1, corresponds to the surface $\mu(\mathbf{s}, t)$: rather, it is a form of residual dependence reflecting the fact that on any given day, neighbouring locations tend to be affected by the same weather system(s).

These modelling requirements are further complicated by constraints on data availability, with at best partial overlap between the records at different stations, few if any complete cases for the multivariate modelling, and often (as here) no observations at any location for which simulation is required. Any widely-applicable modelling approach must therefore be able to work with incomplete data. Moreover, observations can be missing for various reasons. In weather records, missing observations often occur completely at random (in the terminology of Little and Rubin, 2002), for example if equipment fails for reasons unrelated to the weather. Missingness can be highly informative, however: we have encountered daily precipitation records with all non-zero weekend values missing (there is an explanation for this, but no space for it here!). In situations involving substantial amounts of missing data, it is therefore critical to have a good understanding of the missingness mechanism and, if this is potentially informative, to

adopt analysis approaches that reduce or eliminate any associated bias.
A final consideration is computational efficiency, both when fitting and when simulating models. For model fitting, this is important because the datasets involved are often large (the present dataset has around 577 000 cases, which is relatively modest for this kind of problem) and because it is often necessary to explore many alternative models to obtain an adequate representation of the system. Efficient simulation is desirable because many realisations are often required to sample the full range of variability in the process: this is needed in particular when studying risk from events that are relatively rare but have high impact. In this respect, a key challenge is to simulate simultaneously at large numbers of spatial locations. For the Kennet test case illustrated in Figure 1(a), sequences are required at 186 locations: simulation of, for example, a Gaussian field at a few hundred locations is relatively straightforward on a modern laptop computer. However, the ultimate goal in our case study (which was not achieved) was to provide simulations at a $1 \times 1\text{km}^2$ resolution throughout the Thames catchment which has an area of around $10\,000\text{km}^2$. Problems of this size are much more challenging for desktop or laptop computers. Of course, one possibility is to move to parallel, distributed or high-performance computing. Realistically however, such solutions present a barrier to widespread uptake of methodology, particularly in environments (such as in developing countries) where access to such facilities is limited.
We next review some approaches that may be appropriate for use in this kind of situation, including the methodology that was used for the case study in question and is implemented in the R package Rglimclim (Chandler, 2014). We also highlight some of the many remaining open problems.

3 Challenges and possible solutions

The issues highlighted above are all linked: we require models for complex, multivariate, spatiotemporal systems, that are amenable to fast fitting and simulation, and that can be fitted using incomplete data. For convenience, we group the issues under three main headings.

3.1 Multivariate modelling of complex systems

A major challenge in multivariate modelling is the relative scarcity of tractable models for joint distributions outside the Gaussian and multinomial families — see, however, Fahrmeir and Tutz (2001) for some of the available options — and no standard models for situations where, as in our case study, the marginal distributions have different forms.
If all variables have continuous distributions then it is natural to consider transformations to Gaussianity, so as to model the inter-variable dependencies via the covariance structure of the transformed variables. This is

essentially a copula-based modelling approach, which is becoming popular in some application areas (see, for example, Davison et al. 2012; Patton 2012). For binary or discrete variables however, it is less appealing. One possibility is to treat such variables as discretised versions of latent Gaussian quantities and to consider the latent, rather than observed, variables as components of the joint distribution. Inter-relationships among assumed latent variables can be hard to identify, however.

Probit regression is an example of an approach that can be understood via latent variables. This particular technique can, of course, be regarded in an entirely different light as an example of a generalised linear model (GLM) in which the mean of the variable of interest is related to linear combinations of covariates via a link function. In complex multivariate settings, GLMs are most naturally applied by building models for each variable in turn, at each stage conditioning on the variables that have already been modelled. The rationale for this is that any joint density for a random vector $\mathbf{Y} = (Y_1 \ \ldots \ Y_K)^{\mathrm{T}}$ say, can be factorised into a product of conditional densities as $f(\mathbf{y}) = f_1(y_1)f_2(y_2|y_1)\ldots f_K(y_K|y_1,\ldots,y_{K-1})$. In practice, this approach seems particularly suited to the modelling of systems in which each component potentially behaves very differently, because the models for each variable can be constructed separately using whatever distributional assumptions are appropriate. GLMs, and extensions such as generalised additive models (GAMs), offer great flexibility in modelling systematic variation in marginal distributions. Effects such as systematic regional variation in spatial data can be hard to parameterise directly, but can often be represented by including flexible sets of appropriately-chosen basis functions as covariates in the models, as in the spline-based GAMs advocated by Wood (2006) and implemented in the `mgcv` library in R. Further useful features include the use of interaction terms to represent features such as the seasonal variation in temperature-precipitation relationships; and the existence of an established diagnostic toolkit for checking modelling assumptions. For an overview of how GLMs can be used to represent complex structures in space-time data, with a particular focus on climate and weather, see Chandler (2005).

When building a multivariate model using successive conditioning, an obvious question is how to order the variables. Mathematically of course, it makes no difference: the joint density always factorises. In practice however, the models for each variable are at best approximations to the corresponding conditional distributions, whence the quality of the overall model depends on the adequacy of the cumulative approximation. In the first instance therefore, it seems sensible if possible to order the variables so as to reflect any subject-matter understanding of the causal relationships between them. Our experience is that statistical models can perform extremely well if such understanding is incorporated into the model structures (and, conversely, that strange results can occur if this is not done!).

By way of illustration: for the case study in Section 2, a multivariate

TABLE 1. Ordering of variables in successively conditioned multivariate model for Thames catchment, along with distributional choices for each component GLM. See text for description of these distributions.

Variable	Distributional choice
Y_1: air pressure	Normal-heteroscedastic
Y_2: precipitation	Logistic regression for occurrence, gamma for intensity
Y_3: air temperature	Normal-heteroscedastic
Y_4: wind speed	Gamma
Y_5: wet bulb temperature	Normal-heteroscedastic
Y_6: short-wave radiation	Gamma
Y_7: cloud cover	Normal-heteroscedastic after variable transformation

model was developed using successive conditioning, and the ordering of the variables was determined by examining solvers for physical equations of atmospheric dynamics, as used in numerical weather forecasting models (e.g. Skamarock et al. 2008). This led to the ordering shown in Table 1, which also shows our distributional choices for each variable. The 'normal-heteroscedastic' models are Gaussian models in which both the mean and logarithm of the variance are represented as linear functions of covariates (McCullagh and Nelder 1989, Chapter 10; Chandler 2005). The 'gamma' models are standard gamma GLMs with a log link and a constant dispersion parameter. In an approach dating back to Coe and Stern (1982), two models are employed for precipitation: the first determines the probability of non-zero precipitation, while the second determines the distribution of precipitation conditioned on its being non-zero. In the absence of 'standard' GLMs for continuous quantities taking values in $(0, 1)$, an empirical probit transform was applied to the cloud cover data so that a normal-heteroscedastic model could be used. This is arguably slightly unsatisfactory, but our collaborators considered that a highly accurate model for cloud cover was not needed — and, moreover, no other variables depend on it because it is the final variable to be considered.

Although this 'subject-matter-led' choice of variable ordering seems intuitively appealing, it is not always possible or optimal. This is particularly true in situations where 'primary' variables have substantial amounts of missing data, because this potentially limits the number of cases available for modelling other variables as well. In our case study for example, the sparsity of pressure data (see Figure 1(b)) has implications for the modelling of any other variable that depends directly on pressure. A naïve approach, assuming that missingness is uninformative, is to fit models for these other variables using only cases for which all required covariates are

present: this is extremely wasteful of data, however. An alternative is to use spatially aggregated values as covariates (for example spatial averages, or weighted averages where the weights decay with distance from the location of interest), on the grounds that spatial aggregates can be estimated using available data from other locations. This approach has been adopted in our case study because it is practically convenient, makes physical sense and because our simulation experiments indicate that the resulting models are fit for purpose. It is, however, slightly crude and potentially suffers from biases due to the fact that the covariates at each location are merely estimates (possibly with differing precisions, depending on what observations are available) of some underlying quantity. Methods for handling covariate measurement error (e.g. Carroll et al., 2006) could be useful here, although the implementation of these methods seems daunting for such complex models. A final option is to use multiple imputation techniques (see, for example, White and Carlin 2010) to deal with missing covariate data. In the present context, a natural approach when building a model for Y_k is to base imputation around the models that have already been built to define the joint distribution of $Y_1, \ldots, Y_{k-1}$. To date, such approaches have not been adopted widely in the context of spatial or space-time modelling.

3.2 Space-time dependence modelling

The previous section considered inter-variable dependence, along with systematic temporal and spatial variation. A genuinely spatiotemporal model must also represent residual dependence in space and time. If interest lies solely in inference about the model parameters, then a simple approximate representation may be adequate — although, in this case, it is prudent to allow for the mis-specified dependence when carrying out the required inference (see the discussion of composite likelihood techniques below). Moreover, other situations require more careful treatment, as in our case study where the dependence in simulated weather sequences must be realistic.

In the 'classic' hierarchical Gaussian modelling approach, dependence is often modelled via a stationary space-time covariance function for the residual process $\varepsilon(\mathbf{s}, t)$. A limitation of this approach is that space-time covariance models are hard to construct. The most widely-used models are separable (i.e. the space-time covariance function can be written as a product of a purely temporal and a purely spatial component), because few nonseparable covariance models are currently available (for some suggestions see Gneiting, 2002). The class of processes for which separability holds is severely limited, however (Gneiting et al., 2007).

A second limitation is that stationarity assumptions often are not tenable: the strength of dependence can vary over both time and space, and also in response to covariates. Models have been proposed for nonstationary and nonseparable covariance functions (e.g. Fuentes et al. 2008), but there have been few attempts to incorporate structured covariate effects.

A final difficulty with many covariance-based approaches is the computational burden associated with treating any large finite collection $\{Y_i = Y(\mathbf{s}_i, t_i) : i = 1, \ldots, n\}$ as a realisation of a single multivariate Gaussian vector when, as in most realistic space-time settings, n is very large. Without simplifying assumptions such as separability, exact evaluation of a Gaussian likelihood in such situations requires both storage and inversion of an $n \times n$ covariance matrix; and simulation from the corresponding models is similarly challenging. There are, of course, methods that address this issue, often using accurate approximations as in the process convolution approach described in Higdon (2007).

Many of these difficulties can be resolved, or at least alleviated, by exploiting the unidirectional nature of time. Appealing again to the factorisation of a joint density into conditional distributions, temporal dependence can be modelled by conditioning on previous time points. This can be done explicitly, as in the dynamical space-time models of Cressie and Wikle (2011); or by including lagged values of the variables of interest, possibly transformed or spatially aggregated, as additional covariates in GLMs and GAMs. In the latter approach, structured nonstationarity in temporal correlations is handled easily, via interaction terms involving the lagged values. A caveat here is that the interpretation of regression coefficients changes when lagged values are included (Chandler and Scott 2011, Section 3.3.3): this is potentially problematic if the coefficient values themselves are of substantive interest. A second issue is that in general, as with any autoregression, the coefficients of lagged values must be constrained to ensure that the model is dynamically stable: hence interactions involving lagged values must be handled with care, in case changes in the interacting covariates lead to dynamical instability. Large datasets provide partial protection against instability because, if a plausible model is fitted to enough data from a stable system, the estimated parameter values should reflect the dynamics of the observations. It would, nonetheless, be helpful to be able to determine whether, or under what conditions, a given model is dynamically stable: this is not well understood for the kinds of models considered here.

Despite these caveats, our experience is that conditioning on previous time points is a powerful technique for representing complex temporal dependence structures. The remaining challenge is purely spatial, therefore; and issues such as separability no longer arise. In Gaussian settings, the main issues are the modelling of nonstationary covariance structures and the computational burden when working with large numbers of locations. Non-Gaussian settings pose more fundamental problems, however. The common approach, of assuming responses to be conditionally independent given the mean function $\mu(\mathbf{s}, t)$, is rarely justified; and inappropriate modelling based on this assumption can have serious consequences (recall the earlier example of simultaneous flooding on many tributaries of a river).

At some level, residual spatial dependence presents many of the same challenges as multivariate dependence and, for continuous variables, a natural

solution once again is to use Gaussian models after appropriate transformation if necessary. Thus, in our own gamma GLMs (see Table 1), we specify spatial covariance functions for Anscombe residuals as described in Yang et al. (2005): if $Y(\mathbf{s}, t)$ has a gamma distribution with mean $\mu(\mathbf{s}, t)$ then the corresponding Anscombe residual is defined as $[Y(\mathbf{s}, t)/\mu(\mathbf{s}, t)]^{1/3}$.
As with multivariate modelling, the 'transformed Gaussian' approach cannot be applied directly to spatial fields of binary or discrete variables; and, again, one option is to work with latent Gaussian fields. This is relevant to the modelling of precipitation occurrence in our case study. We define a Gaussian process $\{Z(\mathbf{s}, t)\}$, with standard Gaussian marginal distributions, such that precipitation is non-zero at location $\mathbf{s}$ and time t only if $Z(\mathbf{s}, t) > \tau(\mathbf{s}, t)$ where the threshold $\tau(\mathbf{s}, t)$ is determined by the marginal probability of precipitation occurrence: dependence is then induced via the correlation structure of the latent field. This is mathematically convenient and appealing, but it is not the only option. For example, when studying precipitation at many locations in a relatively small area, the locations are often either all wet or all dry: a latent Gaussian approach in this situation can be numerically unstable because the covariance matrix must be close to singular, so that alternative options might be preferable. Yang et al. (2005) characterised such extreme spatial dependence via the distribution of the total number of wet locations in an area, and also reviewed some alternative modelling strategies for dependent binary data.
On the basis of our own experience and priorities, two major current challenges in the area of spatial dependence modelling are: firstly, the representation of structured nonstationarity as already discussed; and secondly the development of tractable and plausible dependence models for non-binary discrete data. This latter challenge is exemplified by current work that aims to build space-time models for dry lightning strikes in Australia, with a view ultimately to determining the potential impacts of climate change upon the frequency of lightning strikes and associated wildfire risk. Lightning strike data take the form of counts, and are phenomenally overdispersed in our experience; we are not aware of any attempts to develop realistic statistical models incorporating spatial and temporal dependence, together with potentially complex covariate dependencies, in this kind of setting.

3.3 Inference and diagnostics

In situations as complex as our case study, model fitting and comparison present their own challenges. Perhaps the 'gold standard' in modern statistical practice is a Bayesian analysis using Markov Chain Monte Carlo (MCMC) techniques which, in principle, can account for arbitrarily complex models along with missing observations, measurement error and other peculiarities of the data. In practice of course, things are rarely so easy: considerable care may be required to design efficient MCMC algorithms

(witness the number of published papers in which an entire section is devoted to the MCMC algorithm description for a specific model), and even efficient algorithms can take tens of minutes to run on large datasets.
We consider the computational cost of MCMC methods to be prohibitive in situations for which a thorough model-building exercise involves fitting and comparing many alternative models. To illustrate: in our case study, the mean and variance components in our final normal-heteroscedastic model for temperature contain 27 and 5 terms respectively. Of the terms in the mean component, one is the intercept; 15 represent systematic regional variation (modelled using two-dimensional basis functions along with site altitude); four represent seasonality via a Fourier expansion; one represents dependence on large-scale atmospheric temperatures (these could reasonably be extracted from a climate model); one represents dependence on a distance-weighted spatial average of wind speed, following the approach outlined in Section 3.1; one represents the previous day's temperature, again via a distance-weighted spatial average; two are interactions representing the seasonally-varying relationship between wind speed and temperature; and the remaining two are interactions representing the seasonally-varying strength of temporal autocorrelation. This model is relatively parsimonious, considering the demands of the application; but this conceals a lot of work to explore dependence on alternative indices of large-scale atmospheric structure, other contemporaneous variables, different transformations of lagged temperature, additional interaction terms and so forth. Given that models are required for six other variables as well (and we have not even discussed the variance component of the temperature model!) the need for efficient fitting methods to support a comprehensive model-building exercise should be clear.
One reason for the widespread use of MCMC methods is that they avoid the need for explicit computation of a likelihood function which, for complex models, is rarely amenable to direct computation. An alternative, allowing fast Bayesian inference for hierarchical models based on (latent) Gaussian fields, uses integrated nested Laplace approximations (Rue et al., 2009; Lindgren et al., 2011); potentially, this can handle extremely large spatial and space-time datasets. Another alternative, this time non-Bayesian, uses so-called composite likelihoods in which estimates are obtained by maximising a (possibly weighted) sum of log-likelihoods derived from low-dimensional subsets of the data. The most common types of composite likelihoods are those formed from univariate marginal distributions (sometimes called 'independence' likelihoods, because they would be genuine likelihoods if the observations were all independent) and those formed from collections of bivariate margins ('pairwise' likelihoods). Under the usual assumptions for likelihood-based inference, along with some unrestrictive additional requirements, composite likelihood estimators are asymptotically normally distributed with covariance matrix of 'sandwich' form, as is standard for estimation based on a mis-specified model (for further details,

see Varin et al. 2011). The components of this covariance matrix can be estimated empirically, if the data can be split into a large number of subsets such that the contributions from different subsets to the composite scores (i.e. gradient vectors of the composite log-likelihood) are uncorrelated. In space-time settings, the composite scores from distinct time points are indeed uncorrelated for any model in which temporal dependence is represented by conditioning on previous values (Chandler et al. 2007).

In practice, independence likelihoods provide a convenient justification for fitting models using standard software routines although standard errors, likelihood ratio tests and other formal methods for model comparison must be modified to account for the neglected dependence; again, see Varin et al. (2011). Of course, dependence parameters do not affect the univariate margins and hence cannot be estimated in this way. These parameters must therefore be estimated separately, for example using nonlinear least-squares to fit correlation functions to residuals from the marginal models: we adopted this approach in our case study. A further disadvantage is that some loss of efficiency is incurred by failing to exploit the full structure of the data, although the precision of the estimators is often high enough for practical purposes when working with large datasets. Pairwise likelihoods overcome both of these problems (there is a growing body of evidence that pairwise likelihood estimators have very high efficiency in many models), and are gaining considerable popularity in some areas such as the modelling of spatial extremes (e.g. Davison et al. 2012).

The methods reviewed so far allow both estimation and formal inference. Care is required, however, when applying formal inference techniques to large and complex datasets. This is because, for systems of the complexity considered here, any model is guaranteed to be wrong; yet many formal frameworks are designed to determine whether one or more of the models under consideration is correct in some sense. Thus, when comparing nested models using a likelihood ratio test, a rejection of the null hypothesis implies that the structure of the simpler model is incorrect: this is not news! Moreover, when working with large datasets, test procedures can often detect small departures from a null hypothesis that are of no practical relevance. Judgement is required, therefore, to decide whether an apparent improvement in model fit is worthwhile in practical terms: if the estimated mean surface $\hat{\mu}(\mathbf{s}, t)$ changes only in the third significant digit at all locations of interest, or if more substantial changes affect just a handful of observations, there is scope for overcoming any ingrained prejudice that a simpler model should be rejected if the associated p-value is less than *(insert your favourite number here)*.

In view of these considerations, it is arguably more important to assess goodness-of-fit than to worry unduly about formal model comparison. Many omnibus statistics, such as the deviance along with various information criteria, are available for this purpose. When modelling complex systems however, any model is guaranteed to exhibit lack of fit somewhere although

it may perform well in other respects: omnibus measures cannot reveal such features. Instead, informal measures may be used to assess a model's ability to reproduce aspects that are relevant for the intended purpose. In our own space-time modelling, we routinely use residual plots to check distributional assumptions along with the reproduction of seasonality, long-term trends, systematic regional variation and residual spatial dependence. Figure 2 shows some examples for the temperature model in our case study: the first two plots show the mean and standard deviations of Pearson residuals for each month of the year, along with guides to indicate what is expected under the model (in the first plot, the dashed lines are 95% intervals around zero, and in the second plot the horizontal line is the anticipated standard deviation). These plots indicate that overall, there is no unexpected seasonal structure in the data. By contrast, the third plot suggests an upward trend in mean residuals over time, possibly indicating a covariate that has been omitted from the model or (just as likely, from our experience of these data) some undocumented change in measurement or recording practice around 1987 or 1988. The fourth plot maps the mean Pearson residuals at each site: circle areas indicate the magnitudes of the residuals, and solid and dashed circles indicate the signs. The lack of clear spatial organisation here suggests that the model captures most of the systematic regional variation. Next is a quantile-quantile (Q-Q) plot revealing some departure from normality in the lower tail (around 0.5%) of the residual distribution; and finally a comparison of empirical and modelled relationships between inter-site distance and residual correlation, where the empirical correlations are shaded according to the number of contributing observations. Here, the fitted model (an isotropic powered exponential correlation function) fits most of the empirical correlations well: an appreciable minority of points lie some way from the fitted curve, but their less intense shading shows that these tend to be associated with smaller sample sizes.

Residual plots can help to reveal lack of fit in a model; but in complex nonlinear multivariate settings they do not necessarily help to decide whether it is fit for purpose. Thus the plots in Figure 2 merely assess the modelled distribution of temperature conditioned upon the previous day's temperatures and the present day's wind speed. The deficiencies revealed there may be entirely unimportant in terms of overall system behaviour; conversely, other small biases may have serious consequences. It is therefore necessary to assess the ability of the entire model to capture application-relevant features of the system. For most models, this cannot be done analytically. Our approach is therefore to identify specific properties that are potentially important in applications, and to use simulations to determine whether the observed values of these properties can credibly be regarded as consistent with the model. Ideally, this comparison should be carried out using observations that were not used in the model-building process.

Figure 3 shows some specimen results for our case study, for the period 2000–2009 (data from 1980–1999 were used for model development). We

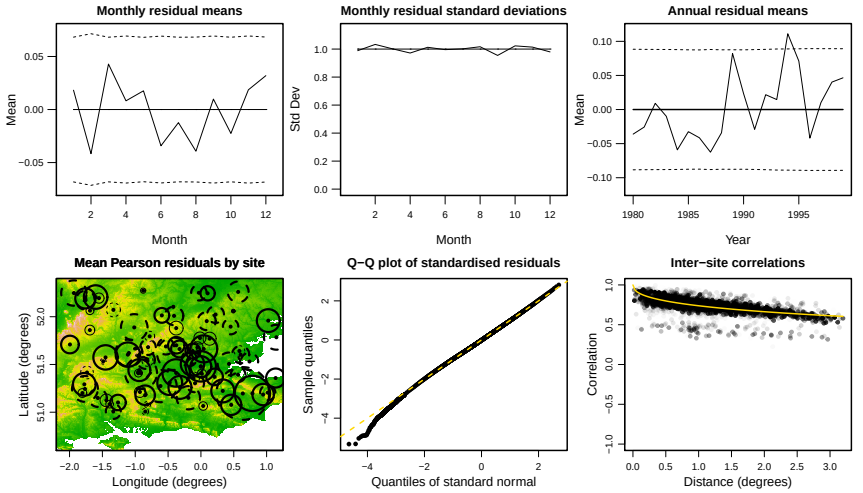

FIGURE 2. Specimen residual plots for Thames catchment temperature model. See text for details.

ran 100 simulations of our full multivariate space-time model, at all 343 locations shown in Figure 1. Each 10-year simulation ran in about nine minutes on a 3.2Ghz personal computer, using Rglimclim under Windows. For each simulation, we calculated several summary statistics that are potentially relevant for the applications that we envisage. Each statistic was computed separately for each month of the year, for the time series formed by averaging the daily values over all of the grid nodes in the Kennet subcatchment (see Figure 1). The results in Figure 3 all relate to temperature. A full assessment would consider other variables and properties as well, together with properties of the series at selected individual locations (spatial averages are considered here because their variability provides a test of the modelled spatial dependence structure). The coloured bands in Figure 3 show percentiles of the simulated distributions of the selected properties. If the model is realistic, the corresponding actual values should appear as samples from these distributions: this can in principle be assessed using techniques such as the probability integral transform (Dawid, 1984), although here we restrict ourselves to a simple visual assessment that the actual values span the full range of the simulations. Missing data contribute to uncertainty about the actual values, however: indeed, in our case study there are no observations at any of the Kennet grid nodes. Our solution is to estimate these missing values using multiple imputation. Thus we run 39 further simulations, this time conditioning throughout upon the available observations: for each quantity, the range from these 39 simulations forms a 95% uncertainty interval for the actual value. All of our spatial dependence models are formulated in such a way that the required conditional distribu-

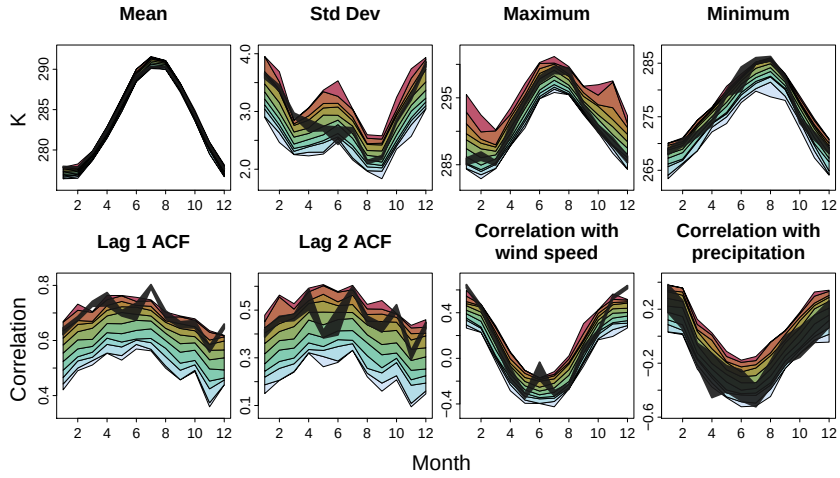

FIGURE 3. Summary statistics for daily temperatures averaged over the Kennet subcatchment. Coloured bands (grey in printed version) indicate the range, 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles of distributions obtained from 100 simulations of our multivariate model. Black bands are 95% uncertainty intervals for the actual values, derived using multiple imputation.

tions can be calculated: this can be very challenging in non-Gaussian settings (we have spent many years searching for imputation-friendly spatial dependence structures for large binary datasets; some ideas are reported in Yang et al. 2005 and in Ambrosino et al. 2014). There is, however, scope for improvement: at present, to impute a missing value $Y_k(\mathbf{s}, t)$, we compute the distribution conditional on all observations up to and including time $t-1$, along with all observations for variables $Y_1, \ldots, Y_{k-1}$ on day t. In general, observations for days $t+1$ onwards will also carry information about $Y_k(\mathbf{s}, t)$, as will the observations for $Y_{k+1}, \ldots, Y_K$ on day t. To incorporate this additional information, however, seems formidably difficult except in simple settings such as linear Gaussian systems (for which the Kalman smoother provides the solution).

Figure 3 suggests that the simulations reproduce most properties well, except for the autocorrelations which are slightly underestimated. Notice that temperature minima are well reproduced, despite the lack of fit in the lower tail of the Q-Q plot in Figure 2 (if anything, the simulated minima tend to be slightly too cool, in contrast to what the Q-Q plot would suggest). Notice also the width of the imputation envelope for the correlation between temperature and precipitation: this is due to a shortage of precipitation measurements in and around the Kennet subcatchment for the 2001–2009 period, coupled with relatively short-range spatial dependence in precip-

itation intensities so that imputation based on distant stations is highly uncertain. The imputation envelopes indicate that the actual values for the remaining properties can be inferred fairly precisely, however.

Again on the basis of our own experience, a major current challenge for space-time modellers is to provide generic frameworks that allow fast model fitting and comparison, in order to support structured model-building exercises. A further challenge is to provide guidance on how to choose between competing models that are all imperfect — or, perhaps more realistically, on how to determine whether a specific model, although acknowledged to be imperfect, is fit for its intended purpose. Other communities are already addressing such problems: for example, European climate researchers have recently developed a common framework for the 'validation' of downscaling procedures in the kinds of applications considered here (Maraun et al. 2015). Statisticians potentially have much to contribute to such initiatives.

4 Summary

The issues discussed in this contribution have necessarily been selective, based on our own experience. Nonetheless, we have demonstrated a wide range of opportunities for modellers who want to tackle difficult challenges; and the issues highlighted in our case study are typical of modern scientific investigations that exploit cheap computing capacity and large datasets to study ever more complex systems. At present, there is a clear need for more modelling options for the analysis of non-Gaussian, and particularly discrete-valued, space-time data. There is also scope for the development of intuitive, interpretable, computationally efficient techniques that can be implemented by statistically literate non-specialists: this will encourage the uptake of modern statistical methods in a wide range of application areas, as well as in environments where computing power is limited.

The quest for interpretability and computational efficiency may entail 'short cuts' such as the use of independence likelihoods for inference. In this case the modeller has a responsibility to understand the implications of what has been done, and to ensure that the resulting model is fit for purpose. Less glamorous, but necessary to reduce the risk of producing sophisticated nonsense, is the need for a thorough understanding of the data collection mechanism including the potential for artefacts (e.g. due to well-intentioned 'preprocessing'), errors, inhomogeneities and informative missingness.

We have focused primarily on broad issues, but these give rise to many more specific problems. In no particular order, these include: dealing with informative missingness in space-time datasets; development of Bayesian techniques for mis-specified models, perhaps using ideas such as those developed by Ribatet et al. (2012) for composite likelihoods; model selection based on fitness for purpose; imputation techniques for space-time data; and an investigation into the dynamical stability of complex space-time models. There is enough here to keep us busy for many years to come.

Acknowledgments: The development of Rglimclim, and the case study, were funded by Natural Environment Research Council grant NE/1006656/1.

References

Ambrosino, C., Chandler, R.E., and Todd, M.C. (2014). Rainfall-derived growing season characteristics for agricultural impact assessments in South Africa. *Theoretical and Applied Climatology*, **115**, 411 – 426.

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical Modelling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall / CRC Press.

Carroll, R.J, Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.). Boca Raton: Chapman & Hall / CRC Press.

Chandler, R.E. (2005). On the use of generalized linear models for interpreting climate variability. *Environmetrics*, **16**, 699 – 715.

Chandler, R.E. (2014). *Rglimclim: A multisite, multivariate weather generator based on generalised linear models*. R package version 1.3-0. `http://homepages.ucl.ac.uk/~ucakarc/work/glimclim.html`.

Chandler, R.E., Isham, V.S., Bellone, E., Yang, C., and Northrop, P.J. (2007). Space-time modelling of rainfall for continuous simulation. In: *Statistical Methods for Spatio-Temporal Systems.* (eds. Finkenstädt, B., Held, L., and Isham, V.S.), Boca Raton: Chapman & Hall / CRC Press, 177 – 215.

Chandler, R.E. and Scott, E.M. (2011). *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*. Chichester: Wiley.

Coe, R. and Stern, R.D. (1982). Fitting models to daily rainfall. *Journal of Applied Meteorology*, **21**, 1024 – 1031.

Cressie, N. and Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken: Wiley.

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

Dawid, A.P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society, Series A*, **147**, 278 – 292.

Diggle, P.J. (2007). Spatio-temporal point processes. In: *Statistical Methods for Spatio-Temporal Systems.* (eds. Finkenstädt, B., Held, L., and Isham, V.S.), Boca Raton: Chapman & Hall / CRC Press, 1 – 45.

European Commission (2013). *An EU Strategy on adaptation to climate change*. `http://ec.europa.eu/clima/policies/adaptation`.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). Springer-Verlag, New York.

Finkenstädt, B., Held, L., and Isham, V.S. (eds.) (2007). *Statistical Methods for Spatio-Temporal Systems*. Boca Raton: Chapman & Hall / CRC Press.

Fuentes, M., Chen, L., and Davis, J. (2008). A class of nonseparable and nonstationary spatial temporal covariance functions. *Environmetrics*, **19**, 487 – 507.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590 – 600.

Gneiting, T., Genton, M.G., and Guttorp, P. (2007). Geostatistical space-time models, stationarity, separability and full symmetry. In: *Statistical Methods for Spatio-Temporal Systems.* (eds. Finkenstädt, B., Held, L., and Isham, V.S.), Boca Raton: Chapman & Hall / CRC Press, 151 – 175.

Higdon, D. (2007). A primer on space-time modeling from a Bayesian perspective. In: *Statistical Methods for Spatio-Temporal Systems.* (eds. Finkenstädt, B., Held, L., and Isham, V.S.), Boca Raton: Chapman & Hall / CRC Press, 217 – 279.

Lavers, D.A., Villarini, G., Allan, R.P., Wood, E.F., and Wade, A.J. (2012). The detection of atmospheric rivers in atmospheric reanalyses and their links to British winter floods and the large-scale climatic circulation. *Journal of Geophysical Research*, **117**, D20106, doi:10.1029/2012JD018027.

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

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken: Wiley.

Maraun, D., Wetterhall, F., Ireson, A.M., Chandler, R.E., Kendon, E.J., Widmann, M., Brienen, S., Rust, H.W., Sauter, T., Themeßl, M., Venema, V.K.C., Chun, K.P., Goodess, C.M., Jones, R.G., Onof, C., Vrac, M., and Thiele-Eich, I. (2010). Precipitation downscaling

under climate change – recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, **48**, doi:10.1029/2009RG000314. 34pp.

Maraun, D., Widmann, M., Gutiérrez, J.M., Kotlarski, S., Chandler, R.E., Hertig, E., Wibig, J., Huth, R. and Wilcke, R.A.I. (2015). VALUE: a framework to validate downscaling approaches for climate change studies. *Earth's Future*, **3**, 1 – 14, doi:10.1002/2014EF000259.

Patton, A.J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, **110**, 4 – 18.

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.

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

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O, Barker, D.M., Duda, M.G., Huang, X.-Y., Wang. W., and Powers, J.G. (2008). *A Description of the Advanced Research WRF Version 3*. Technical Note, National Center for Atmospheric Research, Boulder, Colorado, `http://www2.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf`.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5 – 42.

Vedel Jensen, E.B., Jónsdóttir, K.Ý., Schmiegel, J., and Barnorff-Nielsen, O.E. (2007). Spatio-temporal modelling — with a view to biological growth. In: *Statistical Methods for Spatio-Temporal Systems.* (eds. Finkenstädt, B., Held, L., and Isham, V.S.), Boca Raton: Chapman & Hall / CRC Press, 47 – 75.

White, I.R. and Carlin, J.B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, **28**, 2920 – 2931.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman and Hall/CRC Press.

Yang, C., Chandler, R.E., Isham, V., and Wheater, H.S. (2005). Spatial-temporal rainfall simulation using generalized linear models. *Water Resources Research*, **41**, doi:10.1029/2004WR003739.

Bayesian functional data models for coupling high-dimensional LiDAR and forest variables over large geographic domains

Andrew O. Finley[1], Yuzhen Zhou[1], Sudipto Banerjee[2], Bruce D. Cook[3]

[1] Michigan State University, United States
[2] University of California, Los Angeles, United States
[3] NASA's Goddard Space Flight Center, United States

E-mail for correspondence: finleya@msu.edu

Abstract: Recent advances in remote sensing, specifically Light Detection and Ranging (LiDAR) sensors, provide the data needed to quantify forest variables at a fine spatial resolution over large geographic domains. Here, we define a Bayesian functional spatial data model for coupling high-dimensional and spatially indexed LiDAR signals with forest variables. The proposed modeling frameworks explicitly: 1) reduce the dimensionality of signals; 2) propagate uncertainty in parameters through to prediction, and; 3) acknowledge and leverage spatial dependence among the derived regressors and model residuals to meet statistical assumptions and improve prediction. Gaussian Processes (GPs) are applied in model components where spatial dependence can be used to improve inference. Fitting such models requires matrix operations whose complexity increase in cubic order with the number of spatial locations—resulting in a computational bottleneck. The dimensionality of the problem is tackled by replacing the GP with its low rank process counterpart. A non-separable spatial covariance function is used to capture within and among signal dependence. The proposed framework is illustrated using LiDAR and spatially coinciding forest inventory data collected on the Penobscot Experimental Forest, Maine.

Keywords: Spatial Gaussian process; Low rank; Predictive process; Forestry.

1 Introduction

Linking geo-referenced forest inventory with air- and space-borne Light Detection and Ranging (LiDAR) datasets via regression models offers an

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

attractive approach to mapping forest above-ground biomass (AGB) at stand, regional, continental, and global scales. LiDAR data have shown great potential for use in estimating spatially explicit forest variables, including AGB, over a range of geographic scales, see, e.g., Babcock et al. (2013), Finley et al. (2011), Iqbal et al. (2013), Muss (2011), and Næsset (2011). Encouraging results from these and many other studies have spurred massive investment in new LiDAR sensors, sensor platforms, as well as extensive campaigns to collect field-based calibration data. For example, NASA Goddard's LiDAR, Hyper-spectral, and Thermal (G-LiHT) imager is an air-borne platform developed, in part, to examine how future space-originating LiDAR may be combined with field-based validation measurements to build predictive models for AGB and other forest variables (Cook et al., 2013).

To effectively extract information from these high-dimensional massive datasets, we need a modeling framework to capture within and among LiDAR signal/forest variable association within and across locations. However, the computational complexity of such models increases in cubic order with the number of spatial locations, dimension of the LiDAR signal, and the number of forest variables—a characteristic common to multivariate spatial process models. Here, we propose a modeling framework that explicitly: 1) reduces the dimensionality of signals; 2) propagates uncertainty in data and parameters through to prediction, and; 3) acknowledges and leverages spatial dependence among the derived regressors and model residuals to meet statistical assumptions and improve prediction.

The subsequent sections are organized as follows. The motivating data and associated notation are described in Section 2. Section 3 offers a review on low rank Gaussian predictive process and modified Gaussian predictive process models. The joint model that couples the LiDAR signal and the forest variables is proposed in Section 4. Spatial predictions for forest variables are derived in Section 5. Finally, Section 6 provides some initial data analysis results and indications for future work.

2 Motivating data

Let $D \subset \mathbb{R}^2$ be a spatial domain and let $\mathbf{s}$ be a generic location in D. An outcome variable at location $\mathbf{s}$ is denoted by $y(\mathbf{s})$, which in our study is the AGB at a given geo-referenced forest inventory plot. High-dimensional LiDAR signals are also observed within D. Specifically, at location $(\mathbf{s}, x)$, the variable $z(\mathbf{s}, x)$ denotes the strength of LiDAR signal, where $x \in [0, M]$ is the height from the ground with M the maximum height. In our setting, both y and z are observed at a set of locations $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. For each location $\mathbf{s}_i$, z is measured at height $x_1, \ldots, x_{n_x}$.

The motivating data come from the Penobscot Experimental Forest (PEF), Maine. Here, the LiDAR $z(\mathbf{s}, x)$ were measured using the G-LiHT sensor

(Cook et al., 2013) at 26,286 spatial locations and $n_x = 113$ heights above the ground $[0, 33.9]$ meters. An example LiDAR signal is shown in Figure 1. Forest biomass y was measured at $n = 451$ forest inventory locations across the PEF (Figure 1). All n locations where y is observed spatially coincide with a measured LiDAR signal.

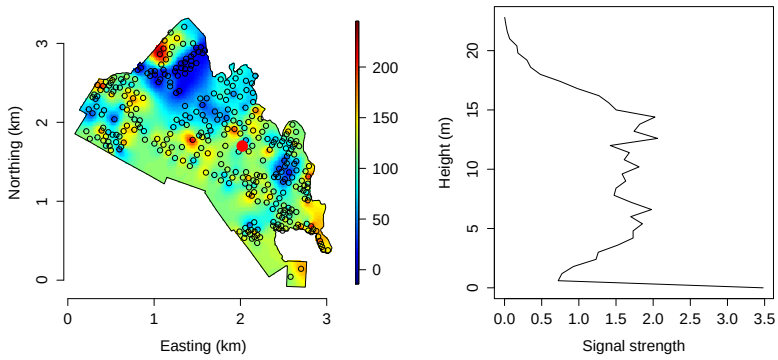

FIGURE 1. Left: Interpolated y, the small points $\circ$ indicate where y's are measured; Right: The LiDAR signal $z(\mathbf{s}, \cdot)$ measured at the red point $\mathbf{s}$ marked on the left figure.

3 Modified Gaussian predictive process

A Gaussian random field $\{w(\mathbf{s}), \mathbf{s} \in D \subset \mathbb{R}^N\}$ is commonly used to model the residual of spatial point-referenced data. For two, potentially different, locations $\mathbf{s}$ and $\mathbf{s}'$ let $C_w(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) := \text{Cov}(w(\mathbf{s}), w(\mathbf{s}'))$ be the covariance function of $w(\cdot)$ with parameters $\boldsymbol{\theta}$. Assuming data are observed at locations in $\mathcal{S}$, parameter estimation requires the inverse and determinant of the $n \times n$ covariance matrix, e.g., using Cholesky decomposition, which involves $O(n^3)$ flops. When the sample size n is large, such decomposition is prohibitively expensive from a computational standpoint.

To address this issue, Banerjee et al. (2008) introduced the *Gaussian predictive process*, which is a degenerate Gaussian random field obtained by projecting the *parent* random field to a lower-dimensional subspace. Specifically, by choosing a set of "knots" $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_r^*\} \subset D$, they defined the Gaussian predictive process

$$\tilde{w}(\mathbf{s}) = \mathbb{E}(w(\mathbf{s}) | w(\mathbf{s}_1^*), \ldots, w(\mathbf{s}_r^*)). \tag{1}$$

Let $c(\mathbf{s}; \boldsymbol{\theta}) = [C_w(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^r$ and $\mathbf{C}^*(\boldsymbol{\theta}) = [C_w(\mathbf{s}_i^*, \mathbf{s}_j^*)]_{i,j=1}^r$. The covariance function of $\tilde{w}(\mathbf{s})$ can be derived directly from its parent process, that

is

$$\mathrm{Cov}(\tilde{w}(\mathbf{s}), \tilde{w}(\mathbf{s}')) = c^{\mathrm{T}}(\mathbf{s};\boldsymbol{\theta})\mathbf{C}^{*-1}(\boldsymbol{\theta})c(\mathbf{s}';\boldsymbol{\theta}). \tag{2}$$

Decomposing the corresponding covariance matrix requires only $O(nr^2)$ flops with $r << n$.

Finley et al. (2009) pointed out that the predictive process underestimates the variance of the parent random field $w(\mathbf{s})$ since

$$\begin{aligned}&\mathrm{Var}(\mathrm{w}(\mathbf{s})) - \mathrm{Var}(\tilde{\mathrm{w}}(\mathbf{s})) = \mathrm{Var}(\mathrm{w}(\mathbf{s})|\mathrm{w}(\mathbf{s}_1^*), \ldots, \mathrm{w}(\mathbf{s}_{\mathrm{r}}^*))\\ &= C_w(\mathbf{s},\mathbf{s};\boldsymbol{\theta}) - c^{\mathrm{T}}(\mathbf{s};\boldsymbol{\theta})\mathbf{C}^{*-1}(\boldsymbol{\theta})c(\mathbf{s};\boldsymbol{\theta}) \geq 0,\ \forall \mathbf{s} \in D.\end{aligned} \tag{3}$$

As a consequence, the nugget variance in spatial regression model is usually overestimated because it absorbs the variability missed by the predictive process. To remedy this problem, Finley et al. (2009) proposed the *modified Gaussian predictive process* that corrects the variance bias by adding Gaussian noise to the predictive process. Specifically, they defined

$$\ddot{w}(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\epsilon}(\mathbf{s}), \tag{4}$$

where $\tilde{\epsilon}(\mathbf{s}) \overset{ind}{\sim} N(0, C_w(\mathbf{s},\mathbf{s};\boldsymbol{\theta}) - c^{\mathrm{T}}(\mathbf{s};\boldsymbol{\theta})\mathbf{C}^{*-1}(\boldsymbol{\theta})c(\mathbf{s};\boldsymbol{\theta}))$ is a spatially independent Gaussian random field with varying marginal variance. Hence, the modified Gaussian predictive process has the same marginal variance as the parent process.

4 Model

4.1 Modified Gaussian predictive model for z

The signal $z(\mathbf{s}, x)$ can be modeled as follows,

$$z(\mathbf{s}, x) = \mu_z(\mathbf{s}, x; \boldsymbol{\beta}) + u(\mathbf{s}, x) + \epsilon_z(\mathbf{s}, x), \tag{5}$$

where μ_z is the mean function, $u(\mathbf{s}, x)$ is the random effect which is a Gaussian random field on $\mathbb{R}^3$, and $\epsilon_z(\mathbf{s}, x)$ is the nugget effect.

Assume that the nugget effect $\epsilon_z(\mathbf{s}, x) \overset{ind}{\sim} N(0, \tau_z^2(x))$. Namely, the variance of the nugget is independent across locations.

Denote by $C_u(\mathbf{s}, \mathbf{s}', x, x'; \boldsymbol{\theta}_u) := \mathrm{Cov}[u(\mathbf{s}, x), u(\mathbf{s}', x')]$ the covariance function of random effect u. We approximate the parent model by modified Gaussian predictive processes within locations. Assume that $\{x_1^*, \ldots, x_{n_x^*}^*\}$ are the height knots at every location. Let

$$u^*(\mathbf{s}) = (u(\mathbf{s}, x_1^*), \ldots, u(\mathbf{s}, x_{n_x^*}^*))^{\mathrm{T}},$$

whose covariance at location $\mathbf{s}$ is given by

$$V_{u^*(s)}(\boldsymbol{\theta}_u) := \begin{pmatrix} C_u(\mathbf{s},\mathbf{s},x_1^*,x_1^*;\boldsymbol{\theta}_u) & \cdots & C_u(\mathbf{s},\mathbf{s},x_1^*,x_{n_x^*}^*;\boldsymbol{\theta}_u) \\ C_u(\mathbf{s},\mathbf{s},x_2^*,x_1^*;\boldsymbol{\theta}_u) & \cdots & C_u(\mathbf{s},\mathbf{s},x_2^*,x_{n_x^*}^*;\boldsymbol{\theta}_u) \\ \vdots & \ddots & \vdots \\ C_u(\mathbf{s},\mathbf{s},x_{n_x^*}^*,x_1^*;\boldsymbol{\theta}_u) & \cdots & C_u(\mathbf{s},\mathbf{s},x_{n_x^*}^*,x_{n_x^*}^*;\boldsymbol{\theta}_u) \end{pmatrix}, \tag{6}$$

and let

$$f(\mathbf{s}, x; \boldsymbol{\theta}_u) := (C_u(\mathbf{s}, \mathbf{s}, x, x_1^*; \boldsymbol{\theta}_u), \ldots, C_u(\mathbf{s}, \mathbf{s}, x, x_{n_x^*}^*; \boldsymbol{\theta}_u))^{\mathrm{T}}. \quad (7)$$

So the modified Gaussian predictive process model for z within locations is

$$z(\mathbf{s}, x) = \mu_z(\mathbf{s}, x; \boldsymbol{\beta}) + f^{\mathrm{T}}(\mathbf{s}, x; \boldsymbol{\theta}_u) V_{u^*(s)}^{-1}(\boldsymbol{\theta}_u) u^*(\mathbf{s}) + \epsilon_z^*(\mathbf{s}, x), \quad (8)$$

where

$$\epsilon_z^*(\mathbf{s}; x) \overset{ind}{\sim} N(0, \tilde{\sigma}_u^2(\mathbf{s}, x; \boldsymbol{\theta}_u) + \tau_z^2(x)), \quad (9)$$

with

$$\tilde{\sigma}_u^2(\mathbf{s}, x; \boldsymbol{\theta}_u) = C_u(\mathbf{s}, \mathbf{s}, x, x; \boldsymbol{\theta}_u) - f^{\mathrm{T}}(\mathbf{s}, x; \boldsymbol{\theta}_u) V_{u^*(s)}^{-1}(\boldsymbol{\theta}_u) f(\mathbf{s}, x; \boldsymbol{\theta}_u). \quad (10)$$

4.2 Joint model of y and Z

Denote by

$$Z(\mathbf{s}) = \begin{pmatrix} z(\mathbf{s}, x_1) \\ z(\mathbf{s}, x_2) \\ \vdots \\ z(\mathbf{s}, x_{n_x}) \end{pmatrix}, \ \epsilon_Z^*(\mathbf{s}) = \begin{pmatrix} \epsilon_z^*(\mathbf{s}; x_1) \\ \epsilon_z^*(\mathbf{s}; x_2) \\ \vdots \\ \epsilon_z^*(\mathbf{s}; x_{n_x}) \end{pmatrix},$$

and

$$F(\mathbf{s}; \boldsymbol{\theta}_u) = \begin{pmatrix} f^{\mathrm{T}}(\mathbf{s}, x_1; \boldsymbol{\theta}_u) \\ f^{\mathrm{T}}(\mathbf{s}, x_2; \boldsymbol{\theta}_u) \\ \vdots \\ f^{\mathrm{T}}(\mathbf{s}, x_{n_x}; \boldsymbol{\theta}_u) \end{pmatrix}, \ \mu_Z(\mathbf{s}; \boldsymbol{\beta}) = \begin{pmatrix} \mu_z(\mathbf{s}, x_1; \boldsymbol{\beta}) \\ \mu_z(\mathbf{s}, x_2; \boldsymbol{\beta}) \\ \vdots \\ \mu_z(\mathbf{s}, x_{n_x}; \boldsymbol{\beta}) \end{pmatrix}.$$

At location $\mathbf{s}$, we couple the LiDAR signal Z and the above-ground biomass y through the modified predictive processes $u^*(\mathbf{s})$. The joint model for y and Z is

$$\begin{pmatrix} Z(\mathbf{s}) \\ y(\mathbf{s}) \end{pmatrix} = \begin{pmatrix} \mu_Z(\mathbf{s}; \boldsymbol{\beta}) \\ \mu_y(\mathbf{s}; \boldsymbol{\eta}) \end{pmatrix} + \begin{pmatrix} F(\mathbf{s}; \boldsymbol{\theta}_u) V_{u^*(s)}^{-1}(\boldsymbol{\theta}_u) & \mathbf{0} \\ \boldsymbol{\alpha}_u^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} u^*(\mathbf{s}) \\ v(\mathbf{s}) \end{pmatrix} + \begin{pmatrix} \epsilon_Z^*(\mathbf{s}) \\ \epsilon_y(\mathbf{s}) \end{pmatrix}, \forall \mathbf{s} \in D, \quad (11)$$

where $\boldsymbol{\alpha}_u := (\alpha(x_1^*), \ldots, \alpha(x_{n_x^*}^*))^{\mathrm{T}} \in \mathbb{R}^{n_x^*}$, μ_y is the mean function of y, $v(\mathbf{s})$ is the random effect, and ϵ_y is the nugget effect. Assume that v, ϵ_y are independent of u and ϵ_z. $v(\cdot)$ is a Gaussian random field on $\mathbb{R}^2$ and $\epsilon_y(\mathbf{s}) \overset{ind}{\sim} N(0, \tau_y^2)$.

Let $X_Z(\mathbf{s}) \in \mathbb{R}^{p\times n_x}$ and $X_y(\mathbf{s}) \in \mathbb{R}^q$ be the predictors for the signal Z and y respectively. We further specify that $\mu_Z(\mathbf{s};\boldsymbol{\beta}) = X_Z(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta}$ and $\mu_y(\mathbf{s};\boldsymbol{\eta}) = X_y(\mathbf{s})^{\mathrm{T}}\boldsymbol{\eta}$. Then,

$$\begin{pmatrix} Z(\mathbf{s}) \\ y(\mathbf{s}) \end{pmatrix} = \begin{pmatrix} X_Z(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta} \\ X_y(\mathbf{s})^{\mathrm{T}}\boldsymbol{\eta} \end{pmatrix} + \begin{pmatrix} F(\mathbf{s};\boldsymbol{\theta}_u)V^{-1}_{u^*(s)}(\boldsymbol{\theta}_u) & \mathbf{0} \\ \boldsymbol{\alpha}_u^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} u^*(\mathbf{s}) \\ v(\mathbf{s}) \end{pmatrix} + \begin{pmatrix} \epsilon_Z^*(\mathbf{s}) \\ \epsilon_y(\mathbf{s}) \end{pmatrix}, \forall \mathbf{s} \in D. \tag{12}$$

4.3 Specification of the random effect of Z and y

- ***Covariance function for*** $u(\mathbf{s}, x)$***:*** Gneiting (2002) introduced a class of nonseparable stationary covariance function for space-time model on $\mathbb{R}^d \times \mathbb{R}$. Specifically, when $d = 2$,

$$C(\mathbf{s},\mathbf{s}';x,x') = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)(a|x-x'|^{2\kappa}+1)^{\delta+\gamma}} \times \left(\frac{c \parallel \mathbf{s}-\mathbf{s}' \parallel}{(a|x-x'|^{2\kappa}+1)^{\gamma/2}}\right)^{\nu} K_\nu\left(\frac{c \parallel \mathbf{s}-\mathbf{s}' \parallel}{(a|x-x'|^{2\kappa}+1)^{\gamma/2}}\right)$$

  Here, we use a simplified version to model the covariance of $u(\mathbf{s}, x)$ by fixing $\kappa = 1, \nu = \frac{1}{2}, \delta = 0$, i.e.,

$$\begin{aligned} &C_u(\mathbf{s},\mathbf{s}',x,x';\boldsymbol{\theta}_u) \\ :=&\frac{\sigma_u^2}{(a|x-x'|^2+1)^{\gamma}} \exp\left(-\frac{c \parallel \mathbf{s}-\mathbf{s}' \parallel}{(a|x-x'|^2+1)^{\gamma/2}}\right) \end{aligned} \tag{13}$$

  where $\boldsymbol{\theta}_u = (\sigma_u^2, a, \gamma, c), \sigma_u^2, a, c > 0$ and $\gamma \in [0, 1]$.

- ***Covariance function for*** $v(\mathbf{s})$***:*** We employ the exponential covariance function for $v(s)$, i.e.,

$$C_v(\mathbf{s},\mathbf{s}';\boldsymbol{\theta}_v) = \sigma_v^2 \exp\{-\phi_v \parallel \mathbf{s}-\mathbf{s}' \parallel\}, \tag{14}$$

  where $\boldsymbol{\theta}_v = (\sigma_v^2, \phi_v)$.

4.4 Prior specification

Denote by $\boldsymbol{\Omega}$ all the parameters in the model, which is

$$\boldsymbol{\Omega} = \{\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\alpha}_u, \boldsymbol{\theta}, \tau_y^2, \tau_z^2(x_i), i = 1, \ldots, n_x\}. \tag{15}$$

The hierarchical specification is completed by assigning priors to all parameters in $\boldsymbol{\Omega}$ as follows.

$$\begin{aligned}
&\boldsymbol{\beta} \sim N(\mathbf{0}, \text{diag}(10^4)),\ \boldsymbol{\eta} \sim \text{N}(\mathbf{0}, \text{diag}(10^4)),\\
&\alpha(x_i^*) \sim N(0, 10^4), i = 1, \ldots, n_x^*,\\
&\sigma_u^2 \sim IG(2, b_{\sigma_u}), \gamma \sim U(0,1),\ a \sim U(0, a_{max}),\ c \sim U(0, c_{max}),\\
&\sigma_v^2 \sim IG(2, b_{\sigma_v}),\ \phi_v \sim U(-\log(0.05)/d_{s,max}, -\log(0.01)/d_{s,min}),\\
&\tau_z^2(x_i) \sim IG(2, b_{\tau_z}), i = 1, 2, \ldots, n_x,\ \tau_y^2 \sim IG(2, b_{\tau_y}).
\end{aligned} \tag{16}$$

The specification of hyperparameters for ϕ_v follow from Ren and Banerjee (2013), where $d_{s,min}, d_{s,max}$ are minimum and maximum distance across all the locations. $b_{\sigma_u}, b_{\sigma_v}, b_{\tau_z}$, and b_{τ_y} are assigned from the empirical semivariogram (see, e.g., Banerjee et al., 2008).

5 Predictions

In most applied settings, such as that offered by the PEF dataset, the LiDAR signals are available at a fine spatial resolution across the domain and the forest variable of interest is observed on a small subset of locations. In such cases, our interest is in predicting y at all locations where Z is observed. The predictive distribution for y given Z is derived in Section 5.2. In other settings, both y and Z might not be observed at some locations and hence we want to predict them jointly as detailed in Section 5.1.

5.1 Predictions for y and Z at new locations

Assume there are no observations of y and Z at locations $\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_m$, but there are records of predictors X at those locations. Our goal is to find the conditional distribution of y and Z for the unobserved locations given the observed data.

We stack the data first. For all the observations, denote by

$$O = \begin{pmatrix} Z \\ y \end{pmatrix}, X = \begin{pmatrix} X_Z & 0 \\ 0 & X_y \end{pmatrix}, \tag{17}$$

where

$$Z = \begin{pmatrix} Z(\mathbf{s}_1) \\ \vdots \\ Z(\mathbf{s}_n) \end{pmatrix}, y = \begin{pmatrix} y(\mathbf{s}_1) \\ \vdots \\ y(\mathbf{s}_n) \end{pmatrix}, X_Z = \begin{pmatrix} X_Z(\mathbf{s}_1) \\ \vdots \\ X_Z(\mathbf{s}_n) \end{pmatrix}, X_y = \begin{pmatrix} X_y(\mathbf{s}_1) \\ \vdots \\ X_y(\mathbf{s}_n) \end{pmatrix}.$$

Similarly, denote by $\tilde{O}$ all the Z and y and $\tilde{X}$ all the predictors indexed by $\{\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_m\}$. Let $\Sigma_{O|\Omega}$, $\Sigma_{\tilde{O}|\Omega}$ be the covariance matrix of O and $\tilde{O}$

respectively. Meanwhile, let $C_{O,\tilde{O}|\Omega}$ be the cross-covariance matrix between O and $\tilde{O}$.
It is straightforward to check that

$$\begin{bmatrix} O \\ \tilde{O} \end{bmatrix} |\mathbf{\Omega} \sim N\left(\begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \mathbf{b}, \begin{bmatrix} \Sigma_{O|\Omega} & C_{O,\tilde{O}|\Omega} \\ C_{O,\tilde{O}|\Omega}^{\mathrm{T}} & \Sigma_{\tilde{O}|\Omega} \end{bmatrix} \right), \tag{18}$$

where $\mathbf{b} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}}$. Therefore, we can obtain the conditional distribution of $\tilde{O}$ given O and $\mathbf{\Omega}$,

$$[\tilde{O}|O, \mathbf{\Omega}] \sim N(\mu_{\tilde{O}|O}, \Sigma_{\tilde{O}|O}), \tag{19}$$

where

$$\begin{aligned} \mu_{\tilde{O}|O} &= \tilde{X}\mathbf{b} + C_{O,\tilde{O}|\Omega}^{\mathrm{T}} \Sigma_{O|\Omega}^{-1} (O - X\mathbf{b}) \triangleq (\mu_{\tilde{Z}|O}, \mu_{\tilde{y}|O})^{\mathrm{T}}, \\ \Sigma_{\tilde{O}|O} &= \Sigma_{\tilde{O}|\Omega} - C_{O,\tilde{O}|\Omega}^{\mathrm{T}} \Sigma_{O|\Omega}^{-1} C_{O,\tilde{O}|\Omega} \triangleq \begin{pmatrix} \Sigma_{\tilde{Z}\tilde{Z}} & \Sigma_{\tilde{Z}\tilde{y}} \\ \Sigma_{\tilde{Z}\tilde{y}}^{\mathrm{T}} & \Sigma_{\tilde{y}\tilde{y}} \end{pmatrix}. \end{aligned} \tag{20}$$

Usually, Bayesian prediction proceeds by sampling from the posterior predictive distribution

$$p(\tilde{O}|O) = \int p(\tilde{O}|O, \mathbf{\Omega}) p(\mathbf{\Omega}|O) d\mathbf{\Omega}. \tag{21}$$

In this case, for each posterior sample of $\mathbf{\Omega}$, we draw a corresponding $\tilde{O}$ by (19).

5.2 Predictions for y given Z are observed

Assume that there is no observation of y at locations $\tilde{\boldsymbol{s}}_1, \ldots, \tilde{\boldsymbol{s}}_m$, but there are records of the signal z and predictors X at those locations. Our goal is to find the conditional distribution of y given all the observations.
By (19), we can figure out the conditional distribution of $\tilde{y}$ given $\tilde{Z}$, O and $\mathbf{\Omega}$,

$$[\tilde{y}|\tilde{Z}, O, \mathbf{\Omega}] \sim N(\mu_{\tilde{y}|\cdot}, \Sigma_{\tilde{y}|\cdot}), \tag{22}$$

where

$$\begin{aligned} \mu_{\tilde{y}|\cdot} &= \mu_{\tilde{y}|O} + \Sigma_{\tilde{Z}\tilde{y}}^{\mathrm{T}} \Sigma_{\tilde{Z}\tilde{Z}}^{-1} (\tilde{Z} - \mu_{\tilde{Z}|O}) \\ \Sigma_{\tilde{y}|\cdot} &= \Sigma_{\tilde{y}\tilde{y}} - \Sigma_{\tilde{Z}\tilde{y}}^{\mathrm{T}} \Sigma_{\tilde{Z}\tilde{Z}}^{-1} \Sigma_{\tilde{Z}\tilde{y}}. \end{aligned} \tag{23}$$

The Bayesian prediction proceeds by sampling from the posterior predictive distribution

$$p(\tilde{y}|\tilde{Z}, O) = \int p(\tilde{y}|\tilde{Z}, O, \mathbf{\Omega}) p(\mathbf{\Omega}|O, \tilde{Z}) d\mathbf{\Omega} \tag{24}$$

Since we sample parameters $\mathbf{\Omega}$ only given the information of O due to computing burden by adding massive size of $\tilde{Z}$, we approximate the posterior predictive distribution as follows,

$$p(\tilde{y}|\tilde{Z}, O) \approx \int p(\tilde{y}|\tilde{Z}, O, \mathbf{\Omega}) p(\mathbf{\Omega}|O) d\mathbf{\Omega}. \tag{25}$$

In this case, for each posterior sample of $\mathbf{\Omega}$, we draw a corresponding $\tilde{y}$ by (22).

6 Illustrations: forest LiDAR and biomass data analysis

We analyze a large forestry dataset to assess model performance with regard to learning about process parameters and predicting y at new locations. Posterior inference for subsequent analysis were based upon three chains of 30000 iterations (with a burn-in of 5000 iterations). The samplers were programmed in C++ and leveraged Intel's Math Kernel Library's (MKL) threaded BLAS and LAPACK routines for matrix computations. The computations were conducted on a Linux workstation using two Intel Nehalem quad-Xeon processors.

6.1 Data preparation

Because the heights of trees are usually smaller than 22.8m at the observed area, there is no signal when the height is above 22.8m at most locations. We first cut the signal at 22.8m. Then, we coarsen the signal within $[0, 22.8]$m by averaging every two consecutive measurements and use them to fit Z. The dimension of signals at each location is set to $n_x = 39$. In this study, we do not have other covariates to model Z and y. Hence, we only assume the mean of $Z(\mathbf{s})$ and $y(\mathbf{s})$ are $X_Z^{\mathrm{T}}(\mathbf{s})\boldsymbol{\beta} = \{\boldsymbol{\beta}_i\}_{i=1}^{n_x}$ and $X_y^{\mathrm{T}}(s)\boldsymbol{\eta} = \boldsymbol{\eta}$ respectively. In addition, for numerical stability, we scale the magnitude of the biomass y and the signal Z, i.e., $y \to y/100$ and $z \to 100z$.

6.2 Results

For subsequent prediction validation analysis we held out 25% of data by randomly sampling from the 451 spatial locations. The height knots were chosen to be evenly distributed across the height range. We fit the following models using the training data (i.e., remaining $n = 339$ observations) with varying number of height knots: *i*) the joint model with height knots at $\{0, 22.8\}$ ($n_x^* = 2$); *ii*) the joint model with height knots at $\{0, 11.4, 22.8\}$ ($n_x^* = 3$); *iii*) the joint model with height knots at $\{0, 5.4, 11.4, 17.4, 22.8\}$ ($n_x^* = 5$) knots. Candidate models' fit to the observed data was assessed using DIC and associated metrics (Spiegelhalter et al., 2002). Additionally,

following Section 5.2, each candidate model was used to generate posterior predictive distributions for the holdout set locations. Then root mean squared prediction error (RMSPE) for y was computed for each candidate model using the posterior median of each predictive distribution and the holdout data. Posterior predictive distribution 95% credible interval coverage percent (CICP) and width (CIW) for y were also computed. Parameter estimates, fit, and prediction summaries are provided in Table 1.

TABLE 1. Parameter credible intervals, 50% (2.5%, 97.5%) and predictive validation. Where appropriate, estimates that differ from zero are bold.

Parameter	Results for the following numbers of knots		
	$n_x^* = 2$	$n_x^* = 3$	$n_x^* = 5$
$\boldsymbol{\eta}$	**1.04(0.79,1.3)**	**1.03(0.82,1.24)**	**1.05(0.83,1.28)**
$\alpha(0.00)$	**-0.05(-0.07,-0.02)**	**-0.06(-0.1,-0.03)**	0.03(-0.01,0.07)
$\alpha(5.40)$	-	-	**-0.06(-0.12,-0)**
$\alpha(11.4)$	-	**0.23(0.18,0.27)**	**0.06(0,0.11)**
$\alpha(17.4)$	-	-	**0.25(0.2,0.31)**
$\alpha(22.8)$	**0.25(0.19,0.31)**	**0.17(0.03,0.3)**	0.13(-0.01,0.26)
σ_u^2	0.12(0.1,0.13)	0.52(0.47,0.59)	0.66(0.6,0.73)
a	0.41(0.36,0.47)	0.2(0.18,0.22)	0.76(0.7,0.82)
γ	0.99(0.96,1)	1(1,1)	0.99(0.98,0.99)
c	14.55(12.51,17.39)	8.71(7.79,9.7)	7.59(6.99,8.19)
σ_v^2	0.13(0.09,0.2)	0.09(0.06,0.12)	0.08(0.07,0.1)
ϕ_v	2.13(1.46,2.73)	2.11(1.83,2.9)	2.12(1.75,2.6)
p_D	68.94	69.74	66.73
DIC	25824.84	23733.53	16995.04
RMSPE for y	0.383	0.364	0.357
95% CICP of y	91	89	92
95% CIW of y	1.447	1.398	1.434

Parameter estimates of $\boldsymbol{\alpha}$ that differ substantially from zero in Table 1 suggest that we are able to help explain the variability in y using information from Z. Specifically α's corresponding to higher portions of the LiDAR signal have a positive association with AGB and α's corresponding to lower portions of the LiDAR signal are negatively associated with AGB. This makes sense, because in general taller trees (or groups of trees) carry higher values of AGB.

Lower values of DIC suggest that fit to the observed data is improved as more height knots are added. Again this is initiative since more knots results in a better approximation to the observed signal and AGB. Finally, prediction validation analysis suggests the joint model with five knots results in similar RMSPE and mean CI width as those with reduced number of knots.

Future work will focus on further reducing the dimensionality of the problem by developing a doubly predictive process for the LiDAR signals (i.e., knots in both height and space).

References

Babcock, C.R., Matney, J.A., Finley, A.O., Weiskittel, A. and Cook, B.D. (2013). Multivariate spatial regression models for predicting individual tree structure variables using LiDAR data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **6**, 6 – 14.

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.

Cook, B.D., Corp, L.A., Nelson, R.F., Middleton, E.M., Morton, D.C., McCorkel, J.T., Masek, J.G., Ranson, K.J., Ly, V., and Montesano, P.M. (2013). NASA Goddards LiDAR, Hyperspectral and Thermal (G-LiHT) Airborne Imager. *Remote Sensing*, **5**, 4045 – 4066.

Finley, A.O., Banerjee, S., and MacFarlane, D.W. (2011). A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *Journal of the American Statistical Association*, **106**, 31 – 48.

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 – 2884.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, **97**, 590 – 600.

Iqbal, I.A., Dash, J., Ullah, S., and Ahmad, G. (2013). A novel approach to estimate canopy height using ICESat/GLAS data: A case study in the New Forest National Park, UK. *International Jorunal of Applied Earth Observation and Geoinformation*, **23**, 274 – 287.

Muss, J.D., Mladenoff, D.J., and Townsend, P.A. (2011). A pseudo-waveform technique to assess forest structure using discrete lidar data. *Remote Sensing of Environment*, **3**, 824 – 835.

Næsset, E. (2011). Estimating above-ground biomass in young forests with airborne laser scanning. *International Journal of Remote Sensing*, **32**, 473 – 501.

Ren, Q. and Banerjee, S. (2013). Hierarchical factor models for large spatially misaligned data: a low-rank predictive process approach. *Biometrics*, **69**, 19 – 30.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583 – 639.

Recent advances in sparse Bayesian factor analysis

Sylvia Frühwirth-Schnatter[1]

[1] Department of Finance, Accounting, and Statistics, Vienna University of Economics and Business, Austria

E-mail for correspondence: sfruehwi@wu.ac.at

Abstract: Factor analysis is a popular method to obtain a sparse representation of the covariance matrix of multivariate observations. The present contribution reviews some recent research in the area of sparse Bayesian factor analysis that tries to achieve additional sparsity in a factor model through the use of point mass mixture priors. Identifiability issues that arise from introducing zeros in the factor loading matrix are discussed in detail. Common prior choices in Bayesian factor analysis are reviewed and MCMC estimation is briefly outlined. An application to a small-sized psychological data set serves as an illustration.

Keywords: Shrinkage; Identifiability; Model selection; Point mass mixture priors; Fractional priors.

1 Introduction

For many decades, factor analysis has been a popular method to obtain a representation of the covariance matrix $\mathbf{\Omega}$ of correlated, multivariate observations of dimension m in terms of parameters of lower dimension than the number of unconstrained elements $m(m+1)/2$ in $\mathbf{\Omega}$, see e.g. Anderson (2003) for a comprehensive review of factor analysis.
A factor model relates each observation $\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})^{\mathrm{T}}$ in a random sample $\mathbf{y} = \{\mathbf{y}_t, t = 1, \ldots, T\}$ of T observation to a latent r-variate random variable $\mathbf{f}_t = (f_{1t} \cdots f_{rt})^{\mathrm{T}}$, the so-called common factors, through:

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\epsilon}_t, \tag{1}$$

where $\mathbf{\Lambda}$ is the unknown $m \times r$ factor loading matrix with elements λ_{ij}. r is called the number of factors. The common factors are assumed to be

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

orthogonal, i.e.

$$\mathbf{f}_t \sim N_r\left(\mathbf{0}, \mathbf{I}_r\right). \tag{2}$$

A basic assumption in factor analysis is that $\mathbf{f}_t$, $\mathbf{f}_s$, $\boldsymbol{\varepsilon}_t$, and $\boldsymbol{\varepsilon}_s$ are pairwise independent for all t and s. Furthermore, the following assumption is made concerning the idiosyncratic errors $\boldsymbol{\varepsilon}_t$:

$$\boldsymbol{\varepsilon}_t \sim N_m\left(\mathbf{0}, \boldsymbol{\Sigma}_0\right), \qquad \boldsymbol{\Sigma}_0 = \mathrm{Diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right). \tag{3}$$

Assumption (3) implies that conditional on knowing $\mathbf{f}_t$ the m elements of $\mathbf{y}_t$ are independent, hence all dependence among these variables is explained through the common factors. For the basic factor model, assumption (3) together with (2) implies that the observations $\mathbf{y}_t$ arise from a multivariate normal distribution $N_m(\mathbf{0}, \boldsymbol{\Omega})$ with zero mean and a covariance matrix $\boldsymbol{\Omega}$ with the following constrained structure:

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Sigma}_0. \tag{4}$$

The goal of exploratory factor analysis is to estimate the factor structure $\boldsymbol{\Lambda}$, to understand the driving forces behind the correlation between the features observed through y_t. Another motivation for the application of factor models, in particular in economics and finance, is to obtain a sparse representation of the covariance matrix $\boldsymbol{\Omega}$, especially, if m is large, see e.g. Fan et al. (2008) and Forni et al. (2009). Since r is typically smaller than m, the basic factor model (4) yields, by definition a sparse, i.e. low dimensional representation of $\boldsymbol{\Omega}$.

There has been considerable recent research in the area of sparse Bayesian factor analysis that tries to go beyond that the natural sparsity of the basic factor model. Additional sparsity can been achieved through continuous shrinkage priors on the factor loadings, see e.g. Bhattacharya and Dunson (2011). A different strand of literature considers sparse factor models with point mass mixture priors on the factor loadings. To achieve sparsity, a binary indicator δ_{ij} is introduced for each element λ_{ij} of the factor loading matrix which takes the value $\lambda_{ij} = 0$, iff $\delta_{ij} = 0$. This yields an indicator matrix $\boldsymbol{\delta}$ of the same dimension as $\boldsymbol{\Lambda}$. In sparse Bayesian factor analysis, δ_{ij} is typically unknown and has to be inferred from the data. Hence, inference with respect to $\boldsymbol{\delta}$ is considered as a variable selection problem. Important references for the basic factor model include West (2003), Carvalho et al. (2008), and Frühwirth-Schnatter and Lopes (2015), whereas Kaufmann and Schuhmacher (2014) study sparse dynamic factor models.

Conti et al. (2014) purse a related approach for a dedicated factor model, where equation (1) is combined with correlated (oblique) factors,

$$\mathbf{f}_t \sim N_r\left(\mathbf{0}, \mathbf{R}\right), \tag{5}$$

and the factor loading matrix $\boldsymbol{\Lambda}$ has a perfect simple structure, i.e. each observations loads on at most one factor. Under the assumption that $\mathbf{f}_t$ and

$\boldsymbol{\varepsilon}_t$ are independent, this yields following representation:

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}\mathbf{R}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Sigma}_0.$$

Since the factors are correlated, this approach yields a sparse structure for $\boldsymbol{\Lambda}$, but not for $\boldsymbol{\Omega}$.

Among these references, only Conti et al. (2014) as well as Frühwirth-Schnatter and Lopes (2015) explicitly address identifiability issues that may arise in sparse Bayesian factor analysis. Furthermore, these references show how sparse Bayesian factor analysis of an overfitting factor model yields inference with respect to the number of unknown, common factors, which has been known since long to be a very difficult issue. Lopes and West (2004), for instance, were amongst the first ones to formally tackle this issue in a basic factor model.

The present contribution reviews some of the main findings of these papers and provides an illustrative application to a small-sized psychological data set.

2 Identification of the basic factor model

For a given number of factors r, identification means that for any pair $(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_0)$ and $(\tilde{\boldsymbol{\Lambda}}_0, \tilde{\boldsymbol{\Sigma}}_0)$ satisfying (4), i.e.:

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Sigma}_0 = \tilde{\boldsymbol{\Lambda}}_0\tilde{\boldsymbol{\Lambda}}_0^{\mathrm{T}} + \tilde{\boldsymbol{\Sigma}}_0,$$

it follows that $\tilde{\boldsymbol{\Lambda}}_0 = \boldsymbol{\Lambda}$ and $\tilde{\boldsymbol{\Sigma}}_0 = \boldsymbol{\Sigma}_0$.

Well-known identification problems arise for the basic factor model, meaning that further conditions are necessary to achieve identifiability. Many papers reduce identification of factor models to uniqueness of the factor loading. However, uniqueness of the factor loading matrix alone does not imply identification. A rigorous identification of factor models was first offered by Anderson and Rubin (1956). They consider identification as a two-step procedure.

2.1 Two-step identification of factor models

The first step is uniqueness of the variance decomposition, i.e. the identification of the variances $\sigma_1^2, \ldots, \sigma_m^2$ of the idiosyncratic errors in $\boldsymbol{\Sigma}_0$ from the variance decomposition (4) of $\boldsymbol{\Omega}$. More precisely, given any pair $(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_0)$ and $(\tilde{\boldsymbol{\Lambda}}_0, \tilde{\boldsymbol{\Sigma}}_0)$ satisfying (4), under which condition does this imply that

$$\tilde{\boldsymbol{\Sigma}}_0 = \boldsymbol{\Sigma}_0, \qquad \tilde{\boldsymbol{\Lambda}}_0\tilde{\boldsymbol{\Lambda}}_0^{\mathrm{T}} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}.$$

Anderson and Rubin (1956), in Theorem 5.1, provide a sufficient condition for uniqueness of the variance decomposition, using a rank condition on $\boldsymbol{\Lambda}$, the so-called row deletion property:

AR Whenever an arbitrary row is deleted from $\mathbf{\Lambda}$, two disjoint matrices of rank r remain.

This condition implies that $2r + 1 \leq m$, and yields the following upper limit for the number of factor:

$$r \leq \frac{m-1}{2}. \tag{6}$$

For sparse factor models, the factor loading matrix may contain many zeros and, as shown by Frühwirth-Schnatter and Lopes (2015), uniqueness of the variance decomposition may be lost, if too many zeros are introduced. The following necessary conditions derived by Anderson and Rubin (1956) are very illuminating in this respect:

- ▷ Each column of $\mathbf{\Lambda}$ contains *at least 3 non-zero* factor loadings.
- ▷ Each pair of columns contains *at least 5 non-zero* factor loadings.

From these conditions it is immediately apparent that for a sparse Bayesian factor model a minimum number of non-zero elements has to be preserved in each column, despite variable selection, to guarantee identification $\mathbf{\Sigma}_0$. While the paper by Anderson and Rubin (1956) does not provide any counting rule beyond two factors, Frühwirth-Schnatter and Lopes (2015) were able extend this counting rule on the indicator matrix $\boldsymbol{\delta}$ to an arbitrary number of factors. They prove the following sufficient conditions on the indicator matrix $\boldsymbol{\delta}$ for identification of the variance decomposition:

SC For each $q = 1, \ldots, r$ and for each sub matrix consisting of q column of $\boldsymbol{\delta}$, the number of non-zero rows in this sub-matrix is *at least equal to* $2q + 1$.

This conditions extends the 3–5 counting rule of Anderson and Rubin (1956) in a natural way to a general 3–5–7–... rule. Evidently, too many zeros in a sparse factor loading matrix lead to non-identifiability of the variance decomposition, and subsequently to a failure to identify $\mathbf{\Lambda}$. Note that this counting rule for the zeros of $\boldsymbol{\delta}$ is straightforward to check and is easily incorporated into MCMC estimation of sparse factor models.
The identifiability of $\mathbf{\Sigma}_0$ guarantees that the matrix $\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$ is identified. Given identification of $\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$, the second step of identification is uniqueness of the factor loadings, i.e. subsequent identification of $\mathbf{\Lambda}$ from $\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$.
Without imposing further constraints on $\mathbf{\Lambda}$ the model is invariant under any transformation of the form $\tilde{\mathbf{\Lambda}}_0 = \mathbf{\Lambda}\mathbf{P}$ and $\mathbf{f}_t^\star = \mathbf{P}^{\mathrm{T}}\mathbf{f}_t$, where $\mathbf{P}$ is an arbitrary orthogonal matrix of dimension $r \times r$, since evidently,

$$\tilde{\mathbf{\Lambda}}_0\tilde{\mathbf{\Lambda}}_0^{\mathrm{T}} = \mathbf{\Lambda}\mathbf{P}\mathbf{P}^{\mathrm{T}}\mathbf{\Lambda}^{\mathrm{T}} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}. \tag{7}$$

Geometrically, $\mathbf{P}$ corresponds to a rotation of the factors which preserves orthogonality, because $\mathrm{V}(\mathbf{f}_t^\star) = \mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{P}^{-1}\mathbf{P} = \mathbf{I}_r$.

The usual way of dealing with the problem of factor rotation is to constrain $\mathbf{\Lambda}$ in such a way that the only possible rotation in (7) is the identity $\mathbf{P} = \mathbf{I}_r$. For orthogonal factors at least $r(r-1)/2$ restrictions on the elements of $\mathbf{\Lambda}$ are needed to eliminate the rotation indeterminacy. A common constraint is to consider only positive lower triangular (PLT) matrices, i.e. to constrain the upper triangular part of $\mathbf{\Lambda}$ to be zero and to assume that the main diagonal elements of $\mathbf{\Lambda}$ are strictly positive, i.e. $\lambda_{jj} > 0$ for all $j = 1, \ldots, r$. Evidently, for any two PLT matrices $\tilde{\mathbf{\Lambda}}_0$ and $\mathbf{\Lambda}$ with $\tilde{\mathbf{\Lambda}}_0 = \mathbf{\Lambda}\mathbf{P}$, $\mathbf{P} = \mathbf{I}_r$ is the only solution to (7) and $\mathbf{\Lambda}$ is uniquely identified from $\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$. Although the PLT constraint is pretty popular, it is often too restrictive in practice. It induces an order dependence among the responses, making the appropriate choice of the first r response variables an important modeling decision.

Recently, alternative identification strategies have been suggested. Kaufmann and Schuhmacher (2014), for instance, exploit the single value decomposition of $\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$ to achieve identification. Conti et al. (2014) investigate identification of the dedicated factor model (5) and prove that the following condition implies uniqueness of the variance decomposition as well as uniqueness of the factor loading matrix and, consequently, the 0/1 pattern of the indicator matrix $\boldsymbol{\delta}$: the correlation matrix $\mathbf{R}$ is of full rank, i.e. $\operatorname{rg}\mathbf{R} = r$ and each column of $\mathbf{\Lambda}$ contains at least three non-zero loadings.

Frühwirth-Schnatter and Lopes (2015) relax the PLT constraint by allowing $\mathbf{\Lambda}$ to be a generalized lower triangular (GLT) matrix:

GLT $\mathbf{\Lambda}$ is a *generalized lower triangular* matrix, i.e. $l_1 < \cdots < l_r$, where l_j denotes for $j = 1, \ldots, r$ the row index of the top non-zero entry in the jth column of $\mathbf{\Lambda}$, i.e. $\lambda_{l_j,j} \neq 0$ and $\lambda_{ij} = 0, \forall\, i < l_j$.

The GLT condition covers the PLT constraint as that special case where $l_j = j$ for $j = 1, \ldots, r$, but allows for more general forms of triangular matrices. This condition means that the top non-zero entries in the r columns of $\mathbf{\Lambda}$ have increasing row indices $l_1, \ldots, l_r$ with $l_j \geq j$, which need not lie on the main diagonal. This is particularly useful, if the ordering of the response variables is in conflict with the PLT assumption. Since λ_{jj} is allowed to be 0, response variables different from the first r ones may lead the factors. Indeed, for each factor j, the response variable l_j corresponding to the top non-zero element is the leading variable. The GLT condition prevents factor rotation, but not sign switching. Sign switching is resolved by requiring additionally that $\lambda_{l_j,j}$ is positive for each $j = 1, \ldots, r$.

2.2 Overfitting factor models

Assume that the data $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ are generated by model (1), with the corresponding variance decomposition in (4) being unique, however, the true number of factors r is not known. In this case, the usual procedure is to fit a model with $k \geq r$ factors,

$$\mathbf{y}_t = \boldsymbol{\beta}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim N_m\left(\mathbf{0}, \mathbf{\Sigma}\right), \tag{8}$$

where $\boldsymbol{\beta}$ is a unrestricted $m \times k$ coefficient matrix with elements β_{ij} and $\boldsymbol{\Sigma}$ is a diagonal matrix with strictly positive diagonal elements, in the hope to identify r subsequently from the resulting inference.
Instead of the true, unique variance decomposition (4), we are now dealing with the extended variance decomposition

$$\boldsymbol{\Omega} = \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}} + \boldsymbol{\Sigma}. \tag{9}$$

If $k = r$, then it follows immediately from the uniqueness of the variance decomposition (4), that $\boldsymbol{\beta} = \boldsymbol{\Lambda}$, and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$. However, if model (8) is, indeed, overfitting, i.e. $k > r$, then uniqueness of the extended variance decomposition (9) is lost, as noted e.g. by Geweke and Singleton (1980). Identifiability is lost for overfitting factor models, since infinitely many representations $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ exist, that imply the same covariance matrix $\boldsymbol{\Omega}$ as $(\boldsymbol{\Lambda}, \boldsymbol{\Sigma}_0)$.
More specifically, let $\mathbf{D}$ be an arbitrary diagonal matrix of size m with $\mathbf{D} < \boldsymbol{\Sigma}_0$ and $\operatorname{rg} \mathbf{D} \leq k - r$, and let $\mathbf{M}$ be an arbitrary matrix of size $m \times (k - r)$ such that $\mathbf{M}\mathbf{M}^{\mathrm{T}} = \mathbf{D}$. If $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are constructed as,

$$\boldsymbol{\beta} = (\boldsymbol{\Lambda}\ \mathbf{M})\,\mathbf{P}, \qquad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathbf{0}} - \mathbf{M}\mathbf{M}^{\mathrm{T}},$$

where $\mathbf{P}$ is an arbitrary permutation matrix, then obviously,

$$\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}} + \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \mathbf{D} + (\boldsymbol{\Sigma}_0 - \mathbf{D}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Sigma}_0 = \boldsymbol{\Omega}.$$

$\mathbf{M}$ are so-called spurious factors that do not contribute to explaining the correlation in $\mathbf{y}_t$, since $\mathbf{M}\mathbf{M}^{\mathrm{T}} = \mathbf{D}$. For sparse factor models, the spurious factor matrix $\mathbf{M}$ typically contains either zero columns or columns with a single non-zero loading.
Inference for an overfitting factor model is likely to yield factor loading matrices $\boldsymbol{\beta}$ that contain spurious factors. If the spurious factors in $\boldsymbol{\beta}$ are identified, they can be removed by adding $\mathbf{D} = \mathbf{M}\mathbf{M}^{\mathrm{T}}$ to $\boldsymbol{\Sigma}$ and the rank of the remaining columns of $\boldsymbol{\beta}$ is equal to the true number of factors r. Frühwirth-Schnatter and Lopes (2015) exploit this result to determine the number of factors within the framework of sparse Bayesian factor models.

3 Bayesian inference

Bayesian inference is performed in the representation (8) with k potential factors. If the number r of factors is known or estimated by some external criterion, then $k = r$. Otherwise a value $k \leq (m - 1)/2$ is selected.

3.1 Prior specifications

For sparse Bayesian factor models, a joint prior for the $m \times k$ model indicator matrix $\boldsymbol{\delta}$, the variances $\sigma_1^2, \ldots, \sigma_m^2$, and the $m \times k$ coefficient matrix $\boldsymbol{\beta}$

taking the form $p(\boldsymbol{\delta}, \sigma_1^2, \ldots, \sigma_m^2, \boldsymbol{\beta}) = p(\boldsymbol{\delta})p(\sigma_1^2, \ldots, \sigma_m^2)p(\boldsymbol{\beta}|\boldsymbol{\delta}, \sigma_1^2, \ldots, \sigma_m^2)$ is selected.

Typically, a hierarchical point mass mixture priors on the indicators $\boldsymbol{\delta}$ is applied, a popular example being:

$$\begin{aligned} &\Pr(\delta_{ij} = 1|\tau_j) = \tau_j, \qquad \tau_j \sim B(a_0, b_0), \qquad j = 1, \ldots, k, \\ &\Pr(\beta_{ij} = 0|\delta_{ij} = 0) = 1, \end{aligned} \tag{10}$$

where all indicators are assumed to be independent *a priori* given $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)$, and a beta distribution is assumed as prior for τ_j.

The choice of the hyper parameters of this prior is crucial, in particular, if the number of factors is unknown. In this case, the number of potential factors k is typically larger than r, hence the model is overfitting and we need to learn r. This requires the identification of the appropriate number of zero columns through the use of shrinkage priors that imply column sparsity. Hyperparameters that exclude zero columns apriori are prone to overfit the number of factors. Frühwirth-Schnatter and Lopes (2015) show that prior (10) is related to the two-parameter Bayesian nonparametric latent feature model introduced by Ghahramani et al. (2007), and may be rewritten as:

$$\tau_j \sim B(a_0, b_0) = B\left(\gamma\frac{\alpha}{k}, \gamma\right), \tag{11}$$

where k is the number of potential factors. Values of $\gamma < 1$ considerably smaller than 1, e.g. $\gamma = 0.3$, favour zero columns apriori and is the recommended choice, when the number of factors is unknown. Kaufmann and Schuhmacher (2014) and Conti et al. (2014) consider more general version of this prior.

When estimating factor models using classical statistical methods, such as maximum likelihood (ML) estimation, it frequently happens that the optimal solution lies outside the admissible parameter space with one or more of the idiosyncratic variances σ_i^2s being negative. This difficulty became known as Heywood problem. The introduction of a prior for each of the idiosyncratic variances $\sigma_1^2, \ldots, \sigma_m^2$ within a Bayesian framework, typically chosen from the inverted Gamma family, i.e.

$$1/\sigma_i^2 \sim \Gamma(c_0, C_{0i}),$$

naturally avoids negative values for σ_i^2. Nevertheless there exists a Bayesian analogue of the Heywood problem which takes the form of multi-modality of the posterior of σ_i^2 with one mode lying at 0.

Whereas in most papers on Bayesian factor analysis fixed hyperparameters C_{0i} are chosen, Frühwirth-Schnatter and Lopes (2015) select C_{0i} in such a way that Heywood problems are avoided. The resulting prior reads:

$$1/\sigma_i^2 \sim \Gamma(c_0, (c_0 - 1)/(\widehat{\boldsymbol{\Omega}^{-1}})_{ii}), \tag{12}$$

where $\widehat{\boldsymbol{\Omega}^{-1}} = \mathbf{S}_y^{-1}$ is the inverse of the sample covariance matrix $\mathbf{S}_y$.
Finally, conditional on $\boldsymbol{\delta}$ and $\sigma_1^2, \ldots, \sigma_m^2$, a prior has to be formulated for all non-zeros factor loadings. For a given $\boldsymbol{\delta}$, let $\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}$ be the vector of unconstrained elements in the ith row of $\boldsymbol{\beta}$. Since the likelihood function factors into a product over the rows, it is sensible to assume prior independence across rows, where for each row i with at least one non-zero element,

$$\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} | \sigma_i^2 \sim N\left(\mathbf{b}_{i0}^{\boldsymbol{\delta}}, \mathbf{B}_{i0}^{\boldsymbol{\delta}} \sigma_i^2\right). \tag{13}$$

This choice allows the posterior of the diagonal elements β_{ii} to be centered around 0, if the true value is 0. The variance of the prior (13) depends on σ_i^2, because this allows joint drawing of $\boldsymbol{\beta}$ and $\sigma_1^2, \ldots, \sigma_m^2$ and, even more importantly, sampling the model indicators $\boldsymbol{\delta}$ without conditioning on the model parameters during MCMC estimation.
Most papers in Bayesian factor analysis, e.g. Lopes and West (2004), Ghosh and Dunson (2009), and Conti et al. (2014) consider priors where $\mathbf{b}_{i0}^{\boldsymbol{\delta}} = \mathbf{0}$ and $\mathbf{B}_{i0}^{\boldsymbol{\delta}} = A_0 \mathbf{I}$ with fixed values of A_0. Kaufmann and Schuhmacher (2014) apply a hierarchical prior, where the $\mathbf{B}_{i0}^{\boldsymbol{\delta}} = \text{Diag}\left(A_1 \ldots A_k\right)$ and A_j follows an inverted Gamma distribution.
In Frühwirth-Schnatter and Lopes (2015), a fractional prior in the spirit of O'Hagan (1995) is applied which can be interpreted as the posterior of a non-informative prior and a small fraction b of the data. This yields a conditionally fractional prior for the "regression model"

$$\tilde{\mathbf{y}}_i = \mathbf{X}_i^{\boldsymbol{\delta}} \boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} + \tilde{\boldsymbol{\varepsilon}}_i, \tag{14}$$

where $\tilde{\mathbf{y}}_i = (y_{i1} \cdots y_{iT})^{\mathrm{T}}$, $\tilde{\boldsymbol{\varepsilon}}_i = (\varepsilon_{i1} \cdots \varepsilon_{iT})^{\mathrm{T}}$, and $\mathbf{X}_i^{\boldsymbol{\delta}}$ is a regressor matrix constructed from the latent factor matrix $\mathbf{F} = (\mathbf{f}_1 \cdots \mathbf{f}_T)^{\mathrm{T}}$. The fractional prior is defined from (14) as $p(\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} | \sigma_i^2) \propto p(\tilde{\mathbf{y}}_i | \boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}, \sigma_i^2)^b$ and leads to the following normal distribution,

$$\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} | \sigma_i^2 \sim N\left(\mathbf{b}_{iT}, \mathbf{B}_{iT} \sigma_i^2 / b\right), \tag{15}$$

where $\mathbf{b}_{iT}$ and $\mathbf{B}_{iT}$ are the posterior moments under the non-informative prior $p(\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} | \sigma_i^2) \propto c$:

$$\mathbf{B}_{iT} = \left((\mathbf{X}_i^{\boldsymbol{\delta}})^{\mathrm{T}} \mathbf{X}_i^{\boldsymbol{\delta}}\right)^{-1}, \qquad \mathbf{b}_{iT} = \mathbf{B}_{iT} (\mathbf{X}_i^{\boldsymbol{\delta}})^{\mathrm{T}} \tilde{\mathbf{y}}_i.$$

Frühwirth-Schnatter and Lopes (2015) discuss various choices of the fraction b, among them $b_N = 1/(Tm)$ and $b_R = 1/(km)^2$. Typically, values around $b = b_4 = 10^{-4}$ or $b = b_5 = 10^{-5}$ yield sparse solutions.

3.2 MCMC estimation

MCMC sampling for sparse factor models basically consists of the following steps:

(a) Sample the indicator matrix $\boldsymbol{\delta}$ conditional on the factors $\mathbf{f}_1, \ldots, \mathbf{f}_T$ and $\boldsymbol{\tau}$ without conditioning on the model parameters $\boldsymbol{\beta}$ and $\sigma_1^2, \ldots, \sigma_m^2$.

(b) Sample $\boldsymbol{\beta}$, $\sigma_1^2, \ldots, \sigma_m^2$ and $\boldsymbol{\tau}$ jointly conditional on $\boldsymbol{\delta}$ and $\mathbf{f}_1, \ldots, \mathbf{f}_T$.

(c) Sample $\mathbf{f}_1, \ldots, \mathbf{f}_T$ conditional on $\boldsymbol{\beta}$ and $\sigma_1^2, \ldots, \sigma_m^2$.

Steps (b) and (c) are straightforward and are implemented in a similar fashion in the various papers on sparse Bayesian factor analysis. However, considerable variety exists concerning the precise details of Step (a), since sampling the indicator matrix $\boldsymbol{\delta}$ corresponds to navigating through the extremely high dimensional model space corresponding to all feasible values of $\boldsymbol{\delta}$. The choice of an appropriate sampler depends on whether we are dealing with an overfitting model, where $k \geq r$, or a model where the number of factors r is known.

If the number of factors r is unknown, typically an unrestricted scheme is applied where $\boldsymbol{\delta}$ is sampled without imposing any constraints, not even variance identification. As a result, the posterior draws might violate the sufficient condition for variance identification given by **SC**. In this case, the only quantity that can be inferred without caring at all about identification is the marginal covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}} + \boldsymbol{\Sigma}$. Functionals of $\boldsymbol{\Omega}$, such as the trace of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^{-1}$ as well as (log) determinant of $\boldsymbol{\Omega}$ can be useful to assess convergence of the MCMC sampler.

For further posterior inference, is essential to considering only draws for which the variance decomposition is identified. While most papers ignore this important fact, the importance of variance identification for sparse Bayesian factor models has been recognized for the first time by Frühwirth-Schnatter and Lopes (2015) who perform post-processing and Conti et al. (2014) who designed a sampler that forces identification during sampling.

The challenge is to design MCMC samplers, that deliver a high fraction of draws of $\boldsymbol{\delta}$ for which the variance decomposition is identified despite sampling from an overfitting factor model. Naïve unconstrained Gibbs sampling tends to deliver many draws for $\boldsymbol{\delta}$ for which the variance decomposition is not unique. Including birth-and-death steps within unconstrained Gibbs sampling as well as a RJMCMC step that adds and deletes spurious columns during MCMC sampling greatly improves the rate of sampling indicator matrices $\boldsymbol{\delta}$ with unique $\boldsymbol{\Sigma}$.

For draws with unique variance decomposition, $\boldsymbol{\Sigma}$ is uniquely identified. Hence, posterior inference is possible for the idiosyncratic variances $\boldsymbol{\Sigma} = \mathrm{Diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right)$. Functionals of $\boldsymbol{\Sigma}$, such as the trace of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ as well as the (log) determinant of $\boldsymbol{\Sigma}$ can be useful to assess convergence of the MCMC sampler for the variance identified draws. In addition, inference with respect to the proportion of the variance of y_{it} that is explained by

the common factors, also known as communalities R_i^2, is possible:

$$R_i^2 = \sum_{j=1}^{r} R_{ij}^2, \qquad R_{ij}^2 = \frac{\lambda_{ij}^2}{\sum_{j=1}^{r} \lambda_{ij}^2 + \sigma_i^2}. \tag{16}$$

Due to the point mass mixture prior, the posterior draws of $\boldsymbol{\delta}$ contain valuable information concerning the number of factors and sparsity of the indicator matrix, provided that inference is restricted to MCMC draws for which the variance decomposition is unique.

The rank of $\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}$, or more or less equivalently, the number r of non-zero columns of $\boldsymbol{\delta}$ can be considered as a posterior draw of the number of factors r. Based on the posterior distribution $p(r|\mathbf{y})$, which is estimated from the posterior draws, the posterior modes $\tilde{r}$ provides an estimator of the number of factors r. In addition, sparsity of the indicator matrix in terms of the number d of non-zeros elements, i.e.

$$d = \sum_{j=1}^{k} \sum_{i=1}^{m} \delta_{ij}, \tag{17}$$

may be evaluated.

Inference with respect to r and d is valid, regardless whether the rotation problem for $\boldsymbol{\beta}$ has been solved or not. Uniqueness of the variance decomposition, however, is essential. Popular heuristic methods of inferring the number of factors by counting the number non-zero columns from $\boldsymbol{\delta}$, without checking uniqueness of variance decomposition (Bhattacharya and Dunson, 2011) are prone to be biased, typically overfitting the number of factors.

If the number of factors r is assumed to be known, then inference is possible with respect to the structure of factor loadings $\boldsymbol{\Lambda}$. In addition to identification of $\boldsymbol{\Sigma}$, and consequently $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$, this requires identification of $\boldsymbol{\Lambda}$ from $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$. To achieve identification of $\boldsymbol{\Lambda}$, Frühwirth-Schnatter and Lopes (2015) assume that $\boldsymbol{\Lambda}$ is a GLT matrix. During MCMC sampling, an unordered GLT condition is forced on $\boldsymbol{\delta}$ to resolve the rotation problem. In term of $\boldsymbol{\delta}$ this mean, that the leading elements $l_1, \ldots, l_r$ of $\boldsymbol{\delta}$ are not ordered yet, they are only required to be different, i.e. to occupy different rows. Hence, they allow for trivial non-identification, i.e. column and sign switching during MCMC sampling, which is resolved in a post-processing manner, by reordering the columns of $\boldsymbol{\delta}$ such that the permuted leading elements define an ordered GLT condition, i.e. $l_1 < \cdots < l_r$. The sign of the leading element is used to resolve the sign switching problem. The resulting posterior draws need not be variance identified, hence posterior stratification as described above is necessary, before further inference is carried out.

TABLE 1. Prior distribution $p(r)$.

r	0	1	2	3	4	5	6	7
$p(r)$	0.017	0.110	0.258	0.301	0.215	0.083	0.015	0.001

4 Illustrative application

For illustration, the data considered in Press and Shigemasu (1989), Table 1 which are scores on a ten-point scale on $m = 15$ characteristics of $T = 48$ applicants are reanalyzed. Press and Shigemasu (1989) postulate that a 4 factor model is appropriate for these data and we want to compare this finding with sparse Bayesian factor modeling. As in Press and Shigemasu (1989), the data are standardized. Since the number of factors is unknown, an overfitting factor model is applied with $k = 7$ being equal to the maximum number of possible factor given by inequality (6).

The hyperparameters of the prior (11) for indicators are chosen as $\gamma = b_0 = 0.3$ and $\alpha = 3$, which yields $a_0 = 0.1286$. The corresponding prior distribution $p(r)$, reported in Table 1, was determined by simulating $\boldsymbol{\delta}$ from the prior and rejecting all draws that are not variance identified.

To study sensitivity of factor selection with respect to further prior choices, the prior on the idiosyncratic variances where $c_0 = 2.5$ and C_{0i} is selected to avoid a Heywood problem, see (12), is combined with various fractional priors, defined in (15), where $b = b_N = 1.4 \cdot 10^{-3}$, $b = b_R = 9.1 \cdot 10^{-5}$, and $b = b_p = 10^{-p}$ with p varying from 2 to 5. In addition, a conventional prior is considered where $\mathbf{b}_{i0}^{\boldsymbol{\delta}} = \mathbf{0}$, $\mathbf{B}_{i0}^{\boldsymbol{\delta}} = \mathbf{I}$, and $c_0 = 1.1$ and $C_{i0} = 0.055$ (Lopes and West, 2004)).

An unrestricted MCMC scheme is used to obtain $M = 100000$ draws after a burn-in phase of $M_0 = 50000$. To verify convergence, independent MCMC chain were run starting respectively with $r^{(0)} = 1$ and $r^{(0)} = 7$ non-zero columns and functional of $\boldsymbol{\Omega}$ were used to monitor MCMC convergence. The inefficiency factor τ_d of the model size d defined in (17) may be used to evaluate mixing of the MCMC scheme, see Table 2. Apart from $\boldsymbol{\Omega}$, no further parameters are identified from the unrestricted draws. Hence, for further inference, we use only the draws that are variance identified. As reported in Table 2, the fraction p_V of these draws is reasonably high.

For variance identified draws, the number of non-zero columns of $\boldsymbol{\delta}$ may be regarded as a draw of the number of factors r. Table 2 reports the posterior distribution $p(r|\mathbf{y})$ for different priors as well as the posterior expectation $\hat{d} = \mathrm{E}(d|\mathbf{y})$ of the model size d. The prior is very influential on the number of factors and the model size for this data set and the number of estimated factors ranges from 3 to 5. This is not surprising, given the fact that we have only $T = 48$ observations. Fractional priors based on $b = b_5 = 10^{-5}$, $b = b_R$, and $b = b_4 = 10^{-4}$ point at a three factor solution. Larger values of

TABLE 2. Bayesian inference for an unknown number of factors (maximum number of factors $k = 7$) for various fractional priors with $b = b_N = 1.4 \cdot 10^{-3}$, $b = b_R = 9.1 \cdot 10^{-5}$ and $b = b_p = 10^{-p}$, $p = 2, \ldots, 5$; LW: priors as in Lopes and West (2004). Posterior inference for draws $\boldsymbol{\delta}$ from the unrestricted sampler that obey variance identification (p_V fraction in percent). Posterior distribution $p(r|\mathbf{y})$ of the number r of factors (bold number corresponding to the posterior mode $\tilde{r}$), posterior expectation $\hat{d} = \mathrm{E}(d|\mathbf{y})$ of the model size d; fraction p_G (in percent) of the variance identified draws $\boldsymbol{\delta}$ that obey a GLT constraint; inefficiency factor τ_d of the posterior draws of the model size d. Probabilities smaller than 10^{-2} are indicated by ≈0.

	$p(r\|\mathbf{y})$									
	2	3	4	5	6	7	$\hat{d}$	p_V	p_G	τ_d
b_2	0	0	0.08	**0.83**	0.09	≈0	40	48.6	0.6	18.2
b_N	0	≈0	**0.82**	0.18	≈0	0	30	56.9	7.2	20.6
b_3	0	≈0	**0.86**	0.14	≈0	0	29	57.8	9.63	19.7
b_4	0	**0.57**	0.43	≈0	0	0	21	79.2	71.9	32.5
b_R	≈0	**0.63**	0.37	≈0	0	0	21	78.9	74.1	30.3
b_5	0.11	**0.89**	≈0	0	0	0	17	86.7	95.1	14.5
LW	0	0	0.38	**0.49**	0.13	≈0	45	56.1	0.2	27.4

b extract more information from the likelihood than smaller values, which reduces the influence of the sparsity prior $p(\boldsymbol{\delta})$ as b increases, leading to a larger number of estimated factors. The fractional prior with $b = b_2 = 10^{-2}$, for instance, as well as the standard prior (LW) lead to an overfitting model with 5 factors.

As a next step, the factor loadings are identified. To this aim, the GLT constraint is imposed as in Frühwirth-Schnatter and Lopes (2015). One way to impose the GLT constraint is to stratify the posterior draws obeying the variance identification further according to whether the posterior draws $\boldsymbol{\delta}$ display a GLT structure or not. Table 2 report the corresponding fraction p_G (in percent of the variance identified draws) for all priors. Whereas many of the three factor solutions are GLT, only a small fractions of posterior draws display a GLT structure for models with a higher number of factors. Alternatively, after fixing the number of factors r, an MCMC scheme can be applied which forces the GLT constraint during sampling by navigating through the space of all admissible leading elements $(l_1, \ldots, l_r)$. Bayesian inference, based on $M = 50000$ draws after a burn-in phase of $M_0 = 30000$, is reported in Table 3 for the various priors, with r being equal to the posterior mode estimator $\tilde{r}$ reported in Table 2.

The total number N_v of visited models is determined as well as the highest probability model (HPM), i.e. the indicator matrix $\boldsymbol{\delta}_H$ visited most often and its frequency p_H which may be regarded as an estimator of the posterior

TABLE 3. Bayesian inference for a known number of factors r (based on the posterior mode $\tilde{r}$ obtained from Table 2) under the GLT constraint with unknown leading elements $\mathbf{l} = (l_1, \ldots, l_r)$. Total number of visited models N_v; leading elements $\mathbf{l}^\star$ visited most and corresponding frequency p_L (in percent); frequency p_H (in percent), leading elements $\mathbf{l}_H$ and model size d_H of the HPM (model visited most often); model size d_M of the MPM (median probability model) conditional on $\mathbf{l}^\star$.

Prior	r	N_v	$\mathbf{l}^\star$	$100p_L$	$100p_H$	$\mathbf{l}_H$	d_H	d_M
LW	5	27614	1,2,3,4,6	74.8	0.09	1,2,3,4,6	35	37
b_2	5	21565	1,2,3,4,6	83.8	0.16	1,2,3,4,6	31	31
b_N	4	19764	1,2,3,4	91.8	3.7	1,2,3,4	24	24
b_3	4	17619	1,2,3,4	83.6	3.3	1,2,3,4	24	23
b_4	3	4576	1,2,4	86.5	9.0	1,2,4	18	18
b_R	3	8718	1,2,4	91.4	6.9	1,2,4	19	20
b_5	3	2023	1,2,4	78.1	13.9	1,2,4	17	17

probability of this model. In addition, the leading elements $\mathbf{l}_H$ and the model size d_H of the HPM are reported.

Since the sampler is searching in the space of all GLT matrices, further stratification is necessary to identify $\mathbf{\Lambda}$. To this aim, the identifiability constraint $\mathbf{l}^\star = (l_1^\star, \ldots, l_r^\star)$ visited most often together with its frequency p_L is reported in Table 3. Typically, $\mathbf{l}^\star$ coincides with the leading elements $\mathbf{l}_H$ of the HPM. For all 4-factor models, the standard PLT constraint $\mathbf{l}^\star = (1, 2, 3, 4)$ turns out to be the most likely constraint, whereas for all other models GLT constraints are preferred, e.g. the constraint $\mathbf{l}^\star = (1, 2, 4)$ for a 3-factor model. Inference with respect to $\mathbf{\Lambda}$ is based on all posterior draws where the leading elements of $\boldsymbol{\delta}$ coincide with $\mathbf{l}^\star$ and all further Bayesian inference is performed conditional on $\mathbf{l}^\star$.

For illustration, detailed results are discussed for a 3-factor model based on the GLT constraint $\mathbf{l}^\star = (1, 2, 4)$ and the fractional prior $b = b_4 = 10^{-4}$. Table 4 reports the marginal inclusion probabilities $\Pr(\delta_{ij} = 1|\mathbf{y}, \mathbf{l}^\star)$ for all elements of the indicator matrix and displays considerable sparsity, with inclusion probabilities close to 0 for many loadings. The median probability model (MPM) is obtained by setting each indicator $\delta_{ij} = 1$, whenever $\Pr(\delta_{ij} = 1|\mathbf{y}, \mathbf{l}^\star) \geq 0.5$. The model size d_M of the MPM is also reported in Table 3. Under this prior, the median probability model turns out to be identical with the model visited most often.

Finally, Table 5 shows the posterior means of the factor loading matrix, the idiosyncratic variances and the communalities defined in (16), which are obtained by averaging over all draws where the leading elements of $\boldsymbol{\delta}$ coincide with $\mathbf{l}^\star$. Sign switching in the posterior draws of $\mathbf{\Lambda}$ is resolved through the constraint $\lambda_{11} > 0$, $\lambda_{22} > 0$, and $\lambda_{43} > 0$. As expected, non-zero factors loading in Table 4 have relatively high communalities for the different items,

TABLE 4. Inclusion probabilities for the indicator matrix $\boldsymbol{\delta}$ for a 3-factor model with fractional prior $b = b_4 = 10^{-4}$.

i	Item	Factor 1	Factor 2	Factor 3
1	application letter	**1**	0	0
2	appearance	0	**1**	0
3	academic ability	0.01	0.02	0
4	likeability	0.06	0.27	**1**
5	self confidence	**0.96**	**1**	0.01
6	lucidity	0.06	**1**	0.02
7	honesty	0.37	0.12	**1**
8	salesmanship	0.01	**1**	0.01
9	experience	**1**	0.04	0.01
10	drive	0.11	**1**	0.01
11	ambition	0.01	**1**	0.01
12	grasp	0.01	**1**	0.03
13	potential	0.03	**1**	**0.60**
14	keenness to join	0.02	**0.85**	**0.99**
15	suitability	**1**	**0.95**	0.01

TABLE 5. Posterior means of the factor loading matrix, the idiosyncratic variances and the communalities (in percent) for a 3-factor model with $b = b_4 = 10^{-4}$ under the GLT constraint $l_1 = 1$, $l_2 = 2$, $l_3 = 4$. Loadings with absolute value smaller than 10^{-2} are indicated by ≈ 0.

		Parameters				Communalities		
i	Item	λ_{i1}	λ_{i2}	λ_{i3}	σ_i^2	R_{i1}	R_{i2}	R_{i3}
1	application letter	0.66	0	0	0.54	44.7	0	0
2	appearance	≈ 0	0.48	0	0.73	0.08	25.7	0
3	academic ability	≈ 0	≈ 0	≈ 0	0.96	0.08	0.08	0.03
4	likeability	0.02	0.19	0.78	0.17	0.48	13	69.4
5	self confidence	-0.34	0.85	0.11	0.16	10.1	67.3	9.09
6	lucidity	-0.01	0.79	0.11	0.21	0.34	68.8	9.34
7	honesty	-0.15	0.13	0.63	0.46	5.14	7.47	43
8	salesmanship	≈ 0	0.82	0.11	0.16	0.15	74.2	9.99
9	experience	0.79	≈ 0	≈ 0	0.37	62.3	0.15	0.12
10	drive	0.03	0.75	0.1	0.27	0.99	63	8.5
11	ambition	≈ 0	0.8	0.12	0.18	0.23	72	9.6
12	grasp	≈ 0	0.78	0.11	0.22	0.26	68.3	9.2
13	potential	≈ 0	0.74	0.24	0.21	0.29	62.8	13.7
14	keenness to join	≈ 0	0.39	0.51	0.41	0.20	21.5	32.6
15	suitability	0.73	0.32	0.04	0.2	60.5	14.5	1.9

whereas for zero factors loading the communalities are practically 0. The factors have a similar interpretation as in Table 9.6 of Rowe (2003), namely that factor 1 is a measure of position match (application letter, experience, suitability), factor 2 could be described as charisma (lucidity, salesmanship, drive, ambition, grasp, potential), whereas factor 3 is a measure of personality (likeability, honesty, keenness to join).

References

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*. Chichester: Wiley, 3rd ed.

Anderson, T.W. and Rubin, H. (1956). Statistical inference in factor analysis. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. V, 111 – 150.

Bhattacharya, A. and Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291 – 306.

Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, **103**, 1438 – 1456.

Conti, G., Frühwirth-Schnatter, S., Heckman, J.J., and Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics*, **183**, 31 – 57.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, **147**, 186 – 197.

Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2009). Opening the black box: Structural factor models with large cross sections. *Econometric Theory*, **25**, 1319 – 1347.

Frühwirth-Schnatter, S. and Lopes, H. (2015). Parsimonious Bayesian factor analysis when the number of factors is unknown. Research Report, Vienna University of Economics and Business.

Geweke, J.F. and Singleton, K.J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, **75**, 133 – 137.

Ghahramani, Z., Griffiths, T.L., and Sollich, P. (2007). Bayesian nonparametric latent feature models (with discussion and rejoinder). In: *Bayesian Statistics 8*, Oxford: Oxford University Press.

Ghosh, J. and Dunson, D.B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, **18**, 306 – 320.

Kaufmann, S. and Schuhmacher, C. (2014). Bayesian estimation of sparse dynamic factor models with order-independent identification. Research Report, Study Center Gerzensee, Switzerland.

Lopes, H.F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, **14**, 41 – 67.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B*, **57**, 99 – 138.

Press, S.J. and Shigemasu, K. (1989). Bayesian inference in factor analysis. In: *Contributions to Probability and Statistics. Essays in Honor of Ingram Olkin*, New York: Springer-Verlag, 271 – 287.

Rowe, D.B. (2003). *Multivariate Bayesian Statistics*. Boca Raton, FL: Chapman & Hall/CRC.

West, M. (2003). Bayesian factor regression in the large p, small n paradigm. In: *Bayesian Statistics 7*, Oxford: Oxford University Press, 307 – 326.

A general framework for functional regression

Sonja Greven[1]

[1] Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: sonja.greven@stat.uni-muenchen.de

Abstract: We propose a general framework for regression with functional responses and/or covariates. The framework is highly modular, allowing for a) different features of the conditional distribution of the response to be modeled, e.g. the mean or a quantile b) flexible linear or smooth effects of scalar and functional covariates as well as (functional) random effects c) choice of bases for each additive term, including splines and functional principal components d) choice of estimation method, where gradient boosting is well-suited to general loss functions, many covariates and variable selection, while mixed model-based estimation provides likelihood-based inference. Functional responses and covariates can be observed on equal grids for all functions, with missings or even sparsely. Methods are implemented in the R-packages `refund` and `FDboost`, respectively. We discuss two applications on bioprocess monitoring and the feeding behavior of pigs.

Keywords: Additive mixed models; Boosting; Functional data analysis; Functional principal component analysis; Signal regression.

1 Introduction

In recent years, many studies have collected functional data, where observations consist of curves instead of scalars. Some common examples include spectroscopy, where spectra are measured over a range of wavelengths, trajectories over time, and images from medical imaging. Developing methods to analyze such data, i.e. functional data analysis (e.g. Ramsay and Silverman, 2005), is thus becoming increasingly important. Functions are often observed on fine grids in some interval $\mathcal{T} \subset \mathbb{R}$, but sparsely sampled functions with curve-specific grid points and images, where $\mathcal{T} \subset \mathbb{R}^2$ or $\mathcal{T} \subset \mathbb{R}^3$, are also common.

Functional regression is of interest in many applications, including the two applications on bioprocess monitoring and the feeding behavior 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

pigs we discuss in Section 6. It occurs in several flavors as scalar-on-function, function-on-scalar or function-on-function regression, depending on whether the responses and/or covariates are functional. A recent overview on functional regression can be found in Morris (2015). Most previous work has focused on one of these three models, often further restricting the considered model to special cases such as a functional response with a single functional covariate. Among the most general model classes introduced is that developed in Morris and Carroll (2006), Meyer et al. (2015) and further work of this group for Bayesian mean regression of functional responses on equal grids, including random effects, scalar and functional covariates as model terms, and using (mostly) wavelet transformations of responses which are well-suited to spiky data.

This paper aims at introducing a general framework for functional regression that covers all three models and allows for gridded or sparsely observed functional data. In addition to scalar and functional covariates, the framework covers settings with correlated functional data requiring modeling via random effects. Our approach goes beyond mean regression for continuous-valued functional responses and enables e.g. robust or quantile regression using general loss functions as well as mean regression for 'generalized' e.g. Poisson or binomial process responses. We discuss estimation using mixed model inference or gradient boosting.

The idea is to build on recent advances made in flexible models for scalar data, such as additive mixed models or quantile regression, and extend these to functional data. Our aim is to provide a similarly flexible toolbox for functional regression as is available for scalar regression, and to supply an implementation in open source software. Large parts of this paper are based on Scheipl et al. (2014), Brockhaus et al. (2014) and more recent extensions, and are implemented in the R-packages `refund` (Crainiceanu et al., 2015) and `FDboost` (Brockhaus, 2015).

2 The model

In the following we consider data $(Y, X) \subset \mathcal{Y} \times \mathcal{X}$, where $\mathcal{Y}$ is a suitable space for the responses such as the space of square integrable functions $L^2(\mathcal{T}, \mu)$, and $\mathcal{X}$ is a product space of suitable spaces for the covariates. For functional response the domain $\mathcal{T}$ is an interval over the real numbers, $\mathcal{T} = [t_1, t_2]$, with $t_1,\ t_2 \in \mathbb{R}$, and μ is the Lebesgue measure. For scalar response the set $\mathcal{T}$ consists of a single point, $\mathcal{T} = [t, t]$, and μ is the Dirac measure. The spaces in $\mathcal{X}$ are defined analogously for scalar and functional covariates. We assume that Y given X follows a conditional distribution $F_{Y|X}$; the explanatory variables X may be fixed or random. As generic model we establish the following structured additive regression model:

$$\xi(Y|X = x) = h(x) = \sum_{j=1}^{J} h_j(x), \tag{1}$$

where ξ is some transformation function, for instance the expectation, the median or some quantile. For a generalized linear model the transformation function corresponds to the expectation composed with the link function g that connects response and linear predictor, i.e. $\xi = g \circ \mathbb{E}$. Extensions to vector-valued transformation functions that allow Generalised Additive Models for Location Scale and Shape (GAMLSS)-type regression models for functional data (e.g. modeling the mean and variance of a functional response) are discussed in Brockhaus et al. (2015). The linear predictor h is the sum of partial effects h_j which implies additivity. Note, however, that a partial effect h_j can depend on more than one covariate allowing e.g. for interactions. Each effect $h_j(x) \in \mathcal{Y}$ is a real-valued function over $\mathcal{T}$. Please see Table 1 for an overview of some common effects $h_j(x)$ that can be specified within this framework.

TABLE 1. Some possible effects $h_j(x)$ within model (1). Note that all effects can also be chosen to be constant in t.

covariate(s)	type of effect	$h_j(x)(t)$
(none)	smooth intercept	$\alpha(t)$
scalar covariate z	linear effect	$z\beta(t)$
	smooth effect	$\gamma(z,t)$
two scalars z_1, z_2	linear interaction	$z_1 z_2 \beta(t)$
	functional varying coefficient	$z_1 f(z_2,t)$
	smooth interaction	$f(z_1,z_2,t)$
functional covariate $x(s)$	linear functional effect	$\int x(s)\beta(s,t)ds$
	smooth functional effect	$\int F(x(s),s,t)ds$
scalar z and functional $x(s)$	linear interaction	$z\int x(s)\beta(s,t)ds$
	smooth interaction	$\int x(s)\beta(z,s,t)ds$
grouping variable g	functional random intercept	$b_g(t)$
grouping variable g and scalar z	functional random slope	$zb_g(t)$
curve indicator i	curve-specific smooth residual	$e_i(t)$

3 Specification of model terms

To represent the partial effects $h_j(x)(t)$, we use a basis that is tailored to the covariates and effect type at hand. We expand $h_j(x)(t)$ in the row tensor product of marginal basis vectors $\boldsymbol{b}_j$ of length K_j defined on $\mathcal{X} \times \mathcal{T}$ and $\boldsymbol{b}_Y$ of length K_Y defined on $\mathcal{T}$,

$$h_j(x)(t) = (\boldsymbol{b}_j(x,t)^\top \odot \boldsymbol{b}_Y(t)^\top)\boldsymbol{\theta}_j.$$

The row tensor product of an $m \times a$ matrix $\boldsymbol{A}$ and an $m \times b$ matrix $\boldsymbol{B}$ is defined as the $m \times ab$ matrix $\boldsymbol{A} \odot \boldsymbol{B} = (\boldsymbol{A} \otimes \mathbf{1}_b^{\mathrm{T}}) \cdot (\mathbf{1}_a^{\mathrm{T}} \otimes \boldsymbol{B})$, where $\cdot$ denotes element-wise multiplication. Note that for a single (x, t) pair, the row tensor product is actually the same as the Kronecker product. However, when considering the predictor for a vector of responses, the difference between the row tensor product and the Kronecker product becomes important, as then the dependence of the basis $\boldsymbol{b}_j$ on t means that the basis in covariate direction differs between rows. This more general formulation also allows for the grid in t direction to differ between functions. In cases where the basis for the covariate direction can be chosen independent of t and where the grid in t is the same for all curves, this can be simplified to

$$h_j(x)(t) = (\boldsymbol{b}_j(x)^{\mathrm{T}} \otimes \boldsymbol{b}_Y(t)^{\mathrm{T}})\boldsymbol{\theta}_j$$

and the model can be written in the form of a generalized linear array model (Currie et al., 2006), which can be used to make computations more memory and time efficient. For regularization, we use a quadratic penalty in $\boldsymbol{\theta}_j$. Other penalties are possible, but we use a penalty matrix of the form

$$\boldsymbol{P}_{jY} = \lambda_j(\boldsymbol{P}_j \otimes \boldsymbol{I}_{K_Y}) + \lambda_Y(\boldsymbol{I}_{K_j} \otimes \boldsymbol{P}_Y),$$

which allows for anisotropy of smoothness in covariate and t directions.
For illustration, we discuss some of the possible partial effects $h_j(x)(t)$ in more detail. First, consider a scalar covariate z, which is assumed to have a smooth effect on the functional response $Y(t)$ at each point t, i.e. $h_j(x)(t) = \gamma(z, t)$. Then, we can choose $\boldsymbol{b}_j(x)$ as a vector of B-spline basis evaluations in z with marginal penalty matrix $\boldsymbol{P}_j$ a suitable differencing matrix (Eilers and Marx, 1996). $\boldsymbol{b}_Y(t)$ and $\boldsymbol{P}_Y$ can be chosen likewise – or, if the effect $\gamma(z)$ is assumed to be constant over t or the response is scalar, $\boldsymbol{b}_Y(t) \equiv 1$ and $\boldsymbol{P}_Y = 0$. If the effect is assumed to be linear at each t, i.e. $h_j(x)(t) = z\beta(t)$, $\boldsymbol{b}_j(x) = z$ and $\boldsymbol{P}_j = 0$.
Second, consider the historical model where a functional response $Y(t)$ and a functional covariate $x(s)$ are observed over the same time interval $\mathcal{T}$ and $x(s)$ is believed to influence $Y(t)$ linearly in the time window $[t - \delta, t]$ before any given time point t. We can approximate this effect using numerical integration and marginal spline bases $\{\Phi_l^Y(t), l = 1, \ldots, K_Y\}$ and $\{\Phi_k^j(s), k = 1, \ldots, K_j\}$,

$$\begin{aligned} h_j(x)(t) = \int_{t-\delta}^{t} x(s)\beta(s,t)ds &\approx \sum_{k=1}^{K_j}\sum_{l=1}^{K_Y}\sum_{r:t-\delta\le s_r\le t} \Delta(s_r)x(s_r)\Phi_k^j(s_r)\Phi_l^Y(t)\theta_{j,kl} \\ &= \sum_{k=1}^{K_j}\sum_{l=1}^{K_Y}\sum_{r=1}^{R} \tilde{x}(s_r,t)\Phi_k^j(s_r)\Phi_l^Y(t)\theta_{j,kl}, \end{aligned} \tag{2}$$

where $\Delta(s)$ are weights of a numerical integration scheme and the transformed covariate observations $\tilde{x}(s_r, t)$ are defined as $\tilde{x}(s_r, t) = I(t - \delta \le$

$s_r \leq t)\Delta(s_r)x(s_r)$, with indicator function I. In this example, the basis $\boldsymbol{b}_j(x,t)$ depends on t and contains the terms $\sum_{r=1}^{R} \tilde{x}(s_r,t)\Phi_k^j(s_r)$, $k = 1,\ldots,K_j$, while $\boldsymbol{b}_Y(t)$ is again a vector of spline basis evaluations in t, $(\Phi_1^Y(t),\ldots,\Phi_{K_Y}^Y(t))^{\mathrm{T}}$. The penalty matrices $\boldsymbol{P}_j$ and $\boldsymbol{P}_Y$ are suitably chosen for the spline bases $\{\Phi_k^j(s),k\}$ and $\{\Phi_l^Y(t),l\}$, e.g. differencing matrices for B-splines.

Third, consider the setting of a functional random intercept $h_j(x)(t) = b_g(t)$ for a grouping variable g, which can be thought of as a random intercept at each point t with an additional smoothness assumption over $\mathcal{T}$. A special case is when the grouping variable g is simply a curve indicator i, in which case the term becomes a smooth residual curve $e_i(t)$. We can again model each random intercept using splines, $h_j(x)(t) = \sum_{l=1}^{K_Y} \Phi_l^Y(t)\theta_{j,gl}$ and the corresponding bases are an indicator vector for the groups g for $\boldsymbol{b}_j(x)$ and again $\boldsymbol{b}_Y(t) = (\Phi_1^Y(t),\ldots,\Phi_{K_Y}^Y(t))^{\mathrm{T}}$. While $\boldsymbol{P}_Y$ is a suitable penalty matrix for the splines in $\boldsymbol{b}_Y(t)$, $\boldsymbol{P}_j$ corresponds to a precision matrix defining the dependence structure between the levels of g. In the simplest case of i.i.d. $b_g(t)$, this is the identity matrix, but $\boldsymbol{P}_Y$ could also correspond e.g. to the precision matrix of a Gaussian Markov random field for spatially correlated curves.

Using splines in $\boldsymbol{b}_Y(t)$ can be computationally expensive if there are many groups g or curves i. Particularly if data sets are large, there is an advantage in using a more parsimonious basis, functional principal components (FPCs). In addition, these provide interpretable information on the main modes of variation in the data. Say we have i.i.d. random intercepts $b_g(t)$, i.i.d. realizations of a mean-zero stochastic process $b(t)$. We can then use the Karhunen-Loeve expansion

$$b_g(t) = \sum_{l=1}^{\infty} \theta_{j,gl}\phi_l(t),$$

truncated at a finite number K_Y in practice, where $\phi_l(t)$ are the eigenfunctions and ν_l the eigenvalues of the covariance operator of $b(t)$, and $\theta_{j,gl} \overset{i.i.d.}{\sim} \mathcal{N}(0,\nu_l)$ if $b(t)$ is Gaussian. This motivates the use of $\boldsymbol{b}_Y(t) = (\phi_1(t),\ldots,\phi_{K_Y}(t))^{\mathrm{T}}$, with $\boldsymbol{b}_j(x)$ as before. Due to the optimal approximation property of the eigenfunctions, K_Y can typically be chosen much smaller than for splines, which reduces computational complexity. $\theta_{j,gl} \overset{i.i.d.}{\sim} \mathcal{N}(0,\nu_l)$ also motivates the choice of $\boldsymbol{P}_Y = \boldsymbol{0}$ and $\boldsymbol{P}_j = \mathrm{diag}(\nu_1,\ldots,\nu_{K_Y})^{-1}$ here, with λ_j fixed to 1. This further reduces computational complexity as there is no need to estimate any smoothing parameters. Greven et al. (2010); Cederbaum et al. (2014); Shou et al. (2014) discuss the estimation of the eigenfunctions and eigenvalues for functional random effects models of varying complexity and grid or sparse functional data. Zipunnikov et al. (2014) discuss the extension to image data. The general idea is, after estimation of the mean structure and centering, to use cross-products

$Y_i(s)Y_j(t)$ as estimators of the overall covariance, then to decompose the covariance into the additive contributions from the random intercepts / slopes, smooth residuals and additional white noise, using a least squares approach. Smoothing of covariances and an eigendecomposition on a grid of values in $\mathcal{T}$ then yields estimates that we can subsequently use in our basis $\boldsymbol{b}_Y(t)$.

4 Estimation

After representation of all model terms $h_j(x)(t)$ using suitable bases, we can estimate model (1) as a penalized regression problem and make use of the recent developments in inferential methods for additive regression. For the observed data $y_i(t_{id})$, let $i = 1, \ldots, N$ index the curves, let $t_{id}, d = 1, \ldots, D_i$ denote the grid for curve i and let $\boldsymbol{Y} = (y_i(t_{id}))_{i=1,\ldots,N,d=1,\ldots,D_i}$ be the vector containing all observations.

4.1 Mixed model-based inference

If our transformation function can be written as $\xi = g \circ \mathbb{E}$ for a link function g, we can write our model for the observed data as

$$g(\mathbb{E}(\boldsymbol{Y})) = \boldsymbol{X}\boldsymbol{\theta}$$

with a suitably concatenated design matrix $\boldsymbol{X}$ containing the entries for $(\boldsymbol{b}_j(x_i, t_{id})^{\mathrm{T}} \odot \boldsymbol{b}_Y(t_{id})^{\mathrm{T}})$ ranging over $i = 1, \ldots, N$ and $d = 1, \ldots, D_i$ in rows, and concatenating design matrices column-wise for the partial effects $h_j(x)(t)$, $j = 1, \ldots, J$. The vector of unknown coefficients $\boldsymbol{\theta}$ contains blocks $\boldsymbol{\theta}_j$ of coefficients for each j.
The penalized log-likelihood for this model is given by

$$l(\boldsymbol{\theta}) - \frac{1}{2}\sum_{j=1}^{J} \boldsymbol{\theta}_j^{\mathrm{T}} \boldsymbol{P}_{jY} \boldsymbol{\theta}_j,$$

where the log-likelihood $l(\boldsymbol{\theta})$ is obtained from the assumed conditional density of $\boldsymbol{Y}$ given $\boldsymbol{X}$. We follow the approach of Wood (2006, 2011, 2014); Wood et al. (2015) for optimization of this penalized log-likelihood; for more details, see Scheipl et al. (2014). We build on the methods for generalized additive mixed models available in the R package `mgcv` (Wood, 2015) for our implementation in the `pffr` function of the R package `refund`. One advantage of this approach is the availability of confidence intervals and tests using mixed model-based likelihood inference methodology.

4.2 Boosting

The idea for estimation using boosting is to represent the estimation problem with an adequate loss function. Common loss functions include the squared error loss for mean regression of normal responses, the absolute loss for median regression, the check function for quantile regression and the negative log-likelihood for responses of the exponential family. To obtain a suitable loss function for functional responses, the (weighted) scalar loss function at each t is integrated over $\mathcal{T}$.

A component-wise boosting algorithm (Bühlmann and Hothorn 2007) is then used to optimize this loss function, exploiting the array structure where possible following Hothorn et al. (2013). The empirical loss is iteratively minimized along the steepest gradient. $\hat{h}(\cdot)$ is updated along an estimate of the negative gradient vector $\boldsymbol{U}$, where $\boldsymbol{U}$ is estimated using so-called base-learners, penalized regression models for partial effects $h_j(\cdot)$ in our context. Then, the best fitting base-learner $j^\star$ is selected and parameters are updated for $h_{j^\star}(\cdot)$. The final $\hat{h}(\cdot)$ is a linear combination of base-learner fits. The number of iterations controls the model complexity (for fixed small step-length). The stopping iteration is chosen by resampling methods on the level of curves. Variable selection results from both early stopping and additional stability selection (Meinshausen and Bühlmann 2010; Shah and Samworth 2013). This approach is implemented in the R package `FDboost` and builds on model-based boosting as implemented in the R package `mboost` (Hothorn et al., 2014). For more details, see Brockhaus et al. (2014).

Advantages of boosting include the possibilities of variable selection during estimation, including being able to cope with more variables than observed curves, and the ability to handle general loss functions, enabling regression models for other features ξ of the conditional distribution than the mean such as the median or quantiles. In addition, the computationally more efficient array structure of the generalized linear array model (Currie et al., 2006) is currently exploited only in our implementation using boosting for estimation. A drawback of boosting is its lack of formal inference, which we address by bootstrapping.

5 Identifiability

In regression with functional responses and/or covariates, identifiability has to be carefully considered. The first point concerns functional response models including at least two $h_j(x)(t)$ varying over t. This is easiest to see when the model includes a smooth intercept $\alpha(t)$. As

$$\alpha(t) + h_j(x)(t) = [\alpha(t) + \bar{h}_j(t)] + [h_j(x)(t) - \bar{h}_j(t)] = \widetilde{\alpha}(t) + \widetilde{h}_j(x)(t)$$

with $\bar{h}_j(t) = \frac{1}{n}\sum_{i=1}^{n} h_j(x_i)(t)$, a constraint on h_j such as $\bar{h}_j(t) = 0$ for all t is needed and can be incorporated by appropriately modifying the design

matrix, cf. Wood (2006, sec. 1.8.1), and Brockhaus et al. (2014) on how to preserve the array structure.
The second point is more serious, as it is less easily remedied. It concerns the regression with functional covariates, see Scheipl and Greven (2012) for a more detailed discussion. Consider for simplicity a linear functional effect, in contrast to (2) integrating over the whole domain $\mathcal{S}$ of $x(s)$, and equal grids of length D for all functional responses. The basis vector $\boldsymbol{b}_j(x)$ then contains the terms $\sum_{r=1}^{R} \Delta(s_r)x(s_r)\Phi_k^j(s_r)$, $k = 1, \ldots, K_j$. It is often the case in practice that the functional covariate $x(s)$ can be well approximated by the first M components of the Karhunen-Loeve-expansion,

$$x_i(s) = \sum_{l=1}^{\infty} \xi_{i,l}\phi_l^X(s) \approx \sum_{l=1}^{M} \xi_{i,l}\phi_l^X(s),$$

as in any case at most $\min(N, D)$ eigenfunctions with non-zero eigenvalues can be estimated from the data. If $M < K_j$ or $\text{rank}(\boldsymbol{\phi}^X\boldsymbol{\Delta}\boldsymbol{\Phi}^j) < K_j$, where $\boldsymbol{\phi}^X = (\phi_l^X(s_r))_{l=1,\ldots,M;r=1,\ldots,R}$, $\boldsymbol{\Delta} = \text{diag}(\Delta(s_1), \ldots, \Delta(s_R))$ and $\boldsymbol{\Phi}^j = (\Phi_k^j(s_r))_{r=1,\ldots,R;k=1,\ldots,K_j}$, there is not sufficient information in $x(s)$ for the K_j spline basis coefficients and the resulting design matrix will be rank-deficient. The penalized likelihood criterion then effectively finds the smoothest solution among the possible solutions of the optimization problem. This solution will be unique as long as the kernel of the penalty does not overlap the kernel of the design matrix. This leads to several practical recommendations. The first is to avoid rank-reducing pre-processings of functional covariates, e.g. pre-smoothing or curve-wise centering, where possible. The second is to compute diagnostic measures to check for any problems in practice; such measures are implemented in the `refund` and `FDboost` packages. And the third is to keep the kernel of the penalty small by using first-order differences or derivatives (penalizing deviations from a constant coefficient function) or by using modified full-rank penalties such as those originally introduced in Marra and Wood (2011) for variable selection.

6 Applications

6.1 The functional historical model for bioprocess monitoring

The aim of this study (Melcher et al., 2014) is the monitoring of the cell dry mass (CDM) during a biotechnological fermentation process for the production of the model protein human Cu/Zn superoxide dismutase (h-SOD). Traditionally the CDM is determined offline in cost- and labour-intensive measurements (i.e., by taking and analyzing a sample), which prevents an early fault diagnosis. On the other hand, various physical process variables (process data (PD): feed-, base- or O_2-consumption etc.)

can be measured easily without significant costs and additional chemical information is obtained online via two-dimensional multi-wavelength fluorescence spectroscopy (Bioview data, BV) and proton-transfer-reaction mass spectrometry (PTR-MS). Thus, the goal is the online prediction of the CDM from these up to 160 continuously measurable signals. Our data set consists of 25 E. coli bacterial fermentations with varying values for the factors temperature, induction strength and growth rate. In Figure 1 the response variable CDM is plotted and as an example for the preprocessed covariates the variables feed consumption and em590.ex550 – one of the variables in the Bioview data – are plotted. As time variable the number of generations of the bacteria is used, computed as generation = exp(feeding time · growth rate).

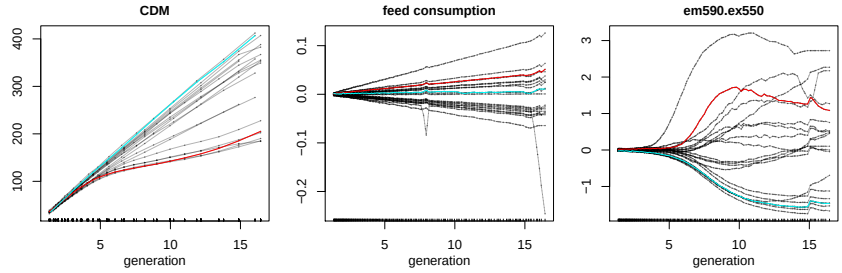

FIGURE 1. In the left plot the response variable CDM is plotted. The other two plots show two functional covariates of the PD and the BV dataset.

As during a fermentation only observations up to the current time-point can be used, we fit historical models using the online measured variables as covariates to predict the CDM,

$$E(Y_i(t)|\boldsymbol{x}_i) = \beta_0(t) + \sum_{j=1}^{J} \int_{t_1}^{t} x_{ji}(s)\beta_j(s,t)\,\mathrm{d}s, \tag{3}$$

where $Y_i(t)$ is the response at generation t, $t \in [t_1, t_2] = [1.2, 16.4]$, for fermentation i, $\boldsymbol{x}_i$ is the vector of covariates, $x_{ji}(s)$ is the observation of fermentation i and functional covariate j at generation s and $\beta_j(s,t)$ is the coefficient surface of covariate j which has triangular support.
We use boosting for estimation, which allows for the number of covariates to far exceed the number of functional observations. We determine the stopping iteration using leave-one-curve-out-cross-validation. Variable selection is crucial in this data set with 25 observations and 160 potential covariates. Using stability selection (Shah and Samworth, 2013) in addition to early stopping results in a model with three variables, see Figure 2. All three variables have a rank deficient design matrix but no kernel overlap with the penalty, indicating that the penalty leads to unique solutions that are chosen on the basis of smoothness as an indicator of higher plausibility. As an

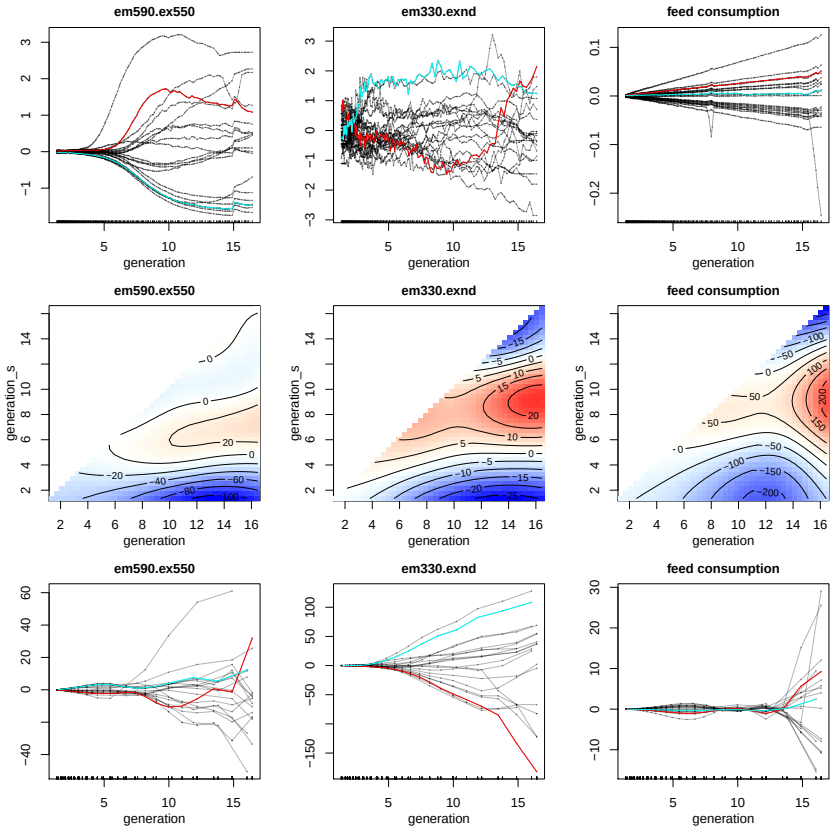

FIGURE 2. The functional covariates x_j (top panel), the estimated coefficient surfaces $\hat{\beta}_j(s,t)$ (middle panel, with the x-axis giving the t-direction and the y-axis the s-direction) and the partial effects $\int_{t_1}^{t} x(s)\hat{\beta}_j(s,t)\,ds$ (bottom panel) are depicted for the historical model using the PD and BV variables that are selected by stability selection. The estimated coefficient functions are colored in red for positive effects, blue for negative effects and white for zero. Note that the scale in each plot is different.

example we interpret the estimated coefficient surface of em330.exnd as it is the effect inducing the highest partial effects, see Figure 2 bottom center. Because of the estimated negative coefficient surface for low s, smaller observations for em330.exnd in the beginning are connected to higher values of CDM later, most strongly for time-points around generation 13. High values for em330.exnd in the middle of the fermentations are associated with higher values of CDM subsequently. At the end of the fermentations the effect is negative again. Looking at the estimated coefficients, a lag model with fixed lag δ does not seem adequate as it would set the coefficient surface to zero for $s < t - \delta$, resulting in a coefficient band under the

diagonal. When we select the model not based on stability selection but on optimal prediction (results not shown), we obtain a model with 17 variables and considerably lower root functional mean squared error (RFMSE = 8.7) than for the model with only three covariates (RFMSE = 33.0) discussed for simplicity above.

6.2 A generalized functional regression model for pig feeding behavior

In the Pigwise project (Maselyne et al., 2014, Gertheiss et al., 2014), data on pigs' feeding behavior is collected using HF RFID antennas installed over the pig troughs and RFID tags that pigs are equipped with. The resulting data gives binary information on the proximitiy of the pigs to the trough (yes or no) every 10 seconds for 127 pigs over 102 days. Figure 3 shows the data for a single pig. Data for each day can be thought of as binary functional data, i.e. as realizations of binary rather than continuous-valued random processes over the [0h,24h] time window. Our aim is to model the feeding behavior, which might give insights into the ethology of pigs and by distinguishing common and uncommon feeding behavior be useful in monitoring individual pigs' health.

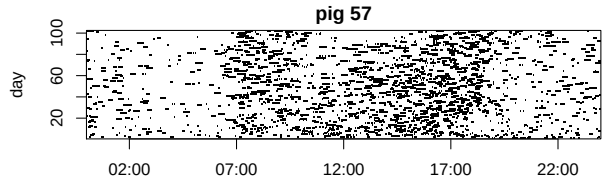

FIGURE 3. Proximity to the trough (dots for yes) for pig 57. The horizontal axis indicates the time of day and the vertical axis gives the age of the pig in days.

For illustration, we here focus on a model for the feeding behavior of one pig only, using pig 57 as shown in Figure 3. To reduce data size while keeping sufficient temporal resolution, we aggregate the binary feeding indicators $\tilde{y}_i(t)$ observed every 10 seconds over 10 minute intervals and assume a binomial conditional distribution with $n = 60$ and probability $p_i(t)$ smooth over t for the resulting processes $Y_i(t)$, $i = 1, \ldots, 102$. In our binomial logit model, we include independent functional random day effects $b_i(t)$ as well as an auto-regressive term assuming that feeding behavior over the three previous hours may influence the current feeding rate,

$$\text{logit}\left(p_i(t)\right) = \beta_0(t) + b_i(t) + \int_{t-3\text{h}}^{t-10\text{min}} y_i(s)\beta(t,s)ds.$$

To enforce similarity of all effects between 0h and 23:59h, we use periodic P-spline bases over t. Fitting this model with five smoothing parameters,

850 spline coefficients and $nD = 9648$ observations takes about 1.5h using `pffr` in `refund`.
The resulting model reasonably reproduces the observed feeding episodes by showing peaks in the estimates $\hat{p}_i(t)$ where spikes occur in $y_i(t)$. This is shown for 12 selected days in Figure 4. Our model explains ca. 41% of the deviance in the training data and yields an average Brier score of ca. 0.019. Prediction for the validation data is more challenging (not shown, Brier score: 0.029) and succeeds better for some days than for others.This is likely at least partly due to the functional random effect $b_i(t)$ being predicted as zero for new observations. If prediction is the focus, a smooth day effect $f(i, t)$ is somewhat more successful at prediction (results not shown) than a random day effect.

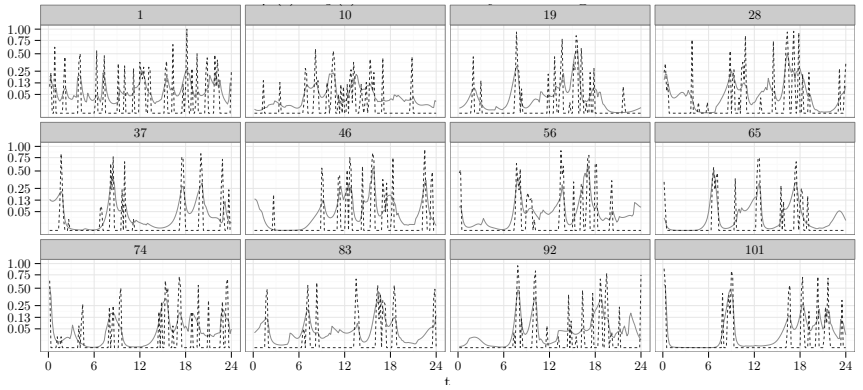

FIGURE 4. Fitted values $\hat{p}_i(t)$ (solid lines) and observed feeding rates $y_i(t)/60$ (dashed lines) for pig 57 for 12 days indicated above the panels.

The estimated additive model terms and estimated confidence bands for intercept and auto-regressive effect are shown in Figure 5. The functional intercept in the top left panel of Figure 5 shows an increase in feeding activity in the morning around 10h and in the evening around 18h. Feeding rates are low during the night from about midnight to 6h. The top right panel of Figure 5 shows the estimated functional random day effects, with some unrealistically large effect sizes for the hours of the night, which are possibly due to the very low probabilities of feeding during that time. The bottom panel of Figure 5 shows the estimated cumulative autoregressive effect. This indicates that feeding in the immediate past is predictive of current feeding especially during times of low average feeding probability. In the early morning hours from ca. 0h to 6h, there is a strong negative association between feeding ('getting up once') and feeding 1-2 hours later ('getting up a second time').

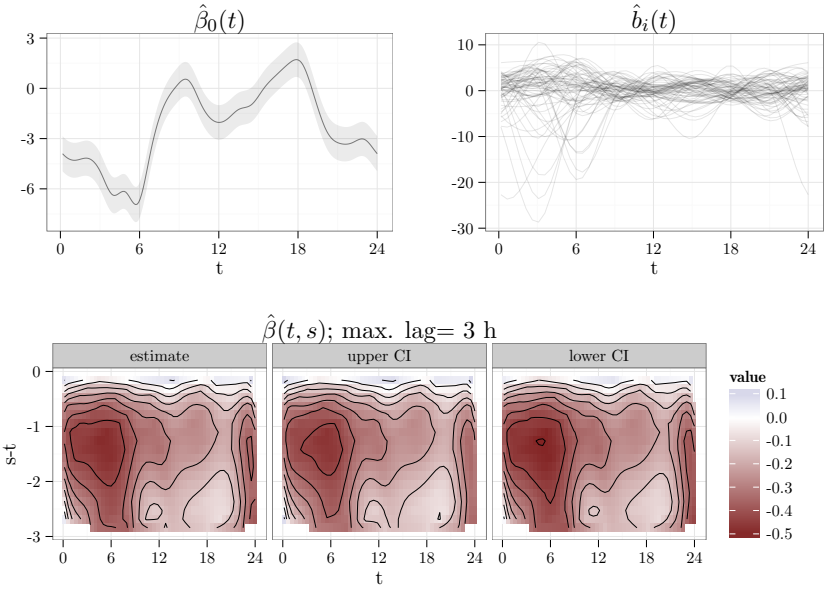

FIGURE 5. Top: Estimated functional intercept and functional random day effects. Bottom: Estimated coefficient surface for the cumulative auto-regressive effect. Confidence intervals are given based on ± 2 the standard errors.

7 Discussion

We provide a model class for functional regression that is highly flexible and modular. The transformation function allows us to choose the feature of the conditional distribution of the scalar or functional response that is to be modeled. We can even extend this to a vector of features to obtain a GAMLSS framework for functional regression, see Brockhaus et al. (2015). The partial predictors $h_j(x)(t)$ allow very flexible models for linear or smooth effects of scalar and functional covariates as well as (functional) random effects. Different bases such as splines and FPC bases can be chosen within this general framework. And finally, the choice of estimation method allows estimation via boosting – which is ideally suited to a large number of covariates and variable selection as well as offering a wide range of loss functions and thus transformation functions – or via mixed model methodology, offering the full range of developments that have been made for generalized additive mixed models in the last years as well as likelihood-based inference.

Acknowledgments: Different parts of this work are joint work with Sarah Brockhaus, Jan Gertheiss, Torsten Hothorn, Friedrich Leisch, Michael Melcher, Fabian Scheipl and Ana-Maria Staicu. S. Greven, S. Brockhaus and F. Scheipl are grateful to the German Research Foundation (DFG) for funding through Emmy Noether grant GR 3793/1-1. We would like to thank Engel Hessel for providing the Pigwise data. The pig feeding data were generated in the framework of the ICT-AGRI era-net project Pigwise "Optimizing performance and welfare of fattening pigs using High Frequent Radio Frequency Identification (HF RFID) and synergistic control on individual level" (Call for transnational research projects 2010). The German contribution was funded by the German Federal Office for Agriculture and Food (BLE).

References

Brockhaus, S. (2015). FDboost: Boosting functional regression models. *R package* version 0.0-8, available at http://cran.r-project.org/web/packages/FDboost/.

Brockhaus, S., Scheipl, F., Hothorn, T., and Greven, S. (2014). The functional linear array model. *Statistical Modelling*, in press, doi:10.1177/1471082X14566913.

Brockhaus, S., Fuest, A., Mayr, A., and Greven, S. (2015). Functional regression models for location, scale and shape with application to stock returns. In: *Proceedings of the 30th International Workshop on Statistical Modelling*, Linz.

Bühlmann, P., and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477 – 505. With discussion.

Cederbaum, J., Greven, S., Pouplier, M., and Hoole, P. (2014). Functional linear mixed model for irregularly spaced phonetics data. In: *Proceedings of the 29th International Workshop on Statistical Modelling*, Göttingen, 75 – 80.

Crainiceanu, C.M., Reiss, P., Goldsmith, J., Huang, L., Huo, L., and Scheipl, F. (2015). refund: Regression with functional data. *R package* version 0.1-1, available at https://github.com/∼refunders/refund.

Currie, I.D., Durban, M., and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, **68**, 259 – 280.

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

Gertheiss, J., Maier, V., Hessel, E.F., and Staicu, A.-M. (2014). Modeling binary functional data with application to animal husbandry. In: *Proceedings of the 29th International Workshop on Statistical Modelling*, Göttingen, 139 – 144.

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

Hothorn, T., Kneib, T., and Bühlmann, P. (2013). Conditional transformation models. *Journal of the Royal Statistical Society, Series B*, **76**, 3 – 27.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2014). mboost: Model-based boosting. *R package* version 2.3-0, available at http://cran.r-project.org/web/packages/mboost/

Marra, G. and Wood, S.N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, **55**, 2372 – 2387.

Maselyne, J., Saeys, W., Ketelaere, B.D., Mertens, K., Vangeyte, J., Hessel, E.F., Millet, S., and Van Nuffel, A. (2014). Validation of a high frequency radio frequency identification (HF RFID) system for registering feeding patterns of growing-finishing pigs. *Computers and Electronics in Agriculture* , **102**, 10 – 18.

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

Melcher, M., Scharl, T., Spangl, B., Luchner, M., Cserjan, M., Bayer, K., Leisch, F., and Striedner, G. (2014). The potential of random forest and neural networks for biomass and recombinant protein modeling in E. coli fed-batch fermentations. Unpublished manuscript.

Meyer, M.J., Coull, B.A., Versace, F., Cinciripini, P., and Morris, J.S. (2015). Bayesian function-on-function regression for multi-level functional data. *Biometrics*, in press, available at http://works.bepress.com/jeffrey_s_morris/52.

Morris, J.S. (2015). Functional regression. *Annual Review of Statistics and its Applications*, **2**, in press, doi: 10.1146/annurev-statistics-010814-020413.

Morris, J.S. and Carroll, R.J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, **68**, 179 – 199.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis (2nd edition)*. New York: Springer.

Scheipl, F., Staicu, A.-M., and Greven, S. (2014). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, in press, doi: 10.1080/10618600.2014.901914.

Scheipl, F. and Greven, S. (2012). Identifiability in penalized function-on-function regression models. *Department of Statistics: Technical Reports*, No. 125, http://epub.ub.uni-muenchen.de/13060/.

Shah, R.D. and Samworth, R.J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society, Series B*, **75**, 55 – 80.

Shou, H., Zipunnikov, V., Crainiceanu, C., Greven, S. (2014). Structured functional principal component analysis. *Biometrics*, in press, doi: 10.1111/biom.12236.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman & Hall/CRC.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B*, **73**, 3 – 36.

Wood, S.N. (2014). General smooth additive modeling. In: *Proceedings of the 29th International Workshop on Statistical Modelling*, Göttingen, 43 – 50.

Wood, S.N. (2015). mgcv: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation. *R package* version 1.8-6, available at http://cran.r-project.org/web/packages/mgcv/.

Wood, S.N., Pya, N., and Säfken, B. (2015). On doubly generalized additive models. Unpublished manuscript.

Zipunnikov, V., Greven, S., Shou, H., Caffo, B., Reich, D., Crainiceanu, C. (2014). Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis. *Annals of Applied Statistics*, **8**, 2175 – 2202.

All statistical models are wrong, some are usable!

John Hinde[1]

[1] Statistics Group, SMSAM, NUI Galway, Ireland

E-mail for correspondence: john.hinde@nuigalway.ie

Abstract: A reflection on 30 years of IWSMs and the development and use of statistcial modelling.

Keywords: Statistical models; IWSM; Random effects; Mixtures

1 Introduction

George Box is famously (mis)-quoted as saying "*all models are wrong, but some are useful*". Well this is certainly true of statistical models, where we may often be interested in description and prediction rather than a necessarily realistic or true model of the underlying process including all features of the system and the data collection. Of course, we want to take into account any specific aspects that we believe may be relevant to the study and attempt to capture both systematic and random features. For the past 30 years, this workshop has been devoted to this task and associated aspects of the modelling process, such as model building, estimation, inference, model selection, model checking, prediction and validation. The range of presentations over this period provide a rich history of the development of statistical modelling and the proceedings volumes are a valuable and fascinating record of emerging trends and the state of the art. What is clear from this large body of work is that, while our models can always be developed, refined and extended, there are now very many usable statistical models out there. The contributors to these meetings have played a major role in making this happen.

The interest in statistical modelling as embodied in this meeting can be considered as going back to the seminal paper of Nelder and Wedderburn (1972) that introduced generalized linear models and provided a unified framework for the analysis of many different types of data, including

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

binary responses, counted proportions, multicategory data, counts, event times, and continuously measured variates. Inherent in the formulation of glms is the separation of the systematic and random parts of the model through the three well known components:

- random component for the response: 1-parameter exponential family
- linear predictor: $\eta = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$
- link function: $g(\mu) = \eta$

In the intervening 40+ years, there have been many extensions of this basic model through generalisations of each component that include

- **response distribution**
  - multivariate vector of responses
  - exponential dispersion models
  - generalized distributions
  - quasi-distributions
  - mixtures
  - joint responses, such as survival time and a longitudinal measure (Rizopoulos, 2012)
- **linear predictor**
  - smooth terms — gams, etc
  - random effects
  - multiple linear predictors — joint modelling of mean and dispersion and possibly other distribution parameters as in `gamlss`.
- **link function**
  - parametric links
  - composite link functions — originally proposed in Thompson and Baker (1981), see also Paul Eilers invited presentation to IWSM2012 in Prague.
  - non-linear glms — see for example the ASA 2007 John M Chambers Statistical Software Award winning R-package `gnm` (Turner and Firth, 2007)

Many of these have featured in IWSM sessions and, of course, the statistical modelling paradigm can be viewed in a wider context of more general data analysis procedures. However, a structured modular view of the modelling process remains important for generalizability.

The 1972 paper was also very strong on applications and this has also always been a feature of IWSMs. The particular application may suggest certain key aspects of the model and the art of statistical modelling is to decide on what is important in any context to provide a usable model that addresses the questions of substantive interest.
Finally, a key aspect in the popularisation of glms was the availability of software, in particular GLIM, which provided an interactive environment for fitting glms in a way that reflected the component structure of the model. The original implementation of GLIM also had various features, or perhaps idiosyncrasies, that reflected John Nelder's own personal perspective on statistical computing, not least of these was the notion of an open, extendible system that enabled the programming of extensions to the standard glms. This idea is now standard in statistical systems and is perhaps most prevalent in R and allows us to make statistical modelling developments and extensions generally available and usable.

2 Workshop history

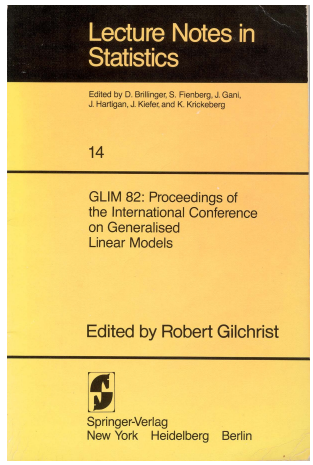

FIGURE 1. Glim 82, Proceedings of the International Conference on Generalised Linear Models.

The origins of the workshop were two conferences on GLIM that were held in London, 1982, see Figure 1, and Lancaster, 1985. At that time, Murray Aitkin was leading a group of us in Lancaster and was a passionate advocate for the popularisation of glms, which ultimately led to the applied book *Statistical Modelling in GLIM* (Aitkin et al. 1989; subsequently updated to R in Aitkin et al. 2009) based on material from short courses that were presented across Europe. These activities prompted considerable interest, particularly in Europe, and from the enthusiasm of statisticians

from Austria, Italy and the UK the first workshop was held in Innsbruck in 1986 with about 30 participants. The meeting was set up with an informal workshop structure and was aimed at applied statisticians, but also embraced methodological, theoretical and computational work. Since then the workshop has maintained much the same structure with no parallel sessions and an emphasis on work that is based on a substantive area with a real motivating application and data.

FIGURE 2. Exeter 1994, Invited speakers and members of the IWSM Board of Trustees.

For many years the IWSM was planned and directed by a Board of Trustees drawn from the founders and active attendees from the early meetings, see Figure 2. In 2003 the Statistical Modelling Society was formed and an Executive Committee, which is elected every two years, took over the running of the IWSM. This has allowed new blood to engage in the planning and management of the meeting while preserving the original structure and distinctive features that make the IWSM different from many other statistics conferences.

Over the years the workshop has moved around much of Europe and made forays to the US and even on one occasion to Australia, see Table 1. The different countries typically have different statistical cultures and approaches to data analysis and modelling that have provided their own flavour to the various meetings. This cross-fertilization has sparked many fruitful collaborations that have featured at subsequent meetings. Indeed, a great strength of the meeting has been the maintenance of the informal workshop-like nature that encourages discussion and collaboration, quite often involving an interaction between students and early-career people with more senior and experienced colleagues.

The international flavour of the meeting can also be seen by looking at

TABLE 1. 30 years of IWSMs.

2015	Linz, Austria	2000	Bilbao, Spain
2014	Göttingen, Germany	1999	Graz, Austria
2013	Palermo, Italy	1998	New Orleans, USA
2012	Prague, Czech Republic	1997	Biel/Bienne, Switzerland
2011	Valéncia, Spain	1996	Orvieto, Italy
2010	Glasgow, Scotland	1995	Innsbruck, Austria
2009	Ithaca, NY, USA	1994	Exeter, UK
2008	Utrecht, The Netherlands	1993	Leuven, Belgium
2007	Barcelona, Spain	1992	Munich, Germany
2006	Galway, Ireland	1991	Utrecht, The Netherlands
2005	Sydney, Australia	1990	Toulouse, France
2004	Florence, Italy	1989	Trento, Italy
2003	Leuven, Belgium	1988	Vienna, Austria
2002	Chania, Greece	1987	Perugia, Italy
2001	Odense, Denmark	1986	Innsbruck, Austria

where invited speakers have been drawn from, see Table 2. Some interesting features emerge from this. Although for many years the main style of statistics presented at the meeting was perhaps in the British empirical modelling tradition, there was always a habit of bringing interested people from the US, where generalised linear models and statistical modelling were slower to develop — the US statistical landscape now looks very different from the more theoretical and abstract emphasis of 30 years ago.

TABLE 2. Invited speakers by country.

USA	49
UK	47
Germany	19
Netherlands	11
France	10
Italy	10
Denmark	9
Australia	6
Spain	6
Switzerland	6
Belgium	5
Nordic	4
Austria	3
E. Europe	2
Rest of World	4

One invited speaker stands out, Murray Aitkin has been invited on 5 separate occasions, albeit representing 4 different countries! This reflects his influence on the meeting from the outset and he will always be remembered for his ability to have a question and helpful comment for almost every presenter.

3 Developing themes

The focus of meetings changes depending on the current trends in statistical modelling and the particular interests and perspective of the local organisers. This is most simply seen in the topics of invited talks, which reflect the development of both methodology, application areas, inferential and computational perspectives. The early workshops maintained their strong link to glms, including sessions devoted to the ongoing developments in GLIM. The analysis of categorical data was also a major theme in early meetings, including notable introductions on graphical models and computational tools for model selection in high-dimensional complex contingency tables. Mixed models also made an early appearance and ideas of smoothing and gams go back to the earliest meetings, but were not as dominant as in more recent years. Indeed, it might be said that P-splines owe themselves in some way to the ISWM with Paul Eilers and Brian Marx forming their collaboration through the meeting. This then grew to a larger group of enthusiasts and papers relating to smoothing and the P-splines methodology have been a backbone of many workshops. Repeated measures and longitudinal data soon became a major theme and have continued to feature in many different contexts, with considerable emphasis on missing data. Bayesian approaches began appearing from the mid-90s and inevitably they have become more prevalent with the wide availability of usable MCMC procedures, although the workshop has never been a particularly Bayesian meeting. Time series methods have never been a major feature of the meeting, although there has been much interest spatial and spatio-temporal modelling, often again linked to smoothing.

Applications have come predominantly from the medical and health sciences, although social science was also very strong in the early years and certainly continues to be present. In recent years there has been the inevitable increase in applications from molecular biology and omics related areas. Ecology and environmental statistics have also appeared at various times. A strength has always been the range of different application areas and the resulting cross-fertilisation of statistical modelling from one area to another.

4 Reproducible research: Statistical Modelling Journal

The style of workshop presentations emphasised the place of a real-world problem and associated data in driving model developments and new methods, along with software and code to implement the ideas. Making all of these components available is an important aspect of reproducible research; a reader should be able to reproduce the results in a paper, explore alternative models, and apply the same procedures to their own data. This ties in completely with the notion of a workshop rather than a conference, as in a workshop we might expect to see the tools and even get to play with them. In the early workshops there were some mini-tutorial sessions that delivered precisely. These soon led to the Sunday Short Courses that were typically intended to give this hands-on experience in some general or developing area of statistical modelling.

For many years, there was also discussion of how to disseminate work from the ISWMs, other than just in the Proceedings Volumes that have very limited circulation. Of course, many presentations from the meetings went on to be published papers in mainstream statistics and subject matter journals and even, in some cases, books. However, it was felt that the style of article that reflected the ethos of statistical modelling did not always find an obvious home in the existing journals. So from the late 1990s we looked into setting up a Statistical Modelling journal that would publish the sort of material that might be presented at the IWSM, although not simply as an outlet for papers from IWSM talks. This eventually came into being in 2001 thanks to the support of the publishers Edward Arnold and the work of the founding editors, John Hinde, Emmanuel Lesaffre, and Brian Marx. The journal remains under the direction of the Statistical Modelling Society, although is now published by Sage, India.

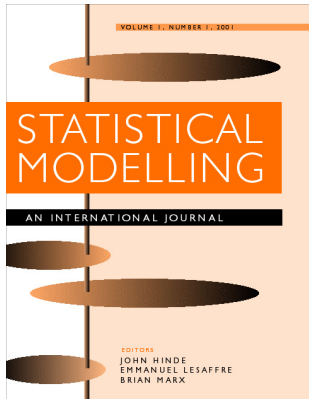

FIGURE 3. Statistical Modelling, first issue of the journal in 2001.

In setting up the journal we were very keen to address the problem that at the time it was often difficult to follow or check work in published papers. Papers rarely contained anything more than summary data and the raw data was typically not readily available. Even when data were available, code for the specific analyses reported in the paper was not and attempts to reproduce analyses were often frustratingly difficult and could frequently lead to somewhat different results. So, to address these issues, from the outset we insisted that authors should submit both their data and code for the work reported in the paper and these are made publicly available through the journal archive website (`www.statmod.org/smij/`). This practice is now much more widespread for both statistics and subject matter journals and is a valuable addition to the printed paper and certainly makes models more usable. New web-based technologies mean that this idea can now be extended to other aspects of the paper, such as Shiny apps to present and explore fitted models.

5 Personal interests

5.1 Random effects and overdispersion

One feature of the simplest generalized linear models, such as the Poisson log-linear model for counts and the binomial logit model for proportion data, is that the response distribution only captures the variability of the response related to the observation, or measurement, process. Other aspects of variability, such as experimental unit or individual heterogeneity and omitted covariates, are not accounted for in the model and unlike in the normal error model there is no single additive error term to absorb these. So, in practice there is typically more variability in the data than predicted by the basic model, a situation often referred to as *overdispersion*. Quasi-glms with an estimated dispersion parameter are one simple way of allowing for this, although it may also be necessary to use more general variance functions than those of the base model.

To allow for more complex sampling designs and to follow the normal mixed model approach it is often more natural to include additional random effects in the linear predictor taking

$$\boldsymbol{\eta} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}.$$

Using a single conjugate random effect at individual observations level can give various standard overdispersion models, such as the negative binomial for count data and the beta-binomial for proportions. Taking $\mathbf{z}$ to have a normal distribution provides the family of generalized linear mixed models. At the GLIM82 conference I presented an approach to fitting a Poisson model with an additional random effect using Gaussian Quadrature and an EM algorithm that was very similar to that used for fitting a finite mixture

model. This same approach allows the fitting of models with an additional single random effect $\mathbf{z}$ that has an unspecified distribution, sometimes referred to as nonparametric maximum likelihood (see Aitkin et al. 2009) and available in the R-package `npmlreg` (Einbeck et al. 2014). Recently, generalized linear models have been extended to incorporate random effects for both individual level variability (overdispersion) and to reflect the data structure, such as clustered sampling and longitudinal observation, as described in Molenberghs et al. (2010) and related papers.

Alternative approaches to handle overdispersion are to move to more general and flexible distributions, although these are typically outside of the exponential family and can lack some of the convenience of glms. For count data these include distributions such as Hermite, Neyman Type A, Polya-Aeppli, generalized Poisson, Poisson-inverse Gaussian, Poisson-inverse gamma, COM-Poisson, etc. The range of possible distributions can be bewildering and is ever growing, although in practice many can give similar fits. The `gamlls` package (Stasinopoulos et al. 2015) provides facilities for fitting many of these, with the additional flexibility to have individual linear predictors for the distribution's parameters.

5.2 Mixture models

Finite mixture models provide useful extensions to standard models that can both account for additional unobserved variability and allow probabilistic clustering of units. A general univariate response mixture model has

$$f(y \mid \Theta, \mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(y \mid \theta_k, \mathbf{x})$$

where the f_k are the individual component densities (often of the same form), the π_k are the component probabilities (with $\sum_k \pi_k = 1$), and θ_k are the component parameter vectors, some of which may be taken to be equal across components. The EM algorithm gives a convenient and conceptually simple way of fitting such models, although because the likelihood function is multimodal there can be problems in finding the global maximum.

One particular simple class of mixture models for count data are the *zero inflated* models, such as the zero-inflated Poisson distribution, a mixture of a degenerate distribution at 0 and a Poisson distribution.

$$\Pr(Y = y) = \begin{cases} \omega + (1-\omega)e^{-\lambda} & y = 0 \\ (1-\omega)\dfrac{e^{-\lambda}\lambda^y}{y!} & y > 0 \end{cases}$$

This simple extension exhibits both zero-inflation and overdispersion. Lambert (1992) considered introducing regression models for both parameters

with appropriate link functions

$$\log(\lambda_i) = \mathbf{x}_i^{\mathrm{T}}\beta \quad \text{and} \quad \log\left(\frac{\omega_i}{1-\omega_i}\right) = \mathbf{z}_i^{\mathrm{T}}\gamma$$

where $\mathbf{x}$ and $\mathbf{z}$ are covariate vectors (that may or may not coincide) and β and γ are vectors of parameters.
Moving away from the Poisson distribution, we can consider zero-inflation in more general count distributions, such as the negative binomial, and again introduce regression models for one, or more, of the model parameters. For a zero-inflated negative binomial distribution we are allowing separate models for the zero-inflation and the overdispersion. However, in practice, it can be hard to separate these effects as the negative binomial distribution can often give a good fit to zero-inflated data.

5.3 Example: Trajan apple cultivation

These data, from Ridout et al. (1988) and also considered in Jansakul and Hinde (2009), relate to a micropropagation experiment of the apple variety Trajan (a 'columnar' variety) using a 4x2 factorial design. Shoot tips of length 1.0-1.5 cm were placed in jars on a standard culture medium. Four different concentrations of growth hormone cytokinin BAP (factor `H`) were added and the jars places at random in one of two growth cabinets, one with 8 hour photoperiod, the other with a 16 hour one (factor `P`). After 4 weeks of culture at a temperature of 22^oC, the number of roots that had formed on each shoot was counted. The raw data shows both overdispersion and, for the longer photoperiod, zero-inflation. Table 3 shows the results of fitting Poisson, Negative binomial and zero-inflated regression models. While the negative binomial model gives a much improved fit, the real issue here is zero-inflation, however once this is taken account of there is still some residual overdispersion that the zero-inflated negative binomial model can pick up as seen in the improved fit. Interestingly, once both zero-inflation and overdispersion are correctly modelling, the only significant effects are due to the different photoperiods, the level of growth hormone is not important.

5.4 Mixtures of mixed models: *Time course microarray*

Time-course gene expression data measure the expression levels for each gene throughout some time interval. The data here are from a study examining the temporal response of yeast to anaerobiosis. Cells were grown aerobically and then at intial experimental time 0 switched from air to nitrogen. Expression values were taken at 14 time points corresponding to 14 generations of anaerobic growth, see Figure 4 for an illustration of some of the data. One general aim is to cluster expression profiles to find groups of co-expressed genes.

TABLE 3. Trajan Apples: model fitting results.

	Models						
	λ	ω	α	$-2\log$L	df	AIC	BIC
Poisson	H*P	0	0	1556.9	262	1572.9	1601.7
	P	0	0	1571.9	268	1575.9	1583.1
Neg-Bin	H*P	0	const	1399.6	261	1417.6	1450.0
	H*P	0	P	1264.6	260	1284.6	1320.6
	H*P	0	H*P	1254.8	254	1286.8	1344.4
	Lin(H)*P	0	P	1270.1	264	1282.1	1303.7
	P	0	P	1272.4	266	**1280.4**	**1294.8**
	P	0	const	1403.9	267	1409.9	1420.7
ZIP	H*P	const	0	1338.0	261	1356.0	1388.4
	H*P	P	0	1244.5	260	1264.5	1300.5
	H*P	H*P	0	1238.2	254	1270.2	1327.8
	Lin(H)*P	P	0	1250.2	264	**1262.2**	1283.8
	P	P	0	1261.3	266	1269.3	**1283.7**
	P	const	0	1355.2	267	1361.2	1372.0
ZINB	H*P	const	const	1324.8	260	1344.8	1380.8
	H*P	P	const	1232.5	259	1254.5	1294.1
	H*P	P	P	1226.3	258	1250.3	1293.5
	H*P	H*P	H*P	1205.6	246	1253.6	1340.0
	Lin(H)*P	P	P	1231.0	262	**1247.0**	1275.8
	P	P	P	1237.7	264	1249.7	**1271.3**
	P	P	const	1243.9	265	1253.9	1271.9
	P	const	const	1336.5	266	1344.5	1358.9
	const	P	const	1257.8	266	1265.8	1280.2

Here, we initially consider representing the observed profiles as observations of underlying smooth functions with measurement error

$$y_j = \underbrace{g(t_j)}_{\text{Signal}} + \underbrace{\varepsilon_j}_{\text{Noise}}.$$

To represent the smooth function we use a pth degree truncated power basis (typically $p = 1$ or 2):

$$g(t_j) = \beta_0 + \beta_1 t_j + \ldots + \beta_p t_j^p + \sum_{\ell=1}^{L} \beta_{1\ell}(t_j - \kappa_\ell)_+^p,$$

the κ_ℓ are the knots and $(t_j - \kappa_\ell)_+ = \max(0, t_j - \kappa_\ell)$. We then implement P-spline smoothing as a linear mixed effects model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

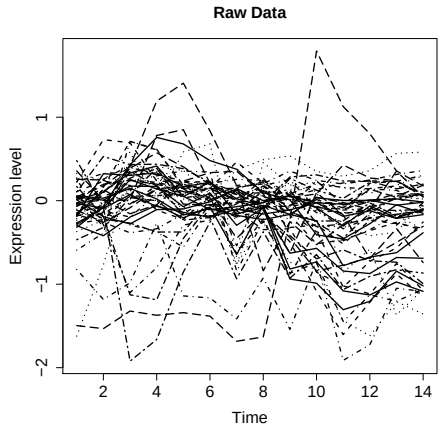

FIGURE 4. Yeast time course course microarray; typical raw data.

$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$. Here $\mathbf{Z}$ contains the P-spline basis functions and $\mathbf{u}$ the smoothing coefficients which are treated as random with $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I})$. Estimates of $\boldsymbol{\beta}$, σ_ε^2, σ_u^2 and $\mathbf{u}$ are determined using (RE)ML and BLUP using standard mixed model software.

For the clustering of genes we want to form groups with similar expression profile patterns over time; the absolute expression level for an individual gene is not of interest, merely how it changes over time. To allow for this we can introduce a random gene specific expression level shift and hence represent the expression level for gene i in cluster c at time j as

$$y_{ij} = \mu_g(t_{ij}) + b_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i,$$

where $b_i \sim N(0, \sigma_{bc}^2)$ represent gene-specific shifts from the mean. For a particular gene cluster c we now have an extended mixed model for the smooth modelling of the cluster mean profile and the gene specific shifts

$$\mathbf{Y}_c = \underbrace{\mathbf{X}_{c,s}\boldsymbol{\beta}_{c,s} + \mathbf{Z}_{c,s}\mathbf{u}_{c,s}}_{\mu_c(t)} + \mathbf{Z}_{c,b}\mathbf{b}_c + \boldsymbol{\varepsilon}_c,$$

where $\mathbf{u}_{c,s} \sim N(\mathbf{0}, \sigma_{uc}^2 \mathbf{I})$, $\mathbf{b}_c \sim N(\mathbf{0}, \sigma_{bc}^2 \mathbf{I})$, $\boldsymbol{\varepsilon}_c \sim N(\mathbf{0}, \sigma_{\varepsilon c}^2 \mathbf{I})$.

Of course, in practice, we do not know the cluster membership, but by using the above mixed model together with a finite mixture model for the clusters we have a model-based clustering procedure for smooth curves. We assume that $\mathbf{y}_i$, the expression profile for gene i, comes from a mixture of C clusters:

$$\mathbf{y}_i \sim \pi_1 N(\mu_1(\mathbf{t}_i), \mathbf{V}_{i1}) + \pi_2 N(\mu_2(\mathbf{t}_i), \mathbf{V}_{i2}) + \ldots + \pi_C N(\mu_C(\mathbf{t}_i), \mathbf{V}_{iC})$$

where

$$\mu_c(\mathbf{t}_i) = \mathbf{X}_{i,s}\boldsymbol{\beta}_{c,s} + \mathbf{Z}_{i,s}\mathbf{u}_{c,s}$$

represents the mean smooth expression profile for cluster c and $\mathbf{V}_{ic} = \sigma^2_{bc} E_{n_i \times n_i} + \sigma^2_{\varepsilon c} \mathbf{I}_{n_i \times n_i}$ gives the nested variance structure for the longitudinal response. Estimation is possible using a standard EM algorithm for mixture models, where the M-Step involves fitting the smoothing mixed model using the probabilistic cluster allocations as weights and the E-step amounts to using Bayes rule to obtain these weights as the (posterior) probability that gene i is from cluster c. This procedure was run over different values of C and using BIC gave the optimal number of clusters as BIC 58. Figure 5 displays six of these clusters and shows some of the different response profiles that are obtained. Individual clusters can be interpreted by using over-represented gene ontology (GO) terms to identify possible biological function.

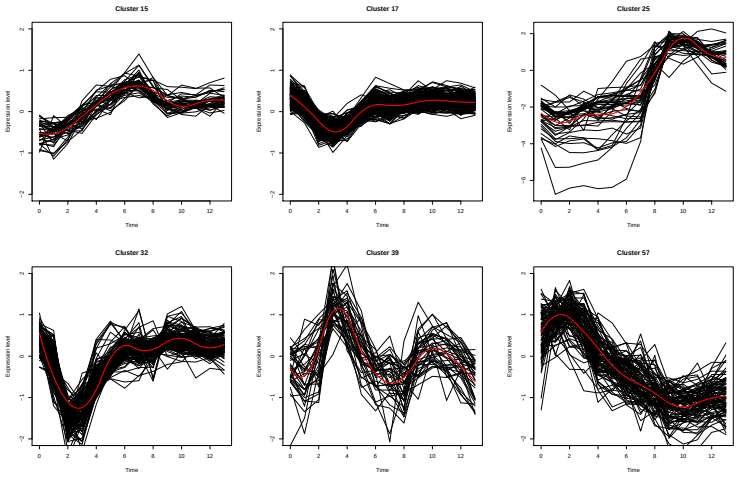

FIGURE 5. Some clusters with over-represented GO functional categories

6 Health warnings

Thirty years of IWSMs have run in parallel with an explosion in the scope, flexibility and power of statistical modelling. The availability of easily adapted computational tools makes it tempting to move on to yet more levels of complexity that may better reflect the real-world situation. The era of *Big Data* also brings both new opportunities and challenges. The potential richness of the data allows us to consider usiing more sophisticated models, although the size of the data can bring its own challenges. Another important aspect is the quality of the data and sadly as data collection procedures become easier (automated) and richer, the quality can frequently suffer. The direct relevance of convenience collected data is also not always

considered — *more* most certainly does not necessarily translate to *better*. Carefully thought out experimental and data collection design for the real questions of interest is often sadly absent. This is perhaps one of the weak areas of IWSMs, we have had very few sessions on this vital first stage aspect of statistical modelling. Also, statistical design of experiments is fast disappearing from many statistics syllabi, not to mention what happens in data science courses.

One other concern that I have is with the quality and reliability of the software that we have for statistical modelling. The R package system and CRAN are undoubtedly a great vehicle for the dissemination of new models, techniques and tools for statistical modelling and there are some truly wonderful packages available to us. However, one weakness in the current setup is the lack of refereeing and testing of these software additions. Certainly, new packages have to work at least on the test data supplied by the authors, but there is no real check on their reliability and robustness in more general use, nor of when they are likely to breakdown. It is quite possible to fit the same model in different packages and obtain rather different results. Novelty and flexibility now seem to be more attractive aims than thorough testing, which is no surprise as there are no obvious academic rewards for the very time-consuming reviewing of software. A definite case of let the user beware!

6.1 Understand your model

A salutary lesson came some years ago in fitting some models to ordered multinomial observations. The date came from an an experiment on biological pest control of the termite *Heterotermes tenuis* (an important pest of sugarcane in Brazil, causing damage of up to 10 metric tonnes/ha/year) using the fungus *Beauveria bassiana* as a possible microbial control. To study the pathogenicity and virulence of 142 different isolates of *Beauveria bassiana* a completely randomized experiment with five replicates of each of the 142 isolates was performed. Solutions of the isolates were applied to groups of $n = 30$ termites kept in Petri-dishes and the mortality in the groups was measured daily for eight days. This resulted in 710 ordered multinomial observations.. Figure 6 shows cumulative mortality plots for a typical sample of the isolates.

For the general setup with I isolates, K replicates, D days, and n insects the responses are

$$\begin{aligned} Y_{ik,j} &= \text{number dead on day } j\text{, isolate } i\text{, replicate } k \\ Y_{ik,D+1} &= \text{number still alive on day } D \end{aligned}$$

Treating these as multinomial responses over the days, $\mathbf{Y}_{ik} \sim \text{Multinomial}(n; \boldsymbol{\pi}_{ik})$ gives a model with mean and variance

$$\text{E}[\mathbf{Y}_{ik}] = n\boldsymbol{\pi}_{ik} \qquad \text{Var}[\mathbf{Y}_{ik}] = n\left[\text{diag}\{\boldsymbol{\pi}_{ik}\} - \boldsymbol{\pi}_{ik}\boldsymbol{\pi}_{ik}^{\text{T}}\right]$$

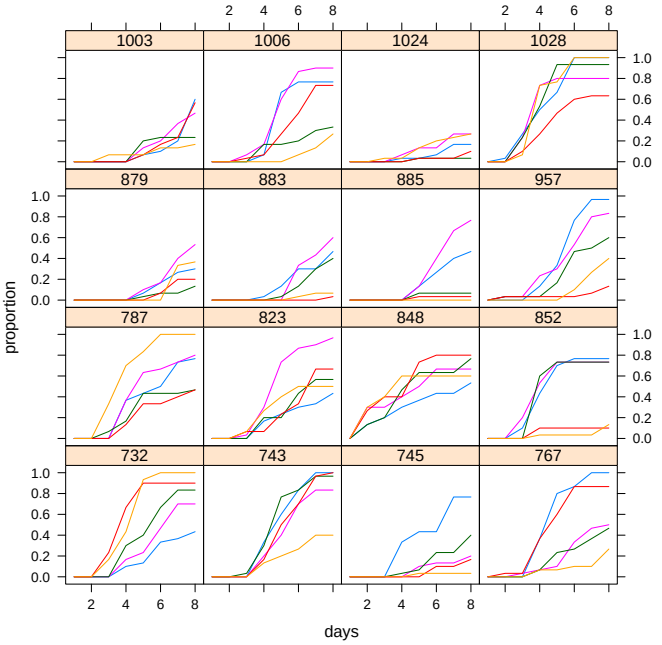

FIGURE 6. Termite Cumulative mortality for a sample of 16 isolates.

Here, it is more natural to work with the cumulative proportions:

$$\mathbf{R}_{ik} = (R_{ik,1}, R_{ik,2}, \ldots, R_{ik,D})^{\mathrm{T}} = \frac{1}{n}\mathbf{L}\mathbf{Y}_{ik}$$

where $\mathbf{L}$ is a lower triangular matrix of 1's. The corresponding mean and variance are

$$\mathrm{E}[\mathbf{R}_{ik}] = \mathbf{L}\boldsymbol{\pi}_{ik} = \boldsymbol{\gamma}_{ik}$$

$$\mathrm{Var}[\boldsymbol{R}_{ik}] = \frac{1}{n}\mathbf{L}[\mathrm{diag}\{\boldsymbol{\pi}_{ik}\} - \boldsymbol{\pi}_{ik}\boldsymbol{\pi}_{ik}^{\mathrm{T}}]\mathbf{L}^{\mathrm{T}} = \mathbf{V}(\boldsymbol{\gamma}_{ik})$$

Modelling γ with a logit link function gives a cumulative logistic model

$$g(\gamma_{ikj}) = \mathrm{logit}(\gamma_{ikj}) = \log\left(\frac{\sum_{s=1}^{j}\pi_{ik,s}}{\sum_{s=j+1}^{\mathrm{D+1}}\pi_{ik,s}}\right) = \eta_{ikj}$$

where the linear predictor can include isolate specific factors, time dependency, etc. A simple baseline models fits an isolate specific linear time effect constant over replicates

$$\eta_{ikj} = \beta_{1i} + \beta_{2i}t_j.$$

From inspection of the data, it is clear that there is considerable additional variation across the replicates, i.e. multinomial overdispersion. This can be handled by allowing variation of the multinomial parameter $\boldsymbol{\pi}$ using a two-stage model where

$$\boldsymbol{Y}_{ik} \mid \boldsymbol{p}_{ik} \sim \text{Multinomial}(n; \boldsymbol{p}_{ik}),$$

$$\boldsymbol{p}_{ik} = (p_{ik,1}, \ldots, p_{ik,\text{D}}, p_{ik,\text{D}+1})^{\text{T}} \text{ follows a Dirichlet distribution.}$$

This results in a Dirichlet-multinomial model for $\mathbf{Y}$ and $\mathbf{R}$ with

$$E[\boldsymbol{R}_{ik}] = \boldsymbol{\gamma}_{ik}$$

and covariance matrix given by

$$\text{Var}[\boldsymbol{R}_{ik}] = \boldsymbol{V}(\boldsymbol{\gamma}_{ik})[1 + \rho_i(n-1)]$$

where ρ_i is an (isolate specific) overdispersion parameter. This is a multivariate generalization of the beta-binomial model.
An alternative would be to incorporate one, or more, random effects in the linear predictor. The simplest random intercept model has

$$g(q_{ikj}) = \eta_{ikj} + \xi_{ik} = \beta_{1i} + \beta_{2i}t_j + \xi_{ik}$$

where ξ_{ik} is a random effect with $E[\xi_{ik}] = 0$, $\text{Var}[\xi_{ik}] = \sigma_i^2$. Taylor series approximations give

$$E[\boldsymbol{R}_{ik}] = E[E(\boldsymbol{R}_{ik}|\boldsymbol{q}_{ik})] = E[\boldsymbol{q}_{ik}] \approx \boldsymbol{\gamma}_{ik}$$

and

$$\text{Var}[\boldsymbol{R}_{ik}] \approx \boldsymbol{V}(\boldsymbol{\gamma}_{ik}) + \left(1 - \frac{1}{n}\right)\sigma_i^2[\boldsymbol{h}'(\boldsymbol{\eta}_{ik})][\boldsymbol{h}'(\boldsymbol{\eta}_{ik})]^{\text{T}}$$

where $\mathbf{h}$ is the inverse link function with derivative $\mathbf{h}'$. This is analagous to Williams's approximate variance function for the logistic-normal distribution.
This model can be extended to include a random slope as well as the intercept

$$g(q_{ikj}) = \beta_{1i} + \xi_{ik} + (\beta_{2i} + \zeta_{ik})t_j = \eta_{ikj} + \xi_{ik} + \zeta_{ik}t_j$$

where the random effects $(\xi_{ik}, \zeta_{ik})^{\text{T}}$ have $E[\xi_{ik}] = E[\zeta_{ik}] = 0$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \nu_i^2 & \lambda_i\nu_i\tau_i \\ \lambda_i\nu_i\tau_i & \tau_i^2 \end{bmatrix}$$

Approximations now give

$$E[\boldsymbol{R}_{ik}] \approx \boldsymbol{\gamma}_{ik}$$

and

$$\begin{aligned}\mathrm{Var}[\boldsymbol{R}_{ik}] \approx \boldsymbol{V}(\boldsymbol{\gamma}_{ik}) \left(1 - \frac{1}{n}\right) \Big\{ & \nu_i^2 [\boldsymbol{h}'(\boldsymbol{\eta}_{ik})][\boldsymbol{h}'(\boldsymbol{\eta}_{ik})]^{\mathrm{T}} \\ & + \tau_i^2 [\boldsymbol{h}'(\boldsymbol{\eta}_{ik}) * \boldsymbol{t}_{ik}][\boldsymbol{h}'(\boldsymbol{\eta}_{ik}) * \boldsymbol{t}_{ik}]^{\mathrm{T}} \\ & + \lambda_i \nu_i \tau_i [\boldsymbol{h}'(\boldsymbol{\eta}_{ik})][\boldsymbol{h}'(\boldsymbol{\eta}_{ik})]^{\mathrm{T}} * [\mathbf{1}\boldsymbol{t}_{ik}^{\mathrm{T}} + \boldsymbol{t}_{ik}\mathbf{1}^{\mathrm{T}}] \Big\}\end{aligned}$$

Estimating these models using a generalized estimating equation (gee) approach for $\boldsymbol{\beta}$ and moment-based estimators for the overdispersion parameters gives some rather surprising results.

- The parameter estimates from all four models are *identical*.
- Sandwich estimator robust standard errors from all four models are *identical*.
- Model based standard errors exhibit simple relationships.

Of course, it is not surprising that there should be close similarities between the results from the models, but that they should be *exactly* the same is more surprising. It turns out to be a consequence of the following results for generalized least-squares. If we have a generalized least square estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{z}$$

and then modify the variance function in the following manner

$$\tilde{\mathbf{V}} = \mathbf{V} + \mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}$$

then the modified estimator

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{V}}^{-1}\mathbf{z}$$

gives identical estimates, $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$, the sandwich variance matrix estimates are identical and the asymptotic covariance matrices are simply related

$$\tilde{M_0} = (\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{V}}^{-1}\mathbf{X})^{-1} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} + \mathbf{D} = M_0 + \mathbf{D}.$$

For the gee fitting of the various overdispersed cumulative multinomial models, the iterative weight matrices

$$\boldsymbol{W}_{ik}^{\mathrm{OD}} = \boldsymbol{W}_i^{\mathrm{OD}} = \boldsymbol{\Delta}_i^{-1} \left[\boldsymbol{V}_i^{\mathrm{OD}}\right]^{-1} \boldsymbol{\Delta}_i^{-1}$$

are related as follows

$$\boldsymbol{\Delta}_i \boldsymbol{V}_i^{\mathrm{OD}} \boldsymbol{\Delta}_i = \boldsymbol{\Delta}_i \boldsymbol{V}_i \boldsymbol{\Delta}_i + \mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}$$

where $\mathbf{D}$ depends on the particular approximate random effects model. That is we are precisely in the above situation, although in a somewhat more complex setting.

So the surprising outcome is simply a feature of the model, the approximation and fitting method and the fact that the data structure is identical for all replicates with only within-individual covariates. In more general models, it seems that similar conclusions remain approximately true in many cases and one can see this in reported results from many mixed and generalized mixed model fits. The moral of this is that as we extend the complexity and sophistication of the models that we use, it can become increasingly difficult to understand precisely what is happening, or what we might expect to see.

7 The future

While it can be hard to summarise the past, it is much harder to predict the future. However, after 30 years of the IWSM it is clear that it has made a great contribution to the statistical modelling community and been influential in making models that are `usable` and thus `useful`. It is also worth reflecting on the influence that the workshop has had in providing a welcoming and receptive forum for several generations of postgraduates, postdocs and younger colleagues. That many of these have continued to partake and organise the meeting is a testament to the usability of this model of a meeting. This is a great reward to those once young people who had the vision to initiate this 30 years ago and all of those who have contributed in many capacities over the years. So one thing is certain, I am sure that the ISWM will continue to play this great role. I look forward to many more years with inspiring and creative presentations on uses and developments of statistical modelling, Here's to another 30 years!

Acknowledgments: Special thanks to my co-workers over many years and I hope many years to come.

References

Aitkin, M.A., Anderson, D.A., Francis, B.F., and Hinde, J.P. (1989). *Statistical Modelling in GLIM*. Clarendon Press, Oxford.

Aitkin, M.A., Francis, B.F., Hinde, J.P., and Darnell, R. (2009). *Statistical Modelling in R*. Oxford University Press.

Coffey, N., Hinde, J., and Holian, E. (2014). Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics and Data Analysis*, **71**, 14–29.

Einbeck, J., Darnell, R., and Hinde, J. (2014). `npmlreg`: Nonparametric maximum likelihood estimation for random effect models, R package version 0.46-1, `CRAN.R-project.org/package=npmlreg`.

Jansakul, N. and Hinde, J.P. (2009). Score tests for extra-zero models in zero-inflated negative binomial models. *Communications in Statistics Simulation and Computation*, **38**, 92 – 108.

Molenberghs, G., Verbeke, G., Demétrio, C., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 32 – 357.

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

Ridout, M.S., Demétrio, C.G.B., and Hinde, J.P. (1998). Models for count data with many zeros. In: *Proceedings of the XIXth International Biometrics Conference*, Cape Town, South Africa, 179 – 192.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*. Boca Raton: Chapman and Hall/CRC.

Stasinopoulos, M., Rigby, R., and Voudouris, V. (2015). `gamlss`: Generalised Additive Models for Location Scale and Shape, R package version 4.3-4, `CRAN.R-project.org/package=gamlss`.

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

Turner, H. and Firth, D. (2007). `gnm`: A Package for Generalized Nonlinear Models. *R News*, **7/2**, 8 – 12.

Part II – Contributed Papers

Bayesian estimation of the tail index of a heavy tailed distribution with censoring

Abdelkader Ameraoui[1], Kamal Boukhetala[1], Jean-François Dupuy[2]

[1] Faculty of mathematics, USTHB, Algiers, Algeria
[2] IRMAR-INSA de Rennes, France

E-mail for correspondence: jean-francois.dupuy@insa-rennes.fr

Abstract: Bayesian estimation of the tail index of a heavy-tailed distribution is addressed when data are rando mly right-censored. Maximum a posteriori and mean posterior estimators are constructed for various prior distributions of the tail index. Properties of these estimators are investigated by simulations.

Keywords: Extreme value; MAP; Mean posterior estimator; Simulations.

1 Introduction

The tail index measures the thickness of the tail of a probability distribution function and thus plays a crucial role for evaluating the risk of occurrence of extremes events, see Beirlant et al. (2004).
Let F be the cumulative distribution function of some non-negative r.v. X. We assume that F is heavy-tailed, that is, there exists a constant $\alpha > 0$ such that

$$\lim_{x\to\infty} \frac{1-F(tx)}{1-F(x)} = t^{-\alpha} \qquad \text{for all } t > 0. \tag{1}$$

If (1) holds, we say that F is regularly varying at infinity with tail index α. Heavy-tailed distributions are useful for investigating phenomena where exceptional values have a significant occurrence frequency.
In this talk, we address Bayesian estimation of α when X is randomly right-censored. Censoring commonly occurs in the analysis of event time data. Under censoring, observations consists of pairs (Z_i, δ_i), $i = 1, \ldots, n$, where $Z_i = \min(X_i, Y_i)$, $\delta_i = 1_{\{X_i \le Y_i\}}$, $1_{\{\cdot\}}$ is the indicator function and Y_i is a random censoring time. Bayesian estimation of α without censoring

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

is addressed by do Nascimento et al. (2012) and So and Chan (2014). But so far, no Bayesian estimator has been proposed when censoring is present. Our work intends to fill this gap.
Based on various priors, we construct several Bayesian estimators of α (Section 2). These estimators are asymptotically normal. We examine their finite-sample performance via simulations (Section 3).

2 The proposed estimators

In Bayesian framework, the unknown α is considered as a r.v. whose density (the prior) represent the initial set of beliefs about α. This prior is updated using the information contained in the data, yielding the posterior density of α. We consider two different priors, namely the non-informative maximal data information (MDI) and Jeffreys priors.

a) Jeffreys prior. Jeffreys prior (Jeffreys, 1961) is proportional to the square root of Fisher's information: $\pi(\alpha) \propto 1/\alpha$. Combining this with the likelihood of α based on the sample of excesses above the $(n-k)$-th order statistic $Z_{n,n-k}$, we define the mean posterior (MPE) and maximum a posteriori (MAP) estimators of α under Jeffreys prior, as:

$$\hat{\alpha}_1 := \sum_{i=1}^{k} \delta_{[n-i+1]} \Big/ \sum_{i=1}^{k} \log\left(\frac{Z_{n,n-i+1}}{Z_{n,n-k}}\right)$$

$$\text{and} \quad \hat{\alpha}_2 := \left(\sum_{i=1}^{k} \delta_{[n-i+1]} - 1\right) \Big/ \sum_{i=1}^{k} \log\left(\frac{Z_{n,n-i+1}}{Z_{n,n-k}}\right),$$

where $\delta_{[n-i+1]}$ is the concomitant value of δ associated with $Z_{n,n-i+1}$.

b) MDI prior. The MDI prior maximizes the average information in the data density relative to that in the prior (Zellner, 1971). In our setting, the MDI prior for α is $\pi(\alpha) \propto \exp\left[\log(\alpha) - \frac{1}{\alpha}\right]$. Combining this with the likelihood of α based on the sample of excesses above $Z_{n,n-k}$, we define the MAP estimator of α under MDI prior, as:

$$\hat{\alpha}_3 := (B + \sqrt{B^2 + 4A}) \Big/ 2A,$$

where $A = \sum_{i=1}^{k} \log\left(\frac{Z_{n,n-k+i}}{Z_{n,n-k}}\right)$, $B = 1 + \sum_{i=1}^{k} \delta_{[n-k+i]}$.

Estimators $\hat{\alpha}_1, \hat{\alpha}_2$, and $\hat{\alpha}_3$ are asymptotically normal (Ameraoui et al., 2014). Next section reports a small part of the results of a comprehensive simulation study conducted by Ameraoui et al. (2014). We also discuss a method for choosing k.

3 Simulation study

A simulation study. We simulate 1000 samples of size $n = 1000$ of independent observations (Z_i, δ_i). The X_i and Y_i are distributed as Fréchet r.v. with cdf $F(x) = \exp(-x^{-\alpha})$ (with $\alpha = 0.5$) and $G(x) = \exp(-x^{-\beta})$ respectively, where β is chosen to yield various values for the proportion p of uncensored observations ($p = 0.9, 0.75, 0.5, 0.25$).
For each sample, we calculate $\hat{\alpha}_\ell$ ($\ell = 1, 2, 3$) for every k. Then for each ℓ, we obtain the averaged (over the 1000 simulated samples) value, empirical bias and RMSE of $\hat{\alpha}_\ell$ at the optimal fraction level $k_{opt} = \arg\min_k \text{RMSE}[\hat{\alpha}_\ell(k)]$. Results are provided in Table 1.

TABLE 1. Results for Fréchet model with $n = 1000$ and $\alpha = 0.5$.

		k_{opt}	Averaged estimator	RMSE	Bias
$p = 0.9$	$\hat{\alpha}_1$	133	0.4701402	0.0498572	-0.02985977
	$\hat{\alpha}_2$	133	0.4663334	0.0519978	-0.03366664
	$\hat{\alpha}_3$	137	0.4804375	0.0439893	-0.01956245
$p = 0.75$	$\hat{\alpha}_1$	131	0.4610741	0.0588477	-0.03892594
	$\hat{\alpha}_2$	131	0.4565036	0.0617286	-0.04349636
	$\hat{\alpha}_3$	131	0.4752784	0.0507112	-0.02472156
$p = 0.5$	$\hat{\alpha}_1$	113	0.4428444	0.0815031	-0.05715557
	$\hat{\alpha}_2$	113	0.4349585	0.0868767	-0.06504148
	$\hat{\alpha}_3$	118	0.4639250	0.0659266	-0.03607495
$p = 0.25$	$\hat{\alpha}_1$	89	0.3979008	0.1343862	-0.10209924
	$\hat{\alpha}_2$	144	0.3704152	0.1468569	-0.12958479
	$\hat{\alpha}_3$	86	0.4656990	0.0900378	-0.03430104

The MAP estimator under MDI prior $\hat{\alpha}_3$ outperforms $\hat{\alpha}_1$ and $\hat{\alpha}_2$ in terms of RMSE and bias for every p (see Ameraoui et al., 2014, for additional results with other values of α, distributions for X and priors for α).

Randomizing the fraction level. The order statistic $Z_{n,n-k}$ is used as threshold for defining excesses in the $\hat{\alpha}_\ell$. In Table 1 the optimal order k (or equivalently, the fraction level k/n) is chosen to minimize the RMSE. This is unfeasible in practice, as the true α is unknown. Hence, we develop an alternative approach for choosing k, whereby we consider the fraction level $t := k/n$ as random.
As k/n lies in $]0, 1[$, we suggest to use a Beta distribution for t, with density:

$$\pi(t) = \frac{1}{B(a,b)} \frac{(t - t_{min})^{a-1}(t_{max} - t)^{b-1}}{(t_{max} - t_{min})^{a+b-1}} 1_{\{t_{min} \leqslant t \leqslant t_{max}\}},$$

where $B(a,b) := \int_0^1 s^{a-1}(1-s)^{b-1}\, ds$ $(a, b > 0)$ and t_{min}, t_{max} are suitable bounds for t. Then the following Monte Carlo approach can be used to infer on α.

First, we simulate $x \sim \mathcal{B}eta(a, b)$ and we calculate $t_x = (t_{max} - t_{min})x + t_{min}$. Then we obtain $k_x = [nt_x]$ (where $[\cdot]$ denotes the integer part) and we calculate the estimate $\hat{\alpha}_\ell$ with $k = k_x$. This procedure is repeated $N = 10000$ times, which yields N Monte Carlo realisations $\hat{\alpha}_\ell^{(1)}, \ldots, \hat{\alpha}_\ell^{(N)}$ of $\hat{\alpha}_\ell$. Finally, α can be estimated by taking the empirical mean, median or mode of these N realisations.
We illustrate this procedure in a short simulation study where $n = 1000$ copies of (Z, δ) are obtained by simulating X and Y from Pareto distributions with tail index $\alpha = 2$ and β respectively, where β is such that $p = 0.25, 0.5, 0.75, 0.9$. Results are reported in Table 2. The empirical mean, median and mode of the estimates $\hat{\alpha}_\ell^{(j)}$, $j = 1, \ldots, N$, provide satisfactory approximations of α, for every p and ℓ.

TABLE 2. Results for the randomization procedure of the fraction level ($\alpha = 2$).

		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$
	Mode	2.037	1.983	2.048
$p = 0.25$	Median	2.074	2.015	2.085
	Mean	2.052	2.014	2.096
	Mode	2.014	1.992	2.027
$p = 0.50$	Median	2.024	2.008	2.034
	Mean	2.002	1.988	2.027
	Mode	1.985	1.975	2.033
$p = 0.75$	Median	2.026	2.015	2.044
	Mean	2.038	2.027	2.054
	Mode	1.973	1.938	1.997
$p = 0.90$	Median	2.044	1.995	2.051
	Mean	2.007	1.998	2.057

Overall, combining the proposed Bayesian estimators with the suggested randomization tool provides a relevant approach for estimating α.

Acknowledgments: Authors acknowledge funding from CNEPRU project MSAR (number B00220120012, Agence Nationale de Développement de la Recherche Universitaire, Algeria) and project SESA (number EDC25895, CNRS - France and DPGRF - Algeria).

References

Ameraoui, A., Boukhetala, K., and Dupuy, J.-F. (2014). Bayesian estimation of the tail index of a heavy tailed distribution under random censoring. *Submitted* and https://hal.archives-ouvertes.fr/hal-01071496.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., De Waal, D., and Ferro, C. (2004). *Statistics of Extremes*. Chichester: John Wiley & Sons Ltd.

do Nascimento, F.F., Gamerman, D., and Lopes, H.F. (2012). A semiparametric Bayesian approach to extreme value estimation. *Statistics and Computing*, **22**, 661 – 675.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.

So, M.K.P. and Chan, R.K.S. (2014). Bayesian analysis of tail asymmetry based on a threshold extreme value model. *Computational Statistics and Data Analysis*, **71**, 568 – 587.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons Ltd.

Robust maximum likelihood estimation of latent class models

Francesco Bartolucci[1], Brian Francis[2], Silvia Pandolfi[1], Fulvia Pennoni[3]

[1] Department of Economics, University of Perugia, Italy
[2] Lancaster University, United Kingdom
[3] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy

E-mail for correspondence: fulvia.pennoni@unimib.it

Abstract: We develop a suitable reweighting approach to deal with outliers when maximum-likelihood estimation is used to estimate latent class models. In such a context, the EM algorithm is used and the presence of singularities and spurious local maxima is common. The proposed method is motivated by an application aimed at finding clusters of offending behaviours.

Keywords: Categorical data; Expectation-Maximization algorithm; Local maxima; Outliers; Trimmed log-likelihood.

1 Introduction

We address the problem of the outliers detection and robust estimation in the context of latent class (LC) models for categorical data. These models, introduced by Lanzarsfeld and Henry (1968), represent a valid tool to explain the association between the categorical variables by assuming the existence of a finite set of latent classes. Maximum likelihood estimates of the model parameters are found by using the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) and inference is based on the solution corresponding to the largest value of the log-likelihood at convergence. As for other finite mixture models (McLachlan and Peel, 2000), strategies to single out the global maximum (e.g. Aitkin et al., 1981; McCutheon, 2002) still need improvements. The likelihood may be multimodal and to deal with this problem, a random rule may be applied for the initialization of the EM algorithm. This method, when repeated a suitable number 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

times, allows us to explore the parameter space adequately, provided that the number of parameters is reduced.
We propose a joint use of the trimmed maximum likelihood approach as that developed by Garcia-Escudero et al. (2014) and of appropriate constraints on the parameters of the latent variable and on the parameters of the conditional distribution of the response variables given the latent variable (Bartolucci et al., 2007; Pennoni, 2014). The aim is to obtain a potential improved fit of the LC model. This may also allows us to cope with the problem of the multimodality of the likelihood function.
We illustrate the proposed method by considering the sample data from 1 in 13 sample of all England and Wales offenders born in 1953 related to the dates of conviction and type of offenses from age 16 up to age 20.

2 Proposal

We denote by $Y_1, \ldots, Y_r$ the categorical response variables with categories labeled from 0 to $c-1$. It includes the case of binary responses when $c = 2$. We suppose the existence of a latent variable U with k levels, $u = 1, \ldots, k$. The model parameters are the conditional probabilities of a single response variable y_j given the latent variable denoted by $\phi_{jy|u}$ and the weights π_u for each latent class. Obviously, we have $\sum_{u=1}^{k} \pi_u = 1$ and $\sum_{y=0}^{c-1} \phi_{jy|u} = 1$. Given a sequence of responses $\boldsymbol{y} = (y_1, \ldots, y_r)$, the conditional distribution of all responses given the latent variable is given by

$$p(\boldsymbol{y}|u) = \prod_{j=1}^{r} \phi_{jy_j|u},$$

and, then, the manifest probability of this sequence is equal to

$$p(\boldsymbol{y}) = \sum_{u} \pi_u p(\boldsymbol{y}|u).$$

The posterior probability that an individual with response vector $\boldsymbol{y}$ belongs to the latent class u is used to construct the allocation rule for each individual to a latent class.
In order to estimate the model parameters on the basis of the observations $\boldsymbol{y}_i$, $i = 1, \ldots, n$, we maximize the weighted log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i} w_i \log p(\boldsymbol{y}_i),$$

where the weight w_i is close to 0 for outliers and θ denotes the vector of the model parameters. This maximization is based on the EM algorithm and uses the weighted complete log-likelihood, which is equal to

$$\ell^*(\boldsymbol{\theta}) = \sum_{i} w_i \left[\sum_{u} z_{iu} \left(\sum_{j} \sum_{y} I(y_{ij} = y) \log \phi_{jy|u} + \log \pi_u \right) \right],$$

where $I(\cdot)$ is the indicator function and z_{iu} is an indicator variable equal to 1 if subject i belongs to latent class u and to 0 otherwise.
The estimation is carried out by using a modified version of the EM algorithm. In the standard case, the EM is performed in the following way. At the E-step, we consider the conditional expected value of the frequency of subjects in each latent class u having value y for the j-th response variable. These are computed at the current value of the parameters. At the M-step, the complete data log-likelihood is maximized by using exact solutions for π_u and for $\phi_{jy|u}$.
The optimal number of latent classes is selected by considering the Bayesian information criterion (BIC, Schwarz, 1978) which involves a penalty for the number of parameters:

$$BIC = -2\hat{\ell} + \log(n)\#par,$$

where, for a given model, $\hat{\ell}$ is the maximum of the log-likelihood and $\#par$ is the number of free model parameters. According to this criterion, the number of classes corresponding to the minimum of the index has to be selected. The estimated proportion of classification error is also considered for each latent class which states how well the latent classes are separated. Here, we propose to use a similar strategy as that developed by Garcia-Escudero et al. (2014); see also Neykov et al. (2007). First, we identify the number of latent classes according to the BIC index. Then, we perform the modified version of the EM algorithm for a large number of random starting values. In the E-step, the observations with the smallest likelihood contribution are tentatively discarded, by setting the corresponding posterior probabilities z_{iu} equal to zero for all $u = 1, \ldots, k$. In the M-step the model parameters are updated on the basis of the selected subsample of observations. After applying the trimmed EM steps, the associated weighted likelihood is evaluated, by setting $w_i = 0$ for the discarded observations and $w_i = 1$ for the selected subsample, until convergence of the algorithm.

3 Application

In criminology research a common task is that of clustering criminal behaviors accounting for their evolution in time. For this aim, LC models may be effectively used. We apply the proposed methodology for the analysis of criminal data referred to males and females offenders in England and Wales. More precisely, the data refer to the 1953 cohort (Francis et al., 2010) and concern 38 binary different indicators of criminal activity. We consider only the young males and females within the cohort age range 16-20 years old, which covers $n = 4558$ cases.
We first fit the LC model for an increasing number of latent classes k from 1 to 11. Table 1 shows the results of this preliminary fitting, in terms of maximum log-likelihood and the corresponding values of the BIC index.

According to this criterion, we select 8 latent classes. The latent classes identified according to the estimated conditional probabilities are the following:

1. shoplifting (9,8%);
2. criminal damage (7,7%);
3. theft by employee: with some fraud and forgery (3,6%);
4. theft (17,7%);
5. versatile of type 2: theft from vehicles, handling and receiving stolen goods, burglary and going equipped (24,9%);
6. versatile of type 1: burglary, commercial burglary theft, criminal damage with some burglary, theft, violence, shoplifting (8,8%);
7. violence (13,4%);
8. fraud and forgery with some theft, handling and receiving stolen goods (14,1%).

We then apply the proposed trimmed estimation strategy, assuming the selected number of latent classes $k = 8$, for different trimming levels, 0.25%, 0.9%, and 2%. Preliminary results are reported in Table 2 for the 0.25% trimming level, which leads to discard 12 observations. For these cases (outlier), the table reports the corresponding residual deviance.
In particular, the first case refers to a subject convicted for violence, burglary, going equipped and robbery. Case 1884 was convicted for violence, robbery and theft. Case 181 was convicted for violence, sexual offenses, handling and receiving stolen goods, criminal damage and perjury/attempting to pervert course of justice. The group cases made of id's 1581, 3255, 4264, 4265 refer to subjects convicted for violence, sexual consensual, burglary (dwelling), going equipped, theft, theft from vehicles, shoplifting.
The group cases made of id's 1566, 1576, 2564 for violence, sexual with above 16 years old, burglary (dwelling), theft, theft from person, theft by employee and criminal damage. Case 3719 was convicted for violence, sexual with above 16 years old, sexual under 16 years old and sexual consensual, burglary (dwelling), burglary (other) and theft from vehicles. Case 1361 was convicted for lethal violence, sexual under 16 years old, going equipped, shoplifting and criminal damage. These are cases mostly characterized by sexual offenses with violence and property offenses. This cluster collects the most dangerous individuals, which were not properly identified in the standard LC model.
Finally, it is worth noting that the results of the trimmed estimation strategy may also be used to define sensible starting values for the EM algorithm in the standard approach, so as to prevent the problem of the multimodality of the model log-likelihood.

TABLE 1. Selection of the number of latent classes for the LC model; k is the number of latent classes, $\hat{\ell}$ is the corresponding maximum log-likelihood, $\#par$ is the corresponding number of parameters, and index BIC is defined in Section 2.

k	$\hat{\ell}$	$\#par$	BIC
1	-23315,36	34	46916,41
2	-23006,21	69	46592,18
3	-22671,11	104	46216,07
4	-22498,11	139	46164,17
5	-22294,99	174	46052,00
6	-22162,30	209	46080,72
7	-22089,01	244	46228,23
8	-21748,02	279	45840,32
9	-21775,68	314	46189,74
10	-21588,08	349	46108,62
11	-21539,43	384	46305,41

TABLE 2. Cases with high log-likelihood contribution.

ID numbe r	$-2\ell^*$(contribution)
707	68,0585
1884	52,8958
181	49,2723
1581	48,3980
3255	48,3979
4264	48,3979
4265	48,3979
3719	46,9193
1566	45,1571
1576	45,1571
2564	45,1571
1361	42,3028

Acknowledgments: We acknowledge the financial support from the grant RBFR12SHVV of the Italian Government (FIRB project "Mixture and latent variable models for causal inference and analysis of socio-economic data"). F. Pennoni also thanks the financial support of the STAR project *"Statistical models for human perception and evaluation"* funded by the University of Naples Federico II.

References

Aitkin M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A*, **144**, 419 – 448.

Bartolucci, F., Pennoni, F., and Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society, Series A*, **170**, 115 – 132.

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

Francis, B., Liu, J., and Soothill, K. (2010). Criminal lifestyle specialization: female offending in England and Wales. *International criminal justice review*, **20**, 188 – 204.

Garcia-Escudero, L., Gordaliza, A., Mayo-Iscar, A. (2014). A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification*, **8**, 27 – 43.

Lazarsfeld, P.F. and Henry, N.W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, **52**, 299 – 308.

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

McCutheon, A.L. (2002). Basic concepts and procedures in single and multiple group latent class analysis. In: *Applied Latent Class Analysis*, Hagenaars, J.A., McCutheon, A.L. (eds.), pp. 56 – 85. Cambridge University Press.

Pennoni, F. (2014). *Issues on the estimation of latent variable and latent class models*. Scholars' Press, Saarbücken.

Schwarz, G. (1978). Estimation the dimension of a model. *The Annals of Statistics*, **6**, 461 – 464.

An extended adjacent categories model accounting for response styles

Moritz Berger[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: moritz.berger@stat.uni-muenchen.de

Abstract: Rating scales used in behavioral research often may not reflect the true opinion of the respondent because of particular response styles. An adjacent categories model is proposed that simultaneously models content related effects and the effect of the response style. The model provides a simple solution to avoid biased estimates that occur if the response style is ignored. A visualization tool is proposed that allows for an easy interpretation of effects.

Keywords: Adjacent categories; Response styles; Rating scale models.

1 Extended adjacent categories model

In behavioural research many studies show that the presence of response styles may affect the response behaviour and therefore the observed ratings, see Messick (1991) and Baumgartner and Steenkamp (2001). Response styles can, for example, differ across nations, ethnicity or educational level. An literature review on response styles in survey research was given by Van Vaerenbergh and Thomas (2013).
Let $Y_i \in 1, \ldots, k$ denote the response on a rating scale and $\mathbf{x}_i$ a vector of explanatory variables. The common adjacent categories model with logit link that links the explanatory variables to the ordinal response is given by

$$\log\left(\frac{\pi_{i,r+1}}{\pi_{i,r}}\right) = \theta_r + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \qquad r = 1, \ldots, k-1,$$

where $\pi_{i,r} = P(Y_i = r|\mathbf{x}_i)$ denotes the conditional probability of response category r. The logits $\log(\pi_{i,r+1}/\pi_{i,r})$ of adjacent categories are determined by a category specific intercept θ_r and a linear term of the explanatory variables, $\mathbf{x}_i^{\mathrm{T}}\beta$, see, for example, Agresti (2009).

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

We consider responses with symmetric categories like 'strongly disagree', 'moderatly disagree', ..., 'moderatly agree', 'strongly agree' and assume response styles that show a tendency to the middle or extreme categories. In the case of an odd number of response categories k there is a natural middle category $m = \lfloor k/2 \rfloor + 1$. With an additional vector of explanatory variables $\mathbf{z}_i$, the proposed extended adjacent categories model that accounts for extreme response styles has the form

$$\begin{aligned} \log\left(\frac{\pi_{i,r+1}}{\pi_{i,r}}\right) &= \theta_r + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma}, \qquad r = 1, \ldots, m-1 \\ \log\left(\frac{\pi_{i,r+1}}{\pi_{i,r}}\right) &= \theta_r + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma}, \qquad r = m, \ldots, k-1 \end{aligned} \tag{1}$$

Explanatory variables $\mathbf{z}_i$ are assumed to determine the response style and can be the same or different from $\mathbf{x}_i$. While $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ represents the common content related effect, $\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma}$ represents the preference of middle or extreme categories. If $\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma}$ is positive the probability for higher categories increases for $r = 1, \ldots, m-1$ but decreases for $r = m, \ldots, k-1$. The extreme value $\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} \to \infty$ yields $\pi_{i,m} \to 1$, which means a strong tendency to the middle category and $\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} \to -\infty$ results in $\pi_{i,2}, \ldots, \pi_{i,k-1} \to 0$, which means a strong tendency to the extreme categories 1 and k.
In the case of an even number of response categories let $m = \lfloor k/2 \rfloor$ define the middle between agreement and disagreement. The model then has the form

$$\begin{aligned} \log\left(\frac{\pi_{i,r+1}}{\pi_{i,r}}\right) &= \theta_r + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma}, \qquad r = 1, \ldots, m-1 \\ \log\left(\frac{\pi_{i,m+1}}{\pi_{i,m}}\right) &= \theta_m + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \\ \log\left(\frac{\pi_{i,r+1}}{\pi_{i,r}}\right) &= \theta_r + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma}, \qquad r = m+1, \ldots, k-1 \end{aligned} \tag{2}$$

The interpretation of effects remains the same as in model (1) for an odd number of categories.

2 Seperation of effects

A very important property of the extended models (1) and (2) is that the content related effects $\boldsymbol{\beta}$ are clearly separated from the response style effects $\boldsymbol{\gamma}$, even if $\mathbf{x}_i = \mathbf{z}_i$. In the simplest case of three response categories k=3 it can be shown that the parameter β_j is given by

$$\exp(2\beta_j) = \frac{\pi_3(x_j+1)/\pi_1(x_j+1)}{\pi_3(x_j)/\pi_1(x_j)}, \tag{3}$$

where $\pi_r(x_j)$ and $\pi_r(x_j + 1)$ denote the probabilities of response category r for the vector of explanatory variables with j-th component x_j and if the j-th component is increased by one unit, while all other components are kept. For the parameter γ_j one obtains

$$\exp(2\gamma_j) = \frac{\pi_2(z_j+1)/\pi_1(z_j+1)}{\pi_3(z_j+1)/\pi_2(z_j+1)} \Big/ \frac{\pi_2(z_j)/\pi_1(z_j)}{\pi_3(z_j)/\pi_2(z_j)} \,. \tag{4}$$

According to equations (3) and (4) the parameter β_j depends only on probabilities for different values of x_j and the parameter γ_j depends only on probabilities for different values of z_j. Thus, parameter β_j is separated from $\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma}$ as well as parameter γ_j is separated from $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. This result also holds in the general case, $k > 3$, and ensures the identifiability of the model.

3 Accuracy of estimates

Although, as shown in the previous section, effects are separated one can not ignore the response style when *estimating* content related effects $\boldsymbol{\beta}$, which are most often the parameters of interest. Simulations showed that ignoring the response style effects γ leads to strongly biased results and poor accuracy of the estimates $\boldsymbol{\beta}$ if the data are generated by the extended adjacent categories model. We consider the case of one standard normal distributed explanatory variable with impact $\beta = 1$ and generate data with 7 categories and intercepts $\theta_r = 0$, $r = 1, \ldots, 6$. Figure 1 shows the mean squared errors, variances and biases of the estimates of β for values $\gamma \in [-2, +2]$ when fitting a simple adjacent categories model (dashed lines) or the extended model (solid lines). The upper panel shows the case where $x = z$. For large absolute values of γ the MSEs strongly increase if the response style is ignored. This is due to the bias, which shows an underestimation of the effect β even for moderated absolute values of γ. If $\beta = -1$, the opposite result is found, namely that the effects are overestimated. In the case where x and z are independently normal distributed (lower panel of Figure 1) the resulting curves show the same pattern. It is seen that biased estimates are present if a variable has two effects, one content related effect and one effect on the response style, and if the response style variable is independent of x.

All previous results show a attenuation of estimated effects, which also appears in random effects models if heterogeneity is ignored, see, for example, Tutz (2012). It is worth noting that there are also scenarios where other phenomena occur. We also simulated data with decreasing intercepts $\theta_1 = 0$, $\theta_2 = -0.4$, …, $\theta_6 = -2$. Therefore, higher categories are generally preferred. If there is in addition a positive effect β and a tendency to extreme categories (γ positive), one obtains a positive bias in the case were $x = z$. This specific combination of parameters results in larger values of β for the simple adjacent categories model.

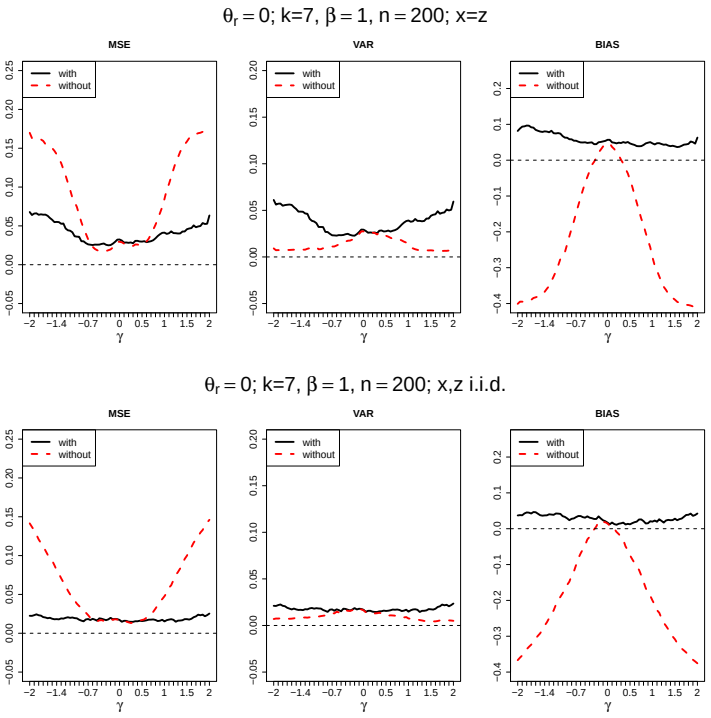

FIGURE 1. MSEs, variances and bias as a function of γ; in the upper panel one has $x = z$, in the lower panel x and z differ and are independent. Dashed lines indicate the model without accounting for the response style, solid lines indicate the model with response style effects.

4 Estimation of parameters and implementation

The proposed model can be estimated and evaluated by embedding the extended adjacent categories model into the framework of multivariate generalized linear models (GLMs). For data $(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, one assumes a multinomial distribution $\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i \sim M(1, \boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i^{\mathrm{T}} = (\pi_{i1}, \ldots, \pi_{ik})$ and $\pi_{ir} = P(Y_i = r|\mathbf{x}_i, \mathbf{z}_i)$. The extended model can then be represented by

$$g(\pi_i) = \mathbf{X}_i\boldsymbol{\delta},$$

where $\mathbf{X}_i$ is a design matrix of values $\mathbf{x}_i$ and $\mathbf{z}_i$, $\boldsymbol{\delta}$ is the total vector of parameters and $g(\cdot)$ is the logit link function. Maximum likelihood estimation and inference for multivariate GLMs can be applied, see, for example, Fahrmeir and Tutz (2001).

For the implementation of the model we make use of the function `vglm()` of the flexible R-package VGAM (Yee, 2010), that allows to estimate so-

called vector generalized linear models. Embedding into VGAM has the advantage of an easy handling and quite fast computation.

TABLE 1. Parameter estimates and standard errors for the student questionnaire.

	Covariates	**Extended Adjacent**		**Adjacent**	
		Estimate	se	Estimate	se
Content related effects	Psychology	0.446	0.187	0.634	0.169
(x-variables)	Physics	0.662	0.182	0.880	0.163
Response style effects	Psychology	0.215	0.231		
(z-variables)	Physics	0.526	0.223		

5 Application

As an example we consider data from a student questionnaire. The objective was to evaluate what effect the expectation of students of getting an appropriate job has on their motivation. The response is the effect on motivation measured on a scale from 1 (often negative) to 5 (often positive). For the analysis we use data from 343 students from subject areas psychology, physics and teaching serving as explanatory variable. The comparison of a simple adjacent categories model and the extended model by a likelihood ratio test ($H_0 : \gamma = 0$) has the χ^2-value of 6.14 on 2 degrees of freedom. Therefore, response style effects should not be neglected. The estimated coefficients and corresponding standard errors for both models with reference category "teaching" are given in Table 1. The estimates of the extended model show that students of psychology and physics see more positive effects on their motivation than students of a teaching profession and that there is a significant tendency to middle categories for students of physics as compared to students of a teaching profession.
Content related effects in Table 1 are considerably larger for the simple adjacent categories model than for the extended model. Thus one might suspect a positive bias in the estimated β-coefficients when ignoring the response style effects. This result corresponds to the findings based on simulated data. In particular when several explanatory variables are included in the model a visualization tool for investigations of effect strengths is useful. In the case of an odd number of response categories (model (1)) the odds between adjacent categories is modified by the factor $\exp(\beta_j)$ if the jth component x_j is increased by one unit. If the jth component z_j is increased by one unit the odds between adjacent categories is modified by the factor $\exp(\gamma_j)$ for categories $1, \ldots, m-1$ and by the factor $-\exp(\gamma_j)$ for categories $m, \ldots, k-1$. If $x_j = z_j$ the effects can be seen by plotting the tupel $(e^{\gamma_j}, e^{\beta_j})$. If a variable is only present in $\mathbf{x}$ or $\mathbf{z}$ the value of the

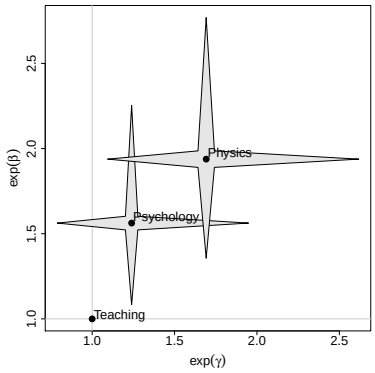

FIGURE 2. Visualization of estimated effects for the student questionnaire.

other component in the tupel is 1. Figure 2 shows the tupel $(e^{\gamma_j}, e^{\beta_j})$ of the extended model for the student questionnaire. For the reference category "teaching" one obtains the values $(e^{\gamma_j}, e^{\beta_j}) = (1, 1)$. We included confidence intervals that are represented by stars, where the horizontal and vertical lengths correspond to the point-wise 0.95 confidence intervals of e^{γ_j} and e^{β_j}, respectively.

References

Agresti, A. (2009). *Analysis of Ordinal Categorical Data, 2nd Edition*. New York: Wiley.

Baumgartner, H. and Steenkamp, J.-B.E.M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, **38**, 143 – 156.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer.

Messik, S. (1991). Psychology and methodology of response styles. *Improving inquiry in social science: A volume in honor of Lee J. Cronbach*, 161 – 200.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press.

Van Vaerenbergh, Y. and Thomas, T.D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, **25**, 195 – 217.

Yee, T. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, **32**, 1 – 34.

Quantile regression: a Bayesian robust approach

Marco Bottone[1], Mauro Bernardi[2], Lea Petrella[1]

[1] Sapienza University of Rome, Italy Department of Methods and Models for Economics, Territory and Finance
[2] University of Padua, Italy

E-mail for correspondence: marco.bottone@uniroma1.it

Abstract: Traditional Bayesian quantile regression relies on the Asymmetric Laplace distribution (ALD) mainly because of its theoretical property of having the τ–level quantile as the natural location parameter. Despite its huge success, the ALD displays medium tails and it is not suitable for data characterized by strong deviations from the Gaussian hypothesis. In this paper, we propose an extension of the Bayesian quantile regression framework that accounts for heavy–tails using the Skew Exponential Power (SEP) distribution having an additional parameter governing the tails–fatness. The linear model is then extended to the class of Generalized Additive Models, where the SEP and the penalized spline functions approximation are combined to robustify the non–linear additive model estimation. Model parameters are estimated using a new adaptive Metropolis–within–Gibbs algorithm. The proposed methodology is validated on simulated and real datasets.

Keywords: Bayesian quantile regression; Skew exponential power; GAM.

1 Introduction

Quantile regression has become a very popular approach to provide a more complete description of the distribution of a response variable conditionally on a set of regressors. Since the seminal work of Koenker and Basset (1978), several papers have been proposed in literature providing contributions to the quantile regression analysis both from a frequentist and a Bayesian points of view. For the former, the estimation strategy relies on the minimization of a given loss function while the Bayesian approach, introduced by Yu and Moyeed (2001), considers the ALD as likelihood function to perform the inference. For a wide and up to date review of the Bayesian

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

quantile regression literature based on the ALD assumption see, e.g., Lum and Gelfand (2012), Sriram et al. (2013) and Bernardi et al. (2015).
Although its nice properties to model conditional quantiles within the Bayesian framework, the ALD displays medium tails which may give misleading informations for extreme quantile confidence levels, in particular when data are characterized by the presence of outliers and heavy tails. The absence of a parameter governing the tails–fatness may influence the final inference. Recently, Wichitaksorn et al. (2014) generalize the classical Bayesian quantile regression using the skewness mechanism of Fernandez and Steel (1998). This class of distributions allow for different degrees of asymmetry of the response variable imposing a given tails–decay. To overcome this drawback, we propose an extension of the Bayesian quantile regression which relies on the Skew Exponential Power (SEP) distribution (see Zhu and Zinde–Walsh, 2009). The SEP distribution shares with the ALD the essential property of having the τ-level quantile as the natural location parameter, but, in addition, it has a parameter governing the decay of the tails. Accounting for the tail behaviour, the proposed SEP distribution allows for to robustify the quantile regression parameter estimation procedure, in particular when outliers or extreme values are observed.
This paper propose a new Bayesian linear quantile regression model based on the SEP as innovation distribution. The linear quantile framework is subsequently generalized to the non linear generalized additive model (GAM) of Hastie and Tibshirani (1986). More precisely, we approximate each additive non linear regression function using the penalized spline (P–Spline) approach. To efficiently estimate the model parameters, we propose an Adaptive Independent Metropolis–within–Gibbs algorithm. The performance of the model in terms of robustness of the delivered estimates are illustrated using both simulated and real datasets.
The paper is organised as follows. Section 2 introduce the quantile regression model based on the SEP distribution, Section 3 proposes a simulation studies and analyses a real dataset, while Section concludes.

2 Model and inference

Zhu and Zinde–Walsh (2009) recently propose a parametrization of the SEP which has the nice property of having the τ–level quantile as natural location parameter. The SEP density, parameterized as in Zhu and Zinde–Walsh (2009), can be written as:

$$f\left(y;\mu,\sigma,\tau,\alpha\right)=\begin{cases}\frac{1}{\sigma}\kappa\left(\alpha\right)\exp\left\{-\frac{1}{\alpha}\left(\frac{\mu-y}{2\tau\sigma}\right)^{\alpha}\right\}, & \text{if} \quad y\leq\mu\\ \frac{1}{\sigma}\kappa\left(\alpha\right)\exp\left\{-\frac{1}{\alpha}\left(\frac{y-\mu}{2(1-\tau)\sigma}\right)^{\alpha}\right\}, & \text{if} \quad y>\mu,\end{cases} \tag{1}$$

where $y\in\Re$, $\mu\in\Re$ is the location parameter, $\sigma\in\Re^{+}$ and $\alpha\in(0,\infty)$ are the scale and shape parameters, respectively, $\tau\in(0,1)$ is the skewness

parameter while $\kappa = \left[2\alpha^{\frac{1}{\alpha}}\Gamma\left(1+\frac{1}{\alpha}\right)\right]^{-1}$ and $\Gamma(\cdot)$ is the complete gamma function. It can be showed that the ALD and the asymmetric Gaussian are particular cases when $\alpha = 1$ and $\alpha = 2$ respectively. Now, suppose to have a random sample of T observations $\{Y_t\}_{t=1}^{\mathrm{T}}$ with the associated set of p covariates $\mathbf{X}_t = (1, X_{t,1}, \ldots, X_{t,p-1})^{\mathrm{T}}$. Let consider the following quantile regression model

$$Y_t = \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\beta}_\tau + \epsilon_t, \qquad \forall t = 1, \ldots, T, \tag{2}$$

where $\boldsymbol{\beta}_\tau = (\beta_{\tau,0}, \beta_{\tau,1}, \ldots, \beta_{\tau,p-1})^{\mathrm{T}}$ is a vector of p unknown regression parameters that depend on τ. Here, ϵ_t, for $t = 1, \ldots, T$, are independent random variables with zero τth quantile and constant variance. To model the τth conditional quantile of the response variable, $Q_\tau(Y_t \mid \mathbf{X}_t = \mathbf{x}_t)$, we make use of the likelihood function given in (1) with the location parameter replaced by the regression function, namely, $\mu = \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\beta}_\tau$, where $\mathbf{x}_t = (1, x_{t,1}, \ldots, x_{t,p-1})^{\mathrm{T}}$ are known realizations of $\mathbf{X}_t$. In what follows, we omit the subscript τ for simplicity. Moreover, in order to generalize the model in equation (2), we consider the GAM extension (Hastie and Tibshirani, 1986), using the likelihood in equation (1), within a semi-parametric Bayesian quantile regression framework:

$$Q_\tau(Y_t \mid \boldsymbol{X}_t = \mathbf{x}_t, \mathbf{Z}_t = \mathbf{z}_t) = \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{J} f_j(z_{j,t}), \tag{3}$$

where $\mathbf{x}_t^{\mathrm{T}}\boldsymbol{\beta}$ is the parametric component while $\mathbf{z}_t = (z_{1,t}, \ldots, z_{J,t})^{\mathrm{T}}$ is an additional set of covariates and each $f_j(z_{j,t})$ is a nonparametric continuous smooth function. To implement the Bayesian analysis we assume that $f_j(z_{j,t})$, can be approximated using a polynomial spline of order d, with $k+1$ equally spaced knots between $\min(z_{j,t})$ and $\max(z_{j,t})$, for $j = 1, \ldots, J$, i.e.

$$Q_\tau(Y_t \mid \mathbf{X}_t = \mathbf{x}_t, \mathbf{Z}_t = \mathbf{z}_t) = \mathbf{x}_t^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{J}\sum_{\nu=1}^{k+d} \theta_{j,\nu} B_{j,\nu}(z_{j,t}), \tag{4}$$

where $B_{j,\nu}(z_{t,j})$ denote B–spline basis functions and $\theta_{j,\nu}$ are the unknown coefficients to be estimated. In order to guarantee that there is enough model flexibility without incurring the overfitting problem, we adopt the penalized spline (P–Spline) approximation assuming that $\theta_{j,\nu} = 2\theta_{j,\nu-1} - \theta_{j,\nu-2} + u_{j,\nu}$ with $u_j \sim \mathcal{N}(0, \phi^2)$. Gaussian priors are chosen for all the regression parameters while an Inverse Gamma prior is used for the penalization parameter ϕ^2. Finally we choose a Beta distribution defined on $(0, 2)$ for the shape parameter α. To sample from the posterior distribution we employ a Metropolis algorithm with an adaptive selection of the proposal parameters. A separate Gibbs step is performed for the penalization parameter in order to improve efficiency.

3 Simulation and empirical studies

To provide a performance comparison between the Bayesian quantile regression approach based on the usual ALD likelihood specification and that obtained by imposing the proposed SEP distribution, we consider both a simulated study and a real data analysis. First, we simulate data from two underlying median curves on $(0, 1)$

$$Wave: \quad f(x) = 4(x - 0.5) + 2\exp\left(-256(x - 0.5)^2\right) \tag{5}$$

$$Doppler: \quad f(x) = (0.2x(1 - 0.2x))^{\frac{1}{2}} \sin(2\pi(1 + \epsilon)/(0.2x + 0.15)). \tag{6}$$

For these two median curves, data are generated by adding a student-t error term with quadratic heterogeneity. Results are reported in Table 1. It is clearly evident that the SEP assumption outperforms the ALD for both the non-linear specifications and all the considered quantile confidence levels τ.

To illustrate the performance of the SEP quantile regression methodology on a real dataset, we analyze the dataset related to the international economic growth model firstly considered by Barro and Sala i-Martin (1995) and extended to the quantile regression framework by Koenker and Machado (1999). To the best of our knowledge, this is the first attempt to propose a Bayesian quantile GAM model in order to study the impact of the covariates on the cross country GDP growth. The dataset contains 161 world nations observed for 13 covariates covering the two periods 1965-75 and 1975-85. Our preliminary investigation (not reported here for conciseness), revealed a noticeable non linear evolution of the " Male secondary school"sp (MSS), " Human Capital"sp (HC) and " Political Instability"sp (PI) variables on the GDP growth rates. For these reason, we model the response function of those variables nonlinearly using the spline approach, while we adopt a linear representation for the remaining variables.

The parameter estimates of the linear covariates are in line with those obtained by Koenker and Machado (1999) and are not reported here. Figure 1 displays the estimated spline functions along with their credible sets, for three quantile levels $\tau = (0.25, 0.5, 0.75)$. The MSS variable has a positive impact on the GDP growth for low ($\tau = 0.25$) and medium ($\tau = 0.5$) quantile confidence levels, while it has a negative impact for the highest quantile level ($\tau = 0.75$). The effect of the HC variable instead is positive and decreasing with the confidence level τ. Those results can be interpreted in view of the fact that, high GDP growth levels are linked with emerging nations where the intangible assets, such as HC and MSS, do not play a central role at their first stage of development. Our results agree with the economic theory and solve the curious results obtained by Koenker and Machado (1999) where the HC variable exhibits a negative impact for all the quantiles. The effect of the political instability on the GDP growth is always slightly decreasing but positive for low and middle quantiles and

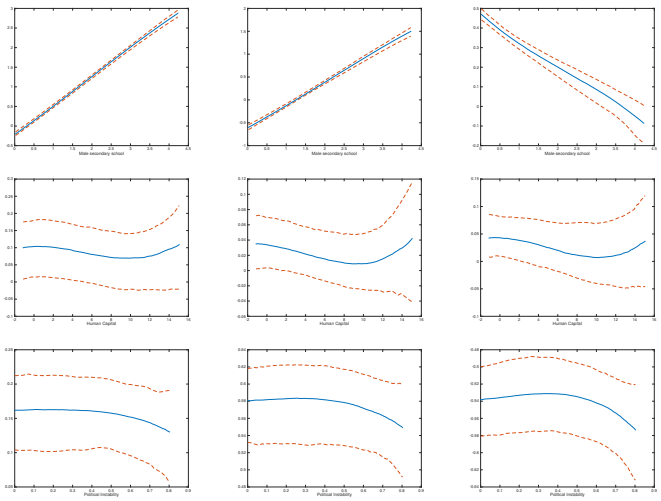

FIGURE 1. Penalized splines with confidence bands for the variables "Male secondary school" (first line), "Human capital" (second line) and "Political Instability" (third line) for three quantile levels (from left to right: $\tau = 0.25, 0.5, 0.75$).

TABLE 1. Average of the Mean Squared Errors and standard errors (in parenthesis) based on 50 repeats for three different quantile levels.

Function	Likelihood	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
Wave	ALD	0.0505	0.0067	0.0752
		(0.1598)	(0.0210)	(0.2377)
	SEP	0.0234	0.0057	0.0305
		(0.0741)	(0.0180)	(0.0965)
Doppler	ALD	0.0018	0.0001	0.0050
		(0.0058)	(0.0006)	(0.0160)
	SEP	0.0011	0.0002	0.0008
		(0.0035)	(0.0006)	(0.0026)

negative for the highest quantile. Even in this case the impact of this variable is different from those obtained by Koenker and Machado (1999), but they adequately explain the effect of the different political systems in emerging and developed countries. Indeed, the emerging nations are in general characterized by authoritarian governments and the presence of political instability usually leads to deep internal crisis. Developed country are instead characterized by democratic governments and the presence of political instability lead to new parliaments or governments and consequently new economic reforms.

4 Conclusion

In this paper, we propose an extension of the Bayesian quantile regression framework that accounts for heavy–tails using the SEP distribution. Our results on simulated and real datasets confirm that modelling the quantile regression error term using the SEP distribution greatly enhances the flexibility of the quantile methodology.

References

Bernardi, M., Gayraud, G., and Petrella, L. (2015). Bayesian tail risk interdependence using quantile regression. *Bayesian Analysis*, (in press).

Barro, R.J. and Sala-i-Martin, X. (1995). Economic Growth. *New York: McGraw-Hill*.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, **1**, 297–310.

Fernandez, C. and Steel, M.F.J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, **93**, 359 – 371.

Koenker, B. and Basset, G. (1978). Regression Quantiles. *Econometrica*, **46**, 33 – 50.

Koenker, B. and Machado, G. (1999). Goodness of fit and related inference precesses for quantile regression. *Journal of the American Statistical Association*, **94**, 1296 – 1310.

Lum, K. and Gelfand, A. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, **7**, 235 – 258.

Sriram, K., Ramamoorthi, R.V., and Ghosh, P. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Analysis*, **2**, 479 – 504.

Wichitaksorn, N., Choy, S.T.B., and Gerlach, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. *Canadian Journal of Statistics*, **42**, 579 – 596.

Yu, K. and Moyeed, R. (2001). Bayesian quantile regression. *Statistics and Probability Letters*, **54**, 437 – 447.

Zhu, D. and Zinde-Walsh, V. (2009). Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics*, **148**, 86 – 99.

Time dependent variables in cure survival models: An application to the timing of a third pregnancy

Vincent Bremhorst[1], Michaela Kreyenfeld[2], Philippe Lambert[1,3]

[1] Université catholique de Louvain, Belgium
[2] Max Planck Institute for Demographic Research, Germany
[3] Université de Liège, Belgium

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

Abstract: When studying the timing of a third pregnancy, classical survival models cannot be used. Indeed, they usually assume that any subject under study will experience the event of interest if she is observed for a sufficiently long time. However, an unknown proportion of mothers will never experience a third pregnancy. Cure survival models extend classical survival models by taking this feature into account. Women and family characteristics such as the marital or the employment status, among others, may vary over time. The inclusion of time dependent variables, influencing simultaneously the probability of observing a third pregnancy and the timing between the second and the third birth for the susceptible mothers will be presented in the specific framework of the promotion time cure model.

Keywords: Cure survival models; Fertility; Time dependent variables.

1 Introduction

When studying the timing of a third pregnancy, classical survival models cannot be used. Indeed, they usually assume that any subject under study will experience the event of interest if she is observed for a sufficiently long time. However, an unknown proportion of mothers will never experience a third pregnancy. Cure survival models extend classical survival models by taking this feature into account. There are two well known families of cure survival models: the mixture cure model (Berkson and Gage, 1952) and the promotion time cure model (Chen et al., 1999). In this work, we shall

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

focus on the second family. This model argues that the failure time of each subject is generated by the minimum of $N \sim P_{ois}(\theta)$ independent latent event times having a common proper distribution $F(t)$ independent of N. One assumes that the latent factors $(Y_1, \ldots, Y_N)$ are independent given N and turn active only at the beginning of the study. One can show that the population survival function is given by

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

Note that $S_p(t)$ is an improper survival function. The probability of being cured differs from 0 and is given by $S_p(+\infty) = \exp(-\theta) = P[N = 0]$.
In this work, an application on data from the German Socio-Economic Panel (Wagner et al., 2007) is presented. The data cover the period 1984-2011. East Germany is, however, only included after 1989. The population under study are mothers with two children. Since women and family characteristics such as marital and employment status can vary over time, regression models enabling to deal with time dependent variables in cure survival studies need to be specified. Unfortunately, the promotion time model in Equation (1) does not allow the inclusion of time dependent variables in the probability of being cured since the latent factors are assumed to be directly active at the beginning of the study. Chi and Ibrahim (2006) relax this assumption by letting latent factors appear at any time during the follow-up. If the failure or censoring occurs at $T = t$, the total number of active latent factors at the end of the follow up is given by $N^* = \int_0^t N(u)du$ where $N(u)$ is the number of latent factors occuring at time u. One assumes that $N(u)$ follows a non-homogeneous Poisson process with intensity function $\lambda(u)$. Using the properties of a Poisson process, $N^* \sim Pois\left(\int_0^t \lambda(u)du = \Lambda(t)\right)$. As in the classical promotion time model, one assumes that only one latent factor needs to be detected for a subject to fail. Then, the population survival function becomes

$$\tilde{S}_p(t) = \exp\left[-\int_0^t \lambda(u)F(t-u)du\right]. \tag{2}$$

Note that if $\int_0^{+\infty} \lambda(u)du$ is bounded, then $\tilde{S}_p(t)$ is an improper survival function and $\tilde{S}_p(+\infty) = \exp\left(-\int_0^{+\infty} \lambda(u)du\right) = P[N^* = 0]$ is the probability of being cured. Otherwise, the survival function in Equation (2) leads to a proper survival function.

2 Regression models

We assume that time dependent covariates ($\boldsymbol{x}(t)$ and $\boldsymbol{z}(t)$) are constant between two successive visits. Following the idea of Chi and Ibrahim (2006), baseline and time dependent covariates ($\boldsymbol{x}_b$ and $\boldsymbol{x}(t)$, respectively) influence

TABLE 1. Numerical results. Estimate of the posterior median, the 95% HPD interval and the posterior standard deviation for each regression parameter.

	Parameter	Estimation	$HPD_{95\%}$	$P[>0]$	sd_{post}
$\theta(\boldsymbol{x}_b)$	Intercept	-1.881	[-2.051; -1.728]	0.000	0.082
	Migrant	0.380	[0.158; 0.627]	0.999	0.120
	Sexcomp (ref: Mixed)				
	Both boys	0.228	[0.035; 0.422]	0.990	0.100
	Both girls	0.281	[0.072; 0.488]	0.995	0.107
	EDU (ref: Voc degree)				
	No degree	0.506	[0.301; 0.703]	1.000	0.104
	Uni degree	0.377	[0.087; 0.663]	0.994	0.148
	Age_std (first birth)	-0.265	[-0.380; -0.152]	0.000	0.059
$F(t\|\boldsymbol{z}_b)$	Migrant	-0.297	[-0.614 ; 0.008]	0.029	0.160
	Age_std (first birth)	0.250	[0.087 ; 0.402]	0.997	0.081

the probability of observing a third pregnancy through a log-link on the intensity function of the non-homogeneous Poisson process.

$$\lambda(t|\boldsymbol{x}_b, \boldsymbol{x}(t)) = \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_b + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}(t)\right). \tag{3}$$

In addition, baseline and time dependent covariates ($\boldsymbol{z}_b$ and $\boldsymbol{z}(t)$, respectively) are supposed to influence the timing between the second and the third birth for the susceptible mothers through the extended Cox model (also known as the Anderson-Gill model)

$$h(t|\boldsymbol{z}_b, \boldsymbol{z}(t)) = R(t)h_0(t)\exp\left[\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{z}_b + \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}(t)\right], \tag{4}$$

where $R(t) = 1$ if the woman is at risk at time t and $R(t) = 0$ otherwise. The logarithm of the baseline hazard ($\log(h_0(t))$ is specified using a linear combination of cubic P-splines. For more details, see Bremhorst and Lambert (2014).

3 Numerical results

We only report on model (1) with baseline covariates. Numerical results on the model including time varying covariates will be reported during the talk. As in Bremhorst and Lambert (2014), a log-link will be used for the mean number of latent factors θ and a flexible Cox proportional hazard model will be specified for the latent distribution $F(t)$.
For the 4200 mothers in the study, baseline informations such as the migration status, the sex composition of prior kids and the age at first birth are available. The education level of the mother is changing over time for

TABLE 2. Posterior probabilily to have a third child for non migrant women with a vocational degree for different combinations of the age of the mothers at first birth and sex composition for the first two kids.

Groups	$Q_{2.5\%}$	$Q_{50\%}$	$Q_{97.5\%}$
Mixed - 21 years	14.4%	17.1%	20.0%
Mixed - 24 years	12.4%	14.5%	16.7%
Mixed - 27 years	10.4%	12.3%	14.4%
Both boys - 21 years	17.6%	21.0%	25.0%
Both boys - 24 years	15.1%	17.9%	21.1%
Both boys - 27 years	12.6%	15.2%	18.1%
Both girls - 21 years	18.2%	22.0%	26.2%
Both girls - 24 years	15.8%	18.7%	22.1%
Both girls - 27 years	13.2%	15.9%	19.1%

some women. However, in a first analysis, we shall only consider its value just after the second birth. Each covariate can influence simultaneously the probability of having a third child and the timing between the second and the third birth for susceptible mothers. Table 1 reports the posterior median, the 95% HPD interval and the posterior standard deviation for each significant effect. These results suggest that age at first birth and migration status have a significant effect on the probability of having a third child and on the timing between the second and the third birth for the susceptible mothers while the sex composition of prior kids and the education level fixed at second birth only have a significant effect on the probability of having a third child. Table 2 shows the posterior probability to have a third child for non migrant women with a vocational degree for different combinations of the age of the mothers at first birth and sex composition for the first two kids. For example, one can see that the probability of having a third child increases when the age at first birth decreases. The left panel of Figure 1 provides the fitted hazard function for susceptible women. One can see that the instantaneous probability of having a third child increases for about 2 years and then tends to stay at that level for the reported time period. The fitted population hazard function (see right panel of Figure 1) suggest that the instantaneous probability of having a third child increases for about 1.5 years and then decreases afterwards.

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

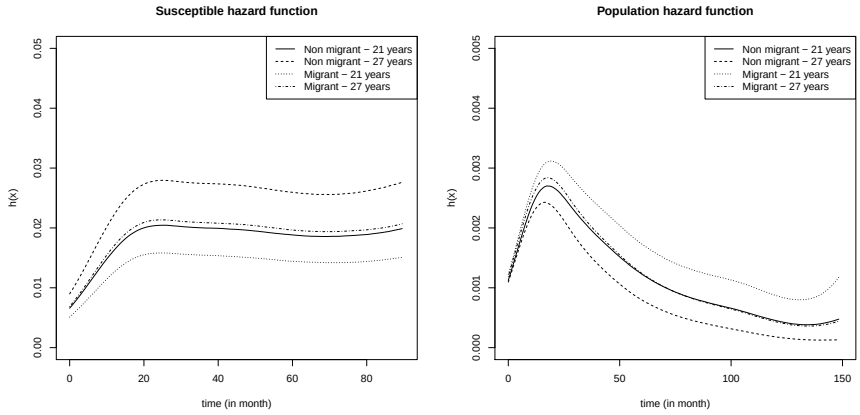

FIGURE 1. Susceptible (left) and population (right) hazard function for different combination of the migration background and of the age at first birth. The value of the other covariates are set to their respective reference level.

References

Berkson, J. and Gage, R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, 501 – 515.

Bremhorst, V. and Lambert, P. (2014). Flexible estimation in cure survival models using Bayesian P-splines. *Computational Statistics and Data Analysis*, http://dx.doi.org/10.106/j.cdsa.2014.05.009.

Chen, M.-H., Ibrahim, J.G., and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**, 909 – 919.

Chi, Y.-Y. and Ibrahim, J.G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, **62**, 432 – 445.

Wagner, G.G., Frick, J.R, and Schupp, J. (2007). The German socio-economic panel study (SOEP): Scope, evolution and enhancements. *Schmollers Jahrbuch*, **127**, 139 – 169.

Functional regression models for location, scale and shape applied to stock returns

Sarah Brockhaus[1], Andreas Fuest[1], Andreas Mayr[2], Sonja Greven[1]

[1] Ludwig-Maximilians-Universität München, Germany
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

E-mail for correspondence: sarah.brockhaus@stat.uni-muenchen.de

Abstract: We propose a generic model class for generalized additive models for location, scale and shape for functional response and/or functional predictors. The models are estimated using a component-wise gradient boosting algorithm, allowing for model selection in high dimensional data settings. The motivating application is a time series on stock returns where expectation and variance should be modeled depending on scalar and functional predictors.

Keywords: Functional data analysis; GAMLSS; Gradient boosting.

1 Introduction

We want to analyze the open-to-close log-returns on the mid-quote for the stocks of Commerzbank traded via the German stock exchange's XETRA system from November 2008 to December 2010. The main interest lies in predicting the variances of the response. As potential predictors one can use lagged response values and the liquidity curves for the cumulative offered and requested number of shares as functions in the relative price averaged per day. See Figure 1 for a descriptive plot of the data. All in all, we want to fit a regression model for the variance and the mean of a scalar response with two functional and several scalar predictors. To this end, we will introduce a very general framework for functional regression models with functional or scalar response including generalized additive models for location, scale and shape (GAMLSS, Rigby, and Stasinopoulos, 2005). The estimation of the models is conducted by a component-wise gradient boosting algorithm, that was introduced by Mayr et al. (2012) for GAMLSS with scalar variables. Boosting can fit high dimensional models allowing for many predictors and variable selection.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

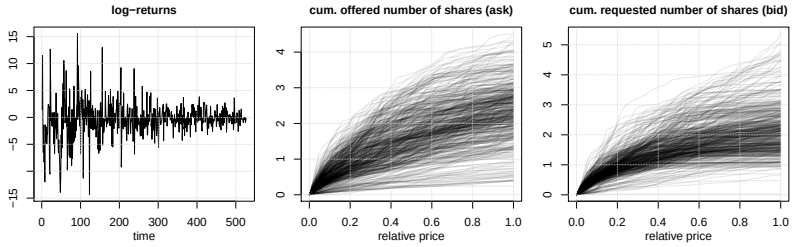

FIGURE 1. The open-to-close log-returns on the mid-quote for the stocks of Commerzbank from November 2008 to December 2010 (left), the cumulative liquidity curves for the offered (middle) and requested number of shares (right) averaged per day.

2 Specification of the model

We assume data $(Y, X) \subset \mathcal{Y} \times \mathcal{X}$ where $\mathcal{Y}$ is the space of square integrable functions $L^2(\mathcal{T}, \mu)$ on a real interval $\mathcal{T}$ with the Lebesgue measure μ. For scalar response we set $\mathcal{Y}$ to $\mathbb{R}$ and use the Dirac measure for μ. $\mathcal{X}$ is a suitable product space of copies of $\mathbb{R}$ and L^2 depending on the scalar and functional predictors. To represent a functional GAMLSS for K distribution parameters, we use the additive regression model

$$\boldsymbol{\xi}^{(k)}(Y|X = x) = h^{(k)}(x) = \sum\nolimits_j h_j^{(k)}(x), \qquad k = 1, \ldots, K, \tag{1}$$

where $\boldsymbol{\xi}^{(k)}$ is a vector of transformation functions indicating the K features of the conditional distribution that are of interest, $h^{(k)}$ are the linear additive predictors and $h_j^{(k)}$ are the corresponding partial effects. To represent a generalized additive model, the model contains only one equation, $K = 1$, and the transformation function is the composition of the expectation $\mathbb{E}$ and the link function g, $\boldsymbol{\xi}^{(1)} = g \circ \mathbb{E}$. Each function in the vector of transformation functions is the composition of a parameter function and a link function. For example, for normally distributed response, the transformation functions can be the expectation, $\boldsymbol{\xi}^{(1)} = \mathbb{E}$, and the variance composed with the natural logarithm as link function, $\boldsymbol{\xi}^{(2)} = \log \circ \mathbb{V}$.
The partial effects $h_j(x)^{(k)}$ are real valued functions and are defined using a basis representation:

$$h_j^{(k)}(x)(t) = \mathbf{b}_{jY}^{(k)}(x, t)^{\mathrm{T}} \boldsymbol{\theta}_j^{(k)}, \qquad j = 1, \ldots, J, \tag{2}$$

where $\mathbf{b}_{jY}^{(k)}(x, t)$ is a matrix of basis evaluations and $\boldsymbol{\theta}_j^{(k)}$ is a vector of coefficients. The effects are regularized by ridge-type penalties.
To improve readability we omit the index (k) in the following representations of the partial effects $h_j^{(k)}$. In order to represent effects (2), we use

the row tensor product $\odot$ of two marginal bases $\mathbf{b}_j : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^{K_j}$ and $\mathbf{b}_Y : \mathcal{T} \to \mathbb{R}^{K_Y}$ multiplied with the coefficient vector $\boldsymbol{\theta}_j \in \mathbb{R}^{K_j K_Y}$,

$$h_j(x)(t) = (\mathbf{b}_j(x,t)^\top \odot \mathbf{b}_Y(t)^\top)\,\boldsymbol{\theta}_j. \tag{3}$$

The row tensor of $\mathbf{A} \in \mathbb{R}^{n \times a}$ and $\mathbf{B} \in \mathbb{R}^{n \times b}$ is the $n \times ab$ matrix $\mathbf{A} \odot \mathbf{B} = (\mathbf{A} \otimes \mathbf{1}_b^\top) \cdot (\mathbf{1}_a^\top \otimes \mathbf{B})$, where $\otimes$ is the Kronecker product and $\cdot$ entry-wise multiplication. As penalty matrix we use $\mathbf{P}_{jY} = \lambda_j(\mathbf{P}_j \otimes \mathbf{I}_{K_Y}) + \lambda_Y(\mathbf{I}_{K_j} \otimes \mathbf{P}_Y)$, where $\mathbf{P}_j \in \mathbb{R}^{K_j \times K_j}$ and $\mathbf{P}_Y \in \mathbb{R}^{K_Y \times K_Y}$ are appropriate penalty matrices for the marginal bases and λ_j, $\lambda_Y \geq 0$ are smoothing parameters. In the special case that the effect of the covariates does not depend on the current index of the response, i.e., $\mathbf{b}_j(x,t) = \mathbf{b}_j(x)$, and the responses are observed on a common grid, the effects (3) can be represented as the Kronecker product of two marginal bases

$$h_j(x)(t) = (\mathbf{b}_j(x)^\top \otimes \mathbf{b}_Y(t)^\top)\,\boldsymbol{\theta}_j, \tag{4}$$

yielding a generalized linear array model, which saves computing time and memory during estimation (Currie et al., 2006). This case was introduced by Brockhaus et al. (2015) for non-vector valued transformation functions, i.e., $K = 1$. Table 1 lists some partial effects h_j that can be fitted within our framework, stating also whether the effect is compatible with the array framework (4). All effects can also be chosen to be constant along t.

TABLE 1. Overview of possible effects that can be represented within the framework for functional regression models, stated in (1) and (3).

covariate(s)	type of effect	$h_j(x)(t)$	array
(none)	smooth intercept	$\beta_0(t)$	yes
scalar variable z_1	linear effect	$z_1\beta(t)$	yes
	smooth effect	$\gamma(z_1, t)$	yes
plus scalar z_2	linear interaction	$z_1 z_2 \beta(t)$	yes
	smooth interaction	$f(z_1, z_2, t)$	yes
grouping variable g	group-specific intercept	$\beta_g(t)$	yes
plus scalar z	group-specific linear effect	$z\beta_g(t)$	yes
fun. variable $x(s)$	linear functional effect	$\int_{\mathcal{S}} x(s)\beta(s,t)ds$	yes
plus scalar z	linear interaction	$z\int x(s)\beta(s,t)ds$	yes
	smooth interaction	$\int x(s)\beta(z,s,t)ds$	yes
fun. variable $x(s)$	concurrent effect	$x(t)\beta(t)$	no
with $\mathcal{S} = \mathcal{T}$	historical effect	$\int_{t_0}^{t} x(s)\beta(s,t)ds$	no
	historical lagged effect	$\int_{t-\delta}^{t} x(s)\beta(s,t)ds$	no

3 Estimation by boosting

The generic model (1) is estimated by optimizing an adequate loss criterion, depending on the transformation functions $\boldsymbol{\xi}^{(k)}$ and on the conditional distribution that is assumed for the response. For functional response, the loss is integrated over the domain of the response (Brockhaus et al., 2015). To fit a GAMLSS the negative log-likelihood is used as loss function (Mayr et al., 2012). For the estimation we use a component-wise gradient boosting algorithm (Bühlmann and Hothorn, 2007), which can estimate models with many covariates (even $p > n$) and inherently does variable selection. Boosting is a step-wise procedure that minimizes the expected loss criterion along the steepest gradient decent. Using equal and rather low degrees of freedom for all base-learners, the model complexity is controlled by the number of boosting iterations. The optimal number of boosting iterations is determined by resampling methods. Stopping the algorithm before convergence achieves variable selection and shrinkage of the parameter effects. Boosting for GAMLSS is implemented in the R add-on package gamboostLSS (Hofner et al., 2015) and base-learners for functional predictors are available in the FDboost package (Brockhaus, 2015).

4 Application to stock returns

We fit a scalar-on-function regression model for mean and variance assuming normally distributed response. More specifically, we estimate the following model for the log-returns y_i depending on lagged response variables and the two functional liquidities $x_{\text{ask},i}$ and $x_{\text{bid},i}$, with $i = 1, \ldots, 531$:

$$y_i | y_{i-1}, \ldots, y_{i-\max(p,q)}, \mathbf{x}_i \sim N(\mu_i, \sigma_i^2),$$

$$\mu_i = h^{(\mu)}(x_i) = \alpha_0 + \sum_{l \in \{\text{bid,ask}\}} \int \alpha_l(s) x_{[l]i}(s) ds + \sum_{j=1}^{p} \alpha_j y_{i-j},$$

$$\log \sigma_i^2 = h^{(\sigma^2)}(x_i) = \beta_0 + \sum_{l \in \{\text{bid,ask}\}} \int \beta_l(s) x_{[l]i}(s) ds + \sum_{j=1}^{q} \beta_j \log y_{i-j}^2,$$

where $\mu_i = \mathbb{E}(Y_i) = \boldsymbol{\xi}^{(1)}(Y_i)$ and $\log \sigma_i^2 = \log \circ \mathbb{V}(Y_i) = \boldsymbol{\xi}^{(2)}(Y_i)$ in model (1). As loss function we use the negative log-likelihood of the normal distribution. R add-on package gamboostLSS (Mayr et al., 2012). For the setup of the effects we use the array representation (4). Because of the scalar response, the basis in t is one, $\boldsymbol{\Phi}_Y(t) = 1$. For the scalar effects we use linear models with the lagged observations as base-learners without penalization. For the effect of a functional covariate $x(s)$ the integral is approximated numerically, $\int x_i(s)\beta(s)ds \approx \sum_{r=1}^{R} \Delta(s_r) x(s_r) \beta(s_r)$, with numerical integration weights $\Delta(s)$, and we compute the basis as

$$\mathbf{b}_j(x(s))^{\mathrm{T}} = [\tilde{x}(s_1) \cdots \tilde{x}(s_R)] \, [\boldsymbol{\Phi}_j(s_1) \cdots \boldsymbol{\Phi}_j(s_R)]^{\mathrm{T}},$$

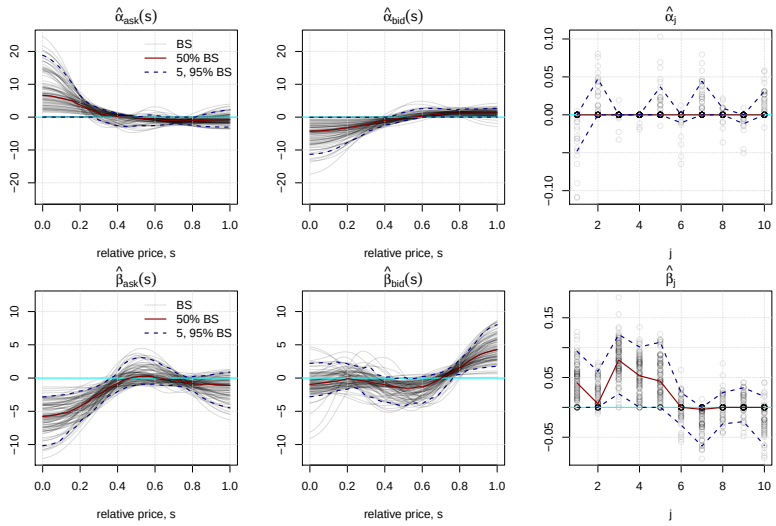

FIGURE 2. Estimated coefficients for μ_i (top panel) and σ_i^2 (bottom panel) in the GAMLSS with the two liquidities as functional covariates and $p = q = 10$ lag variables, computed on 50 block-bootstrap samples (BS).

where $\tilde{x}(s) = \Delta(s)x(s)$ and $\mathbf{\Phi}_j(s)$ is a vector of cubic B-splines evaluated in s. The penalty $\mathbf{P}_j$ is a squared difference matrix of first order. Using $q = p = 10$ lagged variables, we search for the optimal stopping iteration on a two dimensional grid allowing different numbers of boosting iterations for the two parameters. We use 100 fold block-wise bootstrapping (BS) with block length 20 on the first 90% of the time series for the resampling. The last 10% are used to evaluate the model fit (not shown). The estimated coefficients can be seen in Figure 2, where the point-wise 5, 50, and 95% quantiles are marked. The absolute values of the estimated coefficient functions are generally higher for small s, which is sensible, as the bid and ask curves for small s describe the liquidity close to the quoted price. For the effects on the expectation, the estimates of $\hat{\alpha}_{\text{ask}}(s)$ are mostly positive and $\hat{\alpha}_{\text{bid}}(s)$ are mostly negative. Thus, higher liquidity of the ask side and lower liquidity of the bid side tend to be associated with an increase of the expected log-returns. The lagged response values have no influence on the expectation, as $\hat{\alpha}_j$ is virtually always zero. Looking at the model for the variance, the estimated coefficient functions for ask, $\hat{\beta}_{\text{ask}}(s)$, are mostly negative. This means that higher liquidity leads to lower variances, and lower liquidity leads to higher variances. The estimated coefficients for the bid curves, $\hat{\beta}_{\text{bid}}(s)$, are quite close to zero. The logarithm of the lagged squared response values seem to have an influence for close time-points, as many $\hat{\beta}_j$ are greater zero for the first five lags.

5 Conclusions

We provide a general model class to fit GAMLSS for scalar and functional responses and covariates. The framework has a modular structure: (a) the transformation function allows to choose the modeled feature(s) of the conditional response distribution and (b) the additive predictor allows the specification of many covariate effects. The component-wise gradient boosting algorithm used for estimation reflects that generality by (a) optimizing loss functions depending on the transformation function, and (b) providing potentially many base-learners and variable selection. The modular structure invites further extensions such as new loss functions and base-learners.

Acknowledgments: The authors thank Deutsche Börse AG for providing the XETRA historical limit order book data as part of a cooperation with the Chair of Financial Econometrics, LMU München, Germany. The work of Sarah Brockhaus and Sonja Greven was supported by the German Research Foundation through Emmy Noether grant GR 3793/1-1.

References

Brockhaus, S. (2015). FDboost: Boosting functional regression models. R package version 0.0-8. Available at http://CRAN.R-project.org/package=FDboost.

Brockhaus, S., Scheipl, F., Hothorn, T., and Greven, S. (2015). The functional linear array model. *Statistical Modelling*, in press, doi: 10.1177/1471082X14566913.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science*, **22**, 477 – 505.

Currie, I.D., Durban, M., and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, **68**, 259 – 280.

Hofner, B., Mayr, A., Fenske, N., and Schmid, M. (2015). gamboostLSS: Boosting methods for GAMLSS models. R package version 1.1-3. Available at http://CRAN.R-project.org/package=gamboostLSS.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data – a flexible approach based on boosting. *Journal of the Royal Statistical Society, Series C*, **61**, 403 – 427.

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

Regression models with digit preference

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

[1] Institut National d'Études Démographiques, Paris, France
[2] Erasmus University Medical Center, Rotterdam, The Netherlands

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

Abstract: Digit preference (DP) may heavily influence regression outcomes. The pattern of DP and the underlying true distribution can be estimated with a composite link model. These results can be combined with Bayes' rule to assign the probabilities of true values. We replace each original observation (response and covariates) for each of the possible true values and give them weight proportional to their conditional probabilities. Then we fit a model on this expanded data set. We present a simulation study and an application where self-reported weights are regressed on daily minutes of sedentary activity.

Keywords: Composite link model; Conditional distribution; Data augmentation.

1 Introduction

Digit preference, i.e. the tendency to round measurements to pleasing digits, is common in variables collected during surveys, questionnaires, censuses and measurements. Whereas the presence of digit preference (DP) is detectable at the aggregate level, the direct usage of individual data from these misreported variables may jeopardize further analysis. We propose a procedure to correct for digit preference in (non-parametric) regression. The idea is to model DP (Camarda et al., 2008), thus obtaining a matrix of misclassification probabilities and a smooth estimate of the latent true distribution. Bayes' rule then generates for each individual observed value a vector of possible true values and their probabilities. Regression is then done on all possible true values with weights equal to the probabilities.
After an evaluation of the approach by a simulation study, we illustrate the model on data from NHANES (2008), fitting an additive model for self-reported weight with a smooth effect of age, a linear effect of daily sedentary activity and a factor effect of gender.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 The model

We generalize the model presented by Camarda et al. (2008) and we assume that we are studying a weight distribution. DP is interpreted as a misclassification process. Let γ_j be the probability that the real weight is j and let c_{ij} be the probability that it is reported as i. The goal is to estimate the (discrete) distribution $\boldsymbol{\gamma}$ and the matrix $\boldsymbol{C}$ for observed frequencies $\boldsymbol{y}$. This can be done by extending the composite link model (CLM, Thompson and Baker, 1981) with a penalty to force $\boldsymbol{\gamma}$ to be smooth, and by constraining the elements of $\boldsymbol{C}$ towards zero. One approach is to use an L_1 penalty, another is to assume a simple parametric structure. In our application we use both. Details will follow in Section 3. The CLM assumes that the elements of $\boldsymbol{y}$ are Poisson-distributed with expected values $\boldsymbol{\mu} = \boldsymbol{C\gamma}$.
Once we have obtained $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{\gamma}}$ we can apply Bayes' rule to get at the probability that the real weight is j when i was reported. Let r indicate self-reported weight and t true weight. We have that $c_{ij} = P(r = i|t = j)$ and $P(t = j) = \gamma_j$, where $P(\cdot)$ indicates the probability of its argument. It follows that

$$\begin{aligned} P(t = j|r = i) &= \frac{P(r = i|t = j)P(t = j)}{\sum_j P(r = i|t = j)P(t = j)} \\ &= \frac{c_{ij}\gamma_j}{\sum_j c_{ij}\gamma_j} = \frac{c_{ij}\gamma_j}{\mu_i}. \end{aligned}$$

Hence we can connect to every observed weight a vector of possible true weights with corresponding probabilities.
Suppose that we are estimating a simple smooth relationship, $E(z) = f(w)$, where weight w is the explanatory variable showing DP. We augment original observation (w_k, z_k) to get n pseudo-observations, replacing the reported weight by n potential weights and repeating z_k (and covariates) n times. Here n is the length of $, pmb\gamma$. Each pseudo-observations gets weight $P(t = j|r = w_k)$.
If DP is also present in z, we can estimate a model for it too and generate pseudo-observations with their weights. Assuming that DP in z is independent of that in w, we use all possible combinations of exploded w and z, and multiply their associated conditional probabilities. In principle this can lead to an enormous data set, but in practice many conditional probabilities will be zero and the corresponding pseudo-observations can be dropped.

3 A simulation study

To demonstrate the performance of the approach we applied it to two simulated scenarios in which we modified the amount of digit preference in

both response and covariate. Initial data were simulated as follows:

$$z_i \sim \mathcal{N}(\mu_i, 25) \qquad \text{with} \qquad \mu_i = 40 + w_i/2 + 25\sin(w_i/10)$$

with $w_i \sim \Gamma(8, 3)$ and $i = 1, \ldots, 1000$.
Figure 1 shows simulated distributions along with fitted values for both covariate and response. In both cases the assumed digit preference attracted additional observations at each multiplies of 5 from one and two categories to the left and to the right. Two different probabilities are considered: 0.4 and 0.8, which generate left and right panels, respectively. As for the weight distribution presented in the next section, misreporting patterns embody in the $\boldsymbol{C}$ matrix was estimated by weighted least squares, combined with an L_1 penalty.

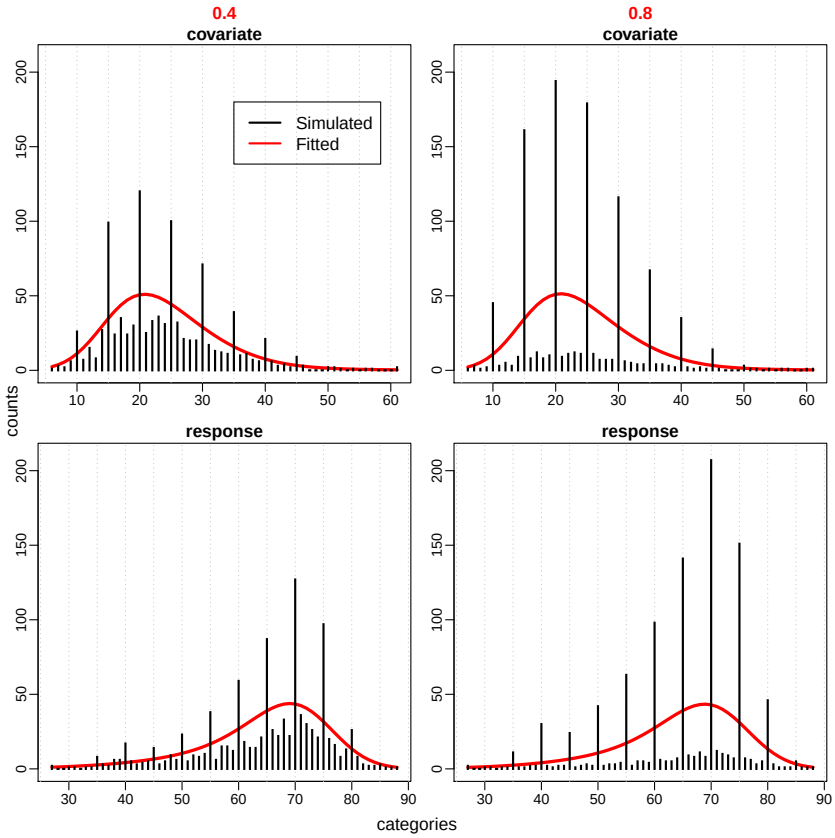

FIGURE 1. Simulated and fitted distribution for covariate (top panels) and response (bottom panels) with different amount of digit preference: 0.4 (left panels) and 0.8 (right panels) as probability to misreported data at each multiply of 5.

We estimate a GAM on both scenarios as well as original and augmented

datasets. Figure 2 presents the outcomes. Differences between left and right panels are not too perceptible in terms of estimated terms, but it is evident the narrowing of the confidence bands in both scenarios when GAM is applied on the augmented dataset corrected for DP. We can also point out a slight increase of the standard errors when moving from a relatively weaker presence of DP (right panel) to a strong one (left panel) for the original regression setting.

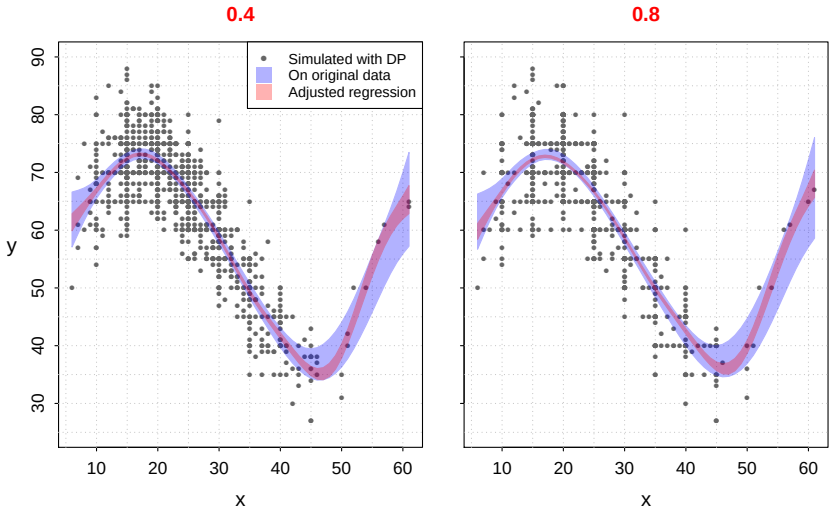

FIGURE 2. Smooth term from a GAM with associated 99% confidence intervals for the simulated covariate. Red and blue areas depict results on original and digit preference adjusted data, respectively. Different amount of digit preference are used: 0.4 (left panels) and 0.8 (right panels) as probability to misreported data at each multiply of 5.

4 An application

We use data from NHANES (2008) and estimate the relationship between self-reported weights and minutes of sedentary activity per day, controlling for gender and age. Figure 3 shows the observed and estimated distributions. Weights are rounded to multiples of 5 and durations to hours.
For weight we estimate the elements of $\boldsymbol{C}$ by weighted least squares, combined with an L_1 penalty, allowing transfers from one and two pounds to the left and to the right. For activity we use decaying exponentials, centered at whole hours. The speed of decay is estimated from the data.
We estimate a GAM for self-reported weight with a smooth influence of age, a linear effect of sedentary activity and sex as a factor.

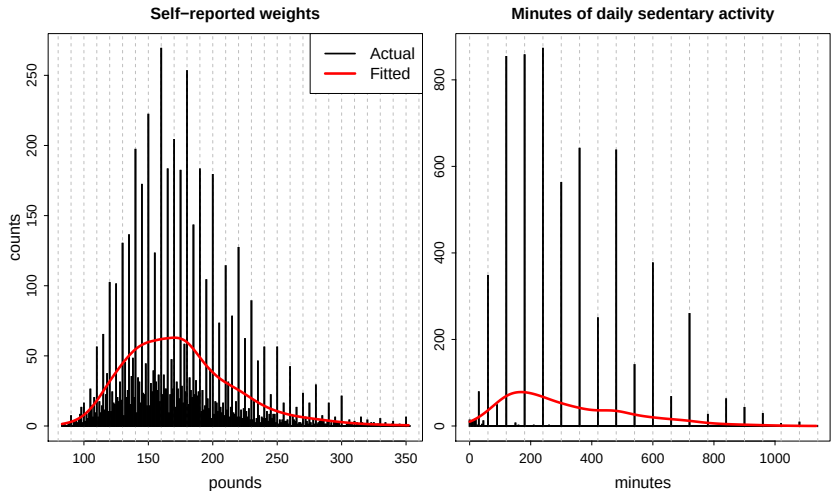

FIGURE 3. Actual and fitted distribution of self-reported weights in pounds (left) and minutes of sedentary activity per day (right). Source: NHANES, 2008.

The functional form for age is in Figure 4. It is evident that accounting for digit preference does not change the expected values of the effects, but shrinks the confidence band. For sedentary activity ($\hat{\beta} = .027$) the standard error drops from 0.0025 to 0.00026. Female weight is on average 28 pounds less than male weight, and the associated standard error drops from 0.98 to 0.1.

5 Conclusions

We have presented a model for applying regression when variables present digit preference and obtained satisfying results.
This approach is not limited to data with digit preference and GAM. We plan to apply it to other regression models and with other types of non-precise variables, e.g. grouped data (Lambert and Eilers, 2009).
A bottleneck is the product of the weights. It can lead to a very large pseudo-data set when the number of explanatory variables with DP grows. It will be worthwhile to develop sampling algorithms that only use a fraction of the possible conditional probabilities.

References

Camarda, C.G., Eilers, P.H.C., and Gampe J. (2008). Modelling general patterns of digit preference. *Statistical Modelling*. **8**, 385 – 401.

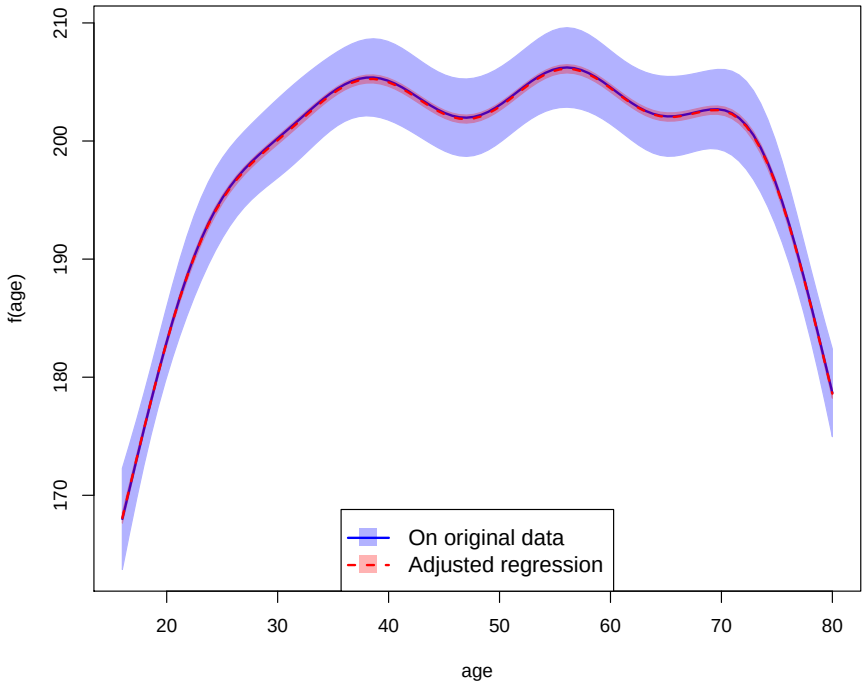

FIGURE 4. Smooth term from a GAM with associated 95% confidence intervals for variable age. Red and blue areas depict results on original and digit preference adjusted data, respectively.

Lambert, P. and Eilers, P.H.C. (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis*, **53**, 1388 – 1399.

NHANES (2008). *National Health and Nutrition Examination Survey*. US National Center for Health Statistics.
Available at `www.cdc.gov/nchs/nhanes.htm`

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

A mixture model for multidimensional ordinal data

Roberto Colombi[1], Sabrina Giordano[2]

[1] Department of Engineering, University of Bergamo, Italy
[2] Department of Economics, Statistics and Finance, University of Calabria, Italy

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

Abstract: In opinion polls, people are requested to express preferences on several aspects related to an item or a service by selecting a category in an ordered scale. We propose a model for such kind of ordinal data. It is a mixture of a uniform distribution and a Sarmanov distribution with CUB marginals (say SCUB). Likeness/repulsion and uncertainty towards every single item – the two components of the respondent's decision in the CUB approach (D'Elia and Piccolo, 2005) – are jointly modelled through the SCUB distribution. As well as specific uncertainty (related to the indecision for single items), the proposed mixture allows us to consider global uncertainty, referred to the respondent's hesitancy in filling up the whole questionnaire. Finally, a real case study highlights the usefulness of the approach.

Keywords: CUB models; Sarmanov distributions; Rating survey.

1 Introduction

Multidimensional ordinal data are often collected by inviting people to express their opinions about topics such as products, brands, services through a category in an ordered list of alternatives (e.g. *strongly disagree*, *disagree*, *neither agree nor disagree*, *agree*, *strongly agree*).
For this kind of categorical data we propose a model which is a mixture of a Sarmanov distribution with CUB marginals (SCUB) and a uniform distribution. The novelty of the proposal is twofold: it generalizes the CUB models (D'Elia and Piccolo, 2005) to the multivariate case and it allows to distinguish two kinds of uncertainty in the answers: global and specific uncertainty. Global uncertainty refers to respondents who are unwilling or hesitant to fill up the whole questionnaire due, for example, to tiredness

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

or lack of involvement, whereas a specific uncertainty is related to the respondent's hesitancy to reply to a single question because of, for example, the partial understanding of the item.
The SCUB distribution is introduced in Section 2 and the mixture model presented in Section 3. An example on real data in Section 4 completes the work.

2 SCUB distribution

The proposed class of SCUB models extends to the multidimensional case the CUB models introduced by D'Elia and Piccolo in 2005 and further developed in the last decade (see Iannario and Piccolo, 2012). According to the CUB approach, the probability of selecting a category x in the ordered list $\{1, \ldots, m\}$ is modelled as a discrete mixture of a (shifted) binomial probability function, $p_x^B(\xi) = \binom{m-1}{x-1}(1-\xi)^{x-1}\xi^{m-x}$, and a uniform probability function $p_x^U = 1/m$

$$p(x; \xi, \pi) = \pi p_x^B(\xi) + (1-\pi)p_x^U, \qquad x = 1, \ldots, m, \tag{1}$$

with $0 < \pi \leq 1$, $0 \leq \xi \leq 1$, with $m > 3$ for the identifiability.
The basic idea of CUB is that when a respondent expresses his/her opinion, with probability π, he/she gives a thoughtful answer, that reflects his/her *feeling* towards the item, and with probability $(1-\pi)$ he/she gives a completely random answer, described by a uniform variable, that reflects his/her *uncertainty*.
The feeling is the personal perception/attitude towards the item (attractiveness or repulsion) while the uncertainty is the indecision intrinsic in any human choice related for example to the lack of self-confidence, limited knowledge, laziness etc.
The CUB approach allows to model parsimoniously the distribution of ordinal data and the parameters are directly related to the concepts of uncertainty/indecision and feeling/preference. In particular, $1-\xi$ increases with the agreement with the item so it gives a feeling measure, while $1-\pi$ raises with the heterogeneity (Tuzt et al., 2014)and it is interpreted as a measure of uncertainty. The higher is π (ξ) the lesser is the contribution of uncertainty (feeling) to the choice.
In our approach, for a vector $(X_1, \ldots, X_v)'$ of v ordinal responses, we introduce a v-variate Sarmanov distribution (Sarmanov, 1966; Lee, 1996) with CUB marginals $p_i(x_i) = p(x_i; \xi_i, \pi_i)$ given by (1), $x_i = 1, \ldots, m_i$, $i = 1, \ldots, v$, whose probability function, at $\mathbf{x} = (x_1, \ldots, x_v)'$, is

$$p_{\mathbf{x}}^{Scub}(\boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\omega}) = \prod_{i=1}^{v} p_i(x_i) \left\{ 1 + \sum_{1 \leq i < j \leq v} \omega_{ij} \varphi_i(x_i) \varphi_j(x_j) \right\}. \tag{2}$$

The $\varphi_i(x_i) = e^{-x_i} - E[e^{-X_i}]$, $i = 1, \ldots, v$, are called mixing functions and the association parameters ω_{ij}, $1 \leq i < j \leq v$, satisfy the conditions $1 + \sum_{1 \leq i < j \leq v} \omega_{ij} \varphi_i(x_i) \varphi_j(x_j) \geq 0$, for all $x_1, \ldots, x_v$. The vectors $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_v)'$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_v)'$, and $\boldsymbol{\omega} = (\omega_{1,2}, \omega_{1,3}, \ldots, \omega_{v-1,v})'$ contain the parameters of feeling, uncertainty and association, respectively.

3 The two-step model

We assume that when a respondent fills up a questionnaire, he/she acts a two-step strategy. At the first step, he/she decides, with probability ϕ, to behave with a propensity to ponder the answers for every item or, with probability $1 - \phi$, to give all the responses at random.
Then, at the second step, each respondent who has decided to reply to the complete questionnaire thoughtfully selects the joint responses $\mathbf{x} = (x_1, \ldots, x_v)'$ according to the SCUB distribution $p_{\mathbf{x}}^{Scub}(\boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\omega})$.
If the respondent is uninterested about the survey, at the second step, he/she opts for a random choice of the answers according to a uniform variable with probability $p_{\mathbf{x}}^{U} = 1/m$, where $m = \prod_{i=1}^{v} m_i$.
Based on the two-step strategy, the observed responses $\mathbf{x} = (x_1, \ldots, x_v)'$ to the v items in the questionnaire are a realization of a random vector $(X_1, \ldots, X_v)'$ whose probability function is defined by the mixture

$$p(x_1, \ldots, x_v) = \phi p_{\mathbf{x}}^{Scub}(\boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\omega}) + (1 - \phi) p_{\mathbf{x}}^{U}. \tag{3}$$

The proposed approach gives rise to two kinds of uncertainty: global and specific uncertainty.
The global uncertainty is referred to respondents who show an a priori refusal of thinking about the whole questionnaire, due for example to the lack of involvement with the topic, willingness to joke, attempt to reduce effort/time, tiredness, laziness and lead people to select all the answers at random. The relevance of global uncertainty behavior depends on $1 - \phi$.
The specific uncertainty is the indecision in evaluating a single item which the respondent considers very unattractive, or partially comprehensible, or time demanding, etc. The relevance of the specific uncertainty is measured by $1 - \pi_i$, $i = 1, \ldots, v$, widely motivated in the literature on CUB models. Subject's characteristics may affect the answers and their effect can be different for every item evaluated. Therefore, the parameters can vary across the n units, so the suffix u, $u = 1, \ldots, n$ is added to the subject dependent quantities.
The effect of subject's covariates $\mathbf{w}_u$ on the feeling parameter of every respondent u, $u = 1, \ldots, n$ for the item i, $i = 1, \ldots, v$, in the questionnaire can be described through the logistic regressions $\log(\frac{\xi_{iu}}{1-\xi_{iu}}) = \mathbf{w}_u' \boldsymbol{\beta_i}$. Analogous logistic models can be introduced for the uncertainty parameters π_u, ϕ_u.

4 Example

As an example we consider a survey conducted at University Federico II of Naples in Italy where 710 interviewees expressed their agreement by selecting a category in the scale of 7 ordered alternatives from *strongly disagree* to *strongly agree*, with the following 4 sentences:

1 - *immigration*: immigrant and Italian workers have the same rights,

2 - *media pluralism*: Italian legislation must guarantee media pluralism,

3 - *vote*: it is a duty to vote, the voters may affect the political decisions,

4 - *ineligibility*: people guilty of corruption must be ineligible in the Parliament.

Education (primary school, secondary school, degree, master), *political orientation* (9 categories from *extreme left* to *extreme right*) and *interest for political broadcasts* (*yes*, *no*) are the covariates.
We fitted the two-step model to the data by taking into account the effects of covariates on the feeling parameters through the logit models

$$\text{logit}(\xi_i) = \beta_{0i} + \beta_{1i} education + \beta_{2i} polorient + \beta_{3i} polbroad,\ i = 1, 2, 3, 4. \tag{4}$$

Remind that, $1 - \xi_i$ increases with the agreement with the item i, while $1 - \pi_i$ rises with the specific uncertainty.
Table 1 displays the ML estimated parameters of the two-step model. We will briefly comment the results.
We found that *political orientation* and the *interest for political broadcasts* are relevant to explain respondents' feeling concerning the first two items *immigration* and *media pluralism*. The coefficients of *political orientation* have positive sign in the logit models (4), for $i = 1, 2$, so we deduce decreasing values of $1 - \xi_i$ moving from *extremely left* to *extremely right* (from 1 to 9) in the personal political ideology. People with a liberal thinking show much more agreement for socio-political matters than conservative people. The covariate *polbroad*, which indicates if the respondent has interest in following the debate on political issues on tv/radio, affects the agreement with the first three statements: *immigration*, *media pluralism* and *vote*. The coefficient of *polbroad* is negative in the logit models (4), for $i = 1, 2, 3$, implying that the feeling towards the first three questions is higher for people who care about political issues treated on tv/radio.
On the other hand, the covariate *education* seems not to have any role in explain feeling for the proposed items. Probably because most of interviewees are similarly qualified. In fact, the level of education of the respondents involved in the survey is high, as about one half of the interviewees has attended a secondary school and about 40% has a university degree.

None of the covariates affects feeling towards the item on *ineligibility*, suggesting a general consensus against corrupted politicians.
Using the estimated parameters $\hat{\alpha}_i$ of Table 1, we get $\hat{\pi}_i = \frac{\exp(\hat{\alpha}_i)}{1+\exp(\hat{\alpha}_i)}$, for $i = 1, 2, 3, 4$. It results that the weights of the feeling component in the CUB models for each item assume quite high values especially with regard to the *ineligibility* question, i.e. $\hat{\pi}_1 = 0.6$, $\hat{\pi}_2 = 0.75$, $\hat{\pi}_3 = 0.69$, and $\hat{\pi}_4 = 0.92$. Being $1 - \pi_i$ a direct measure of indecision in the answers, we deduce that interviewed people share their points of view on the four key socio-political questions with low uncertainty.

TABLE 1. Estimated parameters of the two-step model for data from University of Naples. Logit transformations of specific and global uncertainty parameters are used, i.e. $\alpha_i = \log(\frac{\pi_i}{1-\pi_i})$ for $i = 1, 2, 3, 4$ and $\gamma = \log(\frac{\phi}{1-\phi})$.

Par. Type	Item	Parameter	Estimate	Std error	p-value
feeling	1 immigration	*intercept*	-3.904	0.532	2.1e-13
		education	0.088	0.144	0.541
		pol. orient.	0.499	0.056	$<$ 2e-16
		pol. broad.	-1.573	0.234	2.0e-11
	2 media pluralism	*intercept*	-3.202	0.393	$<$ 4e-16
		education	-0.054	0.111	0.626
		pol. orient.	0.288	0.038	4.0e-14
		pol. broad.	-0.603	0.193	0.002
	3 vote	*intercept*	-2.559	0.541	2.3e-06
		education	0.134	0.165	0.417
		pol. orient.	-0.045	0.061	0.461
		pol. broad.	-0.842	0.285	0.003
	4 ineligibility	*intercept*	-5.405	1.079	5.6e-07
		education	0.271	0.333	0.416
		pol. orient.	0.125	0.123	0.309
		pol. broad.	-0.696	0.553	0.208
specific uncertainty		α_1	0.361	0.134	0.007
		α_2	1.086	0.160	1.1e-11
		α_3	0.786	0.135	7.3e-09
		α_4	2.511	0.260	$<$ 2e-16
association		$\omega_{1,2}$	27.359	6.389	1.8e-05
		$\omega_{1,3}$	8.699	6.0043	0.147
		$\omega_{1,4}$	30.09	4.982	1.5e-09
		$\omega_{2,3}$	8.733	11.508	0.447
		$\omega_{2,4}$	6.562	10.423	0.528
		$\omega_{3,4}$	15.93	8.259	0.053
global uncertainty		γ	1.844	0.161	$<$ 2e-16

Moreover, the weight $1 - \hat{\phi} = \frac{1}{1+\exp(\hat{\gamma})} = 0.14$ provides the measure of global uncertainty in the model for the data at hand. Hence, respondents seem sensible to questions which cover serious socio-political dimensions, so that only a minor part of them remains aloof from such problems and tends to give answers at random.
Table 1 provides also the estimates of the association parameters. Only $\hat{\omega}_{1,2}$ and $\hat{\omega}_{1,4}$ turn out to be significantly different from zero and with positive values. Thus, there exists positive association among the responses given about the issues on *immigration* and *media pluralism* and the questions about *immigration* and *ineligibility*. Respondents who express high agreement with regard to the equality for all people living in Italy no matter what nationality they have, show also similar propensity to demand guarantees for the pluralism of media and for a government free from corruption.

Acknowledgments: This research was partially funded by STAR project (University of Naples Federico II-CUP: E68C13000020003) and by project PRIN-SURWEY (grant 2012F42NS8, Italy).

References

D'Elia, A. and Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**, 917 – 934.

Lee, M.L.T. (1996). Properties and applications of the Sarmanov family of bivariate distributions. *Communications in Statistics - Theory and Methods*, **25**, 1207 – 1222.

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*, 231 – 258. Chichester: J. Wiley & Sons.

Sarmanov, O.V. (1966). Generalized normal correlation and two dimensional Frechet classes. *Doklady (Soviet Mathematics)*, **168**, 596 – 599.

Tutz, G., Schneider, M., Iannario, M., and Piccolo, D. (2014). Mixture models for ordinal responses to account for uncertainty of choice. *Technical Report Number*, **175**, Department of Statistics, University of Munich.

Understanding economic growth regressions using latent class analysis

Jesus Crespo Cuaresma[1], Bettina Grün[2], Paul Hofmarcher[1], Stefan Humer[1], Mathias Moser[1]

[1] Department of Economics, Vienna University of Economics and Business (WU), Austria
[2] Department of Applied Statistics, Johannes Kepler University Linz (JKU), Austria

E-mail for correspondence: paul.hofmarcher@wu.ac.at

Abstract: We propose the use of Latent Class Analysis methods to analyze the covariate inclusion patterns across specifications resulting from Bayesian Model Averaging exercises. Using Dirichlet process clustering, we are able to identify and describe dependency structures among variables in terms of inclusion in the specifications that compose the model space. We apply the method to a dataset of potential determinants of economic growth.

Keywords: Bayesian model averaging; Latent class analysis; Dirichlet processes.

1 Motivation

Bayesian Model Averaging (BMA, Hoeting et al., 1999) has become a popular tool for economic growth applications in economics. Economic growth applications of BMA tend to quantify the relative importance of a given covariate by calculating its so-called posterior inclusion probability (PIP). Such a statistic has become a standard tool in econometric applications of BMA and is routinely used to measure the relative importance of different potential explanatories. While standard PIPs are intuitive measures that provide valuable insights into the overall importance of individual covariates as economic growth determinants, they face a number of shortcomings. The PIP neglects the heterogeneity across typical model specifications and accordingly does not inform about whether the degree of *importance* of the variable is evenly spread across potential specifications or, on the contrary, it is particular to specific combinations of explanatory variables. Previous

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

work assessing joint covariate inclusion in BMA applications has focused on capturing relevant dependency structures using bivariate *jointness* measures, that is, concentrating on the analysis of the joint posterior distribution of the inclusion of pairs of variables over the model space. Here we propose an alternative approach aimed at succinctly and comprehensibly describing the dependency structure across variables in the model space using Latent Class Analysis (LCA, Vermung and Magidson, 2002). In particular, we apply Dirichlet Process Clustering (DPC) in order to carry out inference about the latent classes governing covariance inclusion structures in economic growth regressions.

By applying DPC to the covariate inclusion structure of the models identified by BMA, we are able to capture the dependency patterns across included covariates through a (unobserved) latent variable which induces classes with independent covariate inclusion. Such a setting implies that PIPs within clusters constitute sufficient information to describe the importance of the variables and the differences of PIPs between clusters are representative of the dependencies in the inclusion of a covariate with respect to (all) other variables.

2 Bayesian model averaging and latent class analysis

The standard setting for BMA analysis in the framework of cross-country growth regressions assumes that the growth rate of income per capita (y) can be linearly related to a group of covariates (X_j) chosen from a set of potential growth determinants (X). Assuming that n observations are available, a typical linear regression model (M_j) is given by

$$y|\alpha, \beta_j, \sigma \sim N(\alpha\iota + X_j\beta_j, \sigma^2 I), \tag{1}$$

where ι is a column vector of ones of dimension n. Assuming that a total of K variables are available, inference on a quantity of interest (Δ) is given by

$$p(\Delta|y) = \sum_{j=1}^{2^K} p(\Delta|y, M_j)p(M_j|y), \tag{2}$$

where $p(M_j|y)$ is the posterior model probability. A model profile γ_i, for $i = 1, \ldots, 2^K$, is a K-dimensional vector of ones and zeros indicating the variables which are included in the sampled model i, with elements $\gamma_{ik} = 1$ if variable k is part of model i and $\gamma_{ik} = 0$ otherwise.

We assume that γ_i, the K-dimensional vector summarizing the variable inclusion profile for model i, has elements that arise from a mixture of infinitely, but countably many distributions,

$$p(\gamma_i) = \sum_{c=1}^{\infty} p(g_i = c) \prod_{k=1}^{K} p(\gamma_{ik}|g_i = c), \tag{3}$$

where $p(g_i = c)$ denotes the probability that model i is assigned to cluster c and $p(\gamma_{ik}|g_i = c)$ governs the inclusion probability of the k-th covariate in cluster c. In turn, for our application we use

$$p(\gamma_{ik}|g_i = c) \sim \text{Bern}(\pi_{ck}), \qquad \pi_{ck} \sim \text{Beta}(\delta, \delta),$$
$$p(g_i = c) = V_c \prod_{j=1}^{c-1}(1 - V_j), \qquad V_c \sim \text{Beta}(1, \alpha).$$

Hyperparameters are selected to induce a mixture consisting of only few clusters. The post-processing procedure proposed by Molitor et al. (2010) is employed to identify a suitable partition.

3 Results for economic growth regressions

Ley and Steel (2009) analyzed a dataset which consisted of 67 different determinants for 88 economies in addition to the GDP per capita growth figures and found very weak (bivariate and/or trivariate) jointness in the group of covariates. The DPC procedure in combination with the post-processing procedure was applied to the BMA results fitted to the same dataset using the 500 most frequently sampled models weighted with their posterior model probability covering 40% of the posterior model probability. The procedure resulted in three clusters which covered 21%, 17% and 3% of the posterior model probability.

Figure 1 depicts the PIPs of the variables in all models as well as those derived from the models in the single clusters. The shading indicates the interesting measure *IM* which corresponds to the square root of the mean squared deviation of PIPs across clustering. The results show a large degree of variability in PIPs across clusters for many of the covariates. Remarkable differences in PIPs across the two large clusters can be observed for the `MALFAL66` variable, which presents a much higher PIP in the second cluster, making it the second most important variable for models within that cluster. Such a phenomenon is accompanied by a sizable decrease in PIP for `P60, IPRICE1, TROPICAR, GDPCH60L`, and `DENS65C`.

4 Conclusions

In this contribution we are concerned with covariate inclusion patterns of BMA exercises with large model spaces. Recent research on such *jointness* structures tends to choose a low-dimensional approach to such an analysis and thus concentrates on bivariate or trivariate approaches, by calculating the inclusion relationships of few explaining factors at a time. We propose a novel approach by utilizing LCA techniques and apply DPC to a dataset well known in the BMA growth literature. The clustering method

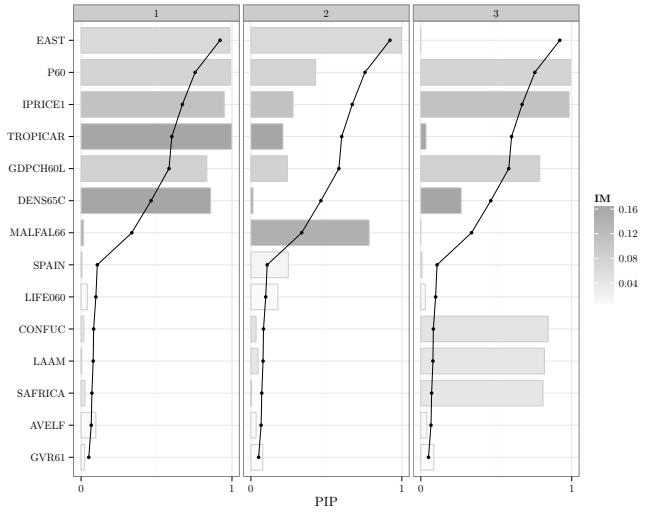

FIGURE 1. PIPs in unclustered BMA (solid line) and by identified cluster (bars).

put forward in our contribution aims at unveiling commonalities in the joint inclusion of variables and thus offering the applied econometrician evidence about the competing structures (as formed by *groups* of variables that appear together) that are covered by the posterior over the model space.

Acknowledgments: This research was supported by the OeNB Jubiläumsfonds: 14663.

References

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, **14**, 382 – 401.

Ley, E. and Steel, M.F. (2009). Comments on 'Jointness of growth determinants'. *Journal of Applied Econometrics*, **24**, 248 – 251.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010). Bayesian profile regression with an application to the national survey of children's health. *Biostatistics*, **11**, 484 – 498.

Vermunt, J.K. and Magidson, J. (2002). Latent class cluster analysis. In: *Applied Latent Class Analysis* Cambridge University Press, Cambridge, 89 – 106.

Application of statistical emulation to an agent-based model: assortative mating and the reversal of gender inequality in education in Belgium

Wim De Mulder[1], André Grow[2], Geert Molenberghs[1], Geert Verbeke[1]

[1] Leuven Biostatistics and Statistical Bioinformatics Centre, Leuven, Belgium
[2] Centre for Sociological Research, Leuven, Belgium

E-mail for correspondence: wim.demulder@cs.kuleuven.be

Abstract: We describe the application of statistical emulation to the outcomes of an agent-based model. The agent-based model simulates the mechanisms that might have linked the reversal of gender inequality in higher education with observed changes in educational assortative mating in Belgium. Using the statistical emulator as a computationally fast approximation to the expensive agent-based model, it is feasible to use a genetic algorithm in finding the parameter values for which the corresponding agent-based model outcome is closest to known empirical output. These optimal parameter values are then interpreted sociologically.

Keywords: Statistical emulation; Gaussian process; Agent-based model; Genetic algorithm.

1 Using agent-based models to simulate assortative mating

1.1 Short introduction to agent-based models

An agent-based model (ABM) is a computational model that simulates the behavior and interactions of autonomous agents. A key feature is that population level phenomena are studied by explicitly modelling the interactions of the individuals in these populations. The systems that emerge from such interactions often are complex and can show regularities that are difficult to anticipate without simulation. For example, with an ABM

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

it has been shown that cultural diversity can persist in a population, even if individuals have a tendency to adapt the cultural traits of those with whom they interact.

1.2 Application of ABM to assortative mating

Educational assortative mating is the sorting of individuals into relations based on educational attainment. Grow et al. (2014) describes an ABM that simulates the mechanisms that might have linked the reversal of gender inequality in higher education with changes in patterns of educational assortative mating. Reversal of such gender inequality refers to the fact that while men have always received more education than women in the past, this imbalance has turned around in large parts of the world. In many countries, women now outperform men in participation and success in higher education. Empirical research shows that this reversal has affected patterns of assortative mating. Earlier, men tended to be similarly or more highly educated than their partners. Today, couples still tend to be similarly educated, but if there is a difference, women tend to be more highly educated. The model studies this association by considering several parameters that describe properties of partner search. For illustration, consider parameters w_s^f and w_s^m that determine the importance that female and male agents attach to the education of prospective partners. The higher their values, the more agents prefer partners with similar educational attainment, and the more willing they become to marry similarly educated agents. We expect that these parameters strongly affect the patterns of educational assortative mating in the agent populations. The simulation outcome considered is the fraction of *hypogamic couples*: couples in which the woman has a higher educational attainment than her partner. We restrict attention to Belgium. The ABM is based on stochastic processes and thus produces a random outcome. We will refer to *the* outcome of the ABM as shorthand for the average outcome over 50 runs. The outcome of the ABM, given w_s^f and w_s^m, is denoted as $f_A(w_s^f, w_s^m)$. The input domain is $[0, 2] \times [0, 2]$.

1.3 Empirical data

To generate realistic agent populations, the ABM is initialized with empirical data from several sources (Grow et al. 2014). The empirical value of the fraction of female hypogamic couples (Section 1.2), denoted as h, is $h = 0.14$ and has been derived from data from the European Social Survey.

1.4 Description of the research question

We consider reverse engineering typical of ABMs: for what parameter values does the ABM produce output that is closest to given empirical output? Applied to our case, this question translates into: for which values of w_s^f

and w_s^m is $|f_A(w_s^f, w_s^m) - h|$ smallest? Obtaining such optimal values provides insight into the processes that might have driven observed population changes and allows validation of the model (e.g., are these parameter values feasible in sociological terms?). See further, Section 3.3.

2 Methodology

2.1 Statistical emulation

Consider an input vector $\mathbf{x}$ and a computer model that maps this input to an output y via some deterministic but possibly unknown function ν, i.e. $y = \nu(\mathbf{x})$. We assume that the considered computer model is highly computationally expensive, so that evaluation of ν in a very large number of different inputs is not feasible. Due to the many nonlinear interactions that are often involved, ABMs are a typical member of the class of highly computationally expensive computer models. Statistical emulation provides an approximation to such a computer model, given training data $(\mathbf{x}_1, \nu(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, \nu(\mathbf{x}_n))$. After training, the resulting so-called emulator is much faster to evaluate than the original computer model. An interesting feature of statistical emulation is that it models the uncertainty in the non-training data points, a feature that is absent in many other approximation techniques such as polynomial approximation. Before training data is presented to the emulator, the output of $\nu(\mathbf{x})$, which is scalar for our ABM case study, is modeled as a Gaussian process with mean $m(\mathbf{x}) = \sum_{i=1}^{q} \beta_i \, h_i(\mathbf{x})$, where β_i represent unknown coefficients, and h_i represent known regression functions. We follow standard practice in emulation and choose them linear. The covariance between the outputs corresponding to inputs $\mathbf{x}$ and $\mathbf{x}'$ is modeled as $\mathrm{Cov}\Big(\nu(\mathbf{x}), \nu(\mathbf{x}')|\sigma^2\Big) = \sigma^2 \, c(\mathbf{x}, \mathbf{x}')$, where σ^2 denotes a variance component and where $c(\mathbf{x}, \mathbf{x}')$ denotes a function that models the correlation between $\nu(\mathbf{x})$ and $\nu(\mathbf{x}')$. We take (in line with standard practice): $c(\mathbf{x}, \mathbf{x}') = \exp\Big[-\sum_i \Big((x_i - x_i')/\delta_i\Big)^2\Big]$, with x_i the ith component of $\mathbf{x}$ and where the δ_i represent parameters, which can be determined by maximizing their posterior distribution $\pi(\delta)$. Notice that the variance component σ^2 in the basic emulation framework is constant.

By applying Bayesian techniques, the prior Gaussian process and the training data can be combined to determine the posterior mean $m^\star(\mathbf{x})$ for any input $\mathbf{x}$, and the posterior standard deviation $\hat{\sigma}\sqrt{c^\star(\mathbf{x}, \mathbf{x})}$, where $\hat{\sigma}$ denotes the estimation of the parameter σ above and where $c^\star(\mathbf{x}, \mathbf{x}')$ denotes the posterior correlation between $\nu(\mathbf{x})$ and $\nu(\mathbf{x}')$. The posterior standard deviation allows to determine a confidence interval around the posterior mean. In this paper we will consider 95% confidence intervals around the posterior mean $m^\star(\mathbf{x})$ which we will denote as $CI(\mathbf{x})$. The application of statistical emulation to ABM is a relatively new research topic, largely neglected in sociological and demographic research. To our knowledge, Bijak

et al. (2013) were the first to illustrate the use of this method in the area of demographic ABMs.

2.2 Genetic algorithms

Genetic algorithms belong to the domain of artificial intelligence and can be used for a.o., function optimization. The optimization is performed via a heuristic search, where a population of candidate solutions is evolved toward better solutions, akin to the biological process of evolution. A main characteristic of genetic algorithms is that they only use function evaluations. This makes them very well suited to solve the optimization problem described in Section 1.4, because an ABM does not provide any other useful information for function optimization than the outcome in given inputs (in particular, information about derivatives is absent). However, since a genetic algorithm considers a large set of candidate solutions, it is typically computationally expensive. This is where the use of the computationally cheap emulator comes in: we will find an approximate solution by replacing the optimization problem by $\arg \min_{w_s^f, w_s^m} |m^\star(w_s^f, w_s^m) - h|$.

To solve this problem, we implemented the genetic algorithm described in Carr (2014, Sec. 2.3). To take into account that $\nu(w_s^f, w_s^m)$ follows a distribution, we use a non-deterministic fitness function FI, namely $FI(w_s^f, w_s^m) = 1/k \sum_{i=1}^{k} |h - z_i|$, where $z_1, \ldots, z_k$ are k values generated from the posterior distribution for $\nu(w_s^f, w_s^m)$, with k a parameter for which the value is chosen by the user. The fitness function measures how good a certain solution is: the lower the fitness value, the better the solution.

3 Results

3.1 Construction of the emulator

The emulator is trained with 100 training data points, for which the inputs are selected according to the Latin hypercube method. We perform a validation of the trained emulator by randomly selecting 5 vectors from the input domain to which the ABM is applied. For these vectors, we also calculate the posterior mean and a 95% confidence interval using the emulator. Finally, the difference, expressed in percentage terms, is calculated between the ABM and the emulator, which we define as $(m^\star(w_s^f, w_s^m) - f_A(w_s^f, w_s^m))/f_A(w_s^f, w_s^m) \times 100\%$. The results are shown in Table 1. The output between the emulator and the ABM is less than 5% in all cases, which we find acceptable. However, four of the five outputs generated by the ABM lie outside the corresponding confidence interval, an indication that the choice of a constant prior variance component might be too restrictive. Some recent research in emulation discusses how to relax this assumption, see e.g. Ba (2012). We plan to consider this and similar research, and to relax the constancy of the variance component, in our future work.

TABLE 1. Validation of the emulator.

w_s^f	w_s^m	$f_A(w_s^f, w_s^m)$	$m^\star(w_s^f, w_s^m)$	% difference	$CI(w_s^f, w_s^m)$
1.398	0.032	0.174	0.169	-2.87%	[0.165,0.172]
1.786	0.68	0.124	0.130	4.84%	[0.126,0.133]
0.716	0.818	0.171	0.170	-0.58%	[0.168,0.172]
0.116	1.358	0.186	0.181	-2.69%	[0.178,0.185]
1.034	1.812	0.127	0.122	-3.94%	[0.119,0.125]

3.2 Application of genetic algorithm

The parameter k, defined in Section 2.2, is chosen as $k = 10$. Other parameters are described in Carr (2014) and are chosen as: mutation rate = 0.2, population size = 500, number of iterations = 100. The initial population is chosen randomly from the input domain, and we run the algorithm 5 times with different initial populations. The results are shown in Table 2. It is clear that there is no global optimum and that all solutions have a very similar and very low fitness value. A highly remarkable observation is that the solutions lie almost exactly on a straight line, the correlation between them being -0.99952.

TABLE 2. Solutions found by the genetic algorithm.

iteration	w_s^f	w_s^m	$FI(w_s^f, w_s^m)$
1	0.873	1.391	0.000216
2	1.176	1.014	0.000181
3	1.026	1.199	0.000206
4	1.394	0.769	0.000225
5	1.306	0.857	0.000200

3.3 Interpretation of the results

The results of the previous section suggest that there is no single optimal value for w_s^f and w_s^m, but that there is an optimal *relationship* between them that is *linear*. Substantively this means that the preferences that the two parameters describe might be partial substitutes in generating the observed level of hypogamy. Thus even if women attach little importance to education (low value of w_s^f) and are thus willing to marry lower educated men, they are unlikely to find lower educated partners who are also willing to marry them if men consider education important (high value of w_s^m).

4 Conclusion

We described the use of statistical emulation and genetic algorithms to find the values of the parameters of an agent-based model that simulates assortative mating. The considered parameters describe the importance that female and male individuals attach to the education of prospective partners. The results indicate that there does not exist a single value for these parameters for which the agent-based model produces output that corresponds to the observed fraction of hypogamic couples. However, there is a single straight line that contains values for the parameters corresponding to the closest match between model output and empirical output.

Acknowledgments: The second author has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. 312290 for the GENDERBALL project.

References

Ba, S. and Joseph, V.R. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, **6**, 1838 – 1860.

Bijak, J., Hilton, J., Silverman, E., and Cao, V.D. (2013). Reforging the wedding ring: exploring a semi-artificial model of population for the United Kingdom with Gaussian process emulators. *Demographic Research*, **29**, 729 – 766.

Carr, J. (2014). An introduction to genetic algorithms. See: karczmarczuk.users.greyc.fr/TEACH/IAD/GenDoc/carrGenet.pdf

Grow, A., Van Bavel, J., and De Hauw, Y. (2014). An agent-based computational model of assortative mating and the reversal of gender inequality in education in Europe. In: *European Population Conference 2014*, Budapest, Hungary.

Signal extraction and power spectral density estimation: A Bayesian semi-parametric approach

Matthew C. Edwards[1], Renate Meyer[1], Nelson Christensen[2]

[1] University of Auckland, Auckland, New Zealand
[2] Carleton College, Minnesota, USA

E-mail for correspondence: matt.edwards@auckland.ac.nz

Abstract: An important goal in many signal processing applications involves extracting a signal buried in noisy data. Often there is much uncertainty around the nature of background noise, and parametric modelling of its power spectral density could result in misleading inferences about the embedded signal. We implement a Bayesian semi-parametric Metropolis-within-Gibbs Markov chain Monte Carlo algorithm to simultaneously estimate the parameters of a signal and characterize the noise power spectral density.

Keywords: Bayesian; Semi-parametric; Power spectral density; Signal extraction.

1 Introduction

Signals in real world applications are always noisy, and the presence of noise can significantly deteriorate the quality of information available in observed data. Common approaches to signal extraction involve constructing filters based on Wiener-Kolmogorov theory, and generally require accurate knowledge about the signal, and the nature of the noise. It is well-known that parametric noise models can result in unfavourable inferences when their underlying assumptions are misspecified. This is particularly noticeable if the noise of a system is poorly understood, or if real observations deviate from the model (see for example, Röver et al., 2011, for applications to gravitational wave data).
Non-parametric methods in time series analysis have been gaining popularity in recent years (see for example Choudhuri et al., 2004, and references

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

therein). Relating to the signal extraction problem, Macaro (2010) implemented a Bayesian non-parametric method with informative priors on the power spectral density (PSD) to identify multiple components in a time series. The method we present in this paper is similar, but is semi-parametric in nature, and we assume an accurate parametric signal model for identification rather than using informative priors. We implement the Bayesian non-parametric PSD estimation method of Choudhuri et al. (2004) to characterize the noise. We include an additional Gibbs component to the Markov chain Monte Carlo (MCMC) sampler to simultaneously estimate the parameters of a signal embedded in this noise.

For this analysis, we assume a length n time-domain data vector $\mathbf{y}$ that can be decomposed into independent and stationary signal and noise components:

$$\mathbf{y} = \mathbf{s}(\boldsymbol{\beta}) + \boldsymbol{\epsilon}(\boldsymbol{\theta}), \tag{1}$$

where $\mathbf{s}$ is a signal parameterized by $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ is noise parameterized by $\boldsymbol{\theta}$. We work in the frequency-domain, using the fast Fourier transform to map these quantities from the time-domain.

2 Methodology

2.1 Non-parametric power spectral density estimation

We briefly introduce the PSD estimation technique of Choudhuri et al. (2004), which uses the Bernstein polynomial prior of Petrone (1999a–b). The Bernstein polynomial prior is a non-parametric prior for a probability density on $[0, 1]$, and is essentially a finite mixture of Beta densities. Since the PSD $f(\lambda)$ is defined on frequencies $\lambda \in (\pi, \pi]$, and not on the unit interval, we reparameterize, such that $f(\pi\omega) = \tau q(\omega)$, for $\omega \in [0, 1]$. Here $\tau = \int_0^1 f(\pi\omega)\mathrm{d}\omega$ is the normalization constant.

To specify a prior on PSD $f(\pi\omega)$, we put a Bernstein polynomial prior on $q(\omega)$. Let $q(\omega) = \sum_{j=1}^{k} G\left(\frac{j-1}{k}, \frac{j}{k}\right] \beta(\omega|j, k-j+1)$, where $G \sim \mathrm{DP}(M, G_0)$ with center measure G_0 and precision parameter $M > 0$, $\beta(\omega|a, b)$ is a Beta density with parameters a and b, and k is a smoothness parameter with discrete probability mass function $p(k) \propto \exp(-\theta_k k^2)$, for $k \in \mathbb{Z}^+$. Let $p(\tau) = \mathrm{IG}(\alpha_\tau, \beta_\tau)$, and assume G, k, and τ are *a priori* independent.

Using the stick-breaking construction of the Dirichlet process by Sethuraman (1994), along with a truncation parameter L, we reparametrize G to $(Z_0, Z_1, \ldots, Z_L, V_1, \ldots, V_L)$ such that

$$G = \left(\sum_{l=1}^{L} p_l \delta_{Z_l}\right) + \left(1 - \sum_{l=1}^{L} p_l\right) \delta_{Z_0}, \tag{2}$$

where $p_1 = V_1$, $p_l = \left(\prod_{j=1}^{l-1}(1 - V_j)\right) V_l$ for $l \geq 2$, $V_l \sim \mathrm{Beta}(1, M)$ for $l = 1, \ldots, L$, and $Z_l \sim G_0$ for $l = 0, 1, \ldots, L$. Note that δ_a is a probability

density, degenerate at a. This yields the prior mixture:

$$f(\pi\omega) = \tau \sum_{j=1}^{k} w_{j,k}\beta(\omega|j, k-j+1), \tag{3}$$

with weights $w_{j,k} = \sum_{l=0}^{L} p_l I\{\frac{j-1}{k} < Z_l \leq \frac{j}{k}\}$ and $p_0 = 1 - \sum_{l=1}^{L} p_l$.
The joint prior is:

$$p(\boldsymbol{\theta}) \propto \left(\prod_{l=1}^{L} M(1-v_l)^{M-1}\right)\left(\prod_{l=0}^{L} g_0(z_l)\right) p(k)p(\tau), \tag{4}$$

where $\boldsymbol{\theta} = (\mathbf{v}, \mathbf{z}, k, \tau)$. This is updated by the Whittle (1957) approximation to the Gaussian likelihood to produce the joint posterior, and is implemented as a Metropolis-within-Gibbs MCMC sampler.

2.2 Parametric signal extraction

To simultaneously extract a signal embedded in noise whilst also characterizing the noise PSD, we include an additional Gibbs component to the MCMC sampler. That is, omitting the conditioning on the data for clarity, we sequentially sample the full set of conditional posterior densities $p(\boldsymbol{\theta}|\boldsymbol{\beta})$ and $p(\boldsymbol{\beta}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\mathbf{v}, \mathbf{z}, k, \tau)$ are the noise parameters, and $\boldsymbol{\beta}$ are the signal parameters.
One could estimate any parametric signal embedded in noise, but we will focus our attention on a linear model for simplicity. Given a *known* PSD (i.e., fixing the most recent MCMC sample of the PSD parameters), and assuming flat priors on the regression coefficients $\boldsymbol{\beta}$, the conditional posterior of $\boldsymbol{\beta}$ is $p(\boldsymbol{\beta}|\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\tilde{\mathbf{X}}'\mathbf{D}^{-1}\tilde{\mathbf{X}})^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\tilde{\mathbf{X}}'\mathbf{D}^{-1}\tilde{\mathbf{y}}$. Here $\tilde{\mathbf{y}}$ is the length n frequency-domain data vector, $\tilde{\mathbf{X}}$ is the $n \times d$ matrix of the d frequency-domain explanatory variables, and $\mathbf{D} = 2\pi \times \mathrm{diag}\,(f(\boldsymbol{\lambda}))$ is the noise covariance matrix.

3 Example

Let the signal of interest be the following damped-sine/sine-Gaussian mixture:

$$\begin{aligned} s(t|\beta_1, \beta_2) &= \beta_1 \left(\exp\left(\frac{-t}{4}\right)(\cos(2\pi t) + \sin(2\pi t))\right) \\ &\quad + \beta_2 \exp\left(\frac{-t^2}{2}\right)\sin(2\pi t), \end{aligned}$$

where $t = \Delta_t, 2\Delta_t, \ldots, n\Delta_t$, $n = 2^{12}$, and $\Delta_t = 1/2^{14}$ is the sampling rate. We embed this signal in AR(1) noise with first-order autocorrelation

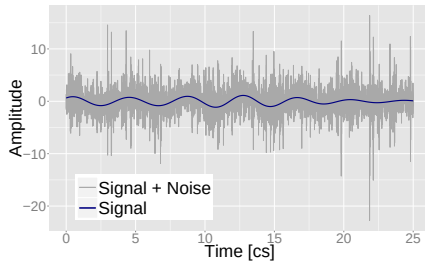

FIGURE 1. Damped-sine/sine-Gaussian mixture embedded in AR(1) noise with $\rho = -0.75$ and Student-t innovations with $\nu = 3$ degrees of freedom.

$\rho = -0.75$, and Student-t innovations with $\nu = 3$ degrees of freedom. The signal is scaled to have a signal-to-noise ratio (SNR) of $\varrho = 50$ (see Edwards et al., 2014, for this calculation). The data set-up can be seen in Figure 1. We use the non-informative prior specification of Choudhuri et al. (2004): $G_0 \sim \text{Uniform}[0, 1], M = 1, \alpha_\tau = \beta_\tau = 0.001, \theta_k = 0.01, L = 20$. The simulations are run for $10,000$ terations, with a burn-in period of $5,000$.

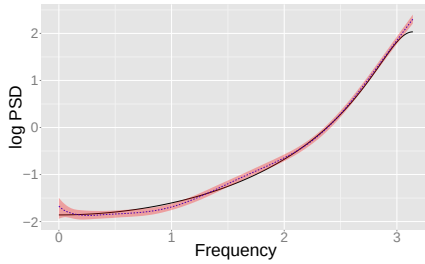

FIGURE 2. Estimated log PSD. 90% credible region (pink) and posterior median (dashed blue) overlaid with true PSD (black).

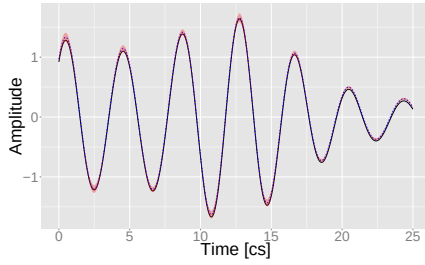

FIGURE 3. Estimated signal. 90% credible region (pink) and posterior mean (dashed blue) overlaid with true signal (black).

Figures 2 and 3 illustrate the estimated log PSD and signal respectively.

The MCMC sampler does well to extract both components. The credible regions generally contain the true PSD and signal, and the point-wise posterior medians are very close to the true values.

4 Discussion and outlook

The approach presented in this paper could be applied to many different fields where one is interested in extracting a signal in noisy data. In fact, one of the key motivations for this research comes from the need for an improved noise model in gravitational wave (GW) data analysis (Röver et al., 2011). The standard GW noise model assumes the noise PSD is known *a priori* and estimated using clean off-source data. Since the PSD for GW detectors is time-varying, off-source characterization of the PSD is an inadequate approach. The noise is also assumed to be normally distributed in each frequency bin, but real data often experience transient non-Gaussian glitches. To mitigate parametric misspecifications, non-parametric methods would be preferable. In a similar fashion to Macaro (2010), one could imagine a fully nonparametric model, where the signal is also modelled non-parametrically. This could be useful for signals with poorly defined parameter spaces, such as those from GW bursts (Edwards et al., 2014).
A well-known property of Bernstein polynomial basis functions is their slow convergence to the function they aim to approximate (Powell, 1981; or Shen and Wu, 2014). This can be a significant computational problem, particularly for poorly behaved noise sources, with steep changes in power, or sharp spectral lines, where the smoothness parameter k is large. A future extension to the current model is to change the basis functions from Bernstein polynomials to normalized B-splines, where convergence is expected to be much faster (Shen and Wu; 2014).

Acknowledgments: Special thanks to the IWSM 2015 Scientific Committee and the Statistical Modelling Society for a generous travel grant.

References

Choudhuri, N., Ghosal, S., and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, **99**, 1050.

Edwards, M.C., Meyer, R., and Christensen, N. (2014). Bayesian parameter estimation of core collapse supernovae using gravitational wave simulations. *Inverse Problems*, **30**, 114008.

Macaro, C. (2010). Bayesian non-parametric signal extraction for Gaussian time series. *Journal of Econometrics*, **157**, 381 – 395.

Petrone, S. (1999a). Random Bernstein polynomials. *Scandinavian Journal of Statistics*, **26**, 373 – 293.

Petrone, S. (1999b). Bayesian density estimation using Bernstein polynomials. *The Canadian Journal of Statistics*, **27**, 105 – 126.

Powell, M.J.D. (1981). *Approximation Theory and Methods*, Cambridge University Press, Cambridge.

Rosen, O., Wood, S., and Stoffer, D.S. (2012). AdaptSPEC: Adaptive Spectral Estimation for Nonstationary Time Series. *Journal of the American Statistical Association*, **107**, 1575.

Röver, C., Meyer, R., and Christensen, N. (2011). Modelling coloured residual noise in gravitational-wave signal processing. *Classical and Quantum Gravity*, **28**, 015010.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639 – 650.

Shen, W. and Ghosal, S. (2014). Adaptive Bayesian procedures using random series priors. arXiv:1403.0625 [math.ST].

Whittle, P. (1957). Curve and periodogram smoothing. *Journal of the Royal Statistical Society, Series B*, **19**, 38 – 63.

Separation of signals with tailored penalties

Paul H.C. Eilers[1,2]

[1] Erasmus MC, Rotterdam, The Netherlands
[2] Carlos III University, Madrid, Spain

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

Abstract: Many signals can be descibed as the sum of components with different properties. It may be possible to characterize each of them by a "tailored" penalty, which gives a low value when applied. This can be exploited to separate the components. The theory and applications with real-world data are presented.

Keywords: Smoothing; L_0 penalty; Difference equations.

1 Introduction

In many applications we encounter signals with a complex structure. An example is a smooth (time) series with a number of jumps and additive noise. We would like to smooth these data, to get rid of most of the noise. But simple smoothing will round off the jumps, which often is undesirable. It can also be the case that we want to remove the jumps, because they are artifacts of the data collection process. Alternatively, we might be interested in the jumps, but the smooth trend was caused by a drifting instrument.
The idea I want to explore here is to model the data as a sum of two (or more) signals and apply well-chosen penalties to each of them. Each penalty is chosen in such a way that it gives a low value for its "own" signal component, but a high value for the other ones. In the case of our example the sum of squares of second differences of the smooth signal will be relatively small, but it will be large for a piece-wise constant signal with jumps. But for a jumpy signal a penalty on the sum of absolute values of first order differences will be small. So we set up an additive model with two components, each with their tailored penalty. We will see that this works. A variation on this theme is a periodic signal (of known frequency) on top of a trend. I will use it to introduce the 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

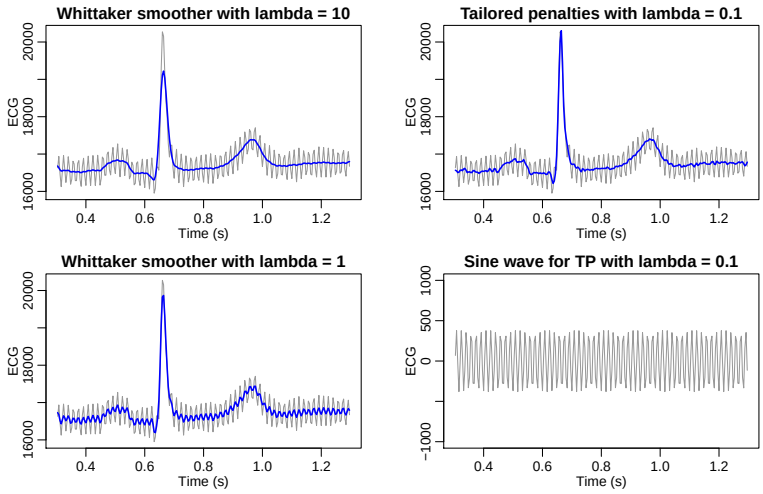

FIGURE 1. ECG signal with 60 Hz periodic interference. Left column: discrete smoothing. Right column: removing the interference with a tailored penalty. In all panels λ refers to penalty on the trend ($\lambda_1 = 10^5$). Signals in gray, fits in blue.

2 Theory

The theory is simple: the model is written as $y = \sum_j v_j + e$ and we minimize the objective function

$$S = \sum_i (y_i - \sum_j v_{ij})^2 + \sum_j P_j(v_j). \tag{1}$$

Here v_j indicates signal component j, $P_j(v_j)$ the penalty on v_j and e_i is the noise. Except for y, one index as in v_j indicates a vector, while a double index as in v_{ij}, indicates element i of vector v_j.

This is an additive model, but with a twist: some penalties do not emphasize smoothness, but other properties of the signal components. An example is shown in Figure 1, showing a very short segment of an electrocardiogram. We have a trend (v_1) plus a sine wave (v_2); the later is a 60 Hz interference. For a pure sine wave with period is p, we have that the difference equation $v_{i-1,2} - 2\phi v_{i,2} + v_{i+1,2} = 0$ holds if $\phi = \cos(2\pi/p)$. We can use it as a tailored penalty. For a smooth v_1 the familiar second order differences $\Delta^2 v_1$ have low values. These two difference equations define the penalties. Their influence is tuned by two parameters, λ_1 and λ_2, with values that are chosen by the user.

Write the penalties as $\lambda_1 ||D_1 v_1||^2$ and $\lambda_2 ||D_2 v_2||^2$, with proper matrices D_1 and D_2. If we define $B = [I : I]$, where I is the identity matrix, we have to solve

$$(B^{\mathrm{T}} B + P)\hat{v} = B^{\mathrm{T}} y, \tag{2}$$

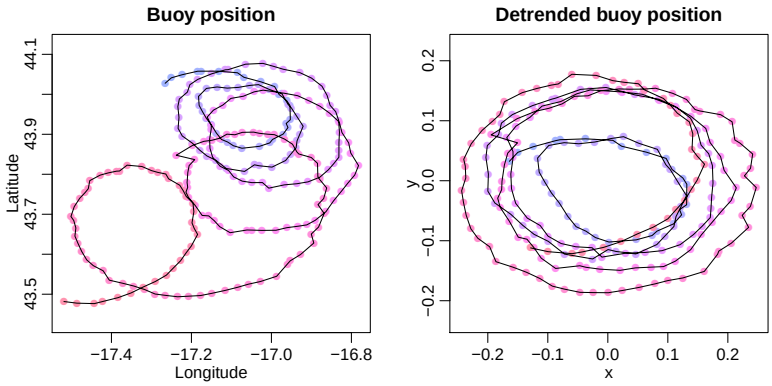

FIGURE 2. Movements of a floating subsurface buoy. Left: observed positions. Right: detrended positions.

where $v^{\mathrm{T}} = [v_1^{\mathrm{T}} : v_2^{\mathrm{T}}]$ and P is block-diagonal matrix, containing the penalty matrices. The system is very sparse.
Figure 1 shows that the classical Whittaker smoother (Eilers, 2003), which is the above model without v_2, has to use a rather strong penalty to get rid of the interference. The peak in the signal gets rounded, which is undesirable. Including a periodic second component in the model, with a tailored penalty, gives better results.
In the example the penalties are quadratic, so the solution is explicit and no iterations are needed. This is not the case when we use an L_1 penalty that involves sums of absolute values. But we can approximate it by repeated use of a weighted quadratic penalty. The same is true for an L_0 penalty. The key is the fact $|x|^p = x^2/|x|^{2-p}$. If have an approximation $\tilde{x}$, we can write $|x|^p \approx x^2/|\tilde{x}|^{2-p} = wx^2$. So we have a weighted square. In practice it is more stable to use $w = (|\tilde{x}|^{2-p} + \beta^2)$, with β a small number.

3 More applications

Figure 2 shows a small part of the movements of a drifting undersea buoy, used to map ocean currents (Lilly and Gascard, 2006). We see a spiraling pattern, on top of a large scale trend. When we plot longitude and latitude separately against time, we get Figure 3. It is of interest to determine the period and the amplitude of the cyclical variations around the trend. In the top panel of Figure 4 we see the reduced phase, computed as the angle of the vector $(-x, y)$. The values are between $-\pi$ and π radians. What we would like to have is a smooth phase vector without the jumps and with a much larger range.
We set up a model with v_1 the smooth phase, giving it the usual second

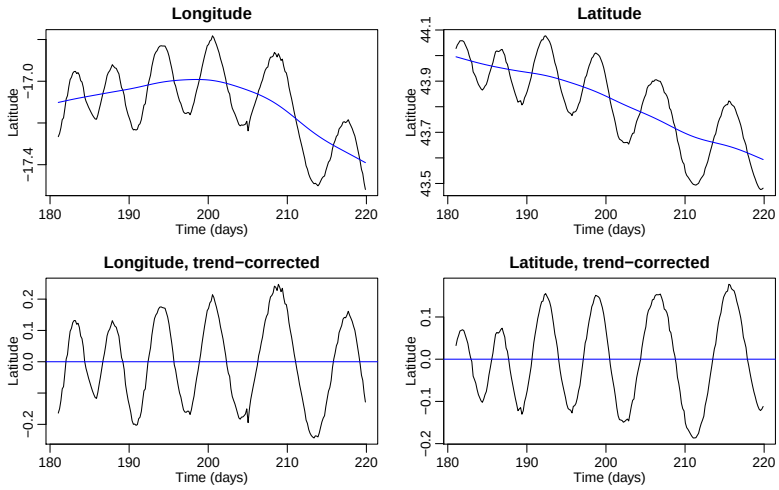

FIGURE 3. Buoy position components as time series, before and after trend removal.

order difference penalty. We give v_2 an (approximate) L_0 penalty on first order differences, to allow jumps (Rippe et al., 2012).

The middle panel of Figure 4 shows the estimated jumps and the estimate of the phase. More useful is the curve of $2\pi\Delta t/\Delta\hat{v}_1$, as it gives the instantaneous period of the signal. It is shown in the bottom panel.

These data were also analyzed by Eilers (2010), fitting amplitude and phase in the model

$$y_i = \alpha_i \cos(\phi_i) + e_i,$$

where α is the amplitude and ϕ is the phase, both assumed to be changing slowly. The model is non-linear in ϕ. In the mentioned paper an ad-hoc method was used, based on zero crossings, to get starting values for the phase. Tailored penalties seem to offer a more principled alternative.

Once the phase is obtained, we can apply a varying-coefficient model to estimate the amplitudes of its sine and cosine. This is more or less standard, so I skip the details.

Another application is the separation of two sine waves with slightly different frequencies, possibly on top of a trend. An example is shown in Figure 5. Quite large λs were used ($\lambda_1 = \lambda_2 = 10^6$). The frequencies differ by ten per cent. Numerical experiments showed that the method works well for much smaller differences, but the values amplitudes of the sine waves get less reliable.

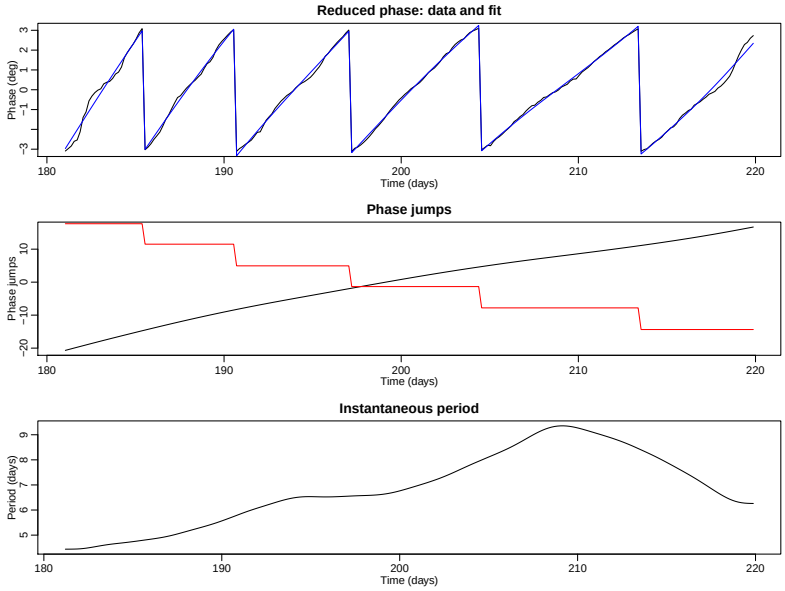

FIGURE 4. Buoy data. Top: reduced phase with jumps. Middle: estimated jumps and continuous phase. Bottom: local period as derived from the continuous phase.

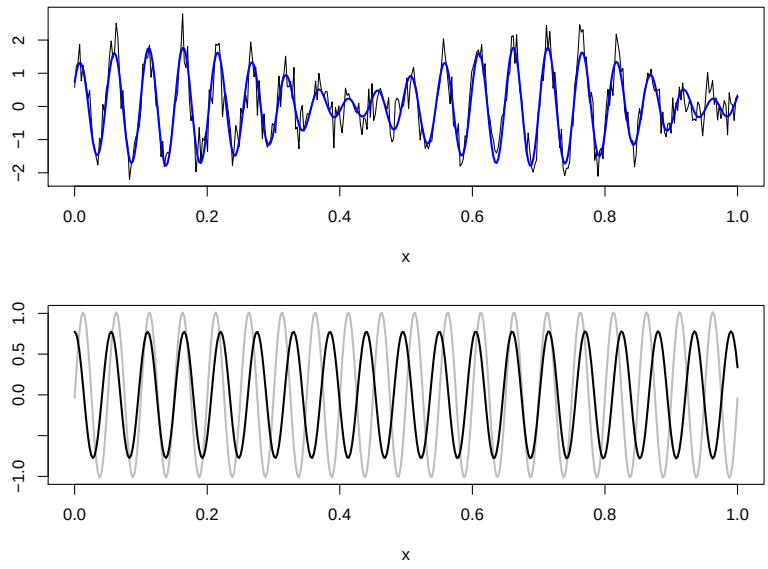

FIGURE 5. Separation of two sinusoidal signals with different frequencies. Top: noisy signal (black) and fit (blue). Bottom: the estimated signal components in grey and black.

4 Discussion

As the examples show, it can be fruitful to model data as a sum of signals that lead to small values of tailored penalties.
A discrete penalty in fact measures how close a difference equation follows fitted values. There is a growing literature that bases penalties on differential equations. Quadratic penalties, measuring sums of squares, are common. As the example of phase estimation showed, it can be advantageous to switch to other norms, depending on the context.
The subject deserves further study. In the case of a sine wave an estimate of the period is needed. The results are not extremely sensitive to this parameter, but it would be pleasant to have a (semi-)automatic procedure. The penalty parameters were set "by hand", using visual inspection of the results of the model. It would be nice if a more objective automatic data-driven procedure would be available. However, we should not be worried too much. For many engineering applications the carpenter's eye of the user will be sufficient. Automatic procedures look attractive but they may make unwarranted assumptions, like no serial correlation in the noise, that can lead to useless results if this is not true.
The present model for a sine wave uses a fixed parameter difference equations. In real life ist frequency may change gradually. Can we develop adaptive penalties to handle this?

Acknowledgments: This research was partly supported by a Chair of Excellence grant from Universidad Carlos III Madrid.

References

Eilers, P.H.C. (2003). A Perfect Smoother. *Analytical Chemistry*, **75**, 3299 – 3304.

Eilers, P.H.C. (2010). The smooth complex logarithm and quasi-periodic models. In: *Statistical Modelling and Regression Structures*, Kneib T. and Tutz (eds.). Springer.

Lilly, J.M. and Gascard, J.-C. (2006). Wavelet ridge diagnosis of time-varying elliptical signals with application to an oceanic eddy. *Nonlinear Processes in Geophysics*, **13**, 467 – 483.

Rippe, R.C.A., Meulman, J.J., and Eilers, P.H.C. (2012). Visualization of genomic changes by segmented smoothing using an L_0 penalty. *PLoS ONE*, **7**, e38230.

Modelling homogeneous and heterogenous populations under joint progressive type-II censoring

Lida Fallah[1], John Hinde[1]

[1] National University of Ireland Galway, Ireland

E-mail for correspondence: l.fallah1@nuigalway.ie

Abstract: We consider the analysis of time to event data from two populations undergoing life-testing under a joint progressive type-II censoring scheme for both homogeneous and heterogeneous situations. We consider maximum likelihood estimation for this complex sampling scenario and its behaviour under different censoring schemes. For heterogeneous populations we consider a mixture model formulation and maximum likelihood estimation using the EM algorithm.

Keywords: EM algorithm; Joint progressive type-II censoring; Lifetime data; Mixture models.

1 Introduction

Comparative life testing is commonly used in an industrial setting to study differences between items produced by different manufacturing lines and similar approaches may be used to compare the survival of subgroups of patients in the medical setting. Here, for simplicity, we limit ourselves to two groups and suppose that we have independent samples of sizes m and n that are put on test or followed up over time. The study is to be terminated as soon as some fixed number, r, of failures are observed, i.e. a type-II censoring scheme. However, to save costs this is combined with joint progressive censoring in which at each observed failure time a random number of units is withdrawn from study, giving what Rasouli and Balakrishnan (2010) refer to as joint progressive type-II (JPTII) censoring. In this paper, Section 2 describes the specification of the model under this censoring scheme and maximum likelihood estimation in the simple case of exponential populations. A simulation study is used to explore the

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

performance of different censoring schemes. In Section 3, we extend the model to heterogeneous groups, to account for individual frailty of the items. Here, for simplicity we use a two-level categorical variable to capture the heterogeneity, giving a two-component mixture for each group (see Kuo and Peng, 2000). This can be thought of as dividing the items into those that are susceptible and non-susceptible to failure. Section 4 is devoted to the the study of censoring in a single mixed population.
The main goal in ongoing work is to discuss maximum likelihood estimation using the EM algorithm, which can be used to handle both the mixture densities and the censoring. We will illustrate this with a small simulation study and an application based on the Boeing 720 air-conditioning failure data from Proschan (1963).

2 Model specification

Suppose that $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are i.i.d. random variables representing the lifetimes from two groups under study with probability density functions $f^{(1)}$ and $f^{(2)}$, respectively. Let $W_1 \leq \cdots \leq W_N$, $N = m + n$, denote the order statistics of the combined set $\{\mathbf{X}, \mathbf{Y}\}$. Under JPTII censoring, at the time of the first failure W_1 (that may be from either X or Y), R_1 units are randomly withdrawn from the remaining $N-1$ surviving units. This process is continued at each subsequent failure time with R_i removed at the ith observed failure. Here, the total number of observed failures r is pre-fixed and for the removals we can write $R_i = S_i + T_i$, $i = 1, \ldots, r$, where S_i and T_i are the number of units withdrawn from the X and Y samples respectively.
The observed data can be summarised as $(\mathbf{Z}, \mathbf{W}, \mathbf{S})$, where Z_i is a binary variable indicating which population the observation is from ($Z_i = 1$ if W_i is from $\mathbf{X}$). Figure 1 shows a schematic picture of this censoring scheme.

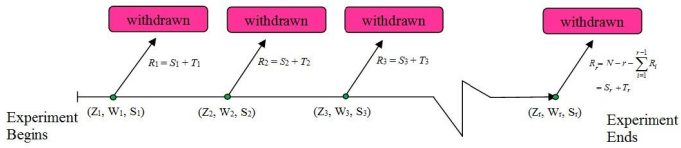

FIGURE 1. Schematic picture of JPTII censoring.

The likelihood function for component parameters Θ given the observed data is

$$L(\Theta|\mathbf{z}, \mathbf{w}, \mathbf{s}) = C \prod_{i=1}^{r} \left[\{f^{(1)}(w_i)\}^{z_i} \{f^{(2)}(w_i)\}^{1-z_i} \right] \{\overline{F}^{(1)}(w_i)\}^{s_i} \{\overline{F}^{(2)}(w_i)\}^{R_i - s_i},$$

where $\overline{F}^{(1)} = 1 - F^{(1)}$, $\overline{F}^{(2)} = 1 - F^{(2)}$ are the survivor functions and C is a normalising constant that depends only on the observed data.
For exponentially distributed lifetimes with $X \sim$ Exponential(θ_1) and $Y \sim$ Exponential(θ_2) the MLEs of θ_1 and θ_2 and the approximate SEs of parameters have been derived by Rasouli and Balakrishnan (2010) as follows:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{r} z_i w_i + \sum_{i=1}^{r} w_i s_i}{\sum_{i=1}^{r} z_i} \tag{1}$$

$$\hat{\theta}_2 = \frac{\sum_{i=1}^{r} (1 - z_i) w_i + \sum_{i=1}^{r} w_i (R_i - s_i)}{r - \sum_{i=1}^{r} z_i} \tag{2}$$

$$\mathrm{SE}(\hat{\theta}_1) = \frac{\sum_{i=1}^{r} z_i w_i + \sum_{i=1}^{r} w_i s_i}{\sqrt{\{\sum_{i=1}^{r} z_i\}^3}}$$

$$\mathrm{SE}(\hat{\theta}_2) = \frac{\sum_{i=1}^{r} (1 - z_i) w_i + \sum_{i=1}^{r} w_i (R_i - s_i)}{\sqrt{\{r - \sum_{i=1}^{r} z_i\}^3}}.$$

2.1 Simulation study

We focus on the above setting with two different exponential populations with parameters $\theta_1 = 60$ and $\theta_2 = 75$ and sample sizes of m and n, respectively, under a JPTII censoring scheme. We conduct a simulation study to evaluate the performance of estimators and compare the average test duration of different situations, taking different choices of $m = 24, 40$ and $n = 27, 100$ and censoring schemes of the form $\mathbf{R} = (a, a, \ldots, a, N - r - \sum_{i=1}^{r-1} r_i)$, with $a = 0$ (standard joint type-II censoring), and $a = 1, 2$.
This is done using a function developed in `R` that (i) simulates samples from the base populations; (ii) applies a JPTII censoring scheme to produce the observed data (Z, W, S); (iii) obtains parameter estimates and standard errors; (iv) evaluates performance in terms of bias, estimated mean-square error, test duration and relative efficiency compared to standard joint type-II censoring.
These results will be discussed to show the possible benefits of joint progressive censoring.

3 Extension to heterogeneous populations

We now assume that each of the populations is given by a two component mixture, representing both susceptible and non-susceptible individuals, i.e.

$$f^{(k)}(t) = \pi_k g_1^{(k)}(t|\theta_{k,1}) + (1 - \pi_k) g_2^{(k)}(t|\theta_{k,2}), \qquad t > 0, \quad k = 1, 2,$$

where $g_j^{(k)}$, $j = 1, 2$, are the individual population component densities with $\theta_{k,1}$ and $\theta_{k,2}$, respectively.

Maximization of the above likelihood is complicated by the presence of the mixture densities and the survivor function for the censored observations, but can proceed using the EM algorithm. We introduce binary latent indicator variables U_i and V_i, $i = 1, \ldots, r$, to identify the components of the observed failures and S_i^* $(0 \leq S_i^* \leq s_i)$, T_i^* $(0 \leq T_i^* \leq t_i)$ to identify the number of censored observations from each component. This allows the complete data likelihood to be written as a product of the component densities:

$$\begin{aligned} L(\Theta) =& C \prod_{i=1}^{r} \Big[\Big\{ \pi_1 g_1^{(1)}(w_i) \Big\}^{U_i z_i} \Big\{ (1-\pi_1) g_2^{(1)}(w_i) \Big\}^{(1-U_i) z_i} \\ & \times \Big\{ \pi_2 g_1^{(2)}(w_i) \Big\}^{V_i (1-z_i)} \Big\{ (1-\pi_2) g_2^{(2)}(w_i) \Big\}^{(1-V_i)(1-z_i)} \\ & \times \Big\{ \pi_1 \overline{G}_1^{(1)}(w_i) \Big\}^{S_i^*} \Big\{ (1-\pi_1) \overline{G}_2^{(1)}(w_i) \Big\}^{(s_i - S_i^*)} \\ & \times \Big\{ \pi_2 \overline{G}_1^{(2)}(w_i) \Big\}^{T_i^*} \Big\{ (1-\pi_2) \overline{G}_2^{(2)}(w_i) \Big\}^{(t_i - T_i^*)} \Big]. \end{aligned}$$

In some simple cases the component survivor functions have a closed form and the EM procedure is a modified simple mixture EM. More generally, the censored observations can be replaced by unobserved latent variables and an extended EM algorithm used, although the E-step for these censoring latent variables can be difficult to evaluate and so a Monte Carlo EM algorithm (MCEM) is required.

4 Study of censoring in a single mixed population

Estimation for the set-up described in Section 3 seems to have some problems, so in order to study this we look at the simpler case of a single mixed two component population. The density function is given by

$$f(t) = \pi g_1(t|\theta_1) + (1-\pi) g_2(t|\theta_2), \qquad t > 0,$$

where g_j, $j = 1, 2$, are the component densities. We assume exponentially distributed component densities and consider the performance of the different sampling schemes. Here again we could use the EM algorithm and the M-step estimates for the θs would be very similar to equations (1) and (2) only now the component labels z_i would be an unobserved latent variable that would be updated in the E-step. Direct optimisation of the three-parameter likelihood function is also possible and we use the `optim` function in `R` to do this.

We use parameters $\theta_1 = 3$, $\theta_2 = 30$ and $\pi = 0.7$ for the mixed exponential population and sample size of $m = 200$ for simulating progressively type-II censored data with a pre-fixed r number of failures, and look at the performance in three different situations.

No censoring: As a special case, if $r = m$ we observe all of the failures and have no censoring. `optim` gives perfect estimates and the profile log-likelihood plots for pairs of parameters show a well behaved log-likelihood with the expected two modes from the possible component label-switching located at the true values.

Type-II censoring: If we terminate the experiment immediately after observing a certain number of failures (say $r = 180$) and allow some censoring to happen at that point, we end up with rather poor estimates. Furthermore, as the mean-lifetimes of the mixture components are made more different, the situation gets worse. This arises from almost all of the observed failure times being allocated to one component and the (terminally) censored values to the other component and this results in computational problems as one the θs tends to ∞. To explore this issue, we consider fitting a cure-fraction model by defining the mixture density and survival functions respectively as

$$f(w_i) = \frac{\pi}{\theta_1} e^{-w_i/\theta_1} \qquad \text{and} \qquad \overline{F}(w_i) = \pi e^{-w_i/\theta_1} + (1 - \pi)$$

with likelihood function

$$L(\Theta|\mathbf{z}, \mathbf{w}, \mathbf{s}) = C \prod_{i=1}^{r} \frac{\pi}{\theta_1} e^{-w_i/\theta_1} \left[\pi e^{-w_i/\theta_1} + (1 - \pi)\right]^{(m-r)}.$$

This model is certainly computationally more stable, however the estimate for π is dominated by the fraction of observed data r/m.

Progressive type-II censoring: From preliminary results, censoring a sensible amount of units progressively after each failure can help improve the parameter estimates in these mixed populations. This is plausible as the censoring scheme is likely to remove some shorter lifetime units and so we get to observe more failure times from the component with the longer mean-lifetime. We again obtain a sensible profile log-likelihood plot, see Figure 2 for the case $r = 180$ with $(\hat{\theta}_1, \hat{\theta}_2, \hat{\pi}) = (2.55, 16.95, 0.73)$, with associated standard errors $(0.41, 3.76, 0.07)$.
From this initial work it is clear that there are problems in fitting mixtures of exponentials to type-II censored data, particularly in obtaining reliable parameter estimates, although there do seem to be advantages in progressive regimes. This will be further explored with different component densities, as also considered in Achcar and Periera (1999), and then applied to the joint setting.

5 Application

The joint progressive censoring approach will be illustrated using data on the failure times of Boeing 720 air-conditioning units from Proschan (1963).

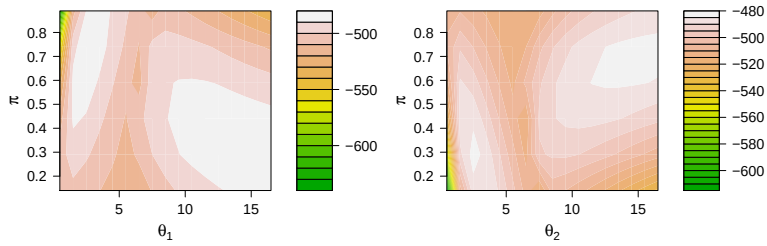

FIGURE 2. Profile log-likelihood plots of (π and θ_1 given θ_2) and (π and θ_2 given θ_1) on the left and right respectively.

References

Achcar, J.A., and Pereira, G.D.A. (1999). Mixture models for type-II censored survival data in the presence of covariates. *Journal of Computational Statistics*, **14**, 233 – 250.

Kuo, L. and Peng, F. (2000). A mixture-model approach to the analysis of survival data. *Generalized Linear Models: A Bayesian Perspective*, (D.K. Dey, S.K. Ghosh, and B.K. Mallick, eds.), 255 – 267. New York: Marcel Dekker.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, **5**, 375 – 383.

Rasouli, A. and Balakrishnan, N. (2010). Exact likelihood inference for two exponential populations under joint progressive type-II censoring. *Journal of Communications in Statistics- Theory and Methods*, **39**, 2172 – 2191.

Generalized INAR models with trend and seasonality for veterinary syndromic surveillance

Amanda Fernández-Fontelo[1], Pere Puig[1], Anna Alba[23]

[1] Dep. de matemàtiques, Universitat Autònoma de Barcelona, 08193, Bellaterra, Barcelona, Spain
[2] Centre de Recerca en Sanitat Animal (CReSA) - Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Campus UAB, 08193 Bellaterra, Barcelona, Spain
[3] Dep. of Veterinary Population Medicine. College of Veterinary Medicine. University of Minnesota, Saint Paul, Minnesota

E-mail for correspondence: amanda@mat.uab.cat

Abstract: Fallen stock data constitutes an important component for animal syndromic surveillance. To build a powerful syndromic surveillance system, the analysis of these data requires to be conducted at different geographical levels over the time. It implies the analysis of some time series with many zeros and low counts. We present here some generalized INAR models which are able to explore data coming from small areas taking into account several trend and seasonality patterns.

Keywords: Fallen stock; Hermite distribution; INAR models; Syndromic surveillance.

1 Introduction

Veterinary syndromic surveillance is based on the continuous collection and analysis of available data on animal health that predate to the diagnosis with the aim of identifying or foreseeing possible disease outbreaks.
Data related to fallen stock is one of these potential sources which can signal a wide range of animal health issues, Alba et al. (2015).
The large-scale analysis of fallen stock is able to detect events which happen in a wide region. Nevertheless, an important event can occur in a small region and be difficult to identify if data is geographically clustered.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Accordingly, it may be necessary the analysis of fallen stock at different geographical levels for the construction of a good surveillance system.
Fallen stock in farms can be captured with the number of carcasses collected during a specific period of time. The analysis of this variable in a small region leads to consider time series with a high number of zeros and low counts. Therefore, discrete or count time series techniques are required.
The goal of this work is to analyze the number of carcasses registered in farms located in several areas at small geographical levels in order to make predictions. Many models for count time series have been considered in the literature, but we focus on INAR models, which can be understood as an extension of the well known AR models, Jung and Tremayne (2006).
This is a solution of the question about small-scale analysis in fallen stock time series data and indeed, this is an important step forward for building a powerful syndromic surveillance system.

2 INAR models with trend and seasonality

A generalized INAR(k) process $\{X_t; t \in \mathbb{Z}\}$ is defined by means of the recurrent equation,

$$X_t = \rho_1(t) \circ X_{t-1} + \rho_2(t) \circ X_{t-2} + \cdots + \rho_k(t) \circ X_{t-k} + W_t, \qquad (1)$$

where $0 \leq \rho_1(t), \rho_2(t), \ldots, \rho_k(t) < 1$ and W_t are the innovations.
W_t can be distributed assuming different types of discrete probabilistic distributions (Poisson, Negative Binomial, Weighted Poisson, Hermite, ...). Nevertheless it can be proved that only in the case of $W_t \sim r$th-Hermite distribution, which includes the Poisson ($r = 1$) and the classical Hermite distribution ($r = 2$), the marginal distribution of X_t belongs to the same family, see Weiß and Puig (2015).
In this paper we focus our attention in the situation where $W_t \sim Hermite$ because of its simplicity. Hence, given parameters a_1 and a_2 of the Hermite distribution, the expression of its probability mass function is

$$P(Y = k) = e^{-(a_1+a_2)} \sum_{j=0}^{\left[\frac{k}{2}\right]} \frac{a_1^{k-2j} a_2^j}{(k-2j)!j!}, \qquad (2)$$

with $k = 0, 1, \ldots$, $a_1, a_2 \geq 0$ and $[\frac{k}{2}]$ is the integer part of $\frac{k}{2}$. The mean and variance can be expressed as $\mu = a_1 + 2a_2$ and $\sigma^2 = a_1 + 4a_2$ respectively, and the dispersion index as $d = \frac{a_1+4a_2}{a_1+2a_2}$. In the case of Poisson distribution $d = 1$. For further details see Kemp and Kemp (1965).
In order to model a possible trend and/or seasonality it is also assumed that $\rho_j(t) = f_j(\boldsymbol{\beta}_j, t)$ for $j = 1, \ldots, k$ and $a_j(t) = g(\boldsymbol{\gamma}_j, t)$ for $j = 1, 2$ can be functions of time, for suitable functions $f(\cdot)$ and $g(\cdot)$, where $\beta_j = \beta_{1j}, \ldots, \beta_{pj}$ and $\gamma_j = \gamma_{1j}, \ldots, \gamma_{pj}$ are also parameters to be estimated.

In our case, the INAR process could remind to a classical AR(k) process with trend and/or seasonality component, but now this is nonlinear due to the $\circ$-operation replacing the usual scalar product used in continuous models. The idea of this operation, called *binomial thinning* or just *thinning*, is to ensure the integer discreteness of the process. It is defined as,

$$\rho_j \circ X_{t-j} = \sum_{i=1}^{X_{t-j}} B_i\left(\rho_j\right), \tag{3}$$

such that $\{B_i(\rho_j)\}$ is a succession of i.i.d random Bernoulli variables with probability of success equal to ρ_j. Note that $\rho_j \circ X_{t-j}|X_{t-j} = x_{t-j}$ is just a binomial random variable, $Bin(x_{t-j}, \rho_j)$.
Parameter ρ_j can be interpreted as the proportion of counts which happen at time $t-j$ and go straight to time t. Otherwise, W_t can be considered as the innovations or new events produced at time t. In general, the interpretation of the parameters ρ_j, $j = 1, \cdots, k$, can be difficult and has to be done specifically for each example of application.

2.1 Parameter estimation and model selection

Parameters $\Theta = (\beta_1, \beta_2, \ldots, \beta_l, \gamma_1, \gamma_2 \ldots \gamma_m)$ in (1) can be estimated by using the maximum likelihood method. Each term of the log-likelihood function depends on the probability function of X_{k+j} conditioned to the k previous observations, $P(X_{k+j} = x_{k+j}|x_k, \ldots, x_j, \Theta)$.
It is immediate to see that this is just the probability function of the sum of k independent Binomial random variables with parameters $(x_{t-i}, \rho_i(t))$ for $i = 1, \ldots, k$, plus and independent Hermite random variable with parameters $a_1(t)$ and $a_2(t)$. Some of the results reported in Moriña et al. (2011) have been used to obtain a closed expression for this probability function, facilitating the estimation of the parameters.
Model selection was made using the AIC criteria, evaluating the statistical significance of the parameters and analyzing its validation. That is, the best model among all considered is one that has all its parameters statistically significant, an AIC small enough and a validation good enough; the final choice is a model with the best balance between the three criteria described.

2.2 Forecasting

We have considered two kinds of quantities to be predicted: the future average behaviour of the series and the crude number of counts.
The point prediction of the average behaviour of the series is computationally simple. Taking expectations in (1), calling $E(X_t) = \mu_t$, we obtain a recurrence relation similar to that described in Moriña et al. (2011). From here, confidence intervals are calculated by using the *delta method*, where

the gradients are estimated by using a first-order approximation of the partial derivates.
The prediction of the crude number of counts is performed by considering the raw distribution of the m-step ahead future observation. It allows to estimate the distribution of the futures X_{n+j} for $j = 1, \dots m$, using the estimated parameters of (1) and the observed values. Once the distributions of X_{n+k} for $k = 1, \dots, m$ are estimated, the predicted interval is obtained from the corresponding quantile.

3 Examples of application

Several researchers remark the pronounced change in the studied series of the number of deaths in dairy cattle farms in 2012 and 2013. It is detected that during these years, many dairy cattle farms closed. A plausible explanation of this decrease would be the abandon of the rural activity in cattle farms due economical crisis of this primary sector. In fact, during 2012 and 2013 the cattle farms under surveillance diminished from 4 to 1, and a unique farm was monitored during this period.
It is decided that an accurate analysis of this time series must be conducted in two different ways. That is, the series will be studied selecting the registers recorded from 2007 to 2011, and using all the information from 2007 to 2013. For this reason, to conduct an accurate analysis the future predictions should be obtained considering that in 2012 and 2013 only one dairy cattle farm is working in the area. However, it is important to study the consistency of the results obtained with and without the information recorded during 2012 and 2013.

3.1 Number of carcasses collected in dairy cattle farms between 2007 and 2011

Data comes from the dairy cattle farms placed in a small region of Catalonia. Bovine carcasses were collected during 2007 to 2011 and aggregated by week. A range between 0 and 7 carcasses were collected. Over the half period the median number of bodies recorded was 0.5, with a weekly average of 1, approximately.
Different INAR models were tested with several patterns of trend and seasonality. An INAR(2) model with Hermite innovations were selected, taking into account the criteria for choosing the most suitable model.

$$\hat{X}_t = 0.067 \circ X_{t-1} + 0.089 \circ X_{t-2} + W_t(\hat{a}_1(t), \hat{a}_2(t)), \tag{4}$$

where $\hat{a}_1(t) = e^{-1.15-0.86\sin\left(\frac{2\pi t}{52}\right)}$ and $\hat{a}_2(t) = 0.226$.
Residuals validation was satisfactory, that is, residuals seemed to be white noise. It was tested by exploring their ACF and PACF profiles. A cross-validation was conducted, and it was also satisfactory because the observed

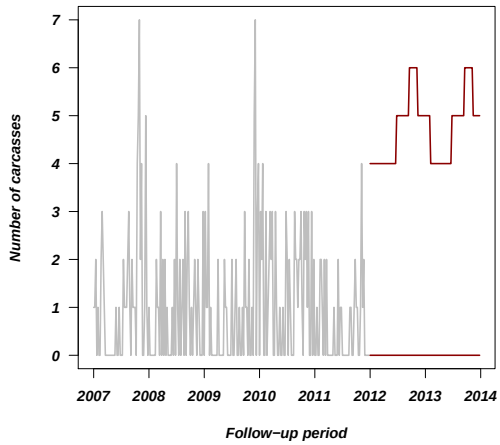

FIGURE 1. Exact 99% confidence interval for the predictions for the number of carcasses recorded during 2012 and 2013 on dairy cattle farms obtained by means of INAR(2) with Hermite innovations.

registers were located inside the confidence intervals. Forecasting were performed for 2012 and 2013 (see Figure 1).

3.2 Number of carcasses collected in dairy cattle farms between 2007 and 2013

Bovine carcasses were collected weekly from 2007 to 2013, and again the number of bovine carcasses were aggregated by week . Here a range between 0 and 8 was collected, and over the half period of study the median number of collection recorded was 0, with a weekly average of 1 carcass.
After exploring different INAR models with several options of trend and seasonality, the best model was an INAR(3) with Poisson innovations, leading to the estimated process,

$$\hat{X}_t = 0.13 \circ X_{t-1} + 0.10 \circ X_{t-2} + 0.09 \circ X_{t-3} + W\left(\widehat{a}_1(t)\right), \qquad (5)$$

where $\hat{a}_1(t) = e^{-0.003-0.004t+0.38\cos\left(\frac{2\pi t}{52}\right)}$.
Again, two model validations were performed. One exploring the ACF and PACF profiles of the model residuals, and the other based on a cross-validation. Residuals seemed to be distributed as white noise according to the ACF and PACF profiles. Cross-validations were dynamic adding new values to the training sample at one step ahead, and predicting future values. Forecasting has also been made for 2014 and 2015 (see Figure 2).

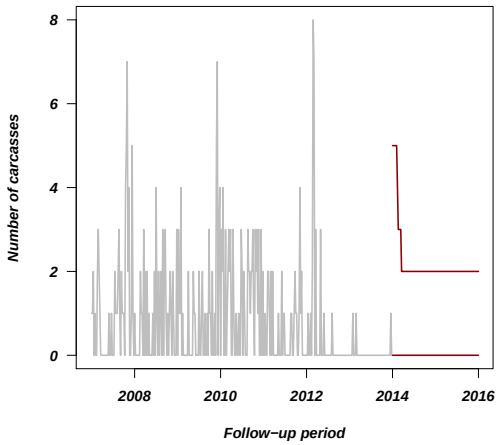

FIGURE 2. Exact 99% confidence interval for the predictions for the number of carcasses recorded during 2014 and 2015 on dairy farms obtained by means of INAR(3) with Poisson innovations.

Acknowledgments: This work was funded by the grant MTM2012-31118 from the Ministry of Economy and Competitiveness of Spain.

References

Alba, A., Dórea, F.C., Arinero, L., Sánchez, J., Cordón, R., Puig, P., and Revie, C.W. (2015). Exploring the surveillance potential of mortality data: nine years of bovine fallen stock data collected in Catalonia (Spain). *Plos One*, accepted.

Jung, R.C. and Tremayne, A.R. (2006). Binomial thinning models for integer time series. *Statistical Modelling*, **6**, 81 – 96.

Kemp, C.D. and Kemp, A.W. (1965). Some properties of the Hermite distribution. *Biometrika*, **52**, 381 – 394.

Moriña, D., Puig, P., Ríos, J., Vilella, A., and Trilla, A. (2011). A statistical model for hospital admissions caused by seasonal diseases. *Statistics in Medicine*, **30**, 3125 – 3136.

Weiß, C.H. and Puig, P. (2015). The marginal distribution of compound Poisson INAR(1) processes. *Stochastic Models, Statistics and Their Applications*, **122**, 351 – 359.

A Bayesian model for the Ebola epidemic in Sierra Leone

Gianluca Frasso[1], Philippe Lambert[12], Wilfried Bonou[1]

[1] Institut des sciences humaines et sociales, Méthodes quantitatives en sciences sociales, Université de Liège, Liège, Belgium.,
[2] Institut de statistique, biostatistique et sciences actuarielles (ISBA), Université catholique de Louvain, Louvain-la-Neuve, Belgium.

E-mail for correspondence: Gianluca.Frasso@ulg.ac.be

Abstract: We propose a Bayesian model for the analysis of the 2014 Ebola outbreak in Sierra Leone. It is based on an extension of the popular compartmental SEIR model specified using a system of differential equations.

Keywords: Bayesian inference; Ebola outbreak; Differential equations; SEIR-D model.

1 Introduction

The 2014 Ebola epidemic in Sierra Leone started in May (WHO Ebola Response Team, 2014). The absence of efficient treatment for Ebola and its large fatality rate make the spread of that infection particularly worrying if the central authorities do not react promptly.
Mathematical models, based on differential equations, are useful tools to describe the epidemic evolution, to measure or to simulate the effects of public health interventions and to forecast future course of the disease in a population.
Inference for such epidemic models is complex: one or more of the state variables involved in the system of ODEs is often not observed and the collected information can be unreliable.
We propose a Bayesian approach to analyze the Ebola epidemic in Sierra Leone by extending the SEIR compartment model (Anderson and May, 1992). The Bayesian framework allows to combine relevant prior information on model parameters with partial reports on the course of the epidemic,

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

and to quantify uncertainty on derived key quantities such as the effective reproduction number (Heffernan et al., 2005).

2 SEIR-D model for the Ebola epidemic

At any time t, an individual belonging to a population of N subjects, can be defined as: *susceptible* (S), i.e. at risk; *exposed* (E), i.e. infected but not yet infectious; *infectious* (I), i.e. able to transmit ebola; *dead* (D) or *recovered* (R). The following system of differential equations describes the epidemic dynamic (lower case letters refer to state proportions):

$$\begin{array}{rcl} s'(t) & = & -\beta\pi_{\mathcal{E}}(t)i(t)s(t) \\ e'(t) & = & \beta\pi_{\mathcal{E}}(t)i(t)s(t) - \sigma e(t) \\ i_r'(t) & = & \sigma(1-\pi_d)e(t) - \gamma_r i_r(t) \\ i_d'(t) & = & \sigma\pi_d e(t) - \gamma_d i_d(t) \\ r'(t) & = & \gamma_r i_r(t) \\ d'(t) & = & \gamma_d i_d(t). \end{array} \tag{1}$$

The state probabilities can be computed by solving the system of ODEs numerically for given values of the unknown parameters $\boldsymbol{\theta} = (\beta, \pi_{\mathcal{E}}(t), \sigma, \pi_d, \gamma_r, \gamma_d)$ and initial conditions.
We let $\pi_{\mathcal{E}}$ change smoothly over time by modeling its logit as a linear combination of a generous number of B-splines: $\log \dfrac{\pi_{\mathcal{E}}(t)}{1-\pi_{\mathcal{E}}(t)} = \sum_k b_k(t)\alpha_k$, penalizing for changes in the successive α_k's (Eilers and Marx, 1996).
The *effective reproduction number* ($\mathcal{R}_e$) (Heffernan et al., 2005) is the expected number of secondary cases caused by an infected individual and indicates the epidemic's potential spread. A value less than one suggests that the disease will die out, whereas it is in an epidemic state when larger than one. It can be computed as:

$$\mathcal{R}_e(t) = \beta\pi_{\mathcal{E}}(t)s(t) \times \left(\frac{1-\pi_d}{\gamma_r} + \frac{\pi_d}{\gamma_d}\right). \tag{2}$$

The first factor corresponds to the expected number of newly infected persons during one unit of time (here: one day) at time t, while the second one is the expectation of the time spent in the infectious state.

3 Analysis of the Sierra Leone Ebola outbreak

The Ministry of Health of Sierra Leone provides, since 18 July 2014, daily summaries on the number of (confirmed) exposed, infectious, recovered or dead persons due to ebola. Sparse and non detailed information is available before that date.

TABLE 1. Prior elicitation on SEIR-D model parameters suggested by WHO Ebola Response Team, 2014.

	Parameter	Mean ± s.d.	n	Our prior
Pr(die)	π_d	307 deaths	445	Beta(308,139)
Incubation (days)	σ^{-1}	9.0 ± 8.1 days	201	$\mathrm{N}\left(9.0, \frac{8.1}{\sqrt{201}}\right)$
Time from symptoms				
– to death	γ_d^{-1}	8.6 ± 6.9 days	128	$\mathrm{N}\left(8.6, \frac{6.9}{\sqrt{128}}\right)$
– to recovery	γ_r^{-1}	17.2 ± 6.2 days	70	$\mathrm{N}\left(17.2, \frac{6.2}{\sqrt{70}}\right)$

3.1 Model parameter inference

Denote by $\mathrm{d}I^+(t)$ the number of new Ebola (infectious) cases over the period $(t, t+\mathrm{d}t)$. Conditionally on the total number of exposed subjects at time t (never exactly known), we assume that $\mathrm{E}(\mathrm{d}I^+(t)|\sigma, \mathrm{E}(t)) = \sigma \mathrm{E}(t)\mathrm{d}t$. Conditionally on the parameters $\boldsymbol{\theta}$, we approximate $\mathrm{E}(t)$ by $Ne(t)$ and so $\mathrm{E}(I^+(t+1)|\boldsymbol{\theta}) = \sigma N e(t)$. If the report of cases at time $t+d$ refers to a $d-$days time span, then we consider it as an aggregation of new cases over d consecutive days:

$$\mathrm{E}(I^{+d}(t)|\boldsymbol{\theta}) = \mathrm{E}(I^+(t)|\boldsymbol{\theta}) + \cdots + \mathrm{E}(I^+(t+d-1)|\boldsymbol{\theta}) = \sigma N(e(t) + \cdots + e(t+d-1)).$$

A working likelihood is obtained by treating the reported daily cases as independent Poisson random variables:

$$\left(\mathrm{d}I^{+d}(t)|\boldsymbol{\theta}\right) \sim \mathrm{Pois}\Big(\sigma N\big(e(t) + \ldots + e(t+d-1)\big)\Big). \tag{3}$$

If overdispersion is suspected, a negative binomial distribution with overdispersion parameter ϕ can be used instead.

Prior distributions of the parameters $\{\pi_d, \sigma, \gamma_d, \gamma_r\}$ (in Table 1) can be elicited using information gathered during the follow-up of Ebola patients in Sierra Leone (WHO Ebola Response Team, 2014).

For identifiability reasons, in the estimation process, β was arbitrarily set to a large (enough) value. A (non-reported) sensitivity analysis confirmed the robustness of the results to that choice. A uniform prior on $(0, 100)$ is considered for the overdispersion parameter ϕ in the distribution for the reported number of new cases.

A prior penalizing changes in successive spline parameters is assumed for $\boldsymbol{\alpha}$ defining $\pi_{\mathcal{E}}$. Following Jullion and Lambert (2007), we take:

$$p(\boldsymbol{\alpha}|\lambda) \propto \exp\left(-\frac{\lambda}{2}\sum_k (D\boldsymbol{\alpha})_k^2\right) = \exp\left(-\frac{\lambda}{2}\boldsymbol{\alpha}^{\mathrm{T}} P \boldsymbol{\alpha}\right), \ (\lambda|\xi) \sim \mathrm{G}\left(\nu/2, \nu\xi/2\right),$$

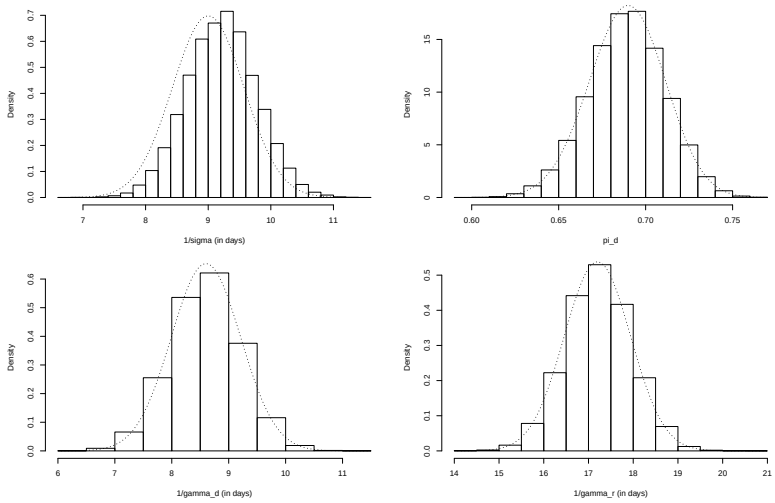

FIGURE 1. Priors (dotted lines) and histograms of the sampled σ^{-1}, π_d, γ_d, γ_r parameters produced by $500k$ iterations of the random-walk Metropolis-within-Gibbs algorithm.

where $\xi \sim \mathrm{G}(\epsilon, \epsilon)$, D and λ denote a 3rd order difference matrix and a roughness penalty parameter with a continuous mixture of gamma priors (with $\epsilon = 10^{-4}$, say) respectively.
The joint posterior is obtained by Bayes' theorem. A Metropolis-within-Gibbs algorithm can be used to sample the posterior. For the Metropolis step we generate normal proposals with adaptive variance as in Haario et al. (2001). Gibbs steps are only possible for the penalty parameters λ and ξ.

3.2 Estimation results

The SEIR-D model was fitted on the confirmed number of new Ebola cases at different calendar times. Data were aggregated on a weekly basis (from Sunday to Saturday) starting on August 12 to account for the irregular timing in the laboratory reports.
A Runge-Kutta scheme has been used to solve the set of ODEs for given values of $\boldsymbol{\theta}$ and initial conditions: $s(0) = 1 - \frac{13}{N}; e(0) = \frac{12}{N}; i_r(0) = 0; i_d(0) = \frac{1}{N}; r(0) = 0; d(0) = 0; N = 6.2 \times 10^6$.
A chain of length $500k$ with a $10k$ burn-in was considered to explore the joint posterior. Not surprisingly, the information on π_d, γ_d and γ_r was found to come solely from their priors (Figure 1). The marginal posterior for σ^{-1} was found slightly left-skewed, but shared a mean and a variance similar to its prior values. The sampled overdispersion parameter of the negative

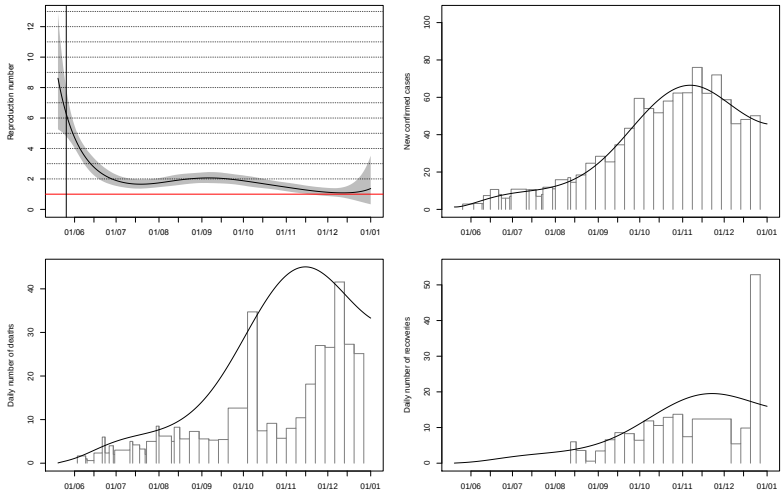

FIGURE 2. Upper left panel: posterior mean and point wise 95% credible intervals for $\mathcal{R}_e(t)$. Upper right panel and lower plots: observed and fitted (solid lines) number of new cases, deaths and recoveries respectively.

binomial distribution for the number of new cases suggests that the Poisson is a legitimate simplification.

The effective reproduction number has been computed at each MCMC iteration. That sample has be used to estimate its posterior mean and pointwise 95% credible intervals for a grid of time points. The estimated $\mathcal{R}_e(t)$ (Figure 2) suggests that, at the end of December 2014, Sierra Leone is most likely still in an epidemic state, although there are strong indications that the disease transmission steadily steps back since mid-September.

The fitted number of new cases shows that the SEIR-D model appropriately describes the dynamic of the disease transmission. The reported death and recovery counts do not appear coherent with the evolution of the confirmed number of new cases and suggest a strong under-reporting.

4 Discussion

We presented a Bayesian approach to analyze the 2014 Ebola epidemic in Sierra Leone based on a SEIR-D system of ODEs.

Our framework enables a coherent combination of the observed data and of prior knowledge on the model parameters. Furthermore, the flexible definition of the probability of infection ensures a very good fit of the number of reported cases even if, the numbers of reported deaths and recoveries seem largely under-reported.

According to the estimated reproduction number, at the end of December

2014, the epidemic appears still in action in Sierra Leone, although there are strong indications that the disease transmission steadily steps back since mid-September.
Some extensions of the discussed framework are possible. If more reliable prior information or data on some extra disease states were available, the SEIR-D model could be enriched by defining a more sophisticated virus transmission dynamic. On the other hand, a spatial description of the virus spread and its connection to neighbor countries would also represent an interesting extension of our framework.

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

References

Anderson, R.M. and May, R.M. (1992). Infectious Diseases of Humans: Dynamics and Control. *Oxford science publications*, **371**, 1481 – 1495.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89 – 101.

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

Heffernan, J.M, Smith, R.J., and Wahl, L.M. (2005) Perspectives on the basic reproductive ratio. *Journal of the Royal Society Interface*, **2**, 281 – 293.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics and Data Analysis*, **51**, 2542 – 2558.

WHO Ebola Response Team (2014). Ebola virus disease in West Africa – The first 9 months of the epidemic and forward projections. *New England Journal of Medicine*, **371**, 1481 – 1495.

Merging parallel MCMC output with yin-yang sampling

Sylvia Frühwirth-Schnatter[1], Alexandra Posekany[1]

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

E-mail for correspondence: alexandra.posekany@wu.ac.at

Abstract: Due to a plethora of available data, researchers in Bayesian computing and machine learning have turned their focus of interest towards methods for handling big data split into subsets, which perform inference independently in parallel and then merge these outputs. Such approaches are of increasing importance in fields reaching from social networks' and search engines' data to research in medical statistics and econometrics. Nowadays, Data too large for a single analysis due to the computational burden have become the main challenge. Finding reasonable approaches for combining independently obtained results, typically samples from posterior distributions, and subsequently obtaining a common result which recovers the joint posterior distribution and resulting posterior estimators or decisions, is far from trivial. We face this challenge by designing the yin-yang sampler which provides a mathematically sound way of merging two samples from posterior distributions based on different subsets of the data. Our idea is to correct for applying the same prior for each subset instead of only once for the full data set. Applying yin-yang sampling steps sequentially or in a pairwise algorithm allows us to recover the full sample's posterior from subsamples' posteriors for any given number of reasonably large subsets which have to contain enough information for sound inference results. For demonstrating our approach, we analyse several simulated sample data sets, including a linear regression example and look into real economic data.

Keywords: Big data analysis; Parallel MCMC; Bayesian computation.

1 Introduction and literature

An increasingly frequently occurring challenge in Bayesian inference is presented by big data, as the available data often surpass the computational infrastructure's capabilities. To deal with these modern day challenges, sev-

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

eral approaches have been proposed focussing on merging sample outputs obtained from posterior inference outcomes computed in parallel. Infrastructure for massively parallel computing, cloud computing and clusters, allow inferring data in many smaller parts, even though a full scale inference of the complete data set is not possible. Here, the 'independent product equation' termed by Wang and Dunson (2013),

$$\begin{aligned} p(\theta|y=(y_1,\ldots,y_S)) &\propto p(\theta|y_1)p(\theta|y_2)\ldots p(\theta|y_S) \\ &\propto p(y_1|\theta)p(\theta)^{\frac{1}{S}}p(y_2|\theta)p(\theta)^{\frac{1}{S}}\ldots p(y_S|\theta)p(\theta)^{\frac{1}{S}} \end{aligned} \tag{1}$$

forms the basis for three approaches by Scott et al. (2013), Neiswanger et al. (2013) and Wang and Dunson (2013) in different respects. Scott et al. (2013) introduced consensus MCMC, a straight-forward approach heuristically motivated for approximately normally distributed data by calculating the iteration-wise mean of the draws generated from parallel MCMC chains on the partition's data subsets ('shards') weighted by the respective sample variances. Neiswanger et al. (2013) introduced the term embarrassingly parallel MCMC, an approach based on estimating the joint posterior of the parameter θ as the product of parallel subset estimates based on data subsets $y_1,\ldots,y_S$ where the prior is adjusted according to (1). Similarly, Wang and Dunson's (2013) Weierstrass rejection sampling method approximates the subset densities applying the Weierstrass transformation.
Analogous to the above mentioned methods, our approach aims for splitting those samples too large for straight-forward analysis into subsamples, performing independent inference on each respective subsample and eventually combining all individual results. Contrary to the above mentioned methods, we correct for this 'overuse' of the prior in a stepwise manner. For demonstrating our approach, we simulated several sampled data sets, including a linear regression example and looked into real economic data.

2 Yin-yang sampling schemes

Both of our approaches start with splitting the data into S smaller data sets $y_1,\ldots,y_S$. Instead of analysing the full data set y, we perform parallel inference on the subsets separately. This step allows for parallelisation, as no information has to be shared between the different inferential algorithms obtaining the subsets' posteriors. The sampling scheme is symmetric in the yin and yang sample by which names we refer to the two samples which are to be merged in this stepwise procedure. Both our approaches, require an estimation of the posterior probability density which we derive from the samples' particles using a standard kernel density estimator. For the yin sample, this estimator is

$$\hat{p}(\theta|y^{in}) = \frac{1}{h_{yin}M_{yin}}\sum_{l=1}^{M_{yin}} K\left(\frac{\theta-\theta_{yin}^{(l)}}{h_{yin}}\right), \tag{2}$$

where $K(\cdot)$ is a symmetric kernel and $h_{yin} > 0$ is the bandwidth. In a similar fashion, we derive the estimator $\hat{p}(\theta|y^{ang})$ for the yang sample.
The idea behind our first approach, the simple yin-yang sampler, is to draw sample indices multinomially. The probabilities of the multinomial distribution depend on the yang sample which is to be merged with the yin sample and vice versa. For this, we apply weights w_l for each observation θ_l^{in} from the yin sample. These weights are derived as the ratio of the posterior density of the yang y^{ang} or yin y^{in} sample estimated over the joint support of both samples and the corresponding prior distribution of this parameter limited to the joint support

$$\hat{w}_l^{in} = \frac{w_l^{in}}{\sum w_l^{in}}, \qquad w_l^{in} = \frac{\hat{p}(\theta_l^{in}|y^{ang})}{p(\theta_l^{in})}.$$

Corresponding weights exist for the yang sample. The required subsamples' posterior distributions are estimated applying kernel density estimation. Based on these weights, the algorithm draws multinomially from the respective sample, yin or yang, and then combines these draws into a new merged sample. The number of draws to include from each subsample depends on its reliability which we currently measure by the sample's variance.
Our second approach utilizes the Metropolis-Hastings acceptance probability to select draws which can be either proposed from the yin or the yang sample. Our algorithm switches between these two proposal distributions with a given probability making it flexible to deal with more or less dissimilar sampling situations. There are four possible cases of moving between the yin and yang sample, either to stay within the same sample, yin or yang, or to move to the other sample, from yin to yang or yang to yin. If we propose a new value from the yin sample, then the acceptance probability $\min(1, \alpha)$ is either

$$\alpha = \frac{\hat{p}(\theta^{new}|y^{ang})p(\theta^{old})}{\hat{p}(\theta^{old}|y^{ang})p(\theta^{new})},$$

when moving from yin to yin or

$$\alpha = \frac{\hat{p}(\theta^{new}|y^{ang})p(\theta^{old})(1 - p_{yin})}{\hat{p}(\theta^{old}|y^{in})p(\theta^{new})p_{yin}},$$

when moving from yang to yin, where p_{yin} is the probability to propose from the yin sample. In analogy to the yin-yang resampling scheme, the probabilities of moving between the samples or staying in the same sample depend on the sample variances.
For both algorithms, we consider two possible implementations. The straightforward idea is to merge two subset posterior samples in the first step. Then, sequentially we take this outcome sample as the yin sample for merging with the next subset posterior's yang sample in the next step. Alternatively, we

consider a treewise merging scheme, combining two samples each in the first sweep and then merging these outcome samples pairwisely. Both schemes perform well; the treewise scenario however is limited to a total number of samples S which is a power of 2. This variant might improve the merging behaviour by offering the possibility to merge samples according to degree of similarity, if the samples are provided in an ordered way e.g. of increasing difference from all outcomes' median or of the posterior sample's spread.
It should be noted that our approach is not limited to working with output from MCMC algorithms. In particular, the simple yin-yang algorithm works with output from any method, providing us with a sample from the posterior distribution e.g. approximate Bayesian Computation or sequential Monte Carlo methods. Unlike the existing approaches discussed in Section 1, we are therefore not limited to a single sampling approach not even a package, as our algorithm is meant for post-processing analysis for any reasonable algorithm's posterior samples.

3 Results and conclusions

A major issue for this algorithm to work is that the support of the samples to be merged should at least slightly overlap. For testing the robustness of this assumption, we simulated various normal and beta distribution scenarios with increasing difference between the means compared to the variances. Figure 1 shows the results for the scenario of means located at 2 and 4 with variance equal to 0.25. Although the overlap of the samples' support is small, we observed that recovering the posterior works fine for this scenario with 8 subsamples out of 10000 iid normal observations. In our simulation studies, we frequently observed that the simple yin-yang sampler

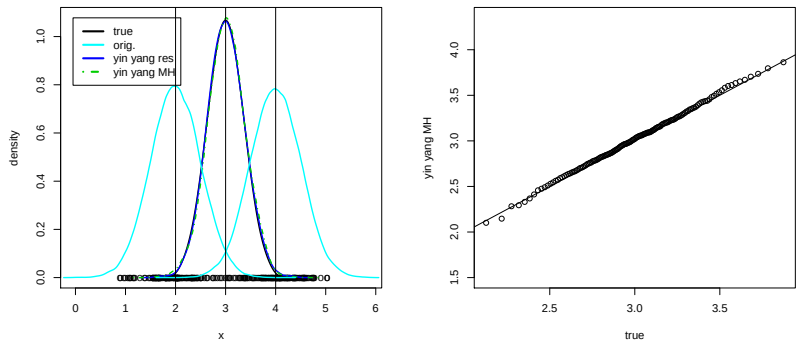

FIGURE 1. Comparison of sampled and true posterior densities for the normal distribution example.

provided results at least as good as the MH yin-yang sampler. In particular, we are interested in the performance of yin-yang sampling for working with MCMC output from a standard implemented Gibbs sampler. Therefore, we chose the regression algorithm from the R package 'MCMCpack' (Martin et al., 2013) representative for a standard tool not handcrafted for a particular situation and perform inference on 8 subsamples. Ex post, the yin-yang sampler merges these outcomes and we compare them against full data outcome. Figure 2 illustrates the algorithm's ability to handle such a scenario which is the target purpose of the algorithm. Here, consensus sampling is not used in its original sense with the 'corrected' prior $p(\theta)^{\frac{1}{S}}$ but with the 'full' prior $p(\theta)$ in order to demonstrate the yin yang sampler's ability to actually account for using the prior S times. As we can observe, the average distribution of all samples has far wider spread and the yin-yang posterior box coincides for all algorithm scenarios with the 25%, 50% and 75% quantile of the 'correct' simulated posterior of the full sample.
In certain cases, splitting the prior distribution according to (1) does not lead to the appropriate joint posterior distribution, even though the inde-

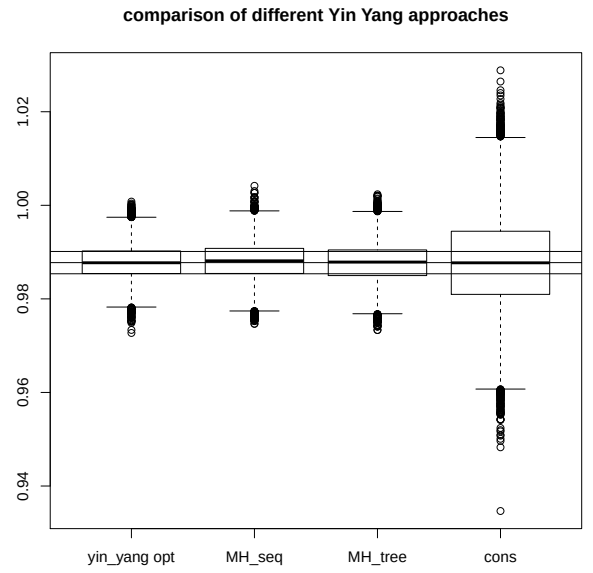

FIGURE 2. Results for the yin-yang sampler, the MH yin-yang sampler in a sequential and treewise merging implementation are compared against consensus sampling with the uncorrected prior, resulting in an incorrectly wide spread. The black lines mark the 25%, 50% and 75% quantile from the 'correct' posterior of the full sample obtained after 20,000 draws.

pendent product equation would suggest otherwise. One such case is the application of shrinkage priors, for which the shrinkage property crucially depends on the choice of hyperparameters which is meddled with, by taking the S-th square root. Mixture distributions present another possibly challenging case, as the availability of samples for each component to estimate is crucial. When 'empty' components with completely distorted posteriors dependent only on the prior information occur, simply averaging over these 'false' posteriors, as consensus MCMC proposes, will distort the whole result, while accounting for the prior can actually capture such cases. Thus, a valuable theoretical property of this algorithm is represented by its wide ability of recovering joint posterior distribution structures. Additionally, we provide theoretical proof that the MH algorithm is sound, even though its input consists of dependent samples from distributions. The main feature which makes our approach very attractive in practice compared to other existing algorithms is that both samplers allow for an easily implementable general purpose post-processing method, which can be applied to any kind of sampling output of Bayesian inference such as MCMC, SMC and ABC, provided by toolboxes like Stan, BUGS or JAGS.

References

Martin, A., Quinn, K., and Park, J. (2013). Markov chain Monte Carlo (MCMC) package. *CRAN repository, R package - version 1.3-3*.

Neiswanger, W., Wang, C., and Xing, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. *arXiv:1311.4780*

Scott, S., Blocker, A., and Bonassi, F. (2013). Bayes and big data: the consensus Monte Carlo algorithm. *Bayes*, **250**.

Wang, X. and Dunson, D. (2013). Parallel MCMC via Weierstrass sampler. *arXiv:1312.4605*.

Hazard modelling for interval-censored data by smoothing within the EM algorithm

Jutta Gampe[1], Hein Putter[2], Paul H.C. Eilers[3]

[1] Max Planck Institute for Demographic Research, Rostock, Germany
[2] Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
[3] Erasmus MC, Rotterdam, The Netherlands

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

Abstract: Flexible hazard modelling from interval-censored survival data is achieved by combining a piecewise constant hazard model and estimation via an EM algorithm that is supplemented by a smoothing step. Smoothness is enforced by a difference penalty on neighboring coefficients of the log-hazard.

Keywords: Interval-censoring; Hazard smoothing; EM algorithm; Penalty; Proportional hazards.

1 Introduction

Interval-censored survival data occur in many applications. They are easily handled in a parametric setting but often a more flexible approach may be preferred. We propose a model based on a piecewise constant hazard, with many short bins, which is supplemented by a roughness penalty on its coefficients to enforce smoothness. To estimate this model from interval-censored (IC) data we use the EM algorithm. In this way we combine an efficient computational approach (Poisson regression) with smoothness restrictions on the (log-)hazard and obtain a simple and stable EM algorithm to impute the unobserved information and to estimate the resulting hazard.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 Smooth hazards, interval-censoring and the EM algorithm

2.1 Smooth hazards

Piecewise constant hazard models have been used to estimate hazards flexibly, and the popularity of this approach is grounded on the relationship with Poisson regression (Holford, 1980), which leads to very efficient computation. However, due to its discontinuous hazard trajectory and the necessity to choose the breakpoints of the step function, alternative solutions that can preserve the advantages are more desirable. Enlarging the number of bins and adding a roughness penalty to enforce smoothness of the hazard is such an alternative.

In the simplest case we observe failure or right-censoring times t_i for $i = 1, \ldots, n$ individuals. We divide the time axis into a relatively large number of m narrow bins of equal length ε and allocate each observation to its proper interval. To estimate a piecewise constant hazard model we need the number of events and the total time under risk for each interval. Therefore we form two $n \times m$ matrices Y and E, which will hold the event and exposure information, respectively. Their rows correspond to the observations and their columns to the time bins. The matrices are filled as follows. For sake of simplicity of notation we re-scale the time axis in units of ε so that the bin length after rescaling is equal to 1. Then, if the event or right-censoring time t_i falls into interval k_i, we set $e_{ij} = 1$ for $j = 1, \ldots, k_i$. This way we represent the exposure to risk of individual i. (As the bins are narrow, setting $e_{ik_i} = 1$ is a negligible over-estimation.)

We set $y_{ik_i} = 1$, if t_i indicates an event, and $y_{ij} = 0$ otherwise. Forming the column sums of Y and E gives the m-dimensional vectors y and e holding the numbers of events and the total exposure times in the m bins.

If we denote by h_j the level of the piecewise constant hazard in bin j, we can estimate these parameters via Poisson regression (Holford, 1980) by assuming that the y_j are Poisson distributed with mean $\mu_j = h_j e_j$. To ensure smoothness of the hazard we impose a difference penalty on the elements of the vector $\eta = \ln \mu$. This leads to a penalized Poisson log-likelihood

$$\ell_{\text{pen}} = \ell - \lambda (D\eta)^{\mathrm{T}} (D\eta)/2 = \sum_{j=1}^{m} (y_j \eta_j - e^{\eta_j}) - \lambda \sum_{j=3}^{m} (\Delta^2 \eta_j)^2/2, \quad (1)$$

where D is a matrix that forms second order differences. λ is a smoothing parameter; increasing its value will give smoother results, because the penalty gets a stronger influence. Taking partial derivatives and setting them equal to zero we obtain the penalized likelihood equations $y - \mu = \lambda P \eta$ with penalty matrix $P = D^{\mathrm{T}} D$. Because $\eta = \ln \mu$ the system is nonlinear in the parameters, but we can linearize it. If $\tilde{\eta}_j$ is an approximate solution

we have $\mu_j \approx \tilde{\mu}_j + \tilde{\mu}_j(\eta_j - \tilde{\eta}_j)$ and therefrom

$$(\tilde{M} + \lambda P)\eta = y - \tilde{\mu} + \tilde{M}\tilde{\eta}, \tag{2}$$

where $M = \mathrm{diag}(\mu)$ and a tilde indicates an approximation to the solution. This linear system has to be solved repeatedly until convergence.
The system (2) is solved for a fixed value of λ, and the optimal value for the smoothing parameter needs to be chosen. To determine the optimal amount of smoothing we minimize the AIC $= Dev + 2\,ED$, where the deviance $Dev = 2\sum_j y_j \ln(y_j/\hat{\mu}_j)$ and the effective dimension of the model is $ED = trace[(\hat{M} + \lambda P)^{-1}\hat{M}]$. AIC is calculated over a sequence of λ-values, on log-scale, and its minimal value is determined.

2.2 Smooth EM algorithm for interval-censored data

For exact failure times and right-censored observations the construction of the matrices Y and E is straightforward. If the data are interval-censored, then the event of individual i is known only to lie in an interval $(l_i, r_i]$, and the likelihood contribution of the individual is given by $S(l_i) - S(r_i)$ where $S(\cdot)$ is the survival function. Consequently, neither the exact event counts y_{ij} nor the exposure contribution e_{ij} are known.
For interval-censored data and a piecewise constant hazard model the EM algorithm can be applied rather easily though (Lindsey and Ryan, 1998). Instead of using the y_{ij} and e_{ij} for each individual i and bin j, they are replaced by the conditional expectations y^*_{ij} and e^*_{ij}, given the data and current values of the hazard parameters $h^c = (h^c_1, \ldots, h^c_m)^{\mathrm{T}}$. If we denote the bins by $I_j = (\tau_{j-1}, \tau_j]$, $j = 1, \ldots, m$, these values are

$$\begin{aligned} y^*_{ij} = E(Y_{ij} \,|\, \mathrm{data}, h^c) &= P(Y_{ij} = 1 \,|\, \mathrm{data}, h^c) \\ &= \begin{cases} \frac{S(\tau_{j-1} \vee l_i) - S(\tau_j \wedge r_i)}{S(l_i) - S(r_i)}, & I_j \cap (l_i, r_i] \neq \emptyset; \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

and

$$e^*_{ij} = E(E_{ij} \,|\, \mathrm{data}, h^c) = \begin{cases} 1, & \tau_j \leq l_i; \\ \frac{S(\tau_{j-1}) - S(r_i)}{S(l_i) - S(r_i)}, & \tau_{j-1} \in (l_i, r_i]; \\ 0, & \tau_{j-1} > r_i. \end{cases}$$

The current hazard parameters h^c enter in the calculation of the survivor function $S(\cdot)$.
From these expected values the expected counts in bin j, $y^*_j = \sum_i y^*_{ij}$, and the expected total time at risk in bin j, $e^*_j = \sum_i e^*_{ij}$, result.
In the M-step the system analogous to (2), with the y^*_j and e^*_j replacing the exact numbers and obvious modification of the remaining terms,

$$(\tilde{M}^* + \lambda P)\eta^* = y^* - \tilde{\mu}^* + \tilde{M}^*\tilde{\eta}^* \tag{3}$$

is solved *once*, and for the obtained update of the hazard parameters h^c the expectations y^*_{ij} and e^*_{ij} are updated as well, before entering the single-iteration M-step again. This EM-cycle is continued until the change in the log-likelihood is below a threshhold. Rather than strictly following the EM algorithm by running the iteratively reweighted least squares (IRWLS) algorithm until convergence for each M-step, we propose to intertwine the E-step and the IRWLS algorithm. We replace the full IRWLS algorithm by a single iteration step. Since each each step in the IRWLS algorithm improves the likelihood, also each step in this adapted EM algorithm improves the likelihood and thus guarantees convergence of the algorithm (at least to a local maximum of the likelihood). The penalization of the parameters not only enforces smoothness of the estimates but also removes disadvantages like slow convergence of the EM algorithm and identifiability problems.
The procedure again hinges on the value of the smoothing parameter λ. Also in this case we run the estimation over a grid of λ-values and choose the optimal smoothing parameter as the value which minimizes the AIC.

2.3 Baseline bases and covariates

So far no covariates were included but generally we will be interested in the impact of covariates on the smooth baseline hazard. Introducing covariates in a proportional hazards setting is rather straightforward in this approach. If the covariates for the n individuals are given in the $n \times K$ matrix X, then the hazard in interval j for individual i is $h_{ij} = \exp(\alpha_j + \sum_k x_{ik}\beta_k)$ and $\mu_{ij} = e_{ij} \exp(\alpha_j + \sum_k x_{ik}\beta_k)$. We can see that all exposures in row i are scaled by $\exp(\sum_k x_{ik}\beta_k)$.
Another possible modification is to express the logarithm of the smooth baseline hazard as a linear combination of B-splines and penalize their coefficients (Eilers and Marx, 1996). If the number of bins is large such a representation can be more parsimonious while preserving sufficient flexibilty. If $B \in \mathbb{R}^{m \times p}$ represents the B-spline basis evaluated at the m bins, we have $\ln h_j = (B\,\alpha)_j$ with $\alpha \in \mathbb{R}^p$. If we choose $B = I$ this reduces to $\ln h_j = \alpha_j$ as before.

3 Application

To illustrate the performance of the smooth EM algorithm we present a simulated data example for which we know the true hazard. The underlying distribution is a mixture of two Gompertz distributions (exponentially increasing hazard $h(t) = a \exp\{bt\}$) with rather different hazard levels. Marginally this leads to a 'wavy' hazard in the population, which is induced by the relatively quicker failure in the high-risk subgroup. With our flexible approach we hope to be able to recover this hazard shape.
Figure 1 shows the results for a sample size $n = 350$. The parameters of the two Gompertz distributions were $(a_1 = 0.015, b_1 = 0.1)$ and $(a_2 =$

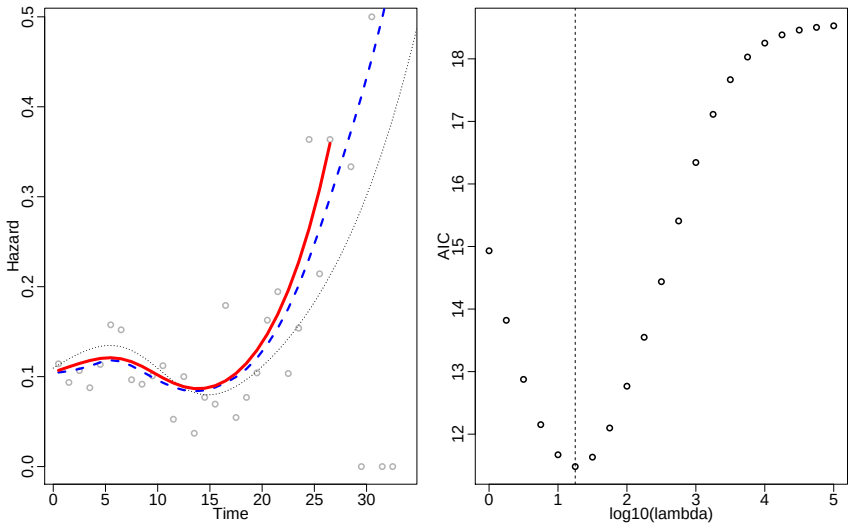

FIGURE 1. Estimates of the hazard (thin dotted line) by piecewise constant model (dots), penalized Poisson regression (dashed line) and from interval-censored data by smooth EM algorithm (thick solid line). Right: AIC profile over λ-grid for smooth EM approach.

$0.15, b_2 = 0.1$), and mixing proportions $\pi_1 = 0.3$ and $\pi_2 = 0.7$. The thin dotted line shows the corresponding true hazard. The time axis was clipped in bins of length $\varepsilon = 1$.

The figure presents three estimates: The hazard that was estimated based on the exact event times, which were aggregated into the bins as described in Section 2.1. The points are the empirical estimates for the simple piecewise constant model without any penalization. The dashed line shows the results from a penalized Poisson model, see (2), with difference penalty of order $d = 2$ and optimal smoothing parameter chosen by minimizing the AIC over a grid.

The exact failure times where then submitted to an observation scheme which creates interval-censored information and which mimics schemes encountered in clinical practice. Follow-up times were scheduled after every five time units, individually jittered by addding a random number from $U[0, 0.8]$. Observations with event times after the last follow-up time were right-censored at last follow-up (13 in this case).

Again the time axis was clipped in bins of length $\varepsilon = 1$, but this time the baseline was expressed via a cubic B-spline basis (size $p = 10$), penalty order $d = 2$. The solid thick line gives the result of the smooth EM algorithm that was applied to these interval- (and right-) censored data. The right panel in Figure 1 shows the AIC over the grid of λ-values and the identified

optimal smoothing parameter for the EM approach.

4 Discussion and outlook

We have proposed an approach to obtain a smooth hazard for interval-censored survival times that is based on a discretization of the time axis. Short bins of equal length and a piecewise constant hazard allow to exploit established links to Poisson regression, while penalization of the coefficients guarantees smoothness and preserves efficient computation. The extension to interval-censored data benefits from a simple E-step, while the penalization removes disadvantages like slow convergence of the EM algorithm and identifiability problems. Covariates can be readily introduced in a proportional hazards setting. Our proposal generalizes the Peto-Turnbull self-consistency algorithm (Meeker and Escobar, 1998).

References

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with splines and penalties (with Discussion). *Statistial Science*, **11**, 187 – 121.

Holford, T.R. (1980). The analysis of rates and of surviorship using log-linear models. *Biometrics*, **36**, 299 – 305.

Lambert, P. and Eilers, P.H.C. (2005). Bayesian proportional hazards model with time-varying regression coefficients: a penalized Poisson regression approach. *Statistical Modelling*, **7**, 239 – 254.

Lindsey, J.C. and Ryan L.M. (1998). Tutorial in biostatistics: Methods for interval-censored data. *Statistics in Medicine*, **17**, 219 – 238.

Meeker, M.Q. and Escobar, L.A. (1998). *Statistical Methods for Reliability Data*. Wiley.

Penalized distributed lag non-linear models

Antonio Gasparrini[1], Ben Armstrong[1], Michael G. Kenward[1]

[1] London School of Hygiene and Tropical Medicine, Keppel Street WC1E 7HT, London, United Kingdom

E-mail for correspondence: antonio.gasparrini@lshtm.ac.uk

Abstract: Distributed lag non-linear models (DLNMs) are a modelling tool for describing non-linear and delayed dependencies. DLNMs defines bi-dimensional exposure-lag-response associations through a tensor product of two sets of basis functions, which specify the standard exposure-response and additional lag-response relationships, respectively. These methods were originally developed in time series analysis, and they have recently been extended to other study designs and data structures. Here we propose an extension of this framework, where the exposure-lag-response surface is modelled through penalized splines within generalized additive models. This development ensures higher flexibility, automatic model selection and improved inferential properties for this modelling class.

Keywords: Distributed lag; Penalized splines; Generalized additive models.

1 Introduction

In time series, dose-response relationships are usually represented by a function $f(x_t)$, which defines the exposure-response dependency between an outcome and a variable x measured at time t. This simple representation does not take any account of potential *lagged* dependencies, which occur when x_t affects the outcome also at future times $t + \ell$, or, from another perspective, when the outcome at time t depends also on $x_{t-\ell}$. This new quantity ℓ, the *lag*, defines a new dimension expressing the temporal structure of the association.

2 Distributed lag linear and non-linear models

Distributed lag non-linear models (DLNMs) are statistical tools for modelling such dependencies (Gasparrini, 2014). DLNMs are specified by a function s that models the dependency in terms of a vector $\mathbf{q}_t = x_{t-\ell_0}, \dots, x_{t-\ell}$

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

of equally-spaced occurrences in the past. This function is defined as:

$$s(x_{t-\ell_0},\dots,x_{t-\ell}) = s(\mathbf{q}_t) = \sum_{\ell=\ell_0}^{L} f \cdot w(x_{t-\ell},\ell). \tag{1}$$

Here the bi-dimensional *exposure-lag-response* function $f \cdot w(x_{t-\ell},\ell)$ is composed by the standard expoure-response function $f(x)$ and an additional lag-response function $w(\ell)$ modelling the lag structure in the space $\boldsymbol{\ell} = \ell_0,\dots,L$, with ℓ_0 and L and minimum and maximum lags. Parameterization of s is obtained by applying known basis transformations to the vectors $\mathbf{q}_t$ and $\boldsymbol{\ell}$, defining basis matrices $\mathbf{R}_t$ and $\mathbf{C}$ with v_x and v_ℓ columns, respectively. By combining them in a special version of tensor product involving vectors of 1's, s is parameterized as:

$$\mathbf{A}_t = (\mathbf{1}_{v_\ell}^{\mathrm{T}} \otimes \mathbf{R}_t) \odot (\mathbf{C} \otimes \mathbf{1}_{v_x}^{\mathrm{T}}), \tag{2}$$

$$s(x_{t-\ell_0},\dots,x_{t-\ell}) = (\mathbf{1}_{L-\ell_0+1}^{\mathrm{T}}\mathbf{A}_t)\boldsymbol{\eta} = \mathbf{w}_t^{\mathrm{T}}\boldsymbol{\eta}. \tag{3}$$

The function s is termed *cross-basis* function. Its shape, degree of flexibility and constraints are defined by the parametric form of the functions f and w, which can be independently chosen between several options such as polynomials, step functions or regression splines. This modelling class includes the standard distributed lag models (DLM) (Almon, 1965) as a special case, where f is a simple linear function.

3 A penalized version

Here we propose an extension of the current DLNM modelling class through penalized splines within generalized additive models (GAM), following a similar developments proposed in simpler DLMs for linear dependencies (Zanobetti, 2000; Obermeier, 2015). The idea is to form a richly parametererized cross-basis and then to penalize the model likelihood $l(\boldsymbol{\eta})$, following:

$$l_p(\boldsymbol{\eta}) = l(\boldsymbol{\eta}) - \frac{1}{2}\boldsymbol{\eta}^{\mathrm{T}}\Big(\lambda_x\left(\mathbf{1}_{v_\ell}^{\mathrm{T}} \otimes \mathbf{S}_x\right) + \lambda_\ell\left(\mathbf{S}_\ell \otimes \mathbf{1}_{v_x}^{\mathrm{T}}\right)\Big)\boldsymbol{\eta}. \tag{4}$$

The penalization is obtained through penalty matrices $\mathbf{S}_x$ and $\mathbf{S}_\ell$, with penalty parameters λ_x and λ_ℓ calibrating the smoothness of the exposure-lag-response function $f \cdot w(x_{t-\ell},\ell)$ in the spaces of x and ℓ, respectively. Different smoothers, penalties and estimation methods are available. Here we propose P-splines with second difference penalties and a REML estimator (Wood 2006 and 2008). The extension is implemented in R by embedding the packages *dlnm* and *mgcv* (Gasparrini, 2011; Wood, 2006).

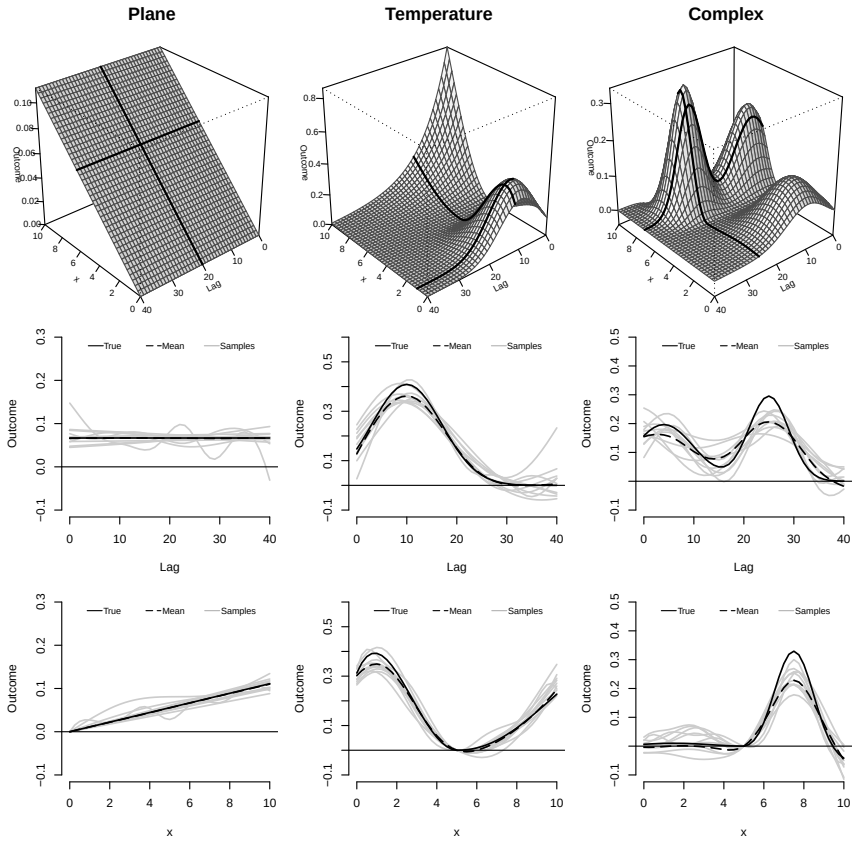

FIGURE 1. Simulated surfaces and fit of penalized DLNMs.

4 Simulation study

To assess the performance and inferential properties of GAM-based version of DLNMs, and to compare it with the standard version using GLM and AIC selection, we performed a simulation with three scenarios of exposure-lag-response surfaces: a simple plane, a shape resembling previously estimated temperature-mortality associations, and a complex surface.
Results are illustrated in Figure 1, displaying the simulated surfaces and the fit from penalized DLNMs, and in Table 1, reporting average computing time and average bias, coverage of nominal 95%CI and root mean square error (RMSE) across the surface. Estimators for penalized DLNMs appear superior to the unpenalized version, with proper coverage and slightly lower error.

TABLE 1. Performance of standard (GLM) and penalized (GAM) DLNMs.

	Plane		Temperature		Complex	
	GLM	GAM	GLM	GAM	GLM	GAM
Time (sec)	4.44	1.00	4.44	0.79	4.48	0.77
Bias	0.00	0.00	0.01	0.01	0.01	0.01
Coverage (%)	0.87	0.94	0.87	0.97	0.81	0.94
RMSE	0.11	0.11	0.18	0.15	0.19	0.18

5 An example

As an illustration, we applied the new penalized DLNM class for modelling temperature-mortality associations using daily time series data from London in 1993 – 2006. The models is specified with P-splines bases with dimension 10 for both spaces. Results are summarized in Figure 2, which displays the estimates for the whole exposure-lag-response risk surface, the lag-response for 29°C, and the overall cumulative exposure-response that represent the net effect cumulated along lags, respectively.

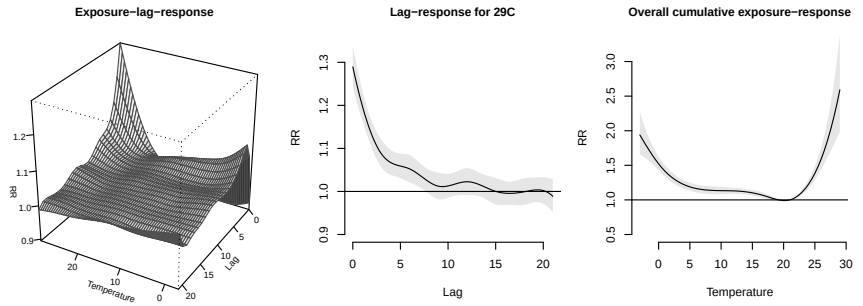

FIGURE 2. Estimated association between outdoor temperature and mortality.

The lag structure of the association varies depending on the temperature, with the exposure to 29°C being associated with a immediate increase in risk, then delayed for 5-10 days (Figure 2, middle panel). The overall cumulative mortality risk has a minimum at approximately 20°C, and increases for both warmer and colder temperatures (Figure 2, right panel).

6 Some conclusions

The penalized version of DLNMs offers higher flexibility, automatic model selection, and improved inferential properties to this modelling class. Further advancements are in development, such as the definition of alternative

smoothers within GAM and the application of additional penalty terms on the lag structure. The recent extension of DLNMs beyond time series data (Gasparrini, 2014) paves the way for original and promising applications of this modelling framework, which unifies and generalizes methods proposed in different research fields (Hauptmann et al., 2000; Berhane et al., 2008; Richardson, 2009; Abrahamowicz et al., 2012; Obermeier et al., 2015). The combination of the R packages *dlnm* and *mgcv* (Gasparrini, 2011; Wood, 2006) provides a well documented, efficient and stable software implementation.

References

Abrahamowicz, M., Beauchamp, M.-E., and Sylvestre, M.-P. (2012). Comparison of alternative models for linking drug exposure with adverse effects. *Statistics in Medicine*, **31**, 1014 – 1030.

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, **33**, 178 – 196.

Berhane, K., Hauptmann, M., and Langholz, B. (2008). Using tensor product splines in modeling exposure-time-response relationships: application to the Colorado Plateau Uranium Miners cohort. *Statistics in Medicine*, **27**, 5484 – 5496.

Hauptmann, M., Wellmann, J., and Kreienbrock, L. (2000). Analysis of exposure-time-response relationships using a spline weight function. *Biometrics*, **56**, 1105 – 1108.

Gasparrini, A. (2011). Distributed lag linear and non-linear models in R: the package dlnm. *Journal of Statistical Software*, **43**, 1 – 20.

Gasparrini, A. (2014). Modeling exposure-lag-response with distributed lag non-linear models. *Statistics in Medicine*, **33**, 881 – 899.

Obermeier, V., Scheipl, F., Wassermann, J., and Küchenhoff, H. (2015). Flexible distributed lags for modelling earthquake data. *Journal of the Royal Statistical Society: Series C*, **64**, 395 – 412.

Richardson, D.B. (2009). Latency models for analyses of protracted exposures. *Epidemiology*, **20**, 395 – 399.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall/CRC.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society: Series B*, **64**, 395 – 412.

Zanobetti, A., Wand, M.P., Schwartz, J., and Ryan, L.M. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, **1**, 279 – 292.

How many people visit YouTube? Imputing missing events in panels with excess zeros

Georg M. Goerg[1], Yuxue Jin[1], Nicolas Remy[1], Jim Koehler[1]

[1] Google, Inc.; United States

E-mail for correspondence: gmg@google.com

Abstract: Media-metering panels track TV and online usage of people to analyze viewing behavior. However, panel data is often incomplete due to non-registered devices, non-compliant panelists, or work usage. We thus propose a probabilistic model to impute missing events in data with excess zeros using a negative-binomial hurdle model for the unobserved events and beta-binomial sub-sampling to account for missingness. We then use the presented models to estimate the number of people in Germany who visit YouTube.

Keywords: Imputation; Missing data; Zero inflation; Panel data.

1 Introduction

Media panels (GfK Consumer Panels, 2013) are used by advertisers to estimate *reach* and *frequency* of a campaign: reach is the fraction of the population that has seen an ad, frequency tells us how often they have seen it (on average). It is important to get good estimates from panel data, as they largely determine the cost of an ad spot on TV or a website.
Naïvely, one would use a sample fraction of the number of non-zero events (website visits, TV spots watched, etc.) per unit time to estimate reach; similarly, for frequency. This, however, suffers from underestimation as panels often only record a fraction of all events due to e.g., non-compliance or work usage. Correcting this bias and imputing missing events has been studied previously (Fader and Hardie, 2000; Yang et al., 2010).
In this work we i) extend the beta-binomial negative-binomial (BBNB) model (Hofler and Scrogin, 2008) with a hurdle component to improve modeling excess zeros in panel data (Section 2); ii) present the maximum likelihood estimator (MLE) and also add prior information on missingness (Section 3); and iii) use the methodology to estimate – from online media

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

panels and internal YouTube log files – how many people in Germany visit YouTube (Section 4).
The proposed methodology can be applied to a great variety of situations where events have been counted – but some are known to be missing.

2 Hierarchical event imputation

Let $N_i \in \{0, 1, 2, \dots\}$ count the true (but unobserved) number of visits by panelist i. The population consists of people who do not visit YouTube at all (with probability $q_0 \in [0, 1]$), and those who visit at least once. If she visits (overcoming the "hurdle" with probability $1 - q_0$), we assume that N_i is distributed according to a shifted Poisson distribution (starting at $n = 1$) with rate λ_i. For model heterogeneity among the population we use a Gamma $\left(r, \frac{q_1}{1-q_1}\right)$ prior for λ_i, with $r > 0$ and $q_1 \in (0, 1)$.
Overall, this yields a shifted negative binomial hurdle (NBH) distribution

$$\mathbb{P}(N = n; q_0, q_1, r) = \begin{cases} q_0, & \text{if } n = 0, \\ (1 - q_0) \cdot \frac{\Gamma(n+r-1)}{\Gamma(r)\Gamma(n)} \cdot (1 - q_1)^r q_1^{n-1}, & \text{if } n \geq 1. \end{cases} \quad (1)$$

We choose a hurdle, rather than a mixture, model for the excess zeros (Hu et al., 2011), since $1 - q_0$ can be directly interpreted as the true – but unobserved – 1+ reach: if an advertiser shows an ad on YouTube they can expect that a fraction of $1 - q_0$ of the population sees it at least once.
Let p_i be the probability a visit of user i is recorded in the panel. Assuming independence across visits the total number of recorded panel events, $K_i \in \{0, 1, 2, \dots\}$, thus follows a binomial distribution, $K_i \sim \text{Bin}(N_i, p_i)$. To account for heterogeneity across the population we assume $p_i \sim \text{Beta}(\mu, \phi)$, with mean μ and precision ϕ (Ferrari and Cribari-Neto, 2004). Here μ represents the expected non-missing rate and ϕ the (inverse) variation across the population. Integrating out p_i gives a Beta-Binomial (BB) distribution,

$$K_i \mid N_i \sim BB(N_i; \mu, \phi). \quad (2)$$

Combining (1) and (2) yields a hierarchical beta-binomial negative-binomial hurdle (BBNBH) imputation model with parameter vector $\theta = (\mu, \phi, q_0, r, q_1)$:

$$N_i \sim NBH(N; q_0, r, q_1) \quad \text{and} \quad K_i \mid N_i \sim BB(K \mid N_i; \mu, \phi). \quad (3)$$

2.1 Joint distribution

The pdf of (2) can be written as

$$g(k \mid n; \mu, \phi) = \binom{n}{k} \frac{\Gamma(k + \phi\mu)\Gamma(n - k + (1 - \mu)\phi)}{\Gamma(n + \phi)} \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1 - \mu))}.$$

For $k = 0$ this reduces to

$$\mathbb{P}(K = 0 \mid N, \mu, \phi) = \frac{\Gamma(n + (1-\mu)\phi)}{\Gamma(n+\phi)} \times \frac{\Gamma(\phi)}{\Gamma(\phi(1-\mu))}. \tag{4}$$

Due to the zero hurdle it is useful to treat $N = 0$ and $N > 0$ separately:

$$\mathbb{P}(N, K) = \mathbb{P}(K \mid N) \cdot \mathbb{P}(N) = BB(k \mid n; \mu, \phi) \cdot NBH(n; q_0, q_1, r) \tag{5}$$

For $n = 0$, (5) is non-zero only for $k = 0$, $\mathbb{P}(N = 0, K = 0) = q_0$, since $\mathbb{P}(K > N) = 0$. For $n > 0$,

$$\begin{aligned}\mathbb{P}(N = n, K = k) =& (1 - q_0)\frac{1}{B(\phi\mu, \phi(1-\mu))}\frac{(1-q_1)^r}{\Gamma(r)} \times \frac{\Gamma(k + \phi\mu)}{\Gamma(k+1)} \\ & \times \frac{\Gamma(n - k + \phi(1-\mu))}{\Gamma(n-k+1)}\frac{\Gamma(n+r-1)}{\Gamma(n+\phi)} q_1^{n-1} \times \frac{\Gamma(n+1)}{\Gamma(n)}.\end{aligned}$$

2.2 Conditional predictive distribution for imputation

The panel records k_i events for panelist i, but we want to know how many events truly occurred. That is, we are interested in (dropping subscript i)

$$\mathbb{P}(N = n \mid K = k) = \frac{\mathbb{P}(K = k \mid N = n)\,\mathbb{P}(N = n)}{\mathbb{P}(K = k)}, \tag{6}$$

To obtain analytical expressions we consider $k = 0$ and $k > 0$ separately:

$k = 0$: Either none truly happened ($n = 0$) or a panelist visited at least once ($n > 0$), but none were recorded. For $n = 0$,

$$\mathbb{P}(N = 0 \mid K = 0) = \frac{q_0}{\mathbb{P}(K = 0)}. \tag{7}$$

For $n > 0$,

$$\begin{aligned}\mathbb{P}(N = n \mid K = 0) =& \frac{1}{\mathbb{P}(K=0)} \times \frac{\Gamma(n + \phi(1-\mu))}{\Gamma(n+\phi)}\frac{\Gamma(\phi)}{\Gamma(\phi(1-\mu))} \\ & \times (1 - q_0)\frac{\Gamma(n+r-1)}{\Gamma(n)}\frac{(1-q_1)^r}{\Gamma(r)} q_1^{n-1},\end{aligned}$$

where the second term comes from (4).

$k > 0$: The zero "hurdle" for N has been surpassed for sure.

$n < k$: By construction of Binomial subsampling

$$\mathbb{P}(N = n \mid K = k) = 0 \text{ for all } n < k. \tag{8}$$

$n \geq k$: Here

$$\begin{aligned}\mathbb{P}(N = n \mid K = k) = n \cdot q_1^{n-1}\frac{\Gamma(n - k + (1-\mu)\phi)}{\Gamma(n-k+1)\Gamma(n+\phi)}\Gamma(n+r-1) \times \\ \left(\sum_{m=0}^{\infty}(m+k)\frac{\Gamma(m + \phi(1-\mu))}{\Gamma(m+1)}\frac{\Gamma(m+k+r-1)}{\Gamma(m+k+\phi)} q_1^{m+k-1}\right)^{-1}.\end{aligned}$$

TABLE 1. MLE of $\theta_{(-\mu)}$ for panel data on YouTube visits in Germany ($\widehat{\mu}_{Logs} = 0.272$).

	Estimate	Std. Err.	t value	$Pr(>\|t\|)$
q_0	0.641	0.016	38.858	0.000
q_1	0.982	0.002	494.105	0.000
r	0.252	0.021	11.811	0.000
ϕ	2.320	0.594	3.907	0.000

3 Parameter estimation

Let $\mathbf{k} = \{k_1, \ldots, k_P\}$ be the number of observed events for all P panelist. Each panelist also has socio-economic indicators such as gender, age, and income. These attributes determine their demographic weight $\tilde{w}_i$, which equals the number of people in the entire population that panelist i represents. Finally, let $w_i = \tilde{w}_i \cdot \left(P / \sum_{i=1}^{P} \tilde{w}_i\right)$ be re-scaled weight of panelist i such that $\sum_{i=1}^{P} w_i$ equals sample size P.

We estimate θ using maximum likelihood (MLE), $\widehat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta; \mathbf{x})$, where the log-likelihood

$$\ell(\theta; \mathbf{x}) = \sum_{\{k | x_k > 0\}} x_k \cdot \log \mathbb{P}\left(K = k; \theta\right), \tag{9}$$

and $\mathbf{x} = \{x_k \mid k = 0, 1, \ldots, \max(\mathbf{k})\}$, where $x_k = \sum_{\{i | k_i = k\}} w_i$ is the total weight of all panelists with k visits.

For deriving closed form expressions of $\mathbb{P}(K = k) = \sum_{n=0}^{\infty} \mathbb{P}(N = n, K = k)$ it is simpler to consider $k = 0$ and $k > 0$ separately:

$$\begin{aligned} \mathbb{P}(K = 0) = q_0 + (1 - q_0) \times & \frac{\Gamma(\phi)}{\Gamma(\phi(1-\mu))} \frac{(1-q_1)^r}{\Gamma(r)} \\ & \times \sum_{n=0}^{\infty} \frac{\Gamma(n + 1 + \phi(1-\mu))}{\Gamma(n+1)} \frac{\Gamma(n+r)}{\Gamma(n+1+\phi)} q_1^n, \end{aligned} \tag{10}$$

and for $k > 0$,

$$\begin{aligned} \mathbb{P}(K = k) = & (1 - q_0)(1 - q_1)^r \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1-\mu))} \frac{1}{\Gamma(r)} \times \frac{\Gamma(k + \mu\phi)}{\Gamma(k+1)} \\ & \times \sum_{m=0}^{\infty} (m + k) \frac{\Gamma(m + \phi(1-\mu))}{\Gamma(m+1)} \frac{\Gamma(m + k + r - 1)}{\Gamma(m + k + \phi)} q_1^{m+k-1}. \end{aligned} \tag{11}$$

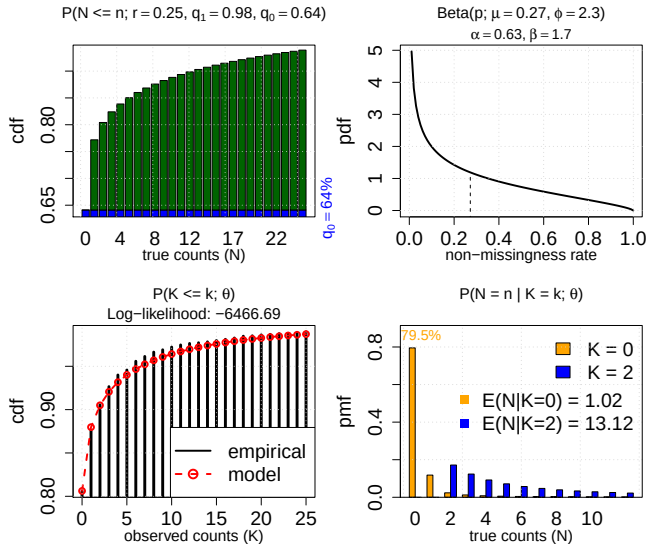

FIGURE 1. Model estimates for: (top left) true counts N_i; (top right) non-missing rate p_i; (bottom left) empirical count frequency and model fit; (bottom right) conditional predictive distributions and expectations.

3.1 Fix expected non-missing rate μ

Usually, researchers must estimate all 5 parameters from panel data. For our application, though, we can estimate (and fix) the non-missing rate μ a-priori as we have access to internal YouTube log files.

Let $\bar{k}_{\tilde{W}} = \sum_{i=1}^{P} \tilde{w}_i k_i$ be the observed panel visits projected to the entire population. Analogously, let $\bar{N}_{\tilde{W}} = \sum_{i=1}^{P} \tilde{w}_i N_i$ be the panel projections of the number of true YouTube visits. While any single N_i is unobservable, we can estimate $\bar{N}_{\tilde{W}}$ by simply counting all YouTube homepage views in Germany from our YouTube log files, yielding $\widehat{\bar{N}}_{\tilde{W}}$. We herewith obtain a plug-in estimate of the non-missing rate, $\widehat{\mu}_{Logs} = \bar{k}_{\tilde{W}} / \widehat{\bar{N}}_{\tilde{W}}$. The remaining 4 parameters, $\theta_{(-\mu)} = (\phi, q_0, r, q_1)$, can be obtained by MLE, $\widehat{\theta}_{(-\mu)} = \arg\max_{\theta_{(-\mu)}} \ell((\widehat{\mu}_{Logs}, \theta_{(-\mu)}); \mathbf{x})$. The overall estimate is $\widehat{\theta} = (\widehat{\mu}_{Logs}, \widehat{\theta}_{(-\mu)})$.

4 Estimating YouTube audience in Germany

Here we use data from a German online panel (GfK Consumer Panels, 2013), which monitors web usage of $P = 6,545$ individuals in October, 2013 (31 days). In particular, we are interested in the probability that an adult in Germany visited the YouTube homepage (`www.youtube.de`). Empirically, $\widehat{\mathbb{P}}(K = 0) = 0.81$, yielding 19% observed 1+ reach. However, we know by

comparison to YouTube log files that the panel only recorded 27.2% of all impressions. We fix the expected non-missing rate at $\widehat{\mu} = 0.272$ and obtain the remaining parameters via MLE (Table 1): Figure 1 shows the model fit for the true, observed, and predictive distribution. In particular, the true 1+ reach is 36% ($\widehat{q}_0 = 0.64$), not 19% as the naïve estimate suggests.

5 Discussion

We introduce a probabilistic framework to impute missing events in count data, including a hurdle component for more flexibility to model lots of zeros. Researchers can use our models to obtain accurate probabilistic predictions of the number of true, unobserved events. We apply our methodology to accurately estimate how many people in Germany visit YouTube.

Acknowledgments: We want to thank Christoph Best, Penny Chu, Tony Fagan, Oli Gaymond, Daniel Meyer, Simon Morris, Raimundo Mirisola, Andras Orban, Simon Rowe, Sheethal Shobowale, Yunting Sun, Wiesner Vos, Xiaojing Wang, and Fan Zhang for constructive discussions and feedback.

References

Fader, P. and Hardie, B. (2000). A note on modelling underreported Poisson counts. *Journal of Applied Statistics*, **27**, 953 – 964.

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

GfK Consumer Panels (2013). *Media Efficiency Panel*.

Hofler, R.A. and Scrogin, D. (2008). *A count data frontier model*. Technical report, University of Central Florida.

Hu, M., Pavlicova, M., and Nunes, E. (2011). Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. *American Journal of Drug and Alcohol Abuse*, **37**, 367 – 75.

Yang, S., Zhao, Y., and Dhar, R. (2010). Modeling the underreporting bias in panel survey data. *Marketing Science*, **29**, 525 – 539.

Regularization in Cox frailty models

Andreas Groll[1], Trevor Hastie[2], Gerhard Tutz[3]

[1] Ludwig-Maximilians-Universität Munich, Department of Mathematics, Theresienstraße 39, 80333 Munich, Germany
[2] University of Stanford, Department of Statistics, 390 Serra Mall, Sequoia Hall, California 94305, United States
[3] Ludwig-Maximilians-Universität Munich, Department of Statistics, Akademiestraße 1, 80799 Munich, Germany

E-mail for correspondence: groll@math.lmu.de

Abstract: In all sorts of regression problems it has become more and more relevant to face high dimensional data with lots of potentially influential covariates. A possible solution is to apply estimation methods that allow to select the relevant covariates. These methods are often based on suitable penalization of the corresponding regression models likelihood function. In this work, a penalization approach for variable selection in a particular regression model for survival analysis is considered, the so-called Cox frailty model. As in many applications the influence of some covariates changes over time, also time-varying effects are considered. A suitable penalization approach then has to cover several model selection issues. Besides, the method incorporates a multiplicative log-normal frailty distribution, resulting in flexible and sparse hazards models for modeling survival data.

Keywords: Variable selection; LASSO; Cox frailty model; Time-varying coefficients; Penalization.

1 Introduction

Among the class of models designed for continuous event times, the proportional hazards models play a major role, in particular the famous Cox model (Cox, 1972). The Cox model assumes the semi-parametric hazard

$$\lambda_i(t|\boldsymbol{x_i}) = \lambda_0(t)\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}),$$

where $\lambda_i(t|\boldsymbol{x}_i)$ is the hazard for observation i at time t, conditionally on the covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir})^{\mathrm{T}}$. $\lambda_0(t)$ is the shared baseline hazard, and

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

$\boldsymbol{\beta}$ the fixed effects vector. Note that in the continuous time case the hazard rate $\lambda_i(t|\boldsymbol{x})$ is defined as $\lambda_i(t|\boldsymbol{x}) = \lim_{\Delta t \to 0} P(t \leq T < \Delta t | T \geq t, \boldsymbol{x})/\Delta t$, representing the instantaneous risk of a transition at time t. Inference is usually based on maximization of the corresponding partial likelihood. This approach allows estimation of $\boldsymbol{\beta}$ while ignoring $\lambda_0(t)$ and performs well in classical problems with more observations than predictors. To combat the $p > n$ problem, Tibshirani (1997) proposed the use of the so-called least absolute shrinkage and selection operator (LASSO) penalty in the Cox model. Since then, several extensions have been proposed, compare Park and Hastie (2007) or Goeman (2010), just to mention two.

2 Cox frailty model with time-varying coefficients

If dependencies within clusters of observations exist or if there is heterogeneity between lusters, these can be captured effectively by frailty models. However, parameter estimation in frailty models is more challenging than in the Cox model, since the corresponding profile likelihood does not have a closed form solution. In the Cox proportional hazards frailty model the hazard rate of the jth subject belonging to subgroup-cluster i, conditionally on the covariates $\boldsymbol{x}_{ij}$ and the shared frailty u_i, is given by

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}, u_i) = u_i \lambda_0(t) \exp(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}),$$

where the frailties u_i, $i = 1, \ldots, n$, are frequently assumed to follow a gamma distribution because of its mathematical convenience.
While there exist several R packages to fit Cox frailty models, for example `frailtypack` (Rondeau et al., 2012) and `survival` (Therneau, 2014), only limited approaches to variable selection are yet available. Though Fan and Li (2002) as well as Androulakis et al. (2012) have contributed considerable works in this context, no software implementation is yet available.
While some multiplicative frailty distributions, such as e.g. the gamma and the inverse Gaussian, have already been extensively studied (cp. Androulakis et al., 2012) and closed form representations of the log-likelihoods are available, in some situations the log-normal distribution is more intuitive and besides, it generally allows for more flexible and complex predictor structures though the corresponding model is computationally more demanding. The conditional hazard function with multiplicative frailties following a multivariate log-normal distribution yields in its general form

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_i) = \lambda_0(t) \exp(\boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathrm{T}} \boldsymbol{u}_i),$$

where the random effects follow a multivariate Gaussian distribution, i.e. $\boldsymbol{u}_i \sim N(\mathbf{0}, \boldsymbol{D}(\boldsymbol{\theta}))$, with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}(\boldsymbol{\theta})$, which is depending on a vector of unknown parameters $\boldsymbol{\theta}$. In this case, a penalized quasi-likelihood (PQL) approach based on Laplace approximation can be

used for estimation, following Breslow and Clayton (1993) in their approach for the generalized linear mixed model (GLMM). In this context, it is especially important to provide effective estimation algorithms, as standard procedures for determination of tuning parameters such as cross validation are usually very time-consuming.

While for Cox frailty models with the simple predictor structure $\eta_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{u}_i$ in the hazard function some solutions have already been given (see e.g. Fan and Li, 2002; and Androulakis et al., 2012), often more complex structures of the linear predictor need to be taken into account. For example, time-varying effects $\gamma_k(t)$ can be incorporated into the linear predictor. For observation i from cluster j, this yields the hazard rate

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_i) = \lambda_0(t) \exp\left(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=1}^{s} z_{ijk}\gamma_k(t) + \boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{u}_i\right).$$

A standard way to estimate the time-varying effects $\gamma_k(t)$ is to expand them in equally spaced B-splines yielding $\gamma_k(t) = \sum_{m=1}^{M_k} \alpha_{k,m} B_{k,m}(t; d)$, where $\alpha_{k,m}$, $m = 1, \ldots, M_k$, denote unknown spline coefficients, which need to be estimated, and $B_{k,m}(t; d)$ is the mth B-spline basis function of the kth time-varying effect of degree d. For a detailed description of B-splines, see for example Wood (2006) and Ruppert et al. (2003).

In general, for the cumulative baseline hazard $\Lambda_0(\cdot)$ often the "least informative" nonparametric modeling is considered. More precisely, with $t_1^0 < \cdots < t_N^0$ denoting the observed event times, the least informative nonparametric cumulative baseline hazard $\Lambda_0(t)$ has a possible jump h_j at every observed event time t_j^0, i.e. $\Lambda_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$. However, the estimation procedure may be stabilized, if instead, similar to the time-varying effects, a semi-parametric baseline hazard is considered, which can be flexibly estimated within the B-spline concept. Then, using the transformation $\gamma_0(t) := \log(\lambda_0(t))$ and setting $z_{ij0} = 1$ for all i, j, we can specify the hazard rate as

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_i) = \exp\left(\eta_{ij}(t)\right), \tag{1}$$

with $\eta_{ij}(t) := \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=0}^{s} z_{ijk}\left(\sum_{m=1}^{M_k} \alpha_{k,m} B_{k,m}(t; d)\right) + \boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{u}_i$. In general, the estimation of parameters in the predictor (1) can be based on Cox's well-known full log-likelihood, which is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i} \delta_{ij}\eta_{ij}(t_{ij}) - \int_0^{t_{ij}} \exp(\eta_{ij}(s))ds, \tag{2}$$

where n denotes the number of clusters, N_i the cluster sizes and the survival times t_{ij} being complete if $\delta_{ij} = 1$ and right censored if $\delta_{ij} = 0$.

3 Penalization

Note that certain questions of model selection are related to the type of predictor (1). In particular, one has to determine which covariates should be included in the model, or, which of the covariates included have a time-varying effect. So our objective is to develop a penalization approach for variable selection in Cox frailty models with time-varying coefficients, such that single varying effects are either included, are included in the form of a constant effect or are totally excluded. These model selection issues can be achieved by incorporating a suitable penalty into the fitting procedure. We propose to subtract the following penalty from the Cox frailty log-likelihood

$$\xi \cdot J_\zeta(\boldsymbol{\alpha}) = \xi\Big(\zeta \sum_{k=1}^{s} \psi_k ||(\vartheta_{k,2}, \ldots, \vartheta_{k,M_k})||_2 + (1-\zeta) \sum_{k=1}^{s} \phi_k ||(\alpha_{k,1}, \ldots, \alpha_{k,M_k})||_2\Big),$$

where $||\cdot||_2$ denotes the L_2-norm, $\xi \geq 0$ and $\zeta \in (0,1)$ are tuning parameters and $\vartheta_{k,l} = \alpha_{k,l} - \alpha_{k,l-1}$. The first term of the penalty controls the smoothness of the time-varying covariate effects, whereby for values of ξ and ζ large enough, all differences $\alpha_{k,l} - \alpha_{k,l-1}$, $l = 2, \ldots, M_k$, are removed from the model, resulting in constant covariate effects. As the B-splines of each varying coefficient sum up to one, a constant effect is obtained, if all spline coefficients are equal. Hence, the first penalty term does not affect the spline's global level. The second term penalizes all spline coefficients belonging to a single time-varying effect in the way of a group LASSO and, hence, controls selection of covariates. Both tuning parameters ξ, ζ should be chosen by an appropriate technique, such as for example by K-fold cross validation. The terms $\psi_k := \sqrt{M_k - 1}$ and $\phi_k := \sqrt{M_k}$ represent weights assigning different amounts of penalization to different parameter groups, relative to the respective group size. Within the estimation procedure, i.e. the corresponding Newton-Raphson algorithm, local quadratic approximations of the penalty term are used, following Oelker and Tutz (2013). Note that the penalty from above may be easily extended by a conventional LASSO penalty for time-constant fixed effects β_k, $k = 1, \ldots, r$.

Besides, as also the baseline hazard in the predictor (1) is considered to be semi-parametric, the penalty from above should be further extended by adding another penalty term to control the roughness of the baseline. If the smooth baseline hazard $\lambda_0(t)$ is twice differentiable, one could for example penalize its second order derivatives, similar to Yu et al. (2012). Alternatively, if $\lambda_0(t)$ is once again expanded in B-spline basis functions, i.e. $\lambda_0(t) = \sum_{m=1}^{M_0} \alpha_{0,m} B_{0,m}(t; d)$, simply the squared differences of adjacent spline weights $\alpha_{0,l}$ and $\alpha_{0,l-1}$, $l = 2, \ldots, M_0$, could be penalized. Hence, beside $\xi \cdot J_\zeta(\boldsymbol{\alpha})$, also the penalty term

$$\xi_0 \cdot J_0(\boldsymbol{\alpha_0}) = \xi_0\Big(\sum_{l=2}^{M_0} (\alpha_{0,l} - \alpha_{0,l-1})^2\Big)$$

has to be subtracted from the Cox frailty log-likelihood, with the vector $\boldsymbol{\alpha_0}^{\mathrm{T}} := (\alpha_{0,1}, \ldots, \alpha_{0,M_0})$ collecting those spline coefficients from $\boldsymbol{\alpha}$ that correspond to the baseline hazard. Although this adds another tuning parameter ξ_0, it turns out that in general it is not worthwhile to select also ξ_0 on a grid of possible values. Note here that we have already obtained similar findings with regard to penalization of the baseline hazard in discrete frailty survival models, see Tutz and Groll (2014). While probably some care should be taken to select ξ and ζ, which determine the performance of the selection procedure, the estimation procedure is already stabilized in comparison to the usage of the least informative nonparametric cumulative baseline hazard $\Lambda_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$ for a moderate choice of ξ_0.
As already mentioned in Section 2, a possible strategy to maximize the full log-likelihood (2) is based on the PQL approach, which was originally suggested for GLMMs by Breslow and Clayton (1993). Typically, the covariance matrix $\boldsymbol{D}(\boldsymbol{\theta})$ of the random effects $\boldsymbol{u}_i$ depends on an unknown parameter vector $\boldsymbol{\theta}$. Hence, the joint likelihood-function can be specified by the parameter vector of the covariance structure $\boldsymbol{\theta}$ and parameter vector $\boldsymbol{\delta}^{\mathrm{T}} := (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{u}^{\mathrm{T}})$. The corresponding *marginal* log-likelihood then yields

$$l^{mar}(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left(\int L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}_i) p(\boldsymbol{u}_i | \boldsymbol{\theta}) d\boldsymbol{u}_i \right),$$

where $p(\boldsymbol{u}_i|\boldsymbol{\theta})$ denotes the density function of the random effects and the quantities $L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}_i) := \prod_{j=1}^{N_i} \exp(\eta_{ij}(t_{ij}))^{\delta_{ij}} \exp\left(- \int_0^{t_{ij}} \exp(\eta_{ij}(s)) ds \right)$ represent the likelihood contributions of single clusters $i, i = 1, \ldots, n$. Approximation along the lines of Breslow and Clayton (1993) yields

$$\begin{aligned} l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \log L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{u}_i) - \frac{1}{2} \boldsymbol{u}^{\mathrm{T}} \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{u} \qquad (3) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{N_i} \left(\delta_{ij} \eta_{ij}(t_{ij}) - \int_0^{t_{ij}} \exp(\eta_{ij}(s)) ds \right) - \frac{1}{2} \boldsymbol{u}^{\mathrm{T}} \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{u}, \end{aligned}$$

the penalty term $\boldsymbol{u}^{\mathrm{T}} \boldsymbol{D}(\boldsymbol{\theta}) \boldsymbol{u}$ resulting from the approximation based on the Laplace method. The PQL approach usually works within the profile likelihood concept. It is distinguished between estimation of $\boldsymbol{\delta}$, given the plug-in estimate $\hat{\boldsymbol{\theta}}$ and resulting in profile likelihood $l^{app}(\boldsymbol{\delta}, \hat{\boldsymbol{\theta}})$, and estimation of $\boldsymbol{\theta}$.

4 Estimation

Estimation is now based on maximization of the penalized log-likelihood, which is obtained by expanding the approximate log-likelihood $l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta})$ from (3) to include the penalty terms $\xi_0 \cdot J_0(\boldsymbol{\alpha_0})$ and $\xi \cdot J_\zeta(\boldsymbol{\alpha})$, i.e.

$$l^{pen}(\boldsymbol{\delta}, \boldsymbol{\theta}) = l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta}) - \xi_0 \cdot J_0(\boldsymbol{\alpha_0}) - \xi \cdot J_\zeta(\boldsymbol{\alpha}).$$

The estimation procedure is based on a conventional Newton-Raphson algorithm, while local quadratic approximations of the penalty term are used, following Oelker and Tutz (2013).
It turns out that the combination of the proposed penalization approach for variable selection in Cox frailty models with time-varying coefficients with the promising class of multivariate log-normal frailties results in very flexible and sparse hazard rate models for modeling survival data.

References

Androulakis, E., Koukouvinos, C., and Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine*, **31**, 2223 – 2239.

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

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

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74 – 99.

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

Oelker, M.-R. and Tutz, G. (2013). A General Family of Penalties for Combining Different Types of Penalties in Generalized Structured Models. *Technical Report*, **139**, Department of Statistics, LMU Munich.

Park, M.Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B*, **19**, 659 – 677.

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

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385 – 395.

Tutz, G. and Groll, A. (2014). Variable Selection in Discrete Survival Models Including Heterogeneity. *Technical Report*, **167**, Department of Statistics, LMU Munich.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

Yu, Z., Lin, X., and Tu, W. (2012). Semiparametric frailty models for clustered failure time data. *Biometrics*, **68)**, 429 – 436.

Network reconstruction with realistic models

Marco Grzegorczyk[1], Andrej Aderhold[2], Dirk Husmeier[2]

[1] Johann Bernoulli Institute, Groningen University, The Netherlands
[2] School of Mathematics and Statistics, Glasgow University, UK

E-mail for correspondence: m.a.grzegorczyk@rug.nl

Abstract: We extend a recently proposed gradient-matching method for inferring interactions in complex systems described by differential equations in various respects: improved gradient inference, evaluation of the influence of the prior on kinetic parameters, comparative evaluation of two model selection paradigms: marginal likelihood versus DIC (divergence information criterion), comparative evaluation of different numerical procedures for computing the marginal likelihood, extension of the methodology from protein phosphorylation to transcriptional regulation, based on a realistic simulation of the underlying molecular processes with Markov jump processes.

Keywords: Gene networks; Semi-mechanistic models; Bayesian model selection.

1 Introduction

A challenging problem for computational statistics is to infer the structure of regulatory networks from postgenomic data. Two approaches can be distinguished. The first paradigm aims to apply generic models like Bayesian networks. The second paradigm is based on mechanistic models and the detailed mathematical description of the underlying interactions with differential equations (DEs). The advantage of this paradigm is a more faithful representation of the interactions. The disadvantage are the substantially higher computational costs of inference. A novel approach, presented by Oates et al. (2014) and termed "chemical model averaging" (CheMa), aims for a compromise between the two paradigms by gradient matching. Given the concentration time series of some quantities ("species") whose interactions are to be inferred, the temporal derivatives of the concentrations are estimated from the data. These derivatives are then matched against those predicted from the DEs by standard statistical techniques.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

The resulting "semi-mechanistic" model is effectively a non-linear regression model, whose computational complexity of inference sits between the two paradigms. We take the work of Oates et al. (2014) further in five respects. We improve the gradient matching by adopting techniques from nonparametric Bayesian statistics with Gaussian processes. We assess the influence of the parameter prior. We carry out a comparative evaluation study to assess two model selection paradigms – the marginal likelihood versus the divergence information criterion. We adapt the method to a new type of application, namely to transcriptional regulation from gene expression data. We evaluate the method in a realistic simulation study based on a stochastic process description of the underlying molecular processes.

2 Methods

A biopathway can be modelled as a system of ordinary DEs:

$$\frac{dx_i}{dt}|_{t=t_j} = c_i - v_{0,i}x_i(t_j) + f_i(\mathbf{x}_i(t_j), \boldsymbol{\theta}), \tag{1}$$

where i is one of N species, $x_i(t_j)$ is the concentration of i at time point t_j $(j = 1, \ldots, T)$, c_i is a baseline production rate, $v_{0,i}$ is a decay rate, $f_i(\cdot)$ is a regulation function, $\boldsymbol{\theta}$ are parameters, and $\mathbf{x}_i(t_j)$ is a vector of concentrations of species that regulate species i. We follow Oates et al. (2014) and estimate the time derivatives $\frac{dx_i}{dt}|_{t=t_j}$ from the observed data D, and treat the problem as non-linear regression with the likelihood:

$$p(D|\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{j=1}^{T} \mathcal{N}(\xi_i(t_j)|f_i(\mathbf{x}_i(t_j), \boldsymbol{\theta}) - v_{0,i}x_i(t_j), \sigma_i^2), \tag{2}$$

where $\xi_i(t_j) = \frac{dx_i}{dt}|_{t=t_j}$, and $\mathcal{N}(\cdot|\mu, \sigma^2)$ is the PDF of a normal distribution. Oates et al. (2014) obtained the temporal derivatives $\xi(t_j)$ with a finite difference quotient (**"numerical gradient"**). We here propose to apply a Gaussian process to smooth interpolation (**"analytical gradient"**). The superiority of this approach was recently established by Aderhold et al. (2014). In the CheMA approach (1) is implemented with $c_i = 0$, and $f_i(\cdot)$ describes Michaelis-Menten kinetics:

$$\xi_i(t_j) = \frac{dx_i}{dt}|_{t=t_j} = -v_{0,i}x_i(t_j) + \sum_{j\in\pi_i} v_{j,i}\frac{I_{j,i}\cdot x_j(t_j) + (1-I_{j,i})\cdot k_{j,i}}{x_j(t_j) + k_{j,i}}, \tag{3}$$

where the sum is over all species $j \in \pi_i$ that are regulators of i. The indicator function $I_{j,i}$ indicates whether species j is an activator or inhibitor. The term $-v_{0,i}x_i(t)$ takes the degradation of $x_i(t)$ into account, while the parameters $v_{j,i}$ and $k_{j,i}$ are the "*maximum reaction rate*" and "*Michaelis-Menten*" parameters. Oates et al. (2004) impose truncated Normal distributions on $k_{j,i}$, $k_{j,i} \sim \mathcal{N}_{\{k_{j,i}\geq 0\}}(1, \nu)$, where $\nu > 0$, use Jeffrey's prior

for σ_i^2, and a truncated **g-prior** on the vector $\mathbf{V}_i$ of maximum reaction rate parameters: $\mathbf{V}_i \sim \mathcal{N}_{\{\mathbf{V}_i \geq 0\}}(\mathbf{1}, T\sigma_i^2(\mathbf{D}_i^{\mathrm{T}}\mathbf{D}_i)^{-1})$, where $\mathbf{D}_i$ is the design-matrix for species i. We show that a truncated **ridge regression prior**: $\mathbf{V}_i \sim \mathcal{N}_{\{\mathbf{V}_i \geq 0\}}(\mathbf{1}, \delta_i^2\sigma_i^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix, and δ_i^2 has an inverse Gamma prior, $\delta_i^2 \sim IG(a_\delta, b_\delta)$, yields a better network reconstruction accuracy. We refer to these methods as **CheMa** (with the g-prior on $\mathbf{V}_i$) and **iCheMa** (improved CheMa with the ridge regression prior on $\mathbf{V}_i$).
To infer the regulator sets π_i of the interaction processes described by (1), we compare the divergence information criterion (DIC) and the marginal likelihood (MLL). DIC is defined as

$$DIC(\pi_i) = 2\log p(D|\overline{\boldsymbol{\theta}}, \pi_i) - 4\int \log p(D|\boldsymbol{\theta}, \pi_i)p(\boldsymbol{\theta}|\pi_i, D)d\boldsymbol{\theta},$$

where $\overline{\boldsymbol{\theta}} = \int \boldsymbol{\theta}p(\boldsymbol{\theta}|\pi_i, D)d\boldsymbol{\theta}$ is the posterior mean of the parameters. The integrals are approximated by sums over parameters sampled from the posterior distribution $p(\boldsymbol{\theta}|G, D)$ with MCMC. The marginal likelihood is

$$p(D|\pi_i) = \int p(D|\boldsymbol{\theta}, \pi_i)p(\boldsymbol{\theta}|\pi_i)d\boldsymbol{\theta}. \tag{4}$$

We compare two methods for approximating (4): Chib's method and thermodynamic integration (TI). **Chib's method** is based on

$$p(D|\pi_i) = \frac{p(D|\boldsymbol{\theta}^\star, G)p(\boldsymbol{\theta}^\star|\pi_i)}{p(\boldsymbol{\theta}^\star|\pi_i, D)}, \tag{5}$$

where the posterior near a selected parameters $\boldsymbol{\theta}^\star$, $p(\boldsymbol{\theta}^\star|G, D)$, is approximated with MCMC. TI is based on the power posteriors:

$$p(\boldsymbol{\theta}|\pi_i, D, \tau) = \frac{p(D|\boldsymbol{\theta}, \pi_i)^\tau p(\boldsymbol{\theta}|\pi_i)}{\int p(D|\boldsymbol{\theta}', \pi_i)^\tau p(\boldsymbol{\theta}'|\pi_i)d\boldsymbol{\theta}'} \tag{6}$$

from which the marginal likelihood is computed via

$$p(D|\pi_i) = \int_0^1 E_{\boldsymbol{\theta},\tau}\big[\log p(D|\boldsymbol{\theta}, \pi_i)\big]d\tau. \tag{7}$$

Here $E_{\boldsymbol{\theta},\tau}[\cdot]$ is an expectation w.r.t. the power posterior in (6). These expectations are computed for various temperatures $0 \leq \tau \leq 1$ with population MCMC, and the integral in (7) is then approximated with the trapezium sum. We choose 10 temperatures $\tau_i = (\frac{i}{9})^m$ $(0 \leq i \leq 9)$, and we vary the exponent $m \in \{4, 8\}$ to obtain TI-4 and TI-8.

3 Data

We generate $T = 240$ data points for four species $x_1, \ldots, x_4$ from iid $\mathcal{N}(0, 1)$ distributions. Subsequently, to obtain non-negative concentrations,

the observations of each individual species are shifted such that the lowest value is equal to 0, before we re-scale the observations of each species to mean 1. With x_1 taking the role of the degradation process and x_2 being an activating regulator ($I_{2,y} = 0$) of a gradient ξ_y, which we here assume to be directly observable, we generate target observations with (3): $\xi_y(t_j) = -v_{0,y}x_1(t_j) + v_{2,y}\frac{x_2(t_j)}{x_2(t_j)+k_{2,y}} + \epsilon_{t_j}$, where $\epsilon_{t_j} \sim \mathcal{N}(0, \sigma^2)$ is additive iid Gaussian noise. We keep $k_{2,y} = 1$ fixed, and vary the rates $(v_{0,y}, v_{2,y}) = \{(1,1),(0.5,1),(1.5,1),(2,1),(0.2,1),(2,0.2),(3,0.1),(0.2,2),(0.1,2)\}$. Our goal is to infer the true regulator set $\pi_y = \{x_2\}$ out of all subsets of $\{x_2, x_3, x_4\}$. The degradation, modelled via x_1, is included in all models. We also use the benchmark data from Aderhold et al. (2014), which contain realistically simulated gene expression time series for genes in the circadian clock of *Arabidopsis thaliana*. We focus on the mRNA data points and those time series generated for the wildtype circadian gene network, shown in the left panel of Figure 2. The molecular interactions were modelled as individual discrete events with a Markov jump process and practically simulated with the Biopepa software, see Aderhold et al. (2014) for details. Finally, we apply the improved CheMa model to Arabidopsis gene expression data which were recently measured under the EU-FP7-funded Timing Metabolism (TiMet) research project. For space restrictions the latter results are not shown in this paper.

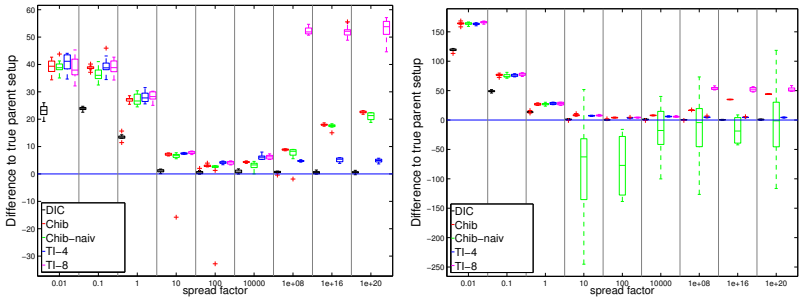

FIGURE 1. **Performance of iCheMa on synthetic data** The box plots show the log marginal likelihood differences between the true regulator set $\pi_y = \{x_2\}$ and an alternative over-complex set that includes one irrelevant regulator. The panels correspond to different prior distributions of the parameters. **Left panel**: the prior variance of $k_{j,i}$ was kept fixed at $\nu = 0.5$, and the hyperparameter δ_i^2 of the truncated ridge regression prior on $\mathbf{V}_i$ was set to *spread factor*. **Right panel**: both hyperparameters δ_i^2 and ν were set to *spread factor*. The box plots show the distributions of the average MLL differences across 9 parameter pairs $(v_{0,y}, v_{2,y})$, computed for 10 independent data instantiations. Positive differences indicate that the true model is favoured. Each panel shows: (i) the DIC difference (DIC), the MLL differences, approximated with (ii) a naive implementation of Chib's method (Chib-naiv), (iii) an improved implementation of Chib's method (Chib), (iv) TI with $m = 4$ (TI-4), and (v) TI with $m = 8$ (TI-8).

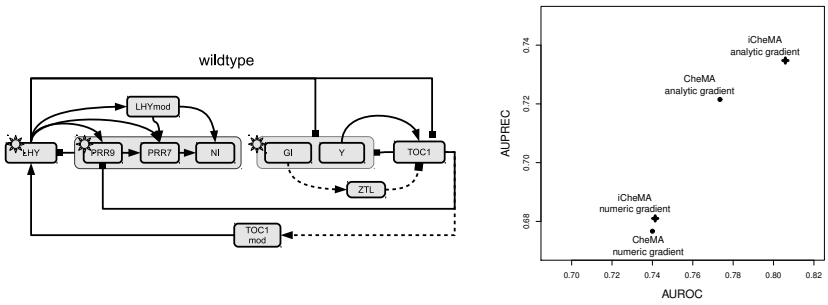

FIGURE 2. Network reconstruction accuracy for the realistic wildtype Arabdidopsis gene network, based on marginal interaction probabilities ("model averaging"). *Left panel:* Hypothetical circadian clock network in *A.thaliana* from Pokhilko et al. (2010). *Right panel:* Mean AUROC and AUPREC scores to quantify the effects of the prior ('CheMa vs. iCheMa') and the gradient type ('numerical vs. analytical') on the network reconstruction accuracy.

4 Results

Synthetic Data: Figure 1 shows the DIC and log marginal likelihood (MLL) differences between the true and an over-complex regulator set for different parameter priors. Each parameter prior was a Gaussian centred on $\mu = 1$, with different variances. For low variances, both DIC and MLL clearly favour the true network, because the prior 'pulls' the spurious interaction parameter from its true value of zero towards a wrong value of $\mu = 1$. As the prior becomes more diffuse, both the DIC and MLL differences become less pronounced, but still select the true model up to spread factors of about 100. As the prior becomes more diffuse, with the spread factor exceeding 100, DIC fails to select the correct model. MLL, on the other hand, starts to increasingly favour the true model as the spread factor further increases beyond 1000. This is a consequence of Lindley's paradox, whereby MLL increasingly penalizes the over-complex model for increasingly vague priors. The left panel of Figure 1 shows that the different ways of computing the MLL give very similar results up to a prior spread factor of about 1e+08. For spread factors exceeding this value, the results differ. The MLL computed with Chib's method monotonically increases, as expected from Lindley's paradox. MLL computed with TI reaches a plateau, with different values obtained for different temperature schemes ($m = 4$ and $m = 8$). This is a numerical discretization error that results from the form of the integrand in (7), which has most of its area concentrated on values near $\tau = 0$. The right panel shows that a naive implementation of Chib's method can run into numerical instabilities, as diffuse prior distributions can yield suboptimal attractor states in parameter configuration

space. We fixed this instability by selecting exclusively pivot parameter vectors $\boldsymbol{\theta}^{\star}$ that are representative for the sampled parameter values. (We never observed unstable results when selecting the parameter vector $\boldsymbol{\theta}^{\star}$ with the highest posterior probability from only the sample phase of the MCMC simulation, rather than the total MCMC trajectory with the burn-in phase included.) With this numerical stabilization, Chib's method, whose numerical complexity compared to TI is lower by a factor of about 10 (as we use $K = 10$ temperatures τ_i for TI), is the favourite method for the realistic network data.

Realistic Network Data: The right panel of Figure 2 shows average AUROC (area under the ROC curve) and AUPREC (area under the precision recall curve) scores, obtained for five realistically simulated time series of the wildtype circadian clock network (shown in the left panel). The proposed analytical gradient (based on a Gaussian process) yields a significantly improved network reconstruction accuracy for both models: (i) CheMa with the truncated g-prior on $\mathbf{V}_i$, as proposed by Oates et al. (2014) and (ii) iCheMa with the truncated ridge regression prior, introduced here. It can also be seen that iCheMa has a slightly better network prediction accuracy than the original CheMa method given the numerical gradient, and a substantially better accuracy for the analytic gradient.

5 Further results

At IWSM 2015 more results will be presented. For space restrictions some results could not be included in this paper. Most importantly, we will (i) show that the iCheMa model is superior to established network reconstruction methods, such as Hierarchical Bayesian regression, Sparse regression with L_1 penalty (Lasso), Sparse regression with L_1 and L_2 penalty (ElasticNet), Sparse regression with change-points (Tesla), Sparse Bayesian regression, Graphical Gaussian models, Bayesian spline autoregression, State-space models, Gaussian processes, and Bayesian networks, and we will (ii) apply the novel iCheMa method to reverse-engineer the circadian clock network in *A. thaliana* from TiMet gene expression time series.

References

Aderhold, A., Husmeier, D., and Grzegorczyk, M. (2014). Statistical inference of regulatory networks for circadian regulation. *Statistical applications in genetics and molecular biology*, **13**, 227 – 273.

Oates, C.J., Dondelinger, F., Bayani, N., Korkola, J., Gray, J.W., and Mukherjee, S. (2014) Causal network inference using biochemical kinetics. *Bioinformatics*, **30**, 468 – 474.

Pokhilko, A., Hodge, S.K., Stratford, K., Knox, K., Edwards, K.D., Thomson, A.W., Mizuno, T., and Millar, A.J. (2010). Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular Systems Biology*, **6**, online article.

Dispersion modelling for Poisson-inverse Gaussian regression

Gillian Heller[1], Dominique-Laurent Couturier[2], Stephane Heritier[3]

[1] Department of Statistics, Macquarie University, Sydney, Australia
[2] Department of Public Health and Primary Care, University of Cambridge, UK
[3] School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

E-mail for correspondence: gillian.heller@mq.edu.au

Abstract: In clinical trials traditionally the effect of a treatment on the mean of a response of interest is hypothesised. The underlying assumption in statistical models for the mean, is that the effect of the treatment on the response distribution is a location shift, with other aspects of the distribution (shape/dispersion/variance) remaining the same. We consider data from a trial in which the response was a count and the treatment was hypothesised to reduce the mean. Inspection of the data revealed that the Poisson-inverse Gaussian (PIG) distribution was appropriate, and that the treatment reduced not only the mean, but also the variability, substantially. The conventional analysis hypothesises a treatment effect on the mean, either adjusted or unadjusted for covariates, while assuming a constant dispersion parameter. On our data, this analysis yields a non-significant treatment effect. We show that if we model a treatment effect on both the mean and dispersion parameters, both effects are highly significant. We show in a simulation study that if a treatment effect exists on the dispersion parameter and is ignored in the modelling, estimation of the treatment effect on the mean can be severely biased. We show further that if we use an orthogonal parametrization of the PIG distribution, estimates of the mean model are robust to misspecification of the dispersion model.

Keywords: Poisson-inverse Gaussian regression; Parameter orthogonality; gamlss; Dispersion modelling; Count data.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

TABLE 1. Summary statistics.

		Treatment	Control
n		105	92
Number of falls:	Mean	3.35	8.65
	Variance	62.0	1388.1
	Maximum	49	358

1 Introduction

This work was motivated by a clinical trial for a drug for the treatment of neurogenic orthostatic hypotension, a condition which affects Parkinson's Disease patients and which is a cause of falling (Rowse, Heritier and Hewitt, 2015). The drug was hypothesised to reduce the incidence of falling in these patients. Patients were randomised to receive the treatment or placebo, and the response was number of falls in the study period. Summary statistics are given in Table 1, which reveals a reduction in the number of falls in the treatment group, as well as a marked reduction in the variance of falls in that group. Candidate distributions for number of falls include the Poisson, and compound Poisson distributions: the negative binomial, Poisson-inverse Gaussian (PIG) and Sichel distributions. Of these, the PIG provided the best marginal fit, and as it was clear that dispersion of falls varied considerably with treatment group, a model for dispersion was considered appropriate.

Regression models which extend modelling to dispersion and other shape parameters, and to distributions beyond the exponential family, have become popular in the last decade (Rigby and Stasinopoulos, 2005; Kneib, 2013). The PIG, parametrized in terms of its mean μ and a dispersion parameter σ, is available as a response distribution in regression software (R package `gamlss`, Stasinopoulos and Rigby, 2007). However, in our study, estimates of the treatment effect on μ were found to be sensitive to specification of the model for σ. A regression model using an alternative parametrization of the PIG, in which the shape parameter (α) is orthogonal to the mean, was considered. For our study, the estimate of the treatment effect on the mean was robust to the model for α. In a simulation study it was confirmed that, for the (μ, σ) parametrization, estimates in the μ model can be severely biased if the σ model is misspecified. However, using the (μ, α) parametrization, μ model estimates are robust to misspecification of the α model. This potentially has implications for regression models for any response distribution, in which the shape parameter(s) are not orthogonal to the mean.

2 Poisson-inverse Gaussian regression

Using the GAMLSS parametrization (Rigby, Stasinopoulos, and Akantziliotou, 2008), the probability function of the PIG is

$$f(y|\mu,\sigma) = \sqrt{\frac{2}{\pi\sigma}}(1+2\mu\sigma)^{\frac{1}{4}} e^{\frac{1}{\sigma}} \frac{\left(\frac{\mu}{\sqrt{1+2\mu\sigma}}\right)^y}{y!} K_{y-0.5}\left(\frac{\sqrt{1+2\mu\sigma}}{\sigma}\right) \qquad (1)$$
$$y = 0, 1, 2, \ldots,$$

where $\mathrm{E}(Y) = \mu$, $\mathrm{Var}(Y) = \mu(1+\sigma\mu)$, and $K_\nu(\cdot)$ is a modified Bessel function of the third kind. The following two PIG regression models were fitted in `gamlss`, including and excluding a treatment effect on the dispersion parameter σ:

Model A (restricted)	Model B (full)
$\log\mu = \beta_0 + \beta_1 x + \log t$	$\log\mu = \beta_0 + \beta_1 x + \log t$
$\log\sigma = \gamma_0$	$\log\sigma = \gamma_0 + \gamma_1 x$

where x is an indicator for treatment and $\log t$ is an offset term for treatment duration t. In Model B, the treatment effect on σ, $\hat{\gamma}_1$, is highly significant ($p = 0.002$). β_1 is the parameter of interest, and in Model A it is non-significant ($p = 0.341$), whereas in Model B it is significant ($p = 0.014$). This difference is troublesome, as we are dealing with a clinical trial.

3 An orthogonal parametrization of the PIG

Stein et al. (1987) proposed an alternative parametrization of the PIG:

$$f(y|\mu,\alpha) = \sqrt{\frac{2\alpha}{\pi}} \exp\left(\sqrt{\mu^2+\alpha^2} - \mu\right) \frac{\left(\frac{\mu\left(\sqrt{\mu^2+\alpha^2}-\mu\right)}{\alpha}\right)^y}{y!} K_{y-0.5}(\alpha) \qquad (2)$$
$$y = 0, 1, 2, \ldots$$

having $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \mu\left(1 + \mu/\left((\mu^2+\alpha^2)^{0.5} - \mu\right)\right)$. This parametrization has the desirable property that μ and α are orthogonal:

$$\mathrm{E}\left(\frac{\partial^2}{\partial\mu\,\partial\alpha}\log f\right) = 0\ ,$$

i.e., the MLEs $\hat{\mu}$ and $\hat{\alpha}$ are asymptotically uncorrelated. We specify the models

Model C (restricted)	Model D (full)
$\log\mu = \beta_0 + \beta_1 x + \log t$	$\log\mu = \beta_0 + \beta_1 x + \log t$
$\log\alpha = \delta_0$	$\log\alpha = \delta_0 + \delta_1 x$

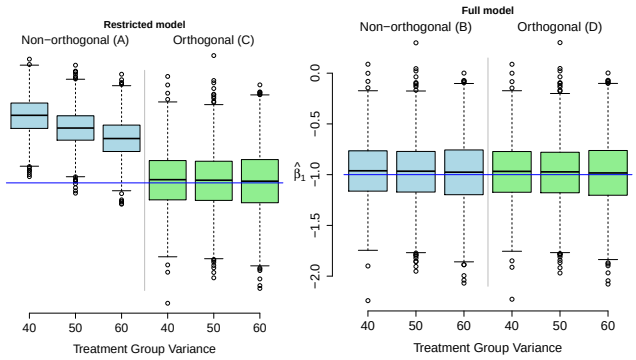

FIGURE 1. Simulation results: boxplots of $\hat{\beta}_1$ estimated under the restricted model (A and C); and the full model (B and D).

A consequence of μ and α being orthogonal, is that the elements of $\boldsymbol{\beta}$ are orthogonal to the elements of $\boldsymbol{\delta}$. MLEs were obtained using the `optim` function in R. The estimates of β_1 are almost the same under models C and D, indicating a strong treatment effect on the number of falls which is robust to specification of the α model. The estimated treatment effects in Models A and B are attenuated by comparison, and quite different from each other.

4 Simulation study

Our simulation study tests estimation under the two parametrizations of the PIG. In order not to bias the comparison towards either model, we specify the dispersion difference between groups on the scale of the variance. One thousand simulated samples of size $n = 125$ per group were generated under each of the the following three conditions, which were chosen to roughly replicate the falls data set: $\beta_1 = -1$; variance: control group 900; treatment group 40, 50, 60. Estimation was carried out under Models A and B for response distribution parametrized as (1); and under Models C and D for parametrization (2). Results are shown in Figure 1. When the full model is specified, $\hat{\beta}_1$ appears to be unbiased, regardless of the parametrization used. However, when the restricted model is specified, severe bias in $\hat{\beta}_1$ is observed when the non-orthogonal parametrization is used; when the orthogonal parametrization is used, $\hat{\beta}_1$ is robust to the model misspecification. In addition, dispersion in $\hat{\beta}_1$ does not appear to be increased by the model misspecification.

5 Discussion

We have demonstrated an important advantage in using an orthogonal parametrization of the PIG response distribution, viz. robustness of the MLEs in the μ model to the dispersion parameter model. This has particular importance in clinical trials, where a dispersion model is not in general specified. These results have implications for regression modelling using any response distribution in which the shape parameter is not orthogonal to the mean. In further work we will investigate inference properties of the MLEs obtained from this model.

References

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling*, **13**, 275 – 303.

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

Rigby, R., Stasinopoulos, D.M., and Akantziliotou, C. (2008). A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. *Computational Statistics & Data Analysis*, **53**, 381 – 393.

Rowse, G., Heritier, S., and Hewitt, L.A. (2015). Falls during a 10-week placebo-controlled study of droxidopa for neurogenic orthostatic hypotension in Parkinsons Disease. *67th American Academy of Neurology (AAN) 2015 Annual Meeting*, Washington D.C.

Stasinopoulos, D.M. and Rigby, R. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1 – 46.

Stein, G.Z., Zucchini, W., and Juritz, J.M. (1987). Parameter estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association*, **82**, 938 – 944.

Clusters with random size: maximum likelihood versus weighted estimation

Lisa Hermans[1], Geert Molenberghs[12], Michael G. Kenward[3], Wim Van der Elst[1], Vahid Nassiri[2], Marc Aerts[1], Geert Verbeke[12]

[1] I-BioStat, Universiteit Hasselt, Belgium
[2] I-BioStat, KU Leuven, Belgium
[3] Dept. of Medical Statistics, London School of Hygiene & Tropical Medicine, UK

E-mail for correspondence: lisa.hermans@uhasselt.be

Abstract: There are many contemporary designs that do not use a random sample of a fixed, *a priori* determined size. In case of informative cluster sizes, the cluster size is influenced by the the cluster's data, but here we cope with some issues that even occur when the cluster size and the data are unrelated. First, fitting models to clusters of varying sizes is often more complicated than when all cluster have the same size. Secondly, in such cases, there usually is no so-called complete sufficient statistic (Molenberghs et al., 2014).

Keywords: Likelihood inference; Pseudo-likelihood; Random cluster size.

1 Introduction

In applied statistics, situations exist where there is no fixed sample size. Molenberghs et al. (2014) provide an overview of various situations. Examples include: sequential trials, incomplete data, censored survival data, etc. Here we focus on hierarchical or clustered data. Random cluster sizes can occur for any outcome type, including continuous data, binary data, counts, and failure times. We will focus on cases where the cluster size is variable but independent of observed and unobserved outcomes. As a simple cluster paradigm, we consider the normal compound-symmetry (CS) model.
Molenberghs et al. (2011) introduced the split-sample methodology, i.e., a form of pseudo-likelihood where a sample is subdivided into subsamples. These subsamples are analyzed as if they were unrelated and afterwards the

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

results are appropriately averaged. There are many options for splitting the data, but here we use splitting along the cluster sizes. For the subsamples, closed-form estimators then exist, whereas they do not in the sample as a whole. A weighted combination of the subgroup-specific estimators is needed. However, Molenberghs et al. (2014) and Hermans et al. (2014) show that there may not be an optimal set of weights, resulting from calculations on incomplete sufficient statistics in the context of weighted averages.

2 Split-sample methods for clusters of variable size

2.1 Compound-symmetry model

Let $\boldsymbol{Y}$ be a vector of length n, following the compound-symmetry normal law $\boldsymbol{Y} \sim N(\mu \mathbf{1}_n, \sigma^2 \boldsymbol{I}_n + d\boldsymbol{J}_n)$. In general, both $\boldsymbol{Y}$ and n are random variables. Let there be a sample of N independent clusters, with K different cluster sizes n_k $(k = 1, \ldots, K)$ with multiplicity c_k. Denote the outcome vector for the ith $(i = 1, \ldots, c_k)$ replicate among the cluster of size n_k by $\boldsymbol{Y}_i^{(k)}$. For a sample with constant cluster size $(K = 1)$, compound-symmetry models allow closed-form solutions for the estimators. These sufficient statistics are complete, the estimator is unique minimum variance unbiased, and the mean parameter estimator and the variance parameter estimator are independent. Hermans et al. (2015) show, based on likelihood calculations, that in case $K \geq 2$ all these results disappear and there is no closed form solution. Likelihood calculations for K cluster sizes with common mean and variance parameter across all clusters, do not lead to explicit solutions, unless the variance components are known or the cluster size is constant. This suggest further study of weighted averages, e.g., of the form

$$\widetilde{\mu} = \sum_{k=1}^{K} a_k \widehat{\mu_k}, \qquad \widetilde{\sigma^2} = \sum_{k=1}^{K} b_k \widehat{\sigma_k^2}, \qquad \widetilde{d} = \sum_{k=1}^{K} g_k \widehat{d_k}, \tag{1}$$

where μ_k, σ_k^2, and d_k are the cluster-specific parameters. This idea is very similar to that in Molenberghs et al. (2011), who splits a sample in subsamples, that are analyzed seperately and than combined in an overall estimator.

2.2 Pseudo-likelihood for split samples

A pseudo-likelihood function is one that replaces a given likelihood function due to computational convenience. The likelihood contribution of a cluster is now a product of contributions for the various sub-vectors. Molenberghs et al. (2011) partitioned a sample in dependent or independent subsamples and used pseudo-likelihood for the fit. Referring to the compound symmetry

model described above, a pseudo-likelihood, for estimating a single vector (μ, σ^2, d) from a dataset divided into K subgroups, each containing c_k replicates, can be written as:

$$p\ell(\boldsymbol{\theta}) = \sum_{k=1}^{K} \ell(\boldsymbol{\theta_k} | \boldsymbol{y}_1^{(k)}, \ldots, \boldsymbol{y}_{c_k}^{(k)}), \tag{2}$$

with $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2, d_k)$. All $\boldsymbol{\theta}_k$ are assumed to be formally different, $\boldsymbol{\theta}$ stacks all vectors $\boldsymbol{\theta}_k$ and the parameter of interest $\boldsymbol{\theta}^*$, is found from an appropriate combination of the $\boldsymbol{\theta}_k$.

2.3 Weighting schemes

Referring to the setting in Section 2.1, note that subjects in different subsamples are allowed to have the same distribution, but that subjects in the same sub-sample must have the same distributions. Consider pseudo-likelihood in the general case (2). Assume that $\boldsymbol{\theta}^*$ is a vector of length p, and that each $\boldsymbol{\theta}_k$ is a separate copy of $\boldsymbol{\theta}^*$. Then $\boldsymbol{\theta}$ is a vector of length $K \cdot p$ and A is a $(p \times K \cdot p)$ matrix. The generic combination rules become:

$$\widetilde{\boldsymbol{\theta}}^* = \sum_{k=1}^{K} A_k \widehat{\boldsymbol{\theta}}_k, \qquad \text{var}(\widetilde{\boldsymbol{\theta}}^*) = \sum_{k=1}^{K} A_k V_k A_k', \tag{3}$$

with $V_k = I_0(\widehat{\boldsymbol{\theta}}_k)^{-1}$. We use the symbol $\widetilde{\boldsymbol{\theta}}^*$ to emphasize that this is not necessarily the maximum likelihood estimator even though, in our formalism, $\widehat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimator when restricting attention to subsample k.

Not every choice of the A_k leads to an unbiased estimator, but to ensure an unbiased expectation of $\widetilde{\boldsymbol{\theta}}^*$, we impose $\sum_{k=1}^{K} A_k = I_p$. Two obvious choices are constant, $A_k = (1/K)I_p$, and proportional weights, $A_k = (c_k/N)I_p$. Constant weights are an intuitive choice when partitioning in sub-samples of equal size, however the latter one is more obvious for sub-samples of varying size. This leads us, using Lagrange multipliers, to the optimal weights, $A_k^{opt} = \left(\sum_{m=1}^{K} V^{-1_m}\right)^{-1} V_k^{-1}$. These then lead to the maximum likelihood estimator. However, not in every case will there be a closed-form solution for V_k and if there are these may depend on unknown parameters. To solve this dilemma, consider first scalar weights by demanding A_k to be diagonal. Each component of $\boldsymbol{\theta}^*$, θ_r^* say, is determined as a linear combination, $\widetilde{\theta}_r^* = \sum_{k=1}^{K} a_{k,r} \widehat{\theta}_{k,r}$, with $A_k = \text{diag}(a_{k,1}, \cdots, a_{k,p})$. The resulting optimum will not necessarily be equal to the MLE, but the weights can be chosen for computational convenience. A second option is iterated optimal weights. The data need to be analyzed only once, to find $\hat{\boldsymbol{\theta}}_k$. From these, an initial estimator for $\boldsymbol{\theta}^*$ is computed using a simple weighting method, e.g., constant or proportional weights. Using $\boldsymbol{\theta}^{*(t)}$ and calculating $V_k^{(t+1)}$, $\boldsymbol{\theta}^{*(t+1)}$ can be

determined as $\boldsymbol{\theta}^{*(t+1)} = \left(\sum_{k=1}^{K}\left[V_k^{(t+1)}\right]^{-1}\right)^{-1}\sum_{k=1}^{K}\left[V_k^{(t+1)}\right]^{-1}\widehat{\boldsymbol{\theta}_k}$. This is repeated until convergence. From this we deduce the approximate optimal weights, a non-iterative approximation.

3 Partitioned-sample analysis for the compound symmetry model

The weights discussed in the previous section can be constructed for this specific case. Due to the independence of the mean and the variance components, the optimal and scalar weights do not make a difference for the mean parameter, but are different for the variance parameters. The weights depend on the parameters, but by plugging in the cluster-size specific mean and variance components, the expressions can be used for approximate weighting. But also the principles of iterated and approximate weights can be applied, as in Section 2.3.
The scalar weights are found to be:

$$a_k = \frac{\frac{c_k n_k}{\sigma^2+n_k d}}{\sum_{m=1}^{K}\frac{c_m n_m}{\sigma^2+n_m d}}, \tag{4}$$

$$b_k = \frac{c_k(n_k-1)}{\sum_{m=1}^{K} c_m(n_m-1)}, \tag{5}$$

$$g_k = \frac{\frac{c_k n_k}{\frac{\sigma^4}{n_k-1}+2\sigma^2 d+n_k d^2}}{\sum_{m=1}^{K}\frac{c_m n_m}{\frac{\sigma^4}{n_m-1}+2\sigma^2 d+n_m d^2}}, \tag{6}$$

with $\sum_{i=1}^{K} a_k = \sum_{i=1}^{K} b_k = \sum_{i=1}^{K} g_k = 1$. These weights again depend on the parameters and they can again be made part of an iterative scheme. Calculations show that the variance of the weighted estimator of the mean equals that of the maximum likelihood, so the weighted split-sample parameter is the maximum likelihood estimator. This is to be expected due to the the independence of the mean estimator from the variance components estimators for a given cluster size. Thus, the optimally weighted estimator and the scalar estimator coincide for the mean. This is not true for the variance components, however.
By approximating these weights for the case where cluster sizes are large, we derive that these weights are almost identical to the proportional weights, which makes them a sensible option for practice.
All this can also be applied when there is only one cluster per sub-sample, a so called cluster-by-cluster analysis. Then, $c_k \equiv 1$, $K \equiv N$, and n_k will no longer be unique. Since we make use of the fact that the cluster size is constant within a stratum, and not that the cluster sizes must be different

TABLE 1. ML and weighted split-sample estimates (standard errors): (a) ML: maximum likelihood; (b) REML: restricted maximum likelihood; (c) Prop.: proportional weights; (d) Equal: equal weights; (e) Appr.sc.: like proportional weights, except that for b_k is used; (f) Scalar: scalar weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights; (g) Opt.: optimal weights, with the sub-sample specific weights plugged in for the parameters figuring in the weights.

Par.	ML	REML	Prop.	Equal	Appr.sc.	Scalar	Opt.
μ	0.89450	0.89450	0.87305	0.85896	0.87305	0.91831	0.91831
σ^2	0.01950	0.01950	0.02200	0.02657	0.01951	0.01930	0.00603
d	0.01449	0.01467	0.00895	0.00858	0.00895	0.00085	0.00220
s.e.(μ)	0.01228	0.01234	0.01035	0.01478	0.01035	0.00761	0.00761
s.e.(σ^2)	0.00087	0.00087	0.00160	0.00431	0.00101	0.00097	0.00040
s.e.(d)	0.00245	0.00248	0.00208	0.00355	0.00208	0.00048	0.00044

between different strata, this is no problem. Result can be combined using again the weighted estimators or a two-stage approach. For the latter one, unbiasedness is not necessarily obtained.

4 Case study: a developmental toxicity study

The chemical compound di(2-ethylexyl)phthalate (DEHP) is used as plasticizer for numerous devices. Due to a possible presence in human and animal tissue, caused by leaks in plastic containers, toxic effects need to be investigated. The study was conducted in timed-pregnant mice during the period of major organogenesis (Tyl et al., 1988). A total of 1082 live fetuses were dissected. Our focus is on the continuous weight outcome. Fetuses are clustered within mothers. The CS model is fitted to the fetal weight outcome to examine the performance of the weighted estimators in Table 1. All split-sample estimators perform well in comparison with the (restricted) maximum likelihood estimators, only the optimal weights give slightly deviating results, which is because uncertainty due to the dependence of the weights on parameters is currently ignored. Using the delta method, this can be rectified. Importantly, the weighted estimators are a magnitude faster than the likelihood-based ones.

5 Concluding remarks

The use of weighted estimators reduced computation time and en enhances computation stability. They are simple to use, especially the proportional weights, and have a high efficiency.

Acknowledgments: Geert Molenberghs, Mike Kenward, Marc Aerts, Geert Verbeke and Wim van der Elst gratefully acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy) and Geert Molenberghs and Geert Verbeke from ExaScience Project.

References

Hermans, L., Molenberghs, G., Aerts, M., Kenward, M.G., and Verbeke, G. (2014). A note on incomplete sufficient statistics. *Submitted for publication.*

Hermans, L., Molenberghs, G., Kenward, M.G, Van der Elst, W., Nassiri, V., Aerts, M., and Verbeke G. (2015). Clusters with random size: maximum likelihood versus weighted estimation. *Working paper.*

Molenberghs, G., Kenward, M.G., Aerts, M., Verbeke, G., Tsiatis, A.A., Davidian, M., Rizopoulos, D. (2014). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, **23**, 11 – 41.

Molenberghs, G., Verbeke, G., and Iddi S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics and Probability Letters*, **81**, 892 – 901.

Tyl, R.W., Price, C.J., Marr, M.C., and Kimmel, C.A. (1988). Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. *Fundamental and Applied Toxicology*, **10**, 395 – 412.

Boosted negative binomial hurdle models for spatiotemporal abundance of sea birds

Benjamin Hofner[1], Adam Smith[2]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] University of Rhode Island, USA

E-mail for correspondence: benjamin.hofner@fau.de

Abstract: Modelling the abundance of sea birds is hampered by various difficulties: Sightings might be scarce (excess of zeros), overdispersion might be present, and the effects of biophysical covariates might be non-linear. Additionally, one should consider the spatiotemporal data structure in the model and apply variable selection to achieve a sparse model representation. We propose a spatiotemporal negative binomial hurdle model that considers all raised issues. Variable selection and model choice are achieved by boosting methods with stability selection.

Keywords: GAMLSS; Variable selection; Spatiotemporal modelling; Hurdle model.

1 Background

Nantucket Sound, Massachusetts, USA, is an important wintering area for seaducks in southern New England. Wind energy development has been fully permitted on 62 km^2 of Horseshoe Shoal in the northwest portion of Nantucket Sound. We conducted 30 aerial strip-transect (180 m wide) surveys throughout 1,100 km^2 of Nantucket Sound (Figure 1) during the winters of 2003–2005 to evaluate sea duck distribution and relative abundance. Here we consider counts only for Common Eider (*Somateria mollissima*). We aggregated eider counts within 2.25km^2 segments in the study area (Figure 1).

2 Modelling approach

We related spatiotemporal variation in sea duck occupancy (presence/ absence) and abundance to potentially relevant biophysical and spatiotemporal covariates. The data shows a high prevalence of zero counts (e.g., 75 % of segments contained no eider observations). Thus, we applied a negative binomial hurdle model that separately modeled the probability 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

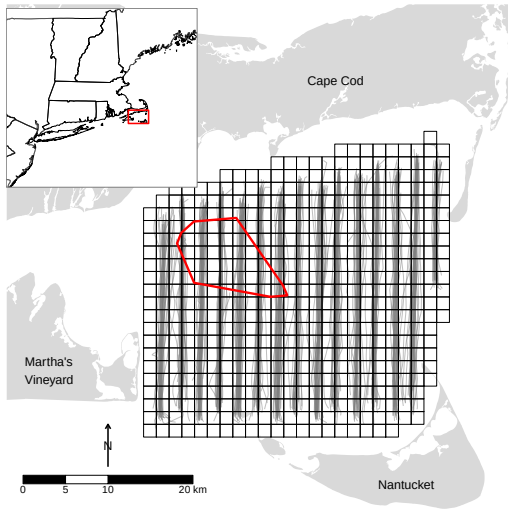

FIGURE 1. Nantucket Sound, Massachusetts, US study area. The grid indicates the extent of the study area and its division into 504 2.25km^2 segments. Gray lines indicate all aerial transects flown over the course of the study. The red polygon indicates the location of permitted wind energy development on Horseshoe Shoal.

occurrence of at least one individual (*"occupancy model"*; logistic regression model) in a given segment and the abundance of eider in that segment conditional on their presence (*"count model"*; truncated negative binomial model). The negative binomial density was specified such that the mean counts are given by $E(Y|x) = \mu$ and the variance as $Var(Y|x) = \mu(1 + \sigma)$ with dispersion parameter σ.

We evaluated biophysical covariates expected to influence the distribution and availability of benthic prey or the distribution, abundance, and movements of eider. Relevant biophysical covariates included, for example, bathymetry (*depth*), distance to nearest land (*dist*), sea floor roughness (SAR), sea bottom sediment grain size (SGS), tendency for summer stratification of the water column (*strat*), sea surface temperature relative to other segments (SST_r), average monthly sea surface temperature (SST_m), water temperature near the sea bottom (SBT), dissolved organic material concentrations (DOM), average epibenthic tidal velocity (ETV_{avg}) and standard deviation of epibenthic tidal velocity (ETV_{sd}). All continuous covariates were standardized, i.e., mean centered and scaled, before entering the model. We specified *a priori* expected time-varying effects for relative sea surface temperature and water depth to allow the effects to vary over time within a given winter. Additionally, we include spatial, temporal and spatiotemporal effects to account for unmeasured heterogeneity.

We used GAMLSS (Rigby and Stasinopoulos, 2005) to model the zero-trun-

cated negative binomial *count model*, i.e., we regressed both the mean μ and the dispersion parameter σ on covariates. The *occupancy model* was based on a generalized (logistic) additive model, where the expected occupancy $E(Y = 1|x) = \pi$ was modeled.

2.1 Boosting

Model-based boosting (Hothorn et al., 2010) was applied to fit all models and to allow for the selection of variables: Boosting aims at minimizing the negative log-likelihood of the model at hand. In each step, all effects are fitted separately to the gradient of the log-likelihood and only the best-fitting effect is selected and updated. For the GAMLSS *count model* we cycled through both parameters, location and dispersion, and for each parameter through all specified effects (Mayr et al., 2012). In both the count and the occupancy model we used smooth P-spline base-learners and applied a model decomposition (Hofner et al., 2011) to allow the boosting algorithm to select the appropriate model complexity.

2.2 Cross-validation

We used 25-fold subsampling to determine the optimal stopping iteration for each model. This is done to prevent overfitting and to minimize the (out-of-bag) prediction error (measured by the negative log-likelihood of the model). Specifically, we randomly drew (without replacement) data sets of size $n/2$ from the original data set. We used these data sets to estimate the model and used the other half of the data to determined the out-of-bag prediction accuracy.

2.3 Stability selection

To obtain even sparser models we used stability selection (Meinshausen and Bühlmann, 2010) to extract variables and effects that were stably selected while controlling the per-family error rate. Stability selection and an improved version (complementary-pairs stability selection; Shah and Samworth, 2013) was recently investigated in the context of boosting models and showed promising results in conjunction with boosted GAMs (Hofner et al., 2015). Here, we applied complementary-pairs stability selection with unimodality assumption, used an upper bound for the per-family error rate of six and allowed each model to select $q = 35$ variables.

3 Results

3.1 Count model

We obtained sparse models that included only a small fraction of the initially specified effects. The model for the mean parameter of the *count*

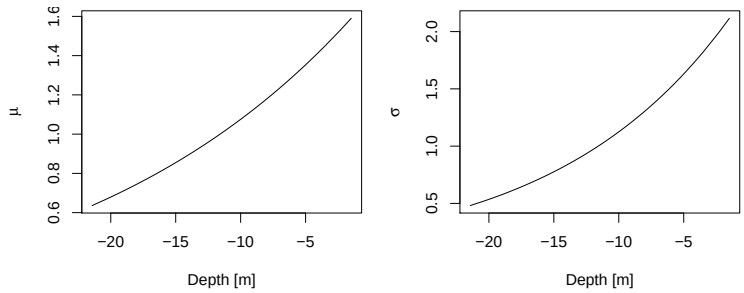

FIGURE 2. Partial effect of depth on the eider count μ and on the dispersion σ. Eider prefer shallower areas but at the same time counts show a greater variability in these areas.

model included categorical and linear effects (i.e., *ferry* and *depth*), and smooth effects f for various biophysical parameters. Temporal, spatial and spatiotemporal effects were not selected.

$$\begin{aligned}\log(\mu) =& \text{ferry} + \text{depth} \\ &+ f(\text{SAR}) + f(\text{SGS}) + f(\text{SST}_r) + f(\text{SBT}) + f(\text{DOM}).\end{aligned}$$

The dispersion of the *count model* included linear and smooth effects for various covariates. Additionally a temporal effect (*y2004*) and a spatial effects was selected. Spatiotemporal effects or time-varying effects were not selected.

$$\begin{aligned}\log(\sigma) =& \text{depth} + \text{SST}_m + \text{y2004} \\ &+ f(\text{DOM}) + f(\text{ETV}_{\text{avg}}) + f(\text{ETV}_{\text{sd}}) + f_{\text{spatial}},\end{aligned}$$

where μ is the (conditional) mean count of eiders, σ is the (conditional) dispersion in the negative binomial model. As an example, the effects of depth on μ and σ are depicted in Figure 2.

3.2 Occupancy model

The binomial logit model for *occupancy* was far more complex and included linear and smooth effects of various biophysical predictors, time-varying effects, spatial effects as well as a spatiotemporal effect:

$$\begin{aligned}\text{logit}(\pi) =& \text{ferry} + \text{SAR} + \text{DOM} + \text{y2005} \\ &+ f(\text{dist}) + f(\text{SGS}) + f(\text{strat}) + f(\text{SST}_m) + f(\text{SBT}) \\ &+ f(\text{time}) + f(\text{depth}, \text{time}) + f_{\text{spatial}} + f_{\text{spatial}} \cdot \text{time},\end{aligned}$$

where π represents the (conditional) occupancy probability. Additionally, the occupancy probability was dependent on the size of the observation

window (i.e., the area of transect surveyed in a given segment). The spatiotemporal effect is displayed in Figure 3. A strong shift of occupancy during the winter, which is not covered by other measured covariates, is apparent.

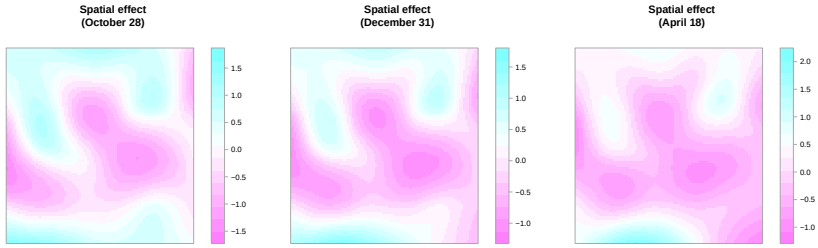

FIGURE 3. Partial spatiotemporal effect of the eider occupancy probability during winter season.

4 Implementation

All models were fitted using the statistical environment R. The *occupancy model* was fitted using the package **mboost** (Hothorn et al., 2010, 2015; Hofner et al., 2014), the *count model* was fitted using the package **gamboostLSS** (Hofner et al., 2015; Mayr et al., 2012) and stability selection was implemented in the package **stabs** (Hofner and Hothorn, 2015; Hofner et al., 2015).

5 Summary

Our boosting approach results in a very flexible model class for count data: It allows to incorporate excess zeros and to model the dispersion. Smooth, spatial and spatiotemporal effects can be easily included. Model choice and variable selection are available within the boosting framework. Additional sparsity with error control can be obtained by applying stability selection.

Acknowledgments: Special thanks to Massachusetts Audubon (particularly T. Allison and S. Perkins) for providing the aerial survey data. J. Osenkowski (Rhode Island Department of Environmental Management) and P. Paton and S. McWilliams (University of Rhode Island) funded A. Smith's contribution.

References

Hofner, B., Boccuto, L., and Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, accepted.

Hofner, B. and Hothorn, T. (2015). stabs: Stability Selection with Error Control, *R package version 0.5-1*, http://cran.r-project.org/package=stabs.

Hofner, B., Hothorn, T., Kneib, T., and Schmid, M. (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, **20**, 956 – 971.

Hofner, B., Mayr, A., Fenske, N., and Schmid, M. (2015). gamboostLSS: Boosting Methods for GAMLSS Models, *R package version 1.1-3*, http://cran.r-project.org/package=gamboostLSS.

Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M. (2014). Model-based boosting in R – A hands-on tutorial using the R package mboost. *Computational Statistics*, **29**, 3 – 35.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based boosting 2.0. *Journal of Machine Learning Research*. **11**, 2109 – 2113.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2015). mboost: Model-Based Boosting. *R package version R package version 2.4-2*. http://cran.r-project.org/package=mboost.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2012). Generalized additive models for location, scale and shape for high-dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society, Series C*, **61**, 403 – 427.

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

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

Shah, R.D. and Samworth, R.J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society, Series B*, **75**, 55 – 80.

Analysis of accelerated failure time data with dependent censoring

Chiu-Hsieh Hsu[1], Jeremy Taylor[2], Chengcheng Hu[1]

[1] University of Arizona, USA
[2] University of Michigan, USA

E-mail for correspondence: pchhsu@email.arizona.edu

Abstract: We consider the situation of estimating the marginal survival distribution from censored data subject to dependent censoring using auxiliary variables. We had previously developed a nonparametric multiple imputation approach using two working proportional hazards models to define an imputing risk set for each censored observation. Here we adapt the method to the situation where the event time and the censoring time follow accelerated failure time models. The Buckley-James estimator is used in the two working models. We discuss why the proposed methods are expected to be robust to misspecification of the link function. In a simulation study, we compare the nonparametric multiple imputation approach with inverse probability of censoring weighted (IPCW) methods, and two parametric multiple imputation methods. The proposed nonparametric imputation method is robust to mis-specification of either one of the two working models and mis-specification of the link function of the two working models. The IPCW method is not robust to mis-specification of the link function of the censoring time model. The parametric imputation methods rely on the specification of the event time model. The approaches are applied to a prostate cancer dataset.

Keywords: Accelerated failure time; Buckley-James estimator; Cox proportional hazards model; IPCW; Multiple imputation.

1 Introduction

We adapt a nonparametric multiple imputation approach we previously developed (Hsu et al., 2006) to handle the case of the data from an AFT model when the goal is estimating the marginal survival function. Specifically, we propose to use two Buckley James estimators, one for the failure time and one for the censoring time, to derive two risk scores using auxiliary variablesto select an imputing risk set for each censored observation. The two

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Buckley James estimators are only used to derive two risk scores to select an imputing risk set. Hence, the approach is expected to be less affected by unstable estimation. The inverse probability of censoring weighted (IPCW) method (Robins and Finkelstein, 2000), where the weight is derived from a model for the censoring times with auxiliary variables as the covariates, and parametric multiple imputation (PMI) methods, where a specific parametric model is used to impute event times for censored observations, are two popular methods for handling censored observations through the use of auxiliary variables in survival analysis. Here, not only will we study the performances of the proposed multiple imputation approaches but will also compare their performances with these two existing popular approaches when the true model for the event time is from an AFT model.

2 Imputation procedures

For each censored observation, we describe the imputation procedures in the following four steps. *Step 1: Estimate the two risk scores on a Bootstrap sample* We first reduce the auxiliary variables to two risk scores. We assume the data arise from an AFT model. Hence, we propose to use two Buckley James estimators, one for the failure time and one for the censoring time, to derive two risk scores on a nonparametric bootstrap sample of the original dataset. Two working PH models will also be fit to the bootstrap sample to calculate the two risk scores to study a robustness property for link function misspecification. *Step 2: Calculate the distance between subjects* For each censored subject in the original dataset, two risk scores are derived using the regression coefficient estimates from the bootstrap sample . The distances between each cenored subject in the original dataset and subjects in the bootstrap sample are calculated using the two risk scores. *Step 3: Define the imputing risk set* For each censored subject, the distance in *Step 2* is then employed to define a set of nearest neighbors consisting of subjects who have longer survival time than the censored subject. *Step 4: Impute a value from the imputing risk set* After the imputing risk set is defined, the Kaplan-Meier imputation (KMI) scheme developed in Taylor et al. (2002) can be easily used. The KMI methods using two Buckley James estimators and two PH models to derive the risk scores are denoted as KMI_{BJ} and KMI_{PH}, respectively. *Step 5: Repeat Steps 1 to 4 independently M times* Each of the M imputed datasets is based on a different Bootstrap sample. Once the M multiply imputed datasets are obtained, we carry out the MI analysis procedure established in Rubin (1987).

3 Data analysis

To demonstrate the MI approach when potential non-proportional hazards exist, baseline PSA value, Age, Gleason score, Total Radiation Dose

and T Stage are treated as time-independent covariates in the two working Buckley James estimators and two working PH models. We also derive estimates from the partially observed (PO) analysis, which is the Kaplan-Meier estimation based on the observed censored event time data, IPCW_{PH}, $\text{IPCW}_{Lognormal}$ and two parametric multiple imputation (PMI) methods ($\text{PMI}_{Lognormal}$ and $\text{PMI}_{Weibull}$), where a parametric model (log-normal or Weibull model) is fitted to the oberved data to impute residual life times for each censored observation. Based on Fig. 1, KMI_{BJ} and KMI_{PH} methods, as well as both PMI and both IPCW methods, produce slightly higher estimated survival at both five and ten years and slightly lower associated estimated standard errors than the PO analysis at five years. Both IPCW methods produce slightly greater survival estimates than the two KMI methods especially at the tail. KMI_{BJ} and KMI_{PH} produce almost identical results for both survival and associated standard error estimates.

4 Simulation study

We perform several simulation studies to investigate the properties of the KMI, IPCW and PMI methods and consider a situation with multiple time-independent prognostic covariates and dependent censoring. We investigate the effects of the magnitude of dependent censoring, sample size, misspecification of one of the two working models and misspecification of the two link functions. For each of independent simulated datasets, there are five hypothetical auxiliary variables and the true failure and censoring time models are from an AFT family. Based on the simulation results, all methods reduced the bias of the standard PO analysis, but the amount of the remaining bias, the efficiency and the validity of the estimated standard errors varied between methods. The performance of the IPCW method depends on whether a correct censoring time model is used to derive the weights, especially when the dependent censoring is strong. In contrast, the KMI methods in which two risk scores are derived from either two working Buckley James estimators or two working PH models can provide reasonable survival estimates for both weak and strong depend censoring and is robust to misspecification of either one of the two working models and is robust to misspecification of the link functions in the failure time and censoring time models. The performance of the parametric multiple imputation approach depends on whether a correct residual time model is used for imputation, especially at the tail area of the survival curve.

5 Discussion

Based on the simulation results, while performing the PMI and IPCW methods, one has to be sure the corresponding model is correct, and specifically requires all aspects of the models including the link functions and

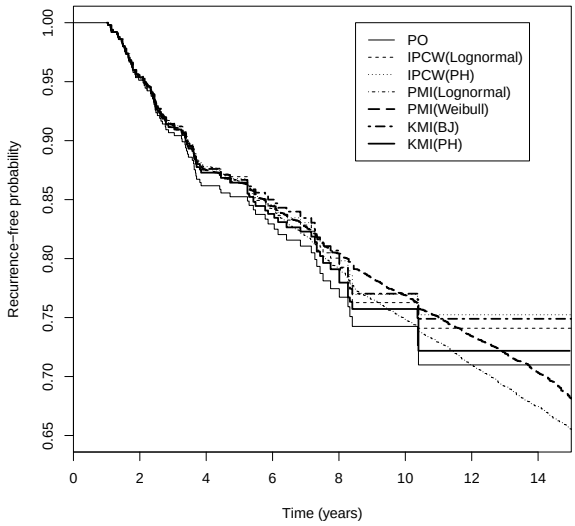

FIGURE 1. Prostate Cancer Study: Recurrence-free curves.

choice of covariates to be correct. In contrast, the nonparametric multiple imputation approach is expected to have weak reliance on the two working models compared to the IPCW method. The simulation study also shows that the two nonparametric multiple imputation approaches produce similar results for both point survival estimates and the associated standard error estimates when the data are from AFT models. This is because the PH model preserves the relative importance of the covariates in the AFT model (Struthers and Kalbfleisch, 1986). This indicates that the multiple imputation approach is robust to misspecification of the link functions of the two working PH models when the data are from AFT models.

References

Hsu, C.H., Taylor, J.M.G., Murray, S., and Commenges, D. (2006). Survival analysis using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine*, **25**, 3503 – 3517.

Robins, J.M. and Finkelstein, D.M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, **56**: 779 – 788.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York. Wiley.

Struthers; C.A. and Kalbfleisch, J.D. (1986). Misspecified proportional hazards models. *Biometrika*, **73**, 363 – 369.

TABLE 1. Monte Carlo results for the marginal survival estimate at the median and 75^{th} percentile survival times, where $\log(T) = 0.10 - 2Z_1 + 0.5Z_2 - 2Z_3 + 2Z_4 + 2Z_5 + \text{Normal}(0, 2^2)$ and $\log(C) = 0.08 - 2.5Z_1 + 0.5Z_2 - 2Z_3 + 2Z_4 + 2Z_5 + \text{Normal}(0, 2^2)$. Censoring rate=50%, the Spearman correlation coefficient between T and C (i.e. ρ)=0.27.

			S(t)=0.25		
			N=200		
Method	Est[a]	bias	SD[b]	SE[c]	CR[d]
FO	0.2510	0.0010	0.0288	0.0306	94.4
PO	0.3049	0.0549	0.0480	0.0475	78.8
KMI_{BJ55}	0.2606	0.0106	0.0472	0.0464	94.8
KMI_{PH55}	0.2601	0.0101	0.0474	0.0461	93.4
KMI_{BJ53}	0.2621	0.0121	0.0481	0.0459	94.2
KMI_{PH53}	0.2600	0.0100	0.0481	0.0458	93.8
KMI_{BJ35}	0.2681	0.0181	0.0487	0.0463	93.4
KMI_{PH35}	0.2684	0.0184	0.0476	0.0465	92.6
$\text{IPCW}_{Lognormal}$	0.2505	0.0005	0.0522	0.0505	93.0
IPCW_{PH}	0.2413	-0.0087	0.0663	0.0420	78.0
$\text{PMI}_{Lognormal}$	0.2490	-0.0010	0.0345	0.0404	97.0
$\text{PMI}_{Weibull}$	0.2308	-0.0192	0.0349	0.0395	92.2
			N=400		
FO	0.2487	-0.0013	0.0218	0.0216	93.2
PO	0.3052	0.0552	0.0337	0.0337	62.6
KMI_{BJ55}	0.2578	0.0078	0.0335	0.0329	94.4
KMI_{PH55}	0.2574	0.0074	0.0333	0.0327	93.4
KMI_{BJ53}	0.2587	0.0087	0.0338	0.0326	93.8
KMI_{PH53}	0.2575	0.0075	0.0335	0.0325	94.0
KMI_{BJ35}	0.2623	0.0123	0.0343	0.0328	93.6
KMI_{PH35}	0.2623	0.0123	0.0338	0.0326	94.4
$\text{IPCW}_{Lognormal}$	0.2499	-0.0001	0.0365	0.0355	94.2
IPCW_{PH}	0.2393	-0.0107	0.0528	0.0333	79.0
$\text{PMI}_{Lognormal}$	0.2492	-0.0008	0.0249	0.0282	96.6
$\text{PMI}_{Weibull}$	0.2311	-0.0189	0.0264	0.0276	89.0

[a]Average of 500 point estimates.
[b]Empirical standard deviation.
[c]Average estimated standard error.
[d]Coverage rate of 500 95% confidence intervals.

Taylor, J.M.G., Murray, S., and Hsu, C.H. (2002). Survival estimation and testing via multiple imputation. *Statistics & Probability Letters*, **58**, 221 – 232.

Testing conditional independence in sets of $I \times J$ tables

Maria Iannario[1], Joseph B. Lang[2]

[1] Department of Political Sciences, University of Naples Federico II, Italy
[2] Department of Statistics and Actuarial Science, University of Iowa, Iowa City

E-mail for correspondence: maria.iannario@unina.it

Abstract: A new testing approach is described for improving statistical tests of independence in sets of tables stratified on one or more relevant factors in case of categorical (nominal or ordinal) variables. Common tests of independence that exploit the ordinality of one of the variables use a restricted-alternative approach. A different, relaxed-null method is presented. Specifically, the M-moment score tests and the Correlation tests are introduced. Using multinomial-Poisson homogeneous modeling theory, it is shown that these tests are computationally and conceptually simple, and can perform better than the omnibus and commonly used Cochran-Mantel-Haenszel statistics.

Keywords: Conditional independence; M-moment score tests; Correlation tests.

1 Introduction

Across many disciplines, it is common to obtain categorical responses for a number of items stratified with respect to one or more other relevant factors. The general goal of a stratified analysis is to better understand the relationship between the response variables by controlling for the potentially confounding effects of factors that are part of the research design, or factors that represent a prespecified poststudy stratification. Stratified analyses can accommodate heterogeneous relationships across strata levels. When one or more of the categorical responses is ordered, it is possible to do a more thorough analysis than that given by the usual tests for contingency tables. For testing conditional independence, it is common to use models such as ordinal loglinear or cumulative logit models, which typically include parameters that measure some shift pattern of association. This model-based approach, which tests independence against a restricted

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

alternative determined by the chosen model, works well when the departure from independence is of the restricted shift form.
Otherwise, it is possible to implement a non-model based alternative, such as the generalization of the Cochran–Mantel–Haenszel statistic (CMH) (Landis et al., 1978) that potentially removes the confounding influence of the explanatory variables that comprise the stratification and provides a gain of power for detecting association.
This paper considers an attractive alternative to the model-based and CMH approaches for stratified nominal-by-ordinal (or ordinal-by-ordinal) tables. Specifically, we generalize the 1- and 2-Moment score tests of Lang and Iannario (2013) to the stratified table setting; we also introduce two correlation score tests. These score tests (i) may have power that nearly matches or exceeds that of the shift tests when the departures from independence are of the shift form, (ii) may have power that can far exceed that of both the shift tests and the omnibus CMH test for substantively interesting departures that are not of the simple shift form, and (iii) in special setting where the correlations are constant across the rows or strata, have power that is higher than the common tests used for detecting linear correlation.
Section 2 reports some notation on the Moment score tests and the Correlation score tests for three-way tables. Section 3 summarizes the main results of a simulation experiment.

2 Methods

Denote the nominal strata variable (or covariate) by Z. We will consider the case where the response variables X and Y are nominal (or ordinal) and ordinal. Using standard distribution theory arguments, let $n = (n_1, \ldots, n_K)^{\mathrm{T}}$ denote the observed frequencies from K tables of dimension $I \times J$. For convenience, we assume that vector n has a product-multinomial distribution with cell probabilities $\pi^{\mathrm{T}} = (\pi_1^{\mathrm{T}}, \ldots, \pi_K^{\mathrm{T}})$, where the kth strata probabilities are $\pi_k^{\mathrm{T}} = (\pi_{k11}, \ldots, \pi_{kIJ})$ and $\pi_{kij} = P(Y = j|X = i, Z = k)$. More generally, the results in this paper hold when n has a multinomial-Poisson distribution with no stratification, or row stratification (see Lang, 2004).
The main goal of the analysis is to test the overall null hypothesis of no treatment effect, i.e. conditional independence $(X \perp Y)|Z$, or in terms of cell probabilities,

$$H_0^{\perp} : \pi_{k1j} = \cdots = \pi_{kIj} \qquad \text{for } j = 1, \ldots, J;\ k = 1, \ldots, K.$$

2.1 The moment score tests

This section presents score tests based on constraints on the conditional moments of the column variable Y. Rather than testing $H_0^{\perp}$ against a restricted alternative, we will use an M-Moment Score statistic to test a

relaxed null $H_0 \supseteq H_0^{\perp}$ against the unrestricted alternative $H_1 =$ not H_0. In particular, we will consider tests of relaxed nulls of the M moment forms

$$H_0^{(M)}: \quad \sum_{k=1}^{K} E(Y^m|X=i,Z=k) = \sum_{k=1}^{K} E(Y^m|X=1,Z=k),$$
$$\text{for } i=2,\ldots,I;\ m=1,\ldots,M$$

and

$$H_0^{(M;s)}: \quad E(Y^m|X=i,Z) = E(Y^m|X=1,Z),$$
$$\text{for } i=2,\ldots,I;\ m=1,\ldots,M.$$

In words, $H_0^{(M)}$ is the hypothesis of no average (over strata) conditional relationship and $H_0^{(M;s)}$ is the hypothesis of no conditional relationship (in any strata) as measured via the first M moments. Notice that when $M=1$, $H_0^{(1)}$ corresponds to the null that the row-mean CMH statistic is tailored to detect departures from. This paper focuses primarily on 1- and 2-Moment score tests of the nulls $H_0^{(1)}$, $H_0^{(1;s)}$, $H_0^{(2)}$, and $H_0^{(2;s)}$. Note that rejection of any of these relaxed nulls implies rejection of conditional independence $H_0^{\perp}$. Also note that $H_0^{(J-1;s)} = H_0^{\perp}$.

As a simulation study suggests, this class of M-Moment score tests can be significantly more powerful than the CMH tests and the omnibus Pearson test of conditional independence. The 1-Moment Score test of $H_0^{(1)}$, which has similar operating characteristics to that of the row-mean CMH statistic, is powerful when there is simply a mean shift from conditional independence. When there is a shift in row dispersion (but not a shift in row mean), the 2-Moment Score tests are generally more powerful than the other available test statistics for detecting departures from conditional independence.

Computing the M-Moment score statistics and their asymptotic distributions is relatively straightforward using ideas in Lang (2004, 2005). For convenience, we give results for nulls of the form $H_0^{(M)}$; the results are analogous for the nulls $H_0^{(M;s)}$. Because we are testing $H_0^{(M)} \supseteq H_0^{\perp}$ against the unrestricted $H_1 =$ not H_0, we know that the M-Moment Score statistic has the Pearson form; specifically,

$$X_M^2 = (n-\hat{\mu}_0)^{\mathrm{T}} D^{-1}(\hat{\mu}_0)(n-\hat{\mu}_0),$$

where $\hat{\mu}_0$ is the maximum likelihood estimate of the the vector of expected table counts under $H_0^{(M)}$ and $D(x)$ is a diagonal matrix with component in x on the diagonal. The model under $H_0^{(M)}$ can be expressed in the multinomial-Poisson homogeneous linear predictor form $L\pi = X\beta$ or homogeneous constraint form $h(\pi) = 0$. ML estimates were obtained using the R program `mph.fit`.

MPH theory, in combination with standard likelihood theory, implies that when $H_0^{\perp}$, and hence $H_0^{(M)}$ holds, and the cell counts are large enough, the M-Moment Score test statistic X_M^2 has an approximate chi-squared distribution. More specifically, $X_M^2 \overset{approx}{\sim} \chi^2_{M(I-1)}$. We reject $H_0^{(M)}$, and hence $H_0^{\perp}$, iff $X_M^2 \geq \chi^2_{\alpha;M(I-1)}$. This test has approximate size α.

2.2 Correlation score tests

When both response variables X and Y are ordinal, we can entertain correlation score tests of conditional independence that are powerful when the conditional correlations are non-zero. As in the previous section, we consider two different relaxed nulls,

$$H_0^{(c)}: \quad \sum_{k=1}^{K} E(XY|Z=k) = \sum_{k=1}^{K} E(X|Z=k)E(Y|Z=k) \quad \text{and}$$
$$H_0^{(c;s)}: \quad E(XY|Z) = E(X|Z)E(Y|Z).$$

In words, $H_0^{(c)}$ states that the average (over strata) conditional correlation is 0 and $H_0^{(c;s)}$ states that the conditional correlations are 0 within each strata. Note that the relaxed null $H_0^{(c)}$ corresponds to the null that the CMH correlation statistic is tailored to detect departures from.

As with the M-Moment score tests of the previous section, the correlation score tests, which test a relaxed null versus an unrestricted alternative have the simple Pearson X^2 form; for example, $X_c^2 = (n-\hat{\mu}_0)^{\mathrm{T}} D^{-1}(\hat{\mu}_0)(n-\hat{\mu}_0)$, where $\hat{\mu}_0$ is the ML estimate of expected counts under the hypothesis $H_0^{(c)}$. The ML estimates were computed using the R program `mph.fit`.

MPH theory implies that when $H_0^{\perp}$, and hence $H_0^{(c)}$ holds, and the cell counts are large enough, the correlation score test statistic X_c^2 has an approximate chi-squared distribution; specifically, $X_c^2 \overset{approx}{\sim} \chi_1^2$. We reject $H_0^{(c)}$, and hence $H_0^{\perp}$, iff $X_c^2 \geq \chi^2_{\alpha;1}$. The same argument implies that under $H_0^{(c;s)}$ the correlation score test statistic $X_{c;s}^2 \overset{approx}{\sim} \chi_K^2$.

3 Simulation study

The simulations experiment is designed to compare the operating characteristics of the 2-and 1-Moment score tests and X_c^2 and $X_{c;s}^2$ tests to several other common tests, under a variety of table probability configurations. In particular, we will compare these tests with the $gCMH$ tests (general association, row-mean and linear association) and the omnibus tests (X^2 and X_s^2) in which the association between ratings is simply expressed using odds ratio and summing log odds ratio of strata (see Iannario and Lang, 2015).

The simulation study generalizes to stratified contingency tables the experiment in Lang and Iannario (2013).
Each $3 \times 7 \times 4$ table in the simulation had row probabilities that were generated discretizing latent Logistic distributions with (non)constant expectations and variances. Starting with Logistic location-shift models, Logistic-scale shift models and both location and scale shift we implement the several tests. Specifically, each row of the table probabilities were generated by means of $\sigma_i V - \beta_i, i = 1, 2, 3$, for some continuous latent variable $V \sim \ Logistic(0, 1)$, where at least two of the σ_i's and/or at least two of the β_i's are distinct. The independence region $H_0^{\perp}$ can be viewed as those tables with ith row distribution that can be viewed as a discretized version of the distribution of $\sigma_i V - \beta_i$, where $\sigma_1 = \cdots = \sigma_r$ and $\beta_1 = \cdots = \beta_r$. We choose cutpoints equal to the $\{1/7, 2/7, \ldots, 6/7\}$ quantiles of the standard logistic distribution. And we consider a variety of (β_i, σ_i) pairs.
Details on latent Logistic location and scale parameters which characterize the *starting* tables are in the working paper of Iannario and Lang (2015); specifically in the case A independence holds, in the cases B and C we consider equal latent means and unequal latent variances; for D-E cases unequal latent means and equal latent variances; finally, in F-G-H-I cases we study unequal latent means and unequal latent variances.
Corresponding to each table (case) we report the Monte Carlo estimate of the probability of rejecting the independence hypothesis for each of the candidate tests, using a nominal 0.05 significance level. These estimates are based on 100 simulations. We used product row multinomial sampling based on sample size $n = 500$ and generate the tables.

TABLE 1. Power of Nominal 0.05 Level Tests of Independence.

	X_2^2	X_1^2	X_c^2	$X_{c;s}^2$	X^2	X_s^2	T_{CMH}	T_{CMH_m}	T_{CMH_c}
A	0.02	0.05	0.03	0.01	0.01	0.05	0.01	0.01	0.04
B	**0.94**	0.01	0.04	0.01	0.38	0.83	0.81	0.01	0.01
C	**0.82**	0.02	0.06	0.08	0.32	0.58	0.62	0.04	0.06
D	0.60	**0.85**	0.69	0.46	0.21	0.44	0.35	0.67	0.69
E	0.45	**0.72**	0.65	0.39	0.08	0.42	0.32	0.39	0.64
F	**0.93**	0.12	0.10	0.05	0.42	0.83	0.75	0.06	0.09
G	0.60	0.11	0.11	0.03	**0.74**	0.48	0.41	0.08	0.12
H	0.91	**0.98**	0.83	0.68	0.56	0.74	0.67	0.96	0.85
I	0.78	**0.83**	0.74	0.56	0.33	0.59	0.56	0.74	0.73

The first row of Table 1 suggests that tests are close to the nominal 0.05, thus independence holds. Rows B-C show that the power is highest for the X_2^2 test, which is not surprising because this test is designed for capturing shift in variances among strata. However, even in location shift cases (D-E), the 2-Moment score test has power that is similar than the power of the

T_{CMH_m} tests, usually implemented for these cases of mean responses difference. For the results related to the mentioned D-E cases it is interesting underline the higher power of the score X_1^2 test. Rows F-I, where there is both location and scale shifts and in which the correlation among strata is different to zero, show that the 2- and 1- Moment score tests have the highest power, sometimes substantially high as in rows H-I.
In sum, it could be argued that the Moment tests have the best operating characteristics overall, for this latent logistic setting. Because they are based on relatively few degrees of freedom, they tend to be reasonably powerful for detecting these shifts by representing an useful alternative for the standard tests.
For the other tests, X_c^2 presents a power close to T_{CMH_c} whereas the $X_{c;s}^2$ has a power close to T_{CMH} by introducing alternative score tests for detecting correlation and general association. Based on other simulations not included in this paper, we found that the comparisons are similar for other table dimensions and sample sizes. In that cases parameters, the number of clusters within each stratum-by-treatment group average cluster size and variability in cluster size (none, moderate, high) were varied. Results do not change also in case of different sample size for rows and strata.
The limit of this proposal concerns sparse data. In this case in fact both the 2-moment score and correlation tests work poorly.
Generally, they present good operating characteristics because they are based on relatively few degrees of freedom and they tend to be reasonably powerful for detecting shifts in data.

Acknowledgments: This research has been supported by FIRB 2012 project (code RBFR12SHVV) and the frame of Shape project in STAR Programme (CUP E68C13000020003).

References

Iannario, M. and Lang, J.B. (2015). Testing conditional independence in sets of $I \times J$ tables by means of moment and correlation score tests with application to HPV vaccine. *Working paper*.

Landis, J.R., Heyman, E.R., and Koch, G.G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review*, **46**, 237 – 254.

Lang, J.B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *Annals of Statistics*, **32**, 340 – 383.

Lang, J.B. (2005). Homogeneous linear predictor models for contingency tables. *Journal of the American Statistical Association*, **100**, 121 – 134.

Lang, J.B. and Iannario, M. (2013). Improved tests of independence in singly-ordered two-way contingency tables. *Computational Statistics & Data Analysis*, **5**, 339 – 351.

Estimation of an MIC distribution using the vertex exchange method

Stijn Jaspers[1], Dankmar Böhning[2], Geert Verbeke[3], Marc Aerts[1]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[2] Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, United Kingdom
[3] Interuniversity Institute for Biostatistics and statistical Bioinformatics, KU Leuven, Leuven, Belgium

E-mail for correspondence: `stijn.jaspers@uhasselt.be`

Abstract: Bacteria that have developed a reduced susceptibility against antimicrobials pose a major threat to the public health. Hence, monitoring their distribution in the general population is of major importance. This monitoring is performed based on minimum inhibition concentration (MIC) values, which are collected through dilution experiments. We present a semi-parametric mixture model to estimate the MIC density on the full continuous scale. A key role in this new approach is put aside for an adjusted version of the Vertex Exchange Method (Böhning, 1986). A data application and simulation study are presented in which the promising behaviour of the new method is illustrated.

Keywords: Antimicrobial resistance; Censoring; Vertex exchange method.

1 Introduction

Antimicrobials are substances used to kill microorganisms or to inhibit their growth. The accidental discovery and isolation of penicillin by Sir Alexander Fleming marks the start of modern day antibiotics. Nevertheless, it soon became clear that bacteria could also develop antibiotic resistance whenever too little penicillin was used or when it was used for a too short period. New antimicrobial agents have been developed ever since, but unfortunately, so has antimicrobial resistance (AMR) (Palumbi, 2001). An excessive and sometimes inappropriate usage of antimicrobials has lead

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

to an increasing amount of bacterial isolates that are able to withstand antimicrobial treatments. Since isolates with a reduced susceptibility to antimicrobials pose a major threat to the public health, it is important to study and monitor their distribution. This may be performed by determining the minimum inhibition concentration (MIC), which is commonly measured via a broth dilution method. A standardized amount of the isolate is exposed to successive two-fold concentrations of the antimicrobial and the MIC is defined as the smallest concentration of the antimicrobial substance that inhibits the visible growth of the microorganism. Figure 1 shows an MIC distribution determined for 5190 isolates of *E. coli* tested for susceptibility against nalidixic acid. Note that, as a result of the dilution type laboratory experiments, MIC data are censored.

Our interest is in identifying the full continuous MIC distribution. In this regard, mixture models are ideally suited as they offer a natural framework for modelling unobserved population heterogeneity. In this paper, we will focus on a semi-parametric mixture model in which the wild-type first component is modelled parametrically, while the second, non-wild-type, component is modelled in a non-parametric manner using an adjusted method of the Vertex Exchange Method.

2 Methodology

With X denoting the MIC value of a certain isolate, the probability density function of interest is represented by

$$f(x;\Phi) = \pi f_1(x;\theta_1) + (1-\pi) f_2(x). \tag{1}$$

As argued above, the first component, f_1, can be assumed to be of a fixed parametric form depending on certain parameters θ_1. In this paper, the primary focus for the first component will be on the log-normal assumption, with θ_1 representing the mean and standard deviation of the first component. The proportion of isolates corresponding to this wild-type component is reflected in the prevalence parameter π. Because less information is present on the second component, a non-parametric basis-function representation of f_2 is given by $f_2(x;\Psi) = \sum_{c=1}^{k} \pi_c g(x;\lambda_c)$, where Ψ denotes the mixing distribution in which weights $\pi_1, \ldots, \pi_k$ are given to support points $\lambda_1, \ldots, \lambda_k$. In this paper, g is assumed to be log-normally distributed and the support points λ_c correspond to the means of these log-normal densities. In order to obtain enough flexibility to estimate the unknown second component, a generous number of log-normal densities will be employed. These densities are located at equidistant support points which are assumed to cover the range of observed MIC values. However, since we do not want to overfit the data, the standard deviations of these basis functions will be kept fixed at a predetermined value. A grid search will be performed

to determine the right amount of smoothing based on the the Akaike Information Criterion (Akaike, 1974): AIC$= -2l + 2K$, with K the number of parameters. In conclusion, the density of the MIC value X is given by $f(x;\Phi) = \pi f_1(x;\theta_1) + (1-\pi)\sum_{c=1}^{k} \pi_c g(x;\lambda_c)$, where $\Phi = (\pi, \theta_1, \Psi)$. Recall that MIC data are typically censored, a direct consequence of the collection using dilution experiments. Suppose therefore that the sample space is partitioned into m mutually exclusive intervals, for which the boundaries are denoted by $a_0, a_1, \ldots, a_m$. The reported data are the number of isolates n_i falling into the interval $]a_{i-1}, a_i]_{i=1,\ldots,m}$, where $n = \sum_{i=1}^{m} n_i$ corresponds to the total number of independently collected MIC values. In what follows, $y_i = (a_i, n_i)$ denotes the observed grouped data. The likelihood to be optimised is expressed by $l(\pi, \theta_1, \Psi|y) = \sum_{i=1}^{m} n_i \log\{F(a_i) - F(a_{i-1})\}$.

2.1 The VEM algorithm

The VEM is a directional derivative based method, used to obtain the NPMLE of a mixing distribution. The main idea of the method is to search in each iteration for the direction that maximizes the log-likelihood increase $\Delta = l(\Psi^{(it+1)}) - l(\Psi^{(it)})$. Once the optimal direction is found, weights are exchanged between the support points that contribute the most and the least to this difference. These points are identified based on the definition of the directional derivative $D(\Psi^{(it)}, \Psi_{\lambda_c})$ of $l(\Psi)$ at $\Psi^{(it)}$ in the direction of Ψ_{λ_c}:

$$\begin{aligned} D(\Psi^{(it)}, \Psi_{\lambda_c}) &= \lim_{s \to 0+} \frac{l\left((1-s)\Psi^{(it)} + s\Psi_{\lambda_c}\right) - l(\Psi^{(it)})}{s} \\ &= \sum_{i=1}^{m} \left[n_i \frac{F(a_i;\lambda_c) - F(a_{i-1};\lambda_c)}{F(a_i;\Psi^{(it)}) - F(a_{i-1};\Psi^{(it)})} \right] - n. \end{aligned} \tag{2}$$

More specifically, after evaluating the directional derivative for each of the k support points λ_c, we can identify $\lambda^- = \text{argmin}_{\lambda_c} D(\hat{\Psi}^{(it)}, \Psi_{\lambda_c})$ and $\lambda^+ = \text{argmax}_{\lambda_c} D(\hat{\Psi}^{(it)}, \Psi_{\lambda_c})$ and update their weights as $\hat{\pi}_{\lambda^-}^{(it+1)} = (1-s^\star)\hat{\pi}_{\lambda^-}^{(it)}$ and $\hat{\pi}_{\lambda^+}^{(it+1)} = s^\star \hat{\pi}_{\lambda^-}^{(it)} + \hat{\pi}_{\lambda^+}^{(it)}$ with the step length $s^\star \in [0,1]$ defined as $s^\star = \text{argmax}_s \left[l(\Psi^{(it+1)}(s)) - l(\Psi^{(it)}) \right]$.

2.2 Back-fitting algorithm

We are required to estimate two sets of parameters. First of all, there are the parameters inherent to the parametric distribution assumed for the first component, i.e. θ_1. In addition, we also need to estimate the weights attached to the distinct components, i.e. $\tilde{\pi} = (\pi, \pi_1, \ldots, \pi_k)$. In this respect, we adopt a back-fitting algorithm, based on the work of Tsonaka et al. (2009). After obtaining initial estimates for the parameters of interest, we iterate between a VEM step to update the mixture weights and a profile likelihood step to update the parameters of the first component.

3 Application to real data

The procedure described in Section 2 is applied to MIC data obtained from the EUCAST website. The data concern the susceptibility of *E. coli* against nalidixic acid. The estimated mean and standard deviation of the wild-type first component are 1.07 (se= 0.10) and 0.51 (se=0.04). The prevalence of wild-type isolates is estimated to be 0.78 (se=0.32).
Figure 1 shows the resulting fit to the observed data on the left, while the posterior probability to belong to the wild-type component in function of the observed MIC value is shown on the right. It is noted that the cut-off value to discriminate between wild-type and non-wild-type isolates is 4 mg/l. This means that isolates which have a MIC value that is higher than 4 mg/l cannot be considered to be wild-type and hence show signs of being resistant to the antimicrobial nalidixic acid.

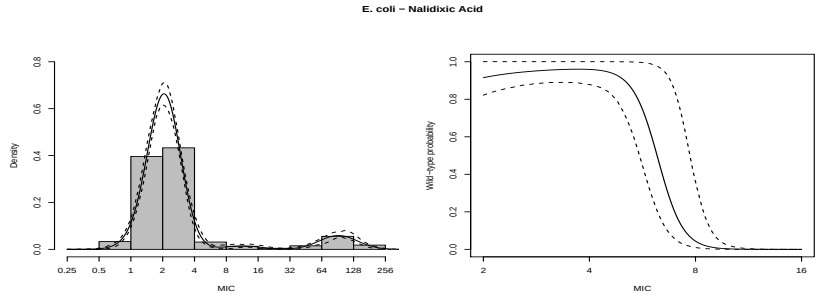

FIGURE 1. Barplot and fitted density (left), posterior wild-type probabilities (right) – source: EUCAST website.

4 Simulation study

Samples are taken from a mixture distribution with two main components. The wild-type component is assumed to be log-normally distributed with mean 2 and standard deviation 0.8. The non-wild-type component is a 50:50 mixture of two log-normal densities with (on the $\log_2$-scale) means equal to 4.5 and 7.5, respectively, and standard deviations equal to 0.7 and 0.6, respectively. The prevalence of wild-type isolates is taken to be 0.6. The developed back-fitting algorithm based on the VEM is compared to two competitors: a first one based on the penalized mixture (PM) approach (Jaspers et al., 2014b) and a second based on the composite link model (CLM; Jaspers et al., 2015). The considered sample sizes are 500, 1000, and 5000.
The performance of the methods is first assessed based on the integrated MSE values. In order to obtain these values, the estimated densities are

evaluated over a grid, x_i, $i = 1, \ldots, I$. The IMSE is the mean of the mean squared error values, which are calculated as follows:

$$\text{MSE}_{\hat{f}(x_i)} = \text{Bias}^2_{\hat{f}(x_i)} + \text{Var}_{\hat{f}(x_i)},$$

with

$$\begin{aligned} \text{Bias}_{\hat{f}(x_i)} &= E[\hat{f}(x_i)] - f(x_i) \\ \text{Var}_{\hat{f}(x_i)} &= E[(\hat{f}(x_i) - f(x_i))^2]. \end{aligned}$$

In addition, we also investigated the classification performance of the methods. Observations are classified into the first component in case the probability

$$\frac{\pi f_1(x;\theta_1)}{\pi f_1(x;\theta_1) + (1-\pi) f_2(x)}$$

exceeds 0.5.

Based on the IMSE values in Table 1, it can be concluded that the semi-parametric mixture model is outperformed by both the back-fitting algorithm and the CLM approach. In addition, the VEM seems to outperform the CLM in terms of these IMSE values. Regarding the miss-classification probabilities, there are less dissimilarities between the methods, with a minor advantage for the CLM.

TABLE 1. Summary of the simulation study. Presented are the integrated mean squared error (IMSE) and the miss-classification probabilities when estimating the mixture distribution using the penalized mixture approach (PM), the back–fitting algorithm (VEM) and the composite link model (CLM).

	IMSE ($\times 10^{-5}$)			MP ($\times 10^{-2}$)		
Sample size	PM	VEM	CLM	PM	VEM	CLM
500	34.086	21.027	29.729	7.011	6.709	6.074
1000	19.218	11.578	17.274	6.682	7.029	6.413
5000	5.675	3.064	4.453	6.648	6.648	6.648

5 Conclusion

The newly developed back-fitting algorithm, with a key-role for the VEM provides an elegant way to estimate an MIC density on the full continuous scale. From this estimate, the prevalence of non-wild-type isolates can be derived. An additional advantage of the proposed method is the ability to perform model-based classification. Simulation results indicate the promising behaviour of the new method.

Acknowledgments: The research of the first author was supported by the Research Foundation Flanders (FWO), grant 11E2913N. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. For the simulations and bootstraps we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government - department EWI.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716 – 723.

Böhning, D. (1986). A vertex-exchange-method in D-optimal design theory. *Metrika*, **33**, 337 – 347.

Jaspers, S., Aerts, M., Verbeke, G., and Beloeil, P.A. (2014a). Estimation of the wild-type minimum inhibitory concentration value distribution. *Statistics in Medicine*, **33**, 289 – 303.

Jaspers, S., Aerts, M., Verbeke, G., and Beloeil, P.A. (2014b). A new semi-parametric mixture model for interval censored data, with applications in the field of antimicrobial resistance. *Computational Statistics and Data Analysis*, **71**, 30 – 42.

Jaspers, S., Lambert, P., and Aerts, M. (2015). A Bayesian approach to the semi-parametric estimation of a MIC distribution. Technical Report, I-Biostat.

Palumbi, S. (2001). Humans as the world's greatest evolutionary force. *Science*, **293(5536)**, 1786 – 1790.

Tsonaka, R., Verbeke, G., and Lesaffre, E. (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics*, **65**, 81 – 87.

Average effect measures for group comparisons in ordinal response models

Maria Kateri[1], Alan Agresti[2]

[1] Institute of Statistics, RWTH Aachen University, Germany
[2] Department of Statistics, University of Florida, USA

E-mail for correspondence: maria.kateri@rwth-aachen.de

Abstract: Ordinal measures of effect size are proposed, which describe in a simple and compact way the difference between two ordered categorical distributions, taking into consideration the effect of explanatory variables. The adjustment for explanatory variables is based on an ordinal model such as a cumulative link model. One "ordinal superiority" measure of this type summarizes the probability that an outcome from one distribution falls above an outcome from the other distribution. For the normal linear latent variable model that implies a probit link function for cumulative probabilities on the ordinal scale, this measure equals $\Phi(\beta/\sqrt{2})$, for standard normal cdf Φ and effect β that is the coefficient of the group indicator variable in the cumulative probit model. Another measure generalizes from binary to ordinal responses the difference of proportions as used to measure an average causal effect.

Keywords: Average causal effect; Cumulative link models; Logit models; Probit models; Stochastic ordering.

1 Introduction

The focus of this contribution is on constructing simple measures of effect summaries for comparing two groups on an ordinal response. Estimating an average causal effect is natural for experimental studies and it has also received much attention for observational studies. The challenge is to propose measures that adjust for other explanatory variables, while remain easy to interpret and are simpler than the effect measures based on a model for the cumulative probabilities.

For subject i, $i = 1, \ldots, n$, let z_i identify the binary group indicator variable, with $z_i = 1$ for group 1 and $z_i = 0$ for group 2, and let x_{ij} denote the value of explanatory variable j, $j = 1, \ldots, p$. For binary responses $\{y_i\}$

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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

and probabilities $\{P(y_i = 1 \mid z_i, x_{i1}, \ldots x_{ip})\}$ for a model such as logistic regression, an average causal effect measure can be defined by

$$\frac{1}{n}\sum_i \big[P(y_i = 1|z_i = 1, x_{i1}, \ldots, x_{ip}) - P(y_i = 1|z_i = 0, x_{i1}, \ldots, x_{ip})\big]. \quad (1)$$

It can be estimated by finding for each observation i the fitted probability of the outcome of interest for the given values of $x_{i1}, \ldots, x_{ip}$ (1) if that observation were in group 1 and (2) if that observation were in group 2, and averaging the differences among all n observations. It is motivated by the concept of the *average causal effect* of Rubin (1974), which is the gold-standard effect summary for group comparisons on a quantitative variable. This summarizes the difference between the overall proportions of "successes" if all subjects in the study were in group 1 compared with all being in group 2. This is a useful summary when this difference is relatively stable across observations, or when the explanatory variable values are randomly chosen from or representative of a population of interest.

In this article, we propose summary effect measures for ordinal response models, one of which simplifies for binary responses to (1), i.e. to the average difference of probabilities. The measures are defined in Section 2. Section 3 presents simple versions of the measures for a normal latent variable model that relates to a model for cumulative probabilities having the probit link. Finally, Section 4 comments on methods of constructing confidence intervals for the measures.

2 Summary ordinal effect measures in presence of explanatory variables

Starting with the case of no explanatory variables, a summary ordinal measure that compares two groups, by summarizing their relative size, is

$$\Delta = P(y_1 > y_2) - P(y_2 > y_1), \quad (2)$$

where y_1 and y_2 denote independent random variables from groups 1 and 2. For binary data with outcomes $(0, 1)$, Δ simplifies to the difference of proportions, $P(y_1 = 1) - P(y_2 = 1)$. If the ordinal y_1 and y_2 are identically distributed, then $\Delta = 0.0$. When $\Delta > 0.0$ (< 0.0), then outcomes of y_1 tend to be larger (smaller) than outcomes of y_2. A related measure that has null value equal to 0.50 rather than 0 is

$$\gamma = P(y_1 > y_2) + \frac{1}{2}P(y_1 = y_2). \quad (3)$$

The measures (2) and (3) are functionally related ($\Delta = 2\gamma - 1$), with γ having range $[0, 1]$ and Δ having range $[-1, 1]$. These "ordinal superiority"

measures are most meaningful when the groups are stochastically ordered. For details, see Agresti (2010, Chap. 2).

Now we generalize these measures for ordinal models that have p other explanatory variables. For each subject i, let $\pi_{1ij} = P(y_i = j|z_i = 1, x_{i1}, \ldots, x_{ip})$ and $\pi_{2ij} = P(y_i = j|z_i = 0, x_{i1}, \ldots, x_{ip})$, $j = 1, \ldots, c$. An ordinal superiority measure for comparing the two groups while adjusting for the explanatory variables is $\Delta = \frac{1}{n}\sum_i \Delta_i$, where

$$\Delta_i = \sum_{j>k} \pi_{1ij}\pi_{2ik} - \sum_{k>j} \pi_{1ij}\pi_{2ik}. \tag{4}$$

With no explanatory variables other than the group indicator, this simplifies to (2). An analog of the ordinal superiority measure (3) is

$$\gamma = \frac{1}{n}\sum_i \gamma_i = \frac{1}{n}\sum_i [\sum_{j>k} \pi_{1ij}\pi_{2ik} + \frac{1}{2}\sum_j \pi_{1ij}\pi_{2ij}]. \tag{5}$$

Corresponding sample values use fitted probabilities for the model.

Like the binary measure of average causal effect, this summary is relevant when the explanatory variable values are randomly chosen or representative for a population of interest. When they are not, it can be more relevant to define the measure at a representative value of interest, such as the sample mean of explanatory variable values. For such a value $\mathbf{x}_0$, one could define

$$\Delta(\mathbf{x}_0) = \sum_{j>k} \pi_{1j}(\mathbf{x}_0)\pi_{2k}(\mathbf{x}_0) - \sum_{k>j} \pi_{1j}(\mathbf{x}_0)\pi_{2k}(\mathbf{x}_0)$$

and

$$\gamma(\mathbf{x}_0) = \sum_{j>k} \pi_{1j}(\mathbf{x}_0)\pi_{2k}(\mathbf{x}_0) + \frac{1}{2}\sum_j \pi_{1j}(\mathbf{x}_0)\pi_{2j}(\mathbf{x}_0),$$

with $\pi_{\ell j}(\mathbf{x}_0) = P(y = j|z = \ell, \mathbf{x}_0)$, $\ell = 1, 2$, $j = 1, \ldots, c$. For either the overall summary versions or the covariate-constrained versions, these ordinal superiority measures are simpler to interpret than the measures that most naturally result from the link function for the model, for link functions such as the logit, probit, or complementary log-log (McCullagh 1980).

Many other ways exist to summarize ordinal effects in comparing two groups, but each has some disadvantage. For example, one could assign ordered scores to the response categories and find the difference of means, in the average causal effect structure, but that requires subjective judgment about distances between categories. Alternatively, measures such as Δ and γ could directly be modeled in terms of the explanatory variables, as in Ryu and Agresti (2008). However, such models are difficult to fit with multiple explanatory variables, and they do not have the corresponding appealing latent variable model (Section 3) that cumulative link models have.

3 Ordinal superiority measures for latent response variables

The normal latent variable model with $y_i^* \sim N(\beta z_i + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, 1)$ implies the cumulative link model

$$\Phi^{-1}[P(y_i \leq j)] = \alpha_j - \beta z_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip} \tag{6}$$

for the observed response, with $\{\alpha_j\}$ being cutpoints on the underlying scale and Φ^{-1} being the probit link function (inverse of the standard normal cdf). We can apply the ordinal superiority measures to this latent variable model. Let y_{i1}^* and y_{i2}^* denote for $z_i = 1$ and $z_i = 0$, respectively, independent underlying latent variables at $(x_{i1}, \ldots, x_{ip})$. We define

$$\gamma^* = \frac{1}{n}\sum_i P(y_{i1}^* > y_{i2}^*), \qquad \Delta^* = \frac{1}{n}\sum_i [P(y_{i1}^* > y_{i2}^*) - P(y_{i2}^* > y_{i1}^*)].$$

For model (6), it can be proved that $\gamma^* = \Phi(\beta/\sqrt{2})$ and $\Delta^* = 2\Phi(\beta/\sqrt{2}) - 1$. The measures discussed here refer to ordinal responses and are thus linked to models for ordinal response variables. Measures of this type may be useful also for other types of response variables in the context of generalized linear models. For example, in case of a quantitative response variable, they are defined based on an ordinary normal linear model.

4 Confidence intervals for ordinal effect measures

For the measure based on the normal latent variable model, simple intervals result directly from ordinary confidence intervals for β for the corresponding cumulative probit model.

Asymptotic confidence intervals for ordinal superiority measures of the type presented in Section 2 are obtained using the delta method, based on the estimated covariance matrix of the ML model parameter estimates. Confidence intervals will be illustrated on classical examples and for alternative choices of ordinal models. It is planned also to develop nonparametric versions of ordinal superiority measures that are less dependent on the assumed structural parametric form for the model.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, (2nd ed.)*. Hoboken: Wiley & Sons, Inc.

McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, **42**, 109 – 142.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688 – 701.

Ryu, E. and Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, **27**, 1703 – 1717.

Geoadditive stochastic frontier analysis

Nadja Klein[1], Helmut Herwartz[2], Thomas Kneib[1]

[1] Chair of Statistics, Georg-August-University Göttingen, Germany
[2] Chair of Econometrics, Georg-August-University Göttingen, Germany

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

Abstract: Stochastic frontier analysis is a common approach for quantifying the efficiency of economic entities when transfering input variables into a desired output. The standard model specification considers a linear production frontier and derives efficiency scores by predicting the distance of the observed performance to this frontier. We embed stochastic frontier analysis in the framework of distributional regression to allow for (i) the inclusion of nonlinear and spatial effects, (ii) the inclusion of covariate effects not only the production frontier but also on the inefficiency scores and (iii) the adaptation of efficient Markov chain Monte Carlo simulation schemes developed for distributional regression. As an empirical illustration, we analyse the efficiency of farms in England and Wales and find that spatial patterns of inefficiency are clearly supported by model diagnostics.

Keywords: Distributional regression; Farm efficiency; Markov chain Monte Carlo simulation; Markov random fields; Spatial heterogeneity.

1 Introduction

The notion of technical efficiency refers to the ability of firms (or other economic entities) to maximise their output given their input variables. In a naive approach, one might consider a linear regression specification for determining the input-output relationship (i.e. the production function). In a second step, the estimated regression line would then be shifted upwards by the largest observed residual to determine the production frontier while the distance between an observation and the production frontier in turn determines the firms inefficiency. This approach is illustrated in Figure 1. To acknowledge potential random perturbations in the achieved production, standard stochastic frontier analysis (SFA) considers the model specification

$$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}, \qquad i = 1, \ldots, n,\, t = 1, \ldots, T_i,$$

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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

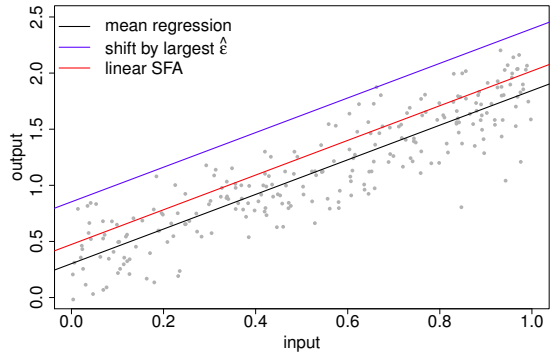

FIGURE 1. Linear input-output relationship (black line) and production frontiers obtained from a naive approach (shift by the largest residual, blue line) and by stochastic frontier analysis (red line).

for output variable y_{it} and input variable x_{it} observed on n firms in a (potentially unbalanced) panel design with the composed error term $\varepsilon_{it} = v_{it} - u_{it}$ consisting of the idiosyncratic error term $v_{it} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_v^2)$ and the inefficiency term $u_{it} \overset{\text{i.i.d.}}{\sim} \mathrm{N}^+(0, \sigma_u^2)$ following a half normal distribution. Firm-specific inefficiencies are then determined by the conditional expectations

$$E(u_{it}|\varepsilon_{it}) \quad \text{or} \quad E(\exp(-u_{it})|\varepsilon_{it})$$

where the latter has the advantage to be bounded in $[0, 1]$. As indicated in Figure 1, this specification shifts the production frontier downwards in acknowledgement of the random component v_{it}.

In our application, we are interested in the efficiency of farms in England and Wales concerned with the production of cereal. In our data set (collected by the Department of Environment, Food and Rural Affairs and National Assembly for Wales, Farm Business Survey, 2006–2007), we have 882 observations on 237 distinct farms between 2003 and 2007. In addition to common covariates such as land size, capital, expenditures for fertilisers, seed cost and crop protection cost, the data set also comprises spatial information in terms of districts.

Standard stochastic frontier models have a number of limitations that prevent their application for this data set:

- All covariates impact only the production function but no firm-specific effects on inefficiency can be included.
- Only linear predictors and in particular no spatial variation are included.
- Inference is challenging in a likelihood-based framework.

We therefore embed geoadditive stochastic frontier analysis in the framework of distributional regression of Klein et al. (2015) and propose a corresponding Bayesian inferential scheme.

2 Geoadditive stochastic frontier analysis

Geoadditive stochastic frontier analysis generalizes the standard stochastic frontier model to

$$y_{it} = \eta_{it}^{(y)} + v_{it} - u_{it}^* \alpha_{it} = \eta_{it}^{(y)} + v_{it} - u_{it}^* \exp\left(\eta_{it}^{(\alpha)}\right),$$

where $\eta_{it}^{(y)}$ is the regression predictor representing the production function, $\eta_{it}^{(\alpha)}$ is the regression predictor comprising (multiplicative) effects on the inefficiencies, $v_{it} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_v^2)$ is the idiosyncratic error and $u_{it} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu^*, (\sigma_u^*)^2)$ represents the base model for the inefficiencies. Note that the scaled inefficiencies $u_{it}^* \alpha_{it}$ again follow a truncated normal distribution with parameters

$$\mu_{it} = \mu^* \alpha_{it} \quad \text{and} \quad (\sigma_u^2)_{it} = (\sigma_u^*)_{it}^2 \, \alpha_{it}$$

such that both the expectation and the variance parameter depend on the inefficiency predictor $\eta_{it}^{(\alpha)}$. Setting $\mu^* = 0$ yields the standard half normal specification albeit with covariate-specific variances. If desired, we could also place regression regression predictors on μ^*, σ_v^2 and $(\sigma_u^*)^2$ to obtain additional flexibility but we will not pursue this route here.
For both predictors in our geoadditive SFA model, we assume a structured additive form

$$\eta = \beta_0 + f_1(z) + \cdots + f_J(z),$$

where $f_1, \ldots, f_J$ are generic functions of the covariate vector z. This allows us to include a variety of effects in the SFA framework, such as nonlinear, smooth effects of continuous covariates, spatial effects or random effects. In any case, each function is approximated in terms of (possibly nonstandard) basis functions such that the function evaluations can be represented as the product of a design matrix Z_j and a vector of regression coefficients γ_j. The latter is assigned a multivariate normal prior

$$p(\beta_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} \gamma_j K_j \gamma_j\right),$$

where τ_j^2 is a prior precision parameter that determines our prior confidence while the precision matrix K_j enforces properties such as smoothness of shrinkage.

To obtain a distributional regression formulation of SFA models, we determine the distribution of the composed error term as

$$p(\varepsilon_{it}) = \frac{\Phi\left(\frac{\sigma_u\sigma_v}{\sqrt{\sigma_u^2+\sigma_v^2}}\left(\frac{\mu_{it}}{\sigma_u^2} - \frac{\varepsilon_{it}}{\sigma_v^2}\right)\right)}{\sqrt{2\pi\,(\sigma_u^2+\sigma_v^2)}\Phi\left(\frac{\mu_{it}}{\sigma_u}\right)} \exp\left(-\frac{(\varepsilon_{it}+\mu_{it})^2}{2\,(\sigma_u^2+\sigma_v^2)}\right),$$

which provides us with with the appropriate likelihood. The (conditional) density of the inefficiencies can then be obtained as

$$p(u_{it}|\varepsilon_{it}) = p(u_{it}, \varepsilon_{it})/p(\varepsilon_{it}),$$

which again gives access to the inefficiency scores defined above in the context of standard SFA models.
Bayesian inference can then proceed along the lines of Bayesian distributional regression by constructing proposal densities for all regression effects based on multivariate normal approximations fitting the mode and curvature at the mode of the log full conditional to obtain a Markov chain Monte Carlo simulation algorithm. More specifically, a typical vector of regression coefficients is proposed from a multivariate normal distribution with expectation and precision matrix

$$\mu_j = P_j^{-1} Z_j' W(\tilde{y} - \eta_{-j}), \qquad P_j = Z_j' W Z_j + \frac{1}{\tau_j^2} K_j,$$

where W is a diagonal matrix of working weights, $\tilde{y}$ is a vector of working observations and η_{-j} is a partial predictor. Bayesian inference also supports model choice via the deviance information criterion (DIC).

3 Farm efficiency analysis

In our application, we consider the following model specification for the production function:

$$\begin{aligned}\eta_{it}^{(y)} &= \beta_0^{(y)} + \beta_1^{(y)} lcapital_{it} + \beta_2^{(y)} lland_{it} \\ &\quad + \beta_3^{(y)} lfert_cost_{it} + \beta_4^{(y)} lseed_cost_{it} + \beta_i^{(y)}\end{aligned}$$

combining effects of capital, land size, fertiliser costs, seed costs (all log-transformed) and an i.i.d. random effect on the farm level. For the inefficiency predictor, we considered five different specifications:

- $m = 1$: Structured and unstructured spatial effects

$$\eta_{it}^{(\alpha)} = f_{str}(dist_i) + f_{unstr}(dist_i) + \delta_1^{(\alpha)} lcrop_prot_{it}.$$

TABLE 1. DIC values for the five different spatial model specifications.

m	1	2	3	4	5
DIC	730	762	791	732	730

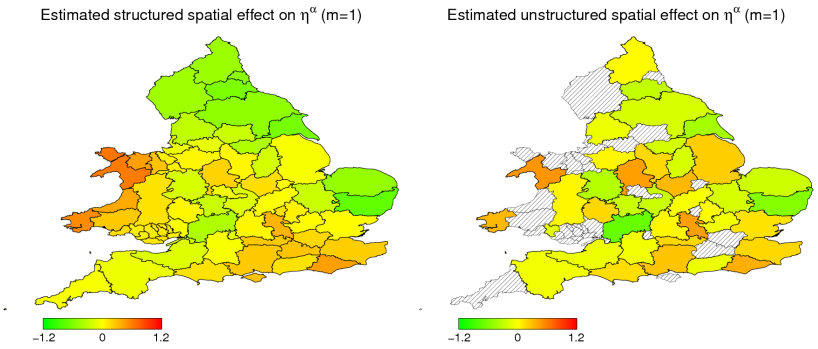

FIGURE 2. Estimated spatially structured (left) and unstructured (right) spatial effect in model $m = 1$.

- $m = 2$: No spatial effect

$$\eta_{it}^{(\alpha)} = \delta_1^{(\alpha)} lcrop_prot_{it}.$$

- $m = 3$: Constant variance

$$\eta_{it}^{(\alpha)} = 1.$$

- $m = 4$: Only spatially structured effects

$$\eta_{it}^{(\alpha)} = f_{str}(dist_i) + \delta_1^{(\alpha)} lcrop_prot_{it}.$$

- $m = 5$: Only spatially unstructured effects

$$\eta_{it}^{(\alpha)} = f_{unstr}(dist_i) + \delta_1^{(\alpha)} lcrop_prot_{it}.$$

with a combination of a spatially correlated effect f_{str}, a spatially uncorrelated effect f_{unstr} and a linear effect of crop protection costs (again log-transformed) as the most complex model specification. The resulting values of the DIC shown in Table 1 strongly support the inclusion of a spatial effect but do not strongly distinguish between spatially structured, spatially unstructured and the composed spatial specification. Due to the additional insights obtained by a decomposition, we show the spatial effects from model 1 in Figure 2.

4 Summary

Embedding stochastic frontier analysis in the framework of distributional regression allows us to flexibly specify regression predictors on various parts of the model and in particular the production function and the inefficiencies. Efficient inference can be implemented by relying on Bayesian Markov chain Monte Carlo simulation schemes with iteratively weighted least squares proposals.

References

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany. *Annals of Applied Statistics*, to appear.

Scale-dependent priors for variance parameters in distributional regression

Nadja Klein[1], Thomas Kneib[1]

[1] Georg-August-Universität Göttingen, Germany

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

Abstract: The selection of appropriate hyperpriors for variance parameters is a sensible topic in all kinds of Bayesian regression models. We consider the special case of structured additive distributional regression where Gaussian priors are used to enforce specific properties such as smoothness or shrinkage on various effect types. Relying on a recently proposed class of priors motivated from a set of simple principles, we come up with a hyperprior structure where prior elicitation is facilitated by assumptions on the scaling of the effect types.

Keywords: Markov chain Monte Carlo simulations; Penalised complexity prior.

1 Introduction

Structured additive regression models are an important model class for regression modelling in various areas of applications. They combine the flexibility of generalised additive models with the inclusion of random effects, spatial components and further types of regression effects. While originally being restricted to responses from the exponential family, structured additive regression has recently been extended to a much broader class of response types known as distributional regression (Klein et. al, 2013).
In these models, it is assumed that the (conditionally) independent response variables y_i, $i = 1, \ldots, n$, given some covariate information $\boldsymbol{\nu}_i$ follow parametric distributions with density $p(y_i|\vartheta_{i1}, \ldots, \vartheta_{iK})$ and distribution parameters ϑ_{ik}, $k = 1, \ldots, K$. Each of the latter is linked to a structured additive predictor η_{ik} via a suitable one-to-one transformation h_k, i.e. $h_k(\eta_{ik}) = \vartheta_{ik}$. Dropping the parameter index k, the predictors are composed additively as

$$\eta_i = \beta_0 + \sum_{j=1}^{J} f_j(\boldsymbol{\nu}_i)$$

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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

where, in turn, each function $f_j(\boldsymbol{\nu}_i)$ is represented by a linear combination of basis functions such that (suppressing i, j) $f(\boldsymbol{\nu}) = \sum_{d=1}^{D} \beta_d B_d(\boldsymbol{\nu})$. Here, $B_d(\boldsymbol{\nu})$, $d = 1, \ldots, D$, is a set of appropriate basis functions while $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)'$ is the vector of basis coefficients to be estimated.
To enforce specific properties such as smoothness we impose partially improper multivariate normal priors

$$p(\boldsymbol{\beta}|\tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}\right)$$

with fixed positive (semi-)definite precision matrix $\mathbf{K}$ and variance parameter τ^2 that inherits the role of the smoothing parameter.
Suitable hyperpriors have then to be augmented to these variance components to complete the Bayesian model specification. While the inverse gamma prior $p(\tau^2) \sim \text{IG}(a, b)$ is a natural, conjugate prior, there has been considerable debate about the suitability of the inverse gamma distribution especially in the context of hierarchical random effects models. As a consequence, several alternatives such as half normal, half Cauchy or (proper) uniform priors for the standard deviation have been suggested as default priors in the literature (Gelman, 2006). Unfortunately, prior elicitation of the hyperparameters of these priors ensuring the propriety of the posterior and justification of the chosen distribution type with respect to axiomatic reasoning are often quite problematic in these cases.
Without relying on a specific modelling context, Simpson et al. (2014) develop a general approach for determining so-called penalised complexity priors reflecting that frequently hyperpriors are desired for parameters governing the deviation of a flexible model from a restrictive base model. We utilise this approach to develop scale-dependent priors for the variance parameters in structured additive distributional regression.

2 Scale-dependent hyperpriors

Omitting the distribution parameter index k as well as the effect index j we assume that $f(\boldsymbol{\nu})$ is one generic function of $\boldsymbol{\nu}$ included in a generic predictor $\boldsymbol{\eta}$.

2.1 Construction

The prior distribution $p(\tau^2|\theta)$ for the smoothing variances τ^2 constructed according to the principled definition of priors developed in Simpson et al. (2014) is a Weibull distribution with shape parameter $\alpha = 1/2$ and scale parameter θ, i.e. $p(\tau^2|\theta) = \frac{1}{2\theta}(\tau^2/\theta)^{-1/2} \exp\left(-(\tau^2/\theta)^{1/2}\right)$, as derived in the following:
1. *Occam's razor.* The Weibull prior invokes the principle of parsimony in the sense that a suitable base model for each effect is preferred as long as the data provide enough evidence for a more complex modelling.

2. *Measure of complexity.* The increased complexity between two models is measured by the unidirectional measure $d(p||p_b) = \sqrt{2\text{KLD}(p||p_b)}$ where $\text{KLD}(p||p_b)$ denotes the Kullback-Leibler divergence (KLD) between the base model represented by density p_b and the alternative represented by density p. Let $\text{N}(\mathbf{0}, \tau^2\mathbf{K}^-)$ denote the flexible model for a vector of regression coefficients and $\text{N}(\mathbf{0}, \tau_b^2\mathbf{K}^-)$ the base model with $\tau_b^2 \to 0$ and $\mathbf{K}^-$ the generalised inverse of $\mathbf{K}$. For $\tau^2 \gg \tau_b^2$ and $\tau_b^2 \to 0$ we then obtain $\text{KLD} \to \frac{\kappa}{2}\frac{\tau^2}{\tau_b^2}$ and hence a distance measure $d(\tau^2) = \sqrt{\frac{\kappa}{\tau_b^2}}(\tau^2)^{1/2}$.

3. *Constant rate penalisation.* This assumption implies an exponential prior on the distance scale $p_d(d) = \lambda \exp(-\lambda d)$ and finally delivers the prior on the original space as

$$p(\tau^2) = \lambda \exp(-\lambda d(\tau^2)) \left| \frac{\partial d(\tau^2)}{\partial \tau^2} \right| = \frac{\lambda}{2} \sqrt{\frac{\kappa}{\tau_b^2}} (\tau^2)^{1/2} \exp\left(-\lambda \sqrt{\frac{\kappa}{\tau_b^2}} (\tau^2)^{1/2} \right).$$

Setting $\theta = (\lambda \sqrt{\frac{\kappa}{\tau_b 2}})^{-1/2}$ gives the prior above which we call scale-dependent hyperprior.

2.2 Choosing the scale parameter – user-defined scaling

The last principle controls the decay-rate $\exp(-\lambda)$ by imposing the condition $\text{P}(g(\tau^2) \leq c) = 1 - \alpha$ for an interpretable transformation g of τ^2 and some user-defined values c and α. Compared to random effects model mostly considered in Simpson et al. (2014) we are interested in relating the scale parameter θ to the functions f rather than directly to the variances τ^2. This is achieved by specifying a certain interval the function f falls into with a high marginal probability

$$\text{P}(|f(x)| \leq c \, \forall x \in \mathcal{D}) \geq 1 - \alpha$$

where $\alpha \in (0, 1)$, $c > 0$ are chosen in advance and $\mathcal{D}$ is the domain of x. To solve the problem above, we use a finite subset $\mathcal{X}_P = \{x_1, \ldots, x_P\}$ of $\mathcal{D}$ together with the Bonferroni inequality to arrive at

$$\text{P}(|f(x_p)| \leq c \, \forall x \in \mathcal{X}_P) \geq 1 - \sum_{p=1}^{P} \text{P}\left(|f(x_p)| \geq c\right).$$

The marginal density of $f(x_p)$ can be obtained by integrating τ^2 out and the optimal scale parameter θ with respect to this criterion is obtained numerically.

Example: Bayesian P-Splines

While for fixed order of the penalty, $\mathbf{K}$ is the always same, the design matrix $\mathbf{Z}$ is dependent on the observations itself as well as the knots and

the number of basis functions. We investigated empirically how much the optimal value θ depends on these factors as well as the loss of efficiency through the inequality when switching from a simultaneous statement to a pointwise approximation and found that the absolute values of θ change only on a very small scale. This allowed us to define default values for different c and α which themselves have been tested in several simulations where we also evaluated the performance compared to usual IG priors. The basic results are summarised in the following:
1. In situations near to the base model, i.e. the true effect is close to linear, the scale-dependent prior delivers smaller credible intervals that still maintain the desired level and can thus be seen as a strong alternative to IG priors in these cases.
2. When the true effect is strongly nonlinear or of a more complex pattern, performance of IG and scale-dependent priors is very similar with respect to bias, mean squared errors and credible intervals.
3. Scale-dependent priors reveal to be advantageous with respect to numerical stability whenever the likelihood is flat such as in a binary model with either high or low success probability.

3 Childhood undernutrition in Zambia

As an illustration, we use data on 4421 children in Zambia. For each child i undernutrition is measured by a Z-score

$$zscore_i = \frac{h_i - m}{s}$$

reflecting the nutritional status of child i with height h_i in the population of interest. The values m and s correspond to the mean height of children and their standard deviation in a suitable reference population of the same age group and gender.
For convenience, we assume a location-scale model, that is, the Z-scores are conditionally normally distributed with

$$\begin{aligned} zscore_i &= \beta_0 + f_1(cage_i) + f_2(mage_i) + f_3(mbmi_i) + f_{spat}(district_i) + \varepsilon_i, \\ \varepsilon_i &\sim \mathrm{N}(0, \sigma_i^2) \\ \log(\sigma_i^2) &= \tilde{\beta}_0 + \tilde{f}_1(cage_i) + \tilde{f}_2(mage_i) + \tilde{f}_3(mbmi_i) + \tilde{f}_{spat}(district_i). \end{aligned}$$

In the equations above, f_1 to f_3 ($\tilde{f}_1$ to $\tilde{f}_3$) are smooth functions of the continuous covariates *cage* (child's age), *mage* (mother's age at birth) and *mbmi* (mother's body mass index), while f_{spat} ($\tilde{f}_{spat}$) represents the spatial effect that was assigned a Markov random field prior, and β_0 ($\tilde{\beta}_0$) is the usual overall intercept. For all nonparametric effects, we use inverse gamma priors with default hyperparameters $a = b = 0.001$ and compare the results to scale-dependent priors.

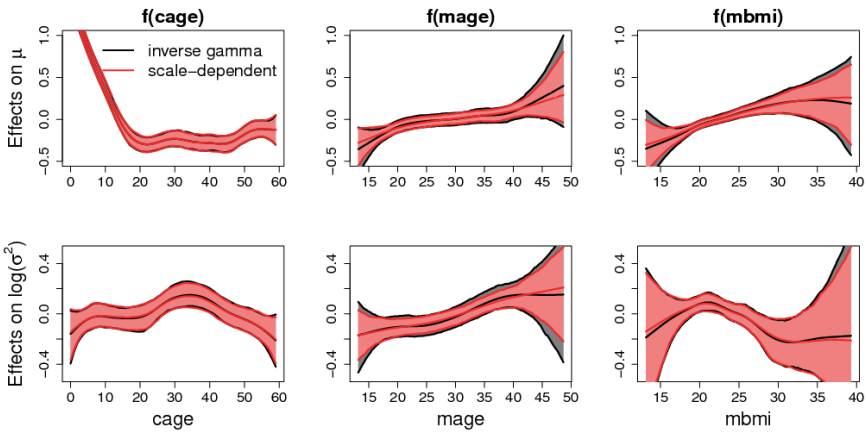

FIGURE 1. Childhood undernutrition. Comparison of estimated nonlinear effects of continuous covariates *cage*, *mage* and *mbmi*. Shown are posterior means and 95% credible intervals on E(*zscore*) (top) and on $\log(\sigma^2)$ (bottom).

Figure 1 is in accordance with the simulations and shows that the effects of *mage* and *mbmi* on the conditional mean are estimated to be closer to a linear effect with smaller credible intervals under the new scale-dependent prior. The spatial effect in Figure 2 is estimated similiar with both hyperpriors which is reasonable in this case since the variable *district* has significant impact on both distribution parameters (based on a 95% credible interval).

References

Gelman, A. (2006). Prior distributions for variance parameters in hierarchichal models. *Bayesian Analysis*, **1**, 515 – 533.

Klein, N., Kneib, T., and Lang, S. (2013). Bayesian Structured Additive Distributional Regression. *Technical report, available at http://eeecon.uibk.ac.at/wopec2/repec/inn/wpaper/2013-23.pdf.*

Simpson, D., Martins, T., Riebler, A., Fuglstad, G., Rue, H., and Sørbye, S. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *Technical report, available at http://arxiv.org/abs/1403.4630.*

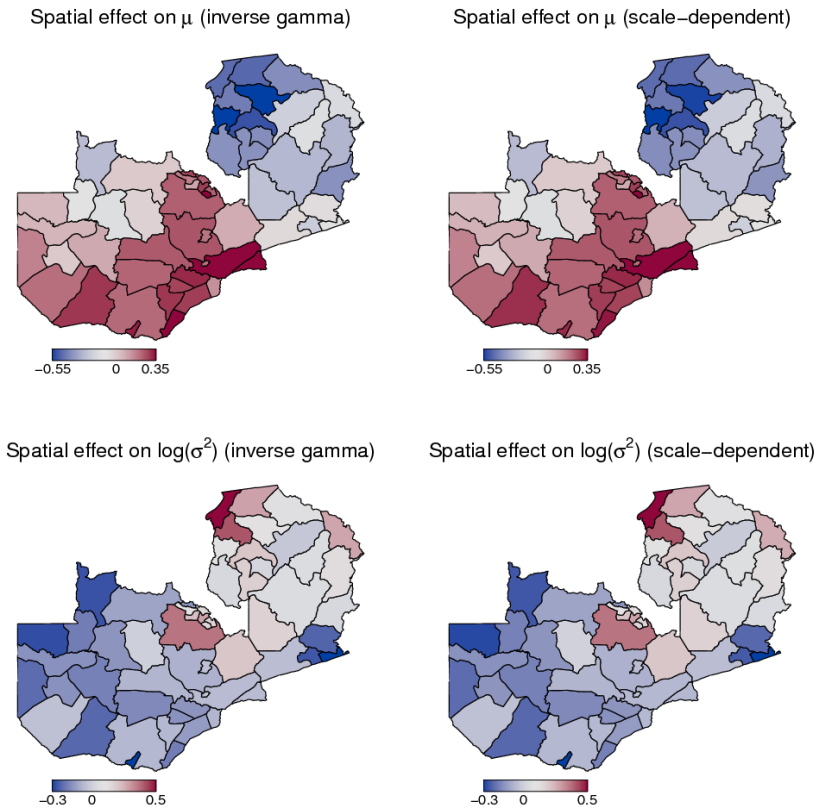

FIGURE 2. Childhood undernutrition. Comparison of estimated spatial effects for the inverse gamma prior (left) and scale-dependent prior (right). Shown are posterior means on E($zscore$) (top) and on $\log(\sigma^2)$ (bottom).

Multiple comparisons of marginal probabilities following GEE estimation

Bernhard Klingenberg[1]

[1] Department of Mathematics & Statistics, Williams College, Williamstown, MA 01267, U.S.A.

E-mail for correspondence: bklingen@williams.edu

Abstract: We consider fitting a simple model for the marginal probabilities of a multivariate binary response using generalized estimating equations (GEE). In many applications, the marginal probabilities refer to different groups (such as treatments, doses of a drug or questions on a survey) and the ultimate goal is to estimate contrasts between them, such as comparing all possible pairs of marginal probabilities or comparing all to a common control. In this article, we use the GEE estimates together with their robust covariance matrix to conduct a post-hoc analysis under strong Type I error control. We also explore a model-free approach that uses a maximum score statistic. This alternative is attractive because Wald-type inference following GEE may be unreliable, as simulations confirm. We illustrate our procedures with binary data from a cross-over trial on air pollution that compares probabilities of distress under 4 exposure levels.

Keywords: All pairwise comparisons; Familywise error rate; Marginal models; Multivariate binary; Simultaneous inference.

1 Introduction

The overwhelming advantage of measuring different treatments on the same subject is that each subject serves as his or her own control, enabling a direct comparison of the treatments and often resulting in smaller standard errors for treatment contrasts. When more than two treatments are involved, usually several contrasts are investigated simultaneously, such as all comparisons to control or all pairwise comparisons among the treatments. For instance, in a cross-over study about the effect of air pollution on subjects with chronic respiratory illness, each subject was exposed to low, medium and high levels of pollution while exercising on a stationary bike. Before the exercise started, an initial distress level was also measured.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

The binary response in each case was whether the subject experienced distress, and the goal was to conduct pairwise comparisons of the probabilities of experiencing distress under the three exposure levels and how these differ from the baseline. Table 1 shows observed counts for the possible response sequences under low, medium and high exposure, separated by the response on the baseline distress levels.

TABLE 1. Counts for binary response sequences for a respiratory cross-over trial. Response sequences refer to no distress (0)/distress (1) for (low, medium, high) levels of exposure, for subjects with no initial distress (base = 0) and initial distress (base = 1).

	(Low, Medium, High)							
	000	001	010	011	100	101	110	111
Base = 0:	0	7	1	29	2	15	1	53
Base = 1:	0	4	0	16	1	6	3	12

The methods presented here equally apply to studies in the social sciences. For instance, the General Social Survey asks the question whether abortion should be legal under each of four conditions: When there is a large chance of a birth defect, when the women is not married, when the pregnancy was the result of a rape or when the women is poor. Then, contrasting the proportions of those that think it should be legal for all 6 possible pairings of conditions is of interest.

Consider, for each subject, a multivariate response denoted by $\mathbf{Y} = (Y_1, \ldots, Y_c)$ consisting of c correlated Bernoulli r.v.'s Y_j with marginal success probability $P(Y_j = 1) = \pi_j$, summarized in the vector $\boldsymbol{\pi}_{\text{marg}} = (\pi_1, \ldots, \pi_c)$. An important question centers around comparing the π_j's with each other (or to a control) inferentially. For instance, for the data in Table 1, we would like to see if there is a significant difference between any of the $c = 4$ exposure levels (baseline, low, medium, high) in terms of the probability of experiencing distress. We will use pairwise differences of marginal probabilities to evaluate such effects.

In Section 1.1 we investigate post-hoc comparisons of marginal probabilities after fitting a marginal model for the multivariate binary response via GEE (Diggle et al., 1994). Since the GEE approach is not likelihood based, we have to rely on Wald-type inference for the pairwise comparisons, which may or may not perform satisfactorily under small to moderate sample sizes. Therefore, in Section 1.2, we look at a model free approach, using a maximum score statistic to conduct simultaneous inference.

We will present multiplicity adjusted P-values for the post-hoc pairwise comparisons. Multiplicity-adjusted refers to the fact that the set of tests we consider (one for each pair of marginal probabilities) controls the familywise error rate (FWER), i.e., the probability of at least one Type I error among all decisions made. This is in contrast to ignoring the multiple inferences,

which would result in an overall error rate that is unknown. Therefore, it is better to control the FWER at some known level α so that precise (asymptotic) error statements can be given.

1.1 Post-hoc comparisons following GEE modeling

Using GEE, we can model the vector of marginal success probabilities as

$$g(\boldsymbol{\pi}_{\text{marg}}) = \mathbf{x}^{\text{T}}\boldsymbol{\beta}$$

without the need to specify a multivariate distribution for $\mathbf{Y}$. Here, g is a specific link function applied elementwise, $\mathbf{x}$ is a vector of subject-specific explanatory variables and $\boldsymbol{\beta}$ is the corresponding parameter vector. In this article, we assume that $\mathbf{x}$ only contains indicators for the c treatments, i.e., $\mathbf{x} = I_c$, with I_c the identity matrix of dimension c. (Note that there is no intercept.) Also, we only consider the identity link function $g(\pi_j) = \pi_j$ since we want to estimate contrasts on the probability scale. Results can be extended to allow other link functions and covariates.
In effect, the marginal model we consider here reduces to $\boldsymbol{\pi}_{\text{marg}} = \boldsymbol{\beta}$. After assuming a specific form for the working correlation matrix, let $\hat{\boldsymbol{\beta}}$ denote the GEE estimator of $\boldsymbol{\beta}$, which is asymptotically multivariate normal (MVN), with estimated variance-covariance matrix $\hat{\Sigma}$ (the robust "sandwich" matrix). With such a simple model, it is not surprising that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\pi}}_{\text{marg}} = (\hat{\pi}_1, \ldots, \hat{\pi}_c)^{\text{T}}$, the vector of the marginal sample proportions, *regardless* of the assumed working correlation structure. Consequently, $\hat{\Sigma}$ also does not depend on the assumed structure for the working correlation. It is the same whether we assume independence, exchangeable, unstructured or any other form for the working correlation between the c binary responses. Let $\mathbf{y}_i$ be the observed vector of the c binary responses for subject i, $i = 1, \ldots, n$, such as $\mathbf{y}_i = (0, 1, 1, 0)^{\text{T}}$ for one of the $n = 150$ subjects in the air pollution study. Then, $\hat{\Sigma}$ is given by

$$\hat{\Sigma} = \widehat{\text{Cov}}(\hat{\boldsymbol{\pi}}_{\text{marg}}) = \widehat{\text{Cov}}\left(\frac{1}{n}\sum_i \mathbf{y}_i\right) = \frac{1}{n^2}\sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^{\text{T}},$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_i \mathbf{y}_i$.
Now, our interest lies in testing multiple contrasts on the parameter vector $\boldsymbol{\beta}$. Let C describe a contrast matrix of dimension $r \times c$, with rows $\{c_l\}_{l=1}^{r}$, such as $c_1 = (1, -1, 0, 0)$ for the first of the $r = 6$ possible pairwise comparisons when considering the $c = 4$ exposure levels in the air pollution study. The global null hypothesis to test these contrasts is $H : C\boldsymbol{\beta} = \mathbf{0}$. Under H, $\hat{\boldsymbol{\theta}} = C\hat{\boldsymbol{\beta}}$ is MVN with mean $\mathbf{0}$ and estimated covariance matrix $\hat{\Gamma} = C\hat{\Sigma}C^{\text{T}}$. Using a quadratic form Wald-type test statistic for H (such as $\hat{\boldsymbol{\theta}}^{\text{T}}\hat{\Gamma}^{-1}\hat{\boldsymbol{\theta}}$) will only indicate whether at least one of the contrasts is significant, but

will not tell which one. This is why we propose $\max_l |z_l|$ as a test statistic of H, where z_l is the lth component of the vector $\mathbf{z} = \text{diag}(\hat{\Gamma})^{-1/2}\hat{\boldsymbol{\theta}}$, or $z_l = c_l^{\mathrm{T}}\hat{\boldsymbol{\pi}}_{\text{marg}}/c_l^{\mathrm{T}}\hat{\Sigma}c_l$. For pairwise comparisons, i.e., when c_l has exactly one element equal to 1, one element equal to -1 and the remaining elements equal to 0, z_l is simply the Wald statistic for comparing two dependent proportions in a 2×2 contingency table formed by cross-classifying the two responses under considerations. We can think of each as the test statistic for the elementary null hypotheses $H_l : c_l^{\mathrm{T}}\boldsymbol{\beta} = 0$, which are related to the global null hypotheses through $H = \bigcap_{l=1}^{r} H_l$.
Because $\mathbf{z}$ is MVN with mean $\mathbf{0}$ and estimated covariance (=correlation) matrix $R = \text{diag}(\hat{\Gamma})^{-1/2}\hat{\Gamma}\text{diag}(\hat{\Gamma})^{-1/2}$, we reject H if $\max_l |z_l| \geq u_{1-\alpha}$, the upper $1-\alpha$ equidistant quantile of a mean-zero MVN r.v. with correlation matrix R. An elementary hypothesis H_l is rejected if $|z_l| \geq u_{1-\alpha}$. By the closure principle (for a description, see Bretz et al., 2010), these decisions strongly (i.e., under any setting of true and false null hypotheses $\{H_l\}$) control the FWER at level α. Note that this setup also guarantees that if the global null hypothesis is rejected, at least one of the elementary null hypotheses is rejected. The multiplicity adjusted P-value for an elementary hypothesis is defined as the smallest significance level for which the elementary hypothesis is rejected when controlling the FWER for all elementary hypothesis at level α. These multiplicity adjusted P-values can be directly compared to α to decide whether to reject any H_l.

1.2 Maximum score statistic

For pairwise comparisons, we noted above that the test statistic z_l for the elementary null hypothesis H_l can be computed from the (marginal) 2×2 table cross-classifying the responses of the two treatments that are compared. Then, an alternative to the Wald test statistic z_l is McNemar's test statistic $s_l = (b-c)/\sqrt{b+c}$, which is the score test. Here, b and c refer to the off-diagonal elements of the 2×2 contingency table. To test the family of elementary null hypotheses H_l, we now use the maximum score statistic $\max_l |s_l|$. Because z_l and s_l are asymptotically equivalent under the (global) null, the vector of score statistics $\mathbf{s} = (s_1, \ldots, s_l)$ is also asymptotically MVN with mean $\mathbf{0}$ and estimated correlation matrix R. Hence, both approaches use the same multivariate reference distribution to compute multiplicity adjusted P-values, but differ in the statistic used. The next section compares the two approaches in terms of actual FWER control and illustrates with the air pollution data.

2 Simulations and example

For the data on air pollution in Table 1, the sample estimates for the marginal probabilities $\pi_j = P(Y_j = 1)$ and odds-ratios $\rho_{jj'} = P(Y_j =$

TABLE 2. Simulation study to estimate FWER and power under various configurations on marginal probabilities and odds ratios, when $n = 100$. Scenario I: $\boldsymbol{\pi}_{\text{marg}} = (0.1, 0.1, 0.1, 0.1)$, $\rho = 3$. Scenario II: $\boldsymbol{\pi}_{\text{marg}} = (0.2, 0.2, 0.2, 0.2)$, $\rho = 3$. Scenario III: $\boldsymbol{\pi}_{\text{marg}} = (0.3, 0.3, 0.3, 0.3)$, $\rho = 3$. Scenario IV: $\boldsymbol{\pi}_{\text{marg}} = (0.2, 0.2, 0.2, 0.3)$, $\rho = 3$. Scenario V: $\boldsymbol{\pi}_{\text{marg}} = (0.2, 0.2, 0.5, 0.5)$, $\rho =$ Table 2. Scenario VI: $\boldsymbol{\pi}_{\text{marg}} =$ Table 2, $\rho =$ Table 2.

	Power GEE				Power Score			
Scenario	**FWER**	H_1	H_2	H_6	**FWER**	H_1	H_2	H_6
I	**0.065**	0.01	0.01	0.01	**0.039**	0.01	0.01	0.01
II	**0.053**	0.01	0.01	0.01	**0.044**	0.01	0.01	0.01
III	**0.053**	0.01	0.01	0.01	**0.043**	0.01	0.01	0.01
IV	**0.031**	0.01	0.01	0.26	**0.023**	0.01	0.01	0.23
V	**0.025**	0.01	0.99	0.01	**0.019**	0.01	0.99	0.01
VI	**-**	0.93	0.99	0.98	**-**	0.92	0.99	0.98

$1, Y_{j'} = 1)P(Y_j = 0, Y_{j'} = 0)/P(Y_j = 1, Y_{j'} = 0)P(Y_j = 0, Y_{j'} = 1)$ are $\hat{\pi}_1 = 0.28, \hat{\pi}_2 = 0.62, \hat{\pi}_3 = 0.77, \hat{\pi}_4 = 0.94, \hat{\rho}_{12} = 0.6, \hat{\rho}_{13} = 0.8, \hat{\rho}_{14} = 0.4, \hat{\rho}_{23} = 0.7, \hat{\rho}_{24} = 0.2$ and $\hat{\rho}_{34} = 2.1$, showing i.) widely varying marginal probabilities and ii.) moderately strong associations among the 4 binary responses.

To generate multivariate binary data $(Y_1, \ldots, Y_c)$ for a simulation study to investigate FWER control and power, we specify various configurations for these marginal probabilities and odds ratios. These, in turn, define a correlation matrix that can be used to generate multivariate binary responses with the given marginal probabilities and odds ratios via truncation of the MVN distribution. (R library "bindata", Leisch et al., 1998.) Under scenarios I, II, and III shown in Table 2, all 4 marginal probabilities were identical, and the response vector had an exchangeable association structure as expressed by a common odds ratio of 3. For these cases, the GEE approach failed to control the FWER at the prescribed 5% level, while the score approach does seem to control it. For scenario IV, the last marginal probability differed from the remaining ones. Under this scenario, a Type I error occurs when any of the true null hypothesis H_1, H_2 or H_4 are rejected. This occurred in 3.1% of the 10,000 simulations with the GEE post-hoc analysis and in 2.3% when using the maximum score approach, both smaller than the nominal 5% FWER. The power of correctly rejecting H_6 under this scenario equals 26% for the GEE and 23% under the score approach. Scenarios V and VI used the data from the pollution study to model the association. Under scenario V, the first two marginal probabilities are identical at 0.2, and the second two marginal probabilities are identical at 0.5, so that hypothesis H_1 and H_6 are true, and H_2, H_3, H_4 and H_5 are false. This leads to FWER's of 2.5% and 1.9%, respectively, and the probability of correctly rejecting H_2 is over 99% for both approaches.In Scenario VI the

data are simulated under the same structure as the observed air pollution data. This was done to evaluate power when $n = 100$, which we see is very hight for all elementary hypothesis.
Finally, we return to the example of the air pollution data. Table 3 shows results for applying the methods of this article. Although both methods provide similar results, based on the preliminary simulation results, the score method has better control of the Type I error probability. We conclude that there is a significant difference (all multiplicity adjusted P-values $<$ 0.05) in the proportion of subjects experiencing distress between any pair of air pollution exposure levels. The probability of at least one Type I error in these conclusions is bounded by 5%.

TABLE 3. Multiplicity adjusted P-values for all pairwise comparisons of the probabilities for distress under the four exposure levels (B=baseline, L=low, M=medium, H=high), using post-hoc GEE analysis ($p_{\text{adj}}^{\text{GEE}}$) and the maximum score statistic ($p_{\text{adj}}^{\text{maxS}}$). Note that the more conservative Bonferroni multiplicity adjustment for the maximum score statistic for comparisons $\pi_L - \pi_M$ is equal to 0.0513, which, under an FWER of $\alpha = 5\%$, would not be significant.

Contrast	Est.	S.E.	z_l	$p_{\text{adj}}^{\text{GEE}}$	s_l	$p_{\text{adj}}^{\text{maxS}}$
$\pi_B - \pi_L$	-0.34	0.057	-5.9	< 0.001	-5.35	< 0.001
$\pi_B - \pi_M$	-0.49	0.051	-9.5	< 0.001	-7.49	< 0.001
$\pi_B - \pi_H$	-0.67	0.043	-15.5	< 0.001	-9.62	< 0.001
$\pi_L - \pi_M$	-0.15	0.054	-2.7	0.0345	-2.63	0.0406
$\pi_L - \pi_H$	-0.33	0.045	-7.1	< 0.001	-6.17	< 0.001
$\pi_M - \pi_H$	-0.18	0.038	-4.8	< 0.001	-4.44	< 0.001

References

Bretz, F., Hothorn, T., and Westfall, P. (2010). *Multiple Comparisons Using R*. Boca Raton: CRC Press.

Diggle, P.J., Liang, K-Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Leisch, F., Weingessel, A., and Hornik, K. (1998). On the generation of correlated artificial binary data. Working Paper Series, SFB Adaptive Information Systems and Modelling in Economics and Management Science, Vienna University of Economics, http://www.wu-wien.ac.at/am.

Beyond beta regression: modelling percentages and fractions in the presence of boundary observations

Ioannis Kosmidis[1], Anyi Zou[23], Kavya Jagan[3]

[1] Department of Statistical Science, University College London, United Kingdom
[2] Department of Statistics, University of Warwick, United Kingdom
[3] former student at University College London

E-mail for correspondence: i.kosmidis@ucl.ac.uk

1 Beta regression and boundary responses

1.1 Beta regression

Bounded-domain continuous response variables are abundant in applications, for example rates, proportions, concentration indices and so on. For the modelling of such response variables, Ferrari and Cribari-Neto (2004) and Smithson and Verkuilen (2006) introduced the beta regression model and illustrated the flexibility it offers in terms of being naturally heteroscedastic and accommodating asymmetries. Specifically, consider a beta-distributed random variable Z with density

$$f(z|\mu,\phi) = \frac{z^{\mu\phi-1}(1-z)^{(1-\mu)\phi-1}}{B(\mu\phi,(1-\mu)\phi)}, \qquad (0<z<1;\, 0<\mu<1,\, \phi>0), \tag{1}$$

where $B(\cdot,\cdot)$ is the beta function. In this parameterization $\mathrm{E}(Z|\mu,\phi)=\mu$ and $\mathrm{var}(Z|\mu,\phi) = \mu(1-\mu)/(1+\phi)$ so that ϕ is a precision parameter. Then, the mean-precision regression model based on the beta distribution assumes that the observations $z_1,\ldots,z_n$ are realisations of independent beta random variables $Z_1,\ldots,Z_n$, where the mean and precision of Z_i are linked to covariates as

$$g_1(\mu_i) = \eta_i = x_i^{\mathrm{T}}\gamma \quad \text{and} \quad g_2(\phi_i) = \zeta_i = w_i^{\mathrm{T}}\delta\,. \tag{2}$$

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

In the above expression, x_i and w_i are p- and q-dimensional vectors of covariates observed along with Z_i $(i = 1, \ldots, n)$, and $\gamma = (\gamma_1, \ldots, \gamma_p)^{\top}$, $\delta = (\delta_1, \ldots, \delta_q)^{\top}$ are the vectors of coefficients associated with the means and the precision parameters, respectively. The functions $g_1(\cdot)$ and $g_2(\cdot)$ are monotone link functions, preferably with the property of mapping the range of μ_i $(0, 1)$ and ϕ_i $(0, \infty)$, respectively, to the real line. Usual choices for $g_1(\cdot)$ are the logit, probit and generally any inverse of a cumulative distribution function, and for $g_2(\cdot)$ the log function. Typically, the coefficients γ and δ are estimated by maximum likelihood and standard inferential procedures, such as likelihood ratio, Wald, and score tests, apply on the grounds of central limit theorems for the derivatives of the log-likelihood.

1.2 Boundary observations

One limitation of beta regression models arises when at least one of the observed responses is either zero or one. In such cases, and regardless of the sample size, the log-likelihood is either $+\infty$ or $-\infty$ or indeterminate. Hence, maximum likelihood estimation is not possible, and the usual inferential procedures cannot be used.

The existing approaches to address this limitation focus on either the adjustment of the boundary observations by small constants (Smithson and Verkuilen, 2006) so that the adjusted responses end up in $(0, 1)$, or the use of a discrete-continuous mixture of a beta distribution and point masses at zero and/or one (for example, Ospina and Ferrari, 2012). The former approach suffers from the arbitrariness of choosing the additive adjustment, and the latter approach requires placing a "special" interpretation at the values of zero and/or one relative to the observations in $(0, 1)$. Specifically, the use of a discrete-continuous mixture is natural in applications, where the occurrence of 0 and/or 1 can be assumed to be associated with a different process than that which generates the observations in $(0, 1)$ (see, for example, the examples in Ospina and Ferrari, 2012), but would not be appropriate if the boundary observations are merely due to a combination of chance and rounding, for example. A classical example of data of the latter type is the incidence of *Rhynchosporium secalis*, or leaf blotch, on 10 varieties of barley tested at 9 sites in a variety trial in 1965, which is given and analysed in Wedderburn (1974). Incidence of leaf blotch is recorded as the percentage leaf area affected and takes a value of exactly zero in 4 out of the 90 instances. McCullagh and Nelder (1989, Section 9.2.4) treat the percentage of leaf area affected as a proportion and use a quasi-likelihood approach to account for the heteroscedasticity in the incidence percentages. The analysis of the leaf blotch incident on barley is revisited here using a fully parametric model which we call extended beta regression, and which is applicable even when boundary observations are present in the data.

2 Extended beta regression

2.1 Extended-domain beta distribution

If a random variable Z has a beta distribution with density as in (1), then for any $e > 0$, the random variable $Y = -e(1 - Z) + (1 + e)Z$ has density $f^*(y|\mu, \phi, e) = f(z(y, e)|\mu, \phi)/(1 + 2e)$ with $-e < y < 1 + e$ and $0 < \mu < 1$, $\phi > 0$, where $z(y, e) = (y + e)/(1 + 2e)$. This is a special case of the density of the beta distribution as defined in Johnson et al. (1995). In that same book the distribution with density (1) is referred to as the "standard beta". In what follows, the distribution with density $f^*(y|\mu, \phi, e)$ will be referred to as the "extended-domain beta distribution" in order to highlight the fact that the domain is extended from $(0, 1)$ to $(-e, 1 + e)$. Furthermore, e will be called "contamination". Density $f^*(y|\mu, \phi, e)$ has similar shape properties to (1) on the extended domain, and reduces to (1) when $e = 0$. The particular property of the extended-domain beta distribution, that also motivates its use here, is that, when $e > 0$ its density at $y = 0$ and $y = 1$ is positive. By construction, $\mathrm{E}(z(Y, e); \mu, \phi) = \mu$, and $\mathrm{var}(z(Y, e)|\mu, \phi) = \mu(1 - \mu)/(1 + \phi)$.

2.2 Extended beta regression with random contamination

If $Y_1, \ldots, Y_n$ are independent random variables distributed according to the extended-domain beta distribution, then the regression model defined in Subsection 1.1 can be extended by letting the means μ_i of $z(Y_i, e_i)$ and the precision parameters ϕ_i be functions of regression parameters and covariates as is done for beta regression in (2) ($i = 1, \ldots, n$). This is the most general formulation, where we allow one contamination per observation. Note that such a model has n contaminations and $p+q$ regression parameters to be estimated from only from n observations. Furthermore, because the range of the observed responses is $[0, 1]$, it would be desirable to keep the endpoints of the extended domain as close as is possible to 0 and 1.
In order to address such considerations, we assume here that the contaminations $e_1, \ldots, e_n$ are independent and identically distributed random variables each with an exponential distribution with a known mean ν. We further assume that $Y_1, \ldots, Y_n$ are conditionally independent given $e_1, \ldots, e_n$, and that, conditionally on e_i, Y_i has an extended-domain beta with mean μ_i and precision ϕ_i that are linked to covariates as in (2).

3 Estimation

3.1 Choice of ν

Estimates for γ and δ can be obtained by maximising the marginal likelihood $\prod_{i=1}^{n} L_i^*(\gamma, \delta; \nu)$, where

$$L_i^*(\gamma, \delta; \nu) = \nu^{-1} \int_0^\infty f^*(y_i|\mu_i, \phi_i, e) \exp(-\nu^{-1} e) de, \tag{3}$$

is the likelihood contribution with $\nu > 0$. Not that for $\nu \to 0$, (3) reduces to the likelihood contribution in a beta regression setting.
The notation used above highlights that the current treatment will have ν to be known. In all from the numerous examples with boundary observations we worked on, the profile likelihood of ν was unbounded and kept increasing as ν was approaching 0. An ad-hoc rule for choosing a value of ν can be constructed by attempting to make the contaminations comparable in size to the accuracy that the responses are reported at. In this respect, we choose to set ν to 10^{-d} where d is the maximum, across observations, of the number of decimal places that each response is reported at, ignoring any trailing zeros. If it were possible to measure the response in infinite precision this ad-hoc rule would reduce the extended beta regression model to a simple beta regression model, admitting that in that case there should be no zeros and ones in the sample.

3.2 Gauss-Laguerre quadrature

The integral in (3) is generally intractable and needs to be approximated. Because of the particular choice of exponentially distributed contaminations, an approximation is

$$\tilde{L}_i^*(\gamma, \delta; \nu) = \sum_{t=1}^{T} W_t f^*(y_i|\mu_i, \phi_i, \nu Q_t), \tag{4}$$

where W_t and Q_t $(t = 1, \ldots, T)$ are weights and abscissas whose calculation and derivation are through Laguerre polynomials (see, for example Abramowitz and Stegun, 1964, 25.4.45). The approximate likelihood to be maximised is then $\prod_{i=1}^{n} \tilde{L}_i^*(\gamma, \delta; \nu)$.
The derivatives of the logarithm of the approximated likelihood with respect to some parameter ω are $\sum_{i=1}^{n} \partial/\partial\omega \log \tilde{L}_i^*(\gamma, \delta; \nu)$ where

$$\frac{\partial}{\partial \omega} \log \tilde{L}_i^*(\gamma, \delta; \nu) = \frac{\sum_{t=1}^{T} W_t \frac{\partial}{\partial \omega} f^*(y_i|\mu_i, \phi_i, \nu Q_t)}{\tilde{L}_i^*(\gamma, \delta, \nu)}.$$

Since μ_i and ϕ_i are not functions of ν, the expressions for the derivatives of $f^*(y_i|\mu_i, \phi_i, \nu Q_t)$ with respect to β and γ are the same to those for the derivatives of $f(z(y_i, \nu Q_t); \mu_i, \phi_i)$, which, in turn, are readily available from the expressions for the derivatives of the log-likelihood in the case of beta regression for responses $z(y_i, \nu Q_t)$ instead of y_i (see, e.g., Grün et al., 2012, for expressions of the latter derivatives).
The approximate likelihood can be maximised using some generic optimisation algorithm, that potentially also makes use of the derivatives. Estimated standard errors for the maximum approximated likelihood estimators $\hat{\gamma}$ and $\hat{\delta}$ can be computed from the hessian of the approximate likelihood. For practical purposes, it is most convenient to approximate this hessian using numerical methods.

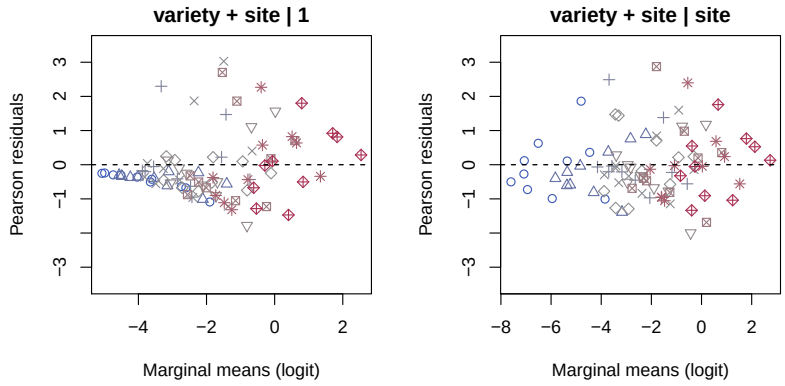

FIGURE 1. The Pearson residuals based on the estimated marginal expectations and variances plotted against the logits of the marginal expectations for constant precision (left) and site-dependent precision (right). The different shapes for the points denote different sites.

4 Incidence of leaf blotch on barley

The data set on the incidence of leaf blotch on barley (see Section 1.2) is used here to fit an extended beta regression model with constant precision and with a linear effect of the 10 varieties and the 9 sites on the logit of the conditional mean of $(y_i + e_i)/(1 + 2e_i)$, where y_i is the observed incidence of leaf blotch. The percentage area affected is reported here in 4 decimal places and, hence, we use a mean $\nu = 0.0001$ for the contamination distribution. The approximate likelihood is calculated with $T = 20$. The marginal expectations and variances for the Y_i can be estimated simply as $-\nu + (1 - 2\nu)\hat{\mu}_i$ and $(1 - 2\nu)^2\hat{\mu}_i(1 - \hat{\mu}_i)/(1 - \hat{\phi}_i)$, respectively, where $\hat{\mu}_i$ and $\hat{\phi}_i$ are the mean and the precision evaluated at $\hat{\gamma}$ and $\hat{\delta}$ $(i = 1, \ldots, n)$. The left plot on Figure 1 shows the Pearson residuals plotted against the logits of the marginal expectations. As is apparent, the variability on this plot seems to change with site. This has also been noted in McCullagh and Nelder (1989, Section 9.2.4) were maximum quasi-likelihood was used with variance function $\mu(1-\mu)$. Extending the model to include a linear effect of the 8 sites on the logarithm of the precision parameters results in the right residual plot on Figure 1. The effect of site-dependent variability remains but is significantly milder than when precision is constant across observations. A parallel improvement is achieved with maximum quasi-likelihood with variance function $\mu^2(1-\mu)^2$ (see McCullagh and Nelder, 1989, Section 9.2.4). A likelihood ratio test for comparing the model with site-dependent precision to that with constant precision results in a likelihood ratio statistic of value 48.226, which is rather extreme when compared to the quan-

tiles of a chi-squared distribution with 8 degrees of freedom. This provides strong evidence against the model with constant precision. Finally, the estimated variety contrasts from the extended beta regression model with site-dependent precision together with their standard errors are shown in the table below:

Variety									
1	2	3	4	5	6	7	8	9	10
0	-0.45	0.13	0.01	0.52	1.05	1.63	2.52	2.18	3.13
(0)	(0.48)	(0.41)	(0.53)	(0.52)	(0.43)	(0.38)	(0.38)	(0.41)	(0.37)

It is clear that varieties 1-4 are the most resistant while 8-10 are the least resistant, which corresponds to the conclusions in past studies.

References

Abramowitz, M. and Stegun, I. (1964). *Handbook of Mathematical Functions (5th ed)*. New York: Dover.

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

Grün, B., Kosmidis, I., and Zeileis, A. (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software*, *48*, 1 – 25.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions (Vol. 2)*. Wiley & Sons.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models (2nd ed)*. London: Chapman and Hall.

Ospina, R. and Ferrari, S.L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*, 1609 – 1623.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*, 54 – 71.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*, 439 – 447.

Density estimation from tabulated summary statistics

Philippe Lambert[1,2], Michel Denuit[2]

[1] Institut des sciences humaines, Université de Liège, Belgium
[2] Institut de statistique, biostatistique et sciences actuarielles (ISBA), Université catholique de Louvain, Belgium

E-mail for correspondence: p.lambert@ulg.ac.be

Abstract: Data on a continuous variable are often summarized by means of histograms or displayed in tabular format: the range of data is partitioned into consecutive interval classes and the number of observations falling within each class is provided to the analyst. Computations can then be carried in a nonparametric way by assuming a uniform distribution of the variable within each partitioning class, by concentrating all the observed values in the center, or by spreading them to the extremities. Smoothing methods can also be applied to estimate the underlying density or a parametric model can be fitted to these grouped data. For insurance loss data, some additional information is often provided about the observed values contained in each class, typically class-specific sample moments such as the mean, the variance or even the skewness. The question is then how to include this additional information in the estimation procedure. The present paper proposes a method for performing density estimation based on such augmented information with an illustration on insurance data.

Keywords: Tabulated summary statistics; Density estimation; Composite link model; P-splines.

1 Introduction and motivation

We propose an efficient nonparametric estimation procedure for the loss density, based on histograms or grouped data, including information about class-specific sample moments. Specifically, the risk analyst has access to a set of observed losses, grouped into consecutive classes (or tranches). Graphically, this corresponds to an histogram. In addition to these grouped data, the average value of the observations in each class is provided, as well as the corresponding variance and skewness. Reinsurers also often display

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

the information about insurance losses in this way. Confidentiality issues may sometimes justify this grouping procedure. We show in this paper how to obtain a smooth, nonparametric density estimate based on this information.

Our starting point is a set of n observations $x_1, \ldots, x_n$ available in tabular form. We consider that these observations are realizations of independent random variables $X_1, \ldots, X_n$ with common distribution function F and density function f. Precisely, the available data points x_i have been partitioned into consecutive class intervals $\mathcal{C}_j = (a_{j-1}, a_j]$, $j = 1, \ldots, J$, also called tranches. In addition to the number n_j of observations belonging to $\mathcal{C}_j$, we also have summary statistics about observations in each class. Specifically, we assume that we know the class-specific means $\{\overline{x}_j\}_{j=1}^J$ as well as sample centered moments m_{kj} $(k = 2, 3)$.

2 Methodology

2.1 Density estimation from grouped data

Description of the model

The Composite Link Model has been proposed by Thompson and Baker (1981) as an extension of the Generalized Linear Model. This elegant framework can be used to estimate a density from grouped data (Eilers, 2007; Lambert and Eilers, 2009). Assume that each class $\mathcal{C}_j$ is divided into finer sub-intervals. The fine grid spacing Δ is taken small enough to give an accurate description of the density f for plotting it or for computing quantiles or other indices accurately. The fine grid consists in a sufficiently large number of grid points partitioning $[a_0, a_J]$ into I consecutive intervals $\mathcal{D}_i = (b_{i-1}, b_i]$ of equal width Δ with mid-point u_i, $i = 1, \ldots, I$. The relationship between the wide bin $\mathcal{C}_j$ and the narrow ones $\mathcal{D}_i$ is coded by the $J \times I$ matrix $C = (c_{ji})$ where $c_{ji} = 1$ if $\mathcal{D}_i \subset \mathcal{C}_j$ and 0 otherwise. We define $\pi_i = \Pr[X_1 \in \mathcal{D}_i] = \int_{b_{i-1}}^{b_i} f(t)dt \approx f(u_i)\Delta$. These probabilities are related to a rich B-spline basis $\{b_k(\cdot)\}_{k=1}^K$ (corresponding to equidistant knots on $[a_0, a_J]$) using a polytomous logistic regression with linear predictors $\eta_i(\theta) = (B\theta)_i$. In this setting, we cannot observe π itself, but only sums over $J < I$ intervals. The probability masses assigned to these intervals are given by $\gamma_j(\theta) = \Pr[X_1 \in \mathcal{C}_j|\theta] = \int_{a_{j-1}}^{a_j} f(t|\theta)dt \approx \sum_{i=1}^I c_{ji}\pi_i(\theta)$.

Inference

The class frequencies $n_1, \ldots, n_J$ have a Multinomial distribution with probability vector γ, yielding the log-likelihood $\ell(\theta) = \sum_{j=1}^J n_j \ln\left(\sum_{i=1}^I c_{ji}\pi_i(\theta)\right)$. Thompson and Baker (1981) showed how to estimate the vector parameter θ in a frequentist setting using the scoring algorithm. Eilers (2007) proposed to put a roughness penalty on θ, to force the estimated density to be smooth. Inference can also be made in a Bayesian setting (Lambert and

Eilers, 2009) with a prior on the spline coefficients to force smoothness (Jullion and Lambert, 2007). Credible intervals on functions of the density (such as moments or quantiles) are readily available using MCMC samples (Lambert and Eilers, 2005, 2006).

2.2 Density estimation for given class-specific sample moments

Density estimation for given class-specific sample means
Let us now further assume that, together with the frequencies n_j, the sample means $\overline{x}_j$ within $\mathcal{C}_j$ are also reported. The likelihood then becomes

$$\Pr\left[N_1 = n_1, \ldots, N_J = n_J|\theta\right] \prod_{j=1}^{J} f_{\overline{X}_j}(\overline{x}_j|n_j,\theta) \propto \prod_{j=1}^{J} \left(\gamma_j^{n_j} f_{\overline{X}_j}(\overline{x}_j|n_j,\theta)\right)$$

with, when n_j is not too small, $(\overline{X}_j|n_j,\theta) \overset{appr}{\sim} \mathcal{N}(\mu_j(\theta), \sigma_j^2(\theta)/n_j)$ where

$$\begin{aligned}
\mu_j(\theta) &= \frac{1}{\gamma_j(\theta)} \int_{a_{j-1}}^{a_j} x\ f(x|\theta)dx \approx \sum_{i=1}^{I} u_i \frac{c_{ji}\pi_i(\theta)}{\gamma_j}, \\
\sigma_j^2(\theta) &= \frac{1}{\gamma_j(\theta)} \int_{a_{j-1}}^{a_j} (x-\mu_j(\theta))^2\ f(x|\theta)dx \approx \sum_{i=1}^{I} (u_i-\mu_j(\theta))^2 \frac{c_{ji}\pi_i(\theta)}{\gamma_j}.
\end{aligned}$$

Then, the log-likelihood becomes

$$\ell(\theta) = \sum_{j=1}^{J} \left(n_j \ln \gamma_j(\theta) - \frac{1}{2} \ln \sigma_j^2(\theta) - \frac{n_j}{2\sigma_j^2(\theta)} \left(\overline{x}_j - \mu_j(\theta)\right)^2 \right).$$

It can be extended to deal with the joint report of the sample mean, variance and skewness within each $\mathcal{C}_j$.
Denote the rth central population moments by

$$\begin{aligned}
\mu_{rj} = \mu_{rj}(\theta) &= \mathrm{E}\left[(X_1 - \mu_{1j}(\theta))^r | X \in \mathcal{C}_j\right] \\
&\approx \frac{1}{\gamma_j(\theta)} \sum_{j=1}^{J} c_{ji}\pi_i(\theta)(u_i - \mu_{1j}(\theta))^r, \qquad r = 2,3,\ldots.
\end{aligned}$$

If M_{rj} denotes the rth sample central moment in $\mathcal{C}_j$, the sample mean, variance and skewness are

$$\overline{X}_j = M_{1j}, \qquad S_j^2 = M_{2j}, \qquad G_j = M_{3j}/M_{2j}^{3/2}.$$

The joint sampling distribution of (M_{1j}, M_{2j}, M_{3j}) can be approximated using a Normal with mean vector $(\mu_{1j}, \mu_{2j}, \mu_{3j})$ and variance-covariance matrix

$$\frac{1}{n_j} \begin{pmatrix} \mu_{2j} & \mu_{3j} & \mu_{4j} \\ \mu_{3j} & \mu_{4j} - \mu_{2j}^2 & \mu_{5j} - \mu_{2j}\mu_{3j} \\ \mu_{4j} & \mu_{5j} - \mu_{2j}\mu_{3j} & \mu_{6j} - \mu_{3j}^2 \end{pmatrix}$$

TABLE 1. Motor insurance claim amounts.

Claim Interval $\mathcal{C}_j$	Freq. n_j	Mean $\overline{x}_j$	Std.dev s_j	Skewness g_j
(0;2 500]	126	1 630	596	-0.38
(2 500;5 000]	285	3 886	681	-0.19
(5 000;10 000]	706	7 458	1 393	-0.001
(10 000;20 000]	875	14 087	2 725	0.37
(20 000;35 000]	265	24 678	3 668	0.81
(35 000;50 000)	25	39 789	3 515	0.58
Total	2 283	11 604	7 428	1.22

provided that n_j is large enough. The same strategy as before can be used to write the log-likelihood. A penalty can be added to force smoothness for the estimated density.

3 Simulation study

The performances of the proposed estimation method have been evaluated by means of extensive simulations with losses distributed according to the Beta, the Gamma, and the LogNormal laws and moderate to large sample sizes n. After having grouped the n simulated values in a limited number of classes, we compute the class-specific sample moments $\overline{x}_j$, s_j^2, and g_j and include them sequentially in the estimation procedure.
The simulation results suggest that, whatever the sample size and the target distribution, the local sample means are (besides the bin frequencies) the most impacting additional pieces of information on the density point estimates. For some density shapes and with smaller sample sizes, the local sample variances were also found to have a significant impact on the results. Local sample skewness did not reveal to be useful.

4 Numerical illustration

The data displayed in Table 1 represent claim amounts in motor insurance, as recorded by a large insurance company operating in the EU. For confidentiality reasons, the amounts have been rescaled. Figure 1 displays the estimated density obtained first from the knowledge of n_j alone and then when $\overline{x}_j$, s_j^2 and g_j become available.
If it is reasonable to believe that the underlying density is unimodal or if the available data strongly support this assumption, it is possible to include this additional information to improve estimation. Based on the histogram, the risk analyst can identify the class where the mode is located. To the

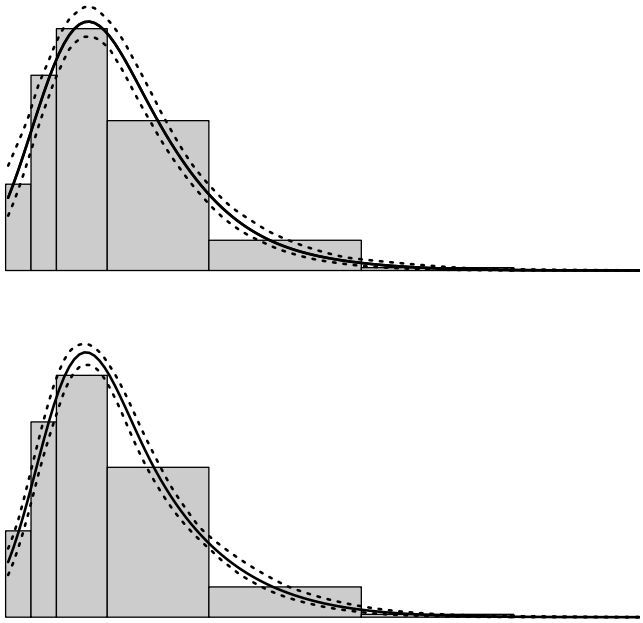

FIGURE 1. Density estimates and credible intervals for motor insurance claim amounts based on n_j alone (top) and on $(n_j, \overline{x}_j, s_j^2, g_j)$ (bottom).

left of this modal class, the density can then be assumed to be increasing whereas to its right it can be assumed to be decreasing. The conditional density, given that the loss belongs to a given class, is thus monotone (and unimodal in the modal class).
In addition to point estimates, credible regions are readily available using MCMC samples generated in the Bayesian framework. They are displayed in Figure 1. The knowledge of local means had a non negligible impact on the shape and the uncertainty of the density estimate on the left hand of the mode, but had a smaller influence elsewhere.

5 Discussion

We have shown how to combine tabulated summary statistics with the observed frequencies to estimate a density from grouped data. For large sample sizes, simulations suggest that the sample means convey most of the useful information. For small sample sizes, simulations suggest that the additional knowledge of the sample variances is relevant.
The method described in this paper is particularly efficient to deal with the central part of the observed losses. The density estimate can then be

extended to the few very large losses by the peak-over-threshold approach from Extreme Value Theory, and extrapolated by means of the Generalized Pareto model.

Acknowledgments: We would like to thank Bernard Lejeune (HEC-ULg) for useful discussions about the Generalized Method of Moments. P. Lambert acknowledges financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

References

Bernard, C., Denuit, M., and Vanduffel, S. (2014). Measuring portfolio risk under partial dependence information. *ISBA Discussion Paper*, 2014/09.

Courtois, C. and Denuit, M. (2007). Local moment matching and s-convex extrema. *ASTIN Bulletin*, **37**, 387 – 404.

Eilers, P.H.C. (2007). Ill-posed problems with counts, the composite link model, and penalized likelihood. *Statistical Modelling*, **7**, 239 – 254.

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

Hürlimann, W. (2008). Extremal moment methods and stochastic orders. *Boletin de la Associacion Matematica Venezolana*, **15**, 153 – 301.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics and Data Analysis*, **51**, 2542 – 2558.

Lambert, P. and Eilers, P.H.C. (2005). Bayesian proportional hazards model with time-varying regression coefficients: A penalized Poisson regression approach. *Statistics in Medicine*, **24**, 3977 – 3989.

Lambert, P. and Eilers, P.H.C. (2006). Bayesian multi-dimensional density estimation with P-splines. *Proceedings of the 21st IWSM*, 313 – 320. Galway.

Lambert, P. and Eilers, P.H.C. (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics and Data Analysis*, **53**, 1388 – 1399.

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

Model-based mixed effects trees for longitudinal and clustered data

Denis Larocque[1], Jeffrey S. Simonoff[2]

[1] HEC Montréal, Montréal, Canada
[2] New York University, New York, USA

E-mail for correspondence: jsimonof@stern.nyu.edu

Abstract: Mixed effects regression trees are a methodology that combines the structure of mixed effects models for longitudinal and clustered data with the flexibility of tree-based estimation methods. In this paper we extend this methodology to allow for linear functions of predictors at terminal nodes, specifically linear function time trends. We show that the method can recover highly nonlinear structure while preserving interpretability, with easy applicability to data with time-varying covariates.

Keywords: Longitudinal and clustered data; Mixed effects; Regression tree.

1 Trees for longitudinal and clustered data

The regression tree is a nonparametric method for estimating a regression function. The regression tree algorithm splits the data set into subsets based on the values of its covariate variables. The process is repeated on each derived subset recursively until halted based on some stopping rule. Typically the estimated response value at each terminal node is the mean response for observations in that node. Efforts to extend regression tree methodology to longitudinal data include Segal (1992), Zhang (1998), De'Ath (2002), and Loh and Zhang (2013), but all of these approaches share the restriction of not allowing time-varying covariates.

Hajjem et al. (2011) and Sela and Simonoff (2012) independently took a mixed-effects models approach and extended regression tree algorithms to the case of clustered data for continuous outcomes (referring to it as a mixed-effects regression tree [MERT] and a random effects-EM [RE-EM] tree, respectively). The essential idea of the approach is to separate the estimation of the fixed effects (in the form of the tree) from random effects

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

(through a mixed-effects model), using CART (Breiman et al., 1984) as the standard regression tree to model the fixed effect. Fu and Simonoff (2015) addressed the tendency of CART to split on variables with more possible split points by basing the tree on the conditional inference tree of Hothorn et al. (2006). A key point is that this approach can easily accommodate time-varying covariates.

A potential weakness of these methods is that they restrict the expected response at each terminal node to be a constant. Eo and Cho (2014) proposed MELT, a regression tree method for longitudinal data that provides an estimated slope at each terminal node for a linear function of time. This method cannot handle time-varying covariates in a direct fashion, and since it does not provide an estimate of the intercept cannot be used for prediction. In this paper we propose a modification of the basic MERT/RE-EM idea that allows for linear functions of predictors at terminal nodes rather than simply mean responses. This is done through the use of model-based partitioning, as is discussed in Zeileis et al. (2008). Bürgin and Ritschard (2015) proposed a similar idea in the context of logistic responses.

2 Proposed method

The proposed method is similar to earlier versions of MERT/RE-EM, with the exception that a linear functional form for the response is provided at each terminal node rather than a constant estimated mean response; we will focus on the most natural situation for longitudinal data, in which the functional form is a linear function of time. The underlying model is

$$y_{it} = Z_{it}\mathbf{b}_i + f\left(x_{it1}, \ldots, x_{itk}, \texttt{time}_{it}\right) + \varepsilon_{it}, \tag{1}$$

where we observe a panel of objects or individuals $i = 1, \ldots, I$ at times $t = 1, \ldots, T_i$, with a vector of covariates $\mathbf{x}_{it}$ and a time value $\texttt{time}_{it}$ associated with the observation (i, t). To account for the differences between objects across time periods, we include a known design matrix (which is actually either a scalar or a row vector) Z_{it}, and a vector of unknown time-constant, object-specific random effects, $\mathbf{b}_i$. Traditional mixed effects models, such as the linear mixed effects model (where $f = X\boldsymbol{\beta}$), assume a parametric form for f, which might be a too restrictive assumption. Instead, a regression tree with linear functions of time at the terminal nodes is used to estimate f. If the random effects were known, (1) implies that we could fit a regression tree to $y_{it} - Z_{it}\mathbf{b}_i$ to estimate f (using the model-based recursive partitioning algorithm of Zeileis et al., 2008). If the population-level effects, f, were known, then we could estimate the random effects using a mixed effects linear model with population-level effects corresponding to the values $f\left(X_i, \texttt{time}_i\right)$. The algorithm proceeds by alternating between estimating the regression tree, assuming that our estimates of the random effects are correct, and estimating the random effects, assuming that the regression tree is correct.

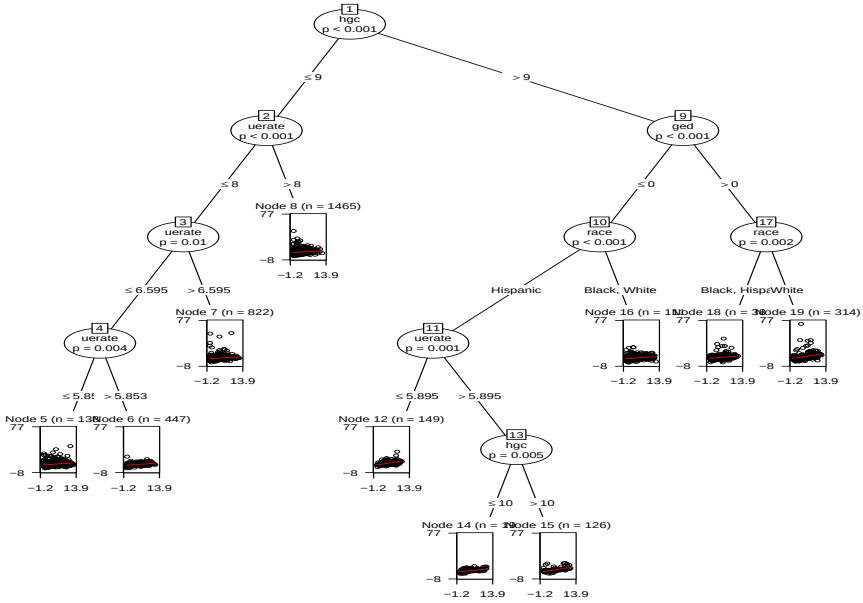

FIGURE 1. Model-based mixed effects tree using time-invariant and time-varying predictors for UCLA wages data.

3 Simulation study

Monte Carlo simulations demonstrate that the model-based mixed effects tree can provide much more accurate estimates of different slopes across nodes compared to MELT when the latter is applicable. The proposed method also provides estimates of the intercept at each node and can be applied when there are time-varying predictors, with similar performance.

4 Analysis of UCLA wages data

Eo and Cho (2014) examined data from the National Longitudinal Survey of Youth (NLSY), focusing on wage data that were also analyzed by Singer and Willett (2003). The data consist of 888 high school dropouts, ages 14-17, with the goal being to model hourly wages (in constant 1990 dollars). Eo and Cho (2014) used only `race` (White, Black, or Hispanic) and `hgc` (highest grade of schooling completed) as predictors because of the restriction of MELT to non-time-varying predictors, and `exper` (duration of work experience in years) as the time variable. Figure 1 gives the model-based mixed effects tree for these data that also takes advantage of other (time-varying) covariates such as `ged` (whether the respondent had

earned a high school equivalency degree) and `uerate` (unemployment rate at the time) are available.
It can be seen that obtaining a GED and the unemployment rate are related to wages throughout the tree. Race matters, with white and Hispanic respondents tending to have a steeper slope of wages with experience when `hgc` > 9, but it is not predictive for `hgc` ≤ 9. Further, and unsurprisingly, a lower unemployment rate is associated with faster growth of wages with experience (being roughly linear in unemployment rate for a lower number of grades completed, as seen in the left branch of the tree), as is a higher grade of school completed. The relationship with earning a GED is more complicated, having a different relationship with the wage growth / experience pattern depending on `hgc`, `race`, and `uerate`.

5 Proposed random forest

A single tree is interesting as it can provide valuable and easy-to-visualize insights about the data. However, a combination of several trees, like a random forest (Breiman, 2001), often has better predictive performance, at the expense of being less easy to interpret. Hajjem et al. (2014) proposed a random forest based on the MERT approach, called "mixed effects random forest" (MERF). Following similar lines, we now describe one possible way to build a random forest in our setting. Instead of estimating f in (1) with a tree, we use a random forest. More specifically, given current estimates of the random effects, we fit a forest to the "centered" observations $y_{it} - Z_{it}\mathbf{b}_i$. This means that a large number of trees are built using bootstrap samples of the centered observations. At each node of any single bootstrap tree, the best splitting variable and best split are chosen among a random subset of the covariates, the size of the subset being a user-defined parameter. The forest estimates of the intercept and slope are taken as the average intercepts and slopes over the trees in the forest. Likewise, predicted random effects are obtained from each tree. The forest-predicted random effects are obtained by averaging those from the underlying trees. These updated random effects are used to update the centered observations, and a new forest can be built with them. This process is repeated until convergence. Upon convergence, estimated forest intercept and slope are available for any individual. Moreover, as in the case when estimating f from a single tree, to obtain a prediction for a new time point, two situations are possible. Firstly, if a predicted random effect is available, then it can be incorporated along with the estimated intercept and slope of this individual to compute the prediction. Secondly, if a predicted random effect is not available, for instance for a new individual not part of the training sample, then only the estimated intercept and slope of this individual are used to compute the prediction, which is then a population level prediction.

TABLE 1. PMSE values for different tree-based methods for UCLA wages data.

Ordinary random forest	19.83
Overfit MOB longitudinal tree (70 terminal nodes)	18.59
MOB longitudinal tree (9 terminal nodes)	17.41
MOB longitudinal random forest with overfit trees	17.18
MOB longitudinal random forest	17.10

6 Revisiting the UCLA wages data

Table 1 summarizes predictive performance in application of the random forest to the UCLA wages data. Lagged values of the time-varying covariates are used to make the example more realistic from a predictive point of view. The table entries represent predictive mean squared error (PMSE) values using the final time point of each individuals as the test set (of size 850) and the rest of the data as the training set (of size 4,664). The ordinary random forest ignoring the longitudinal structure has the worst performance, showing that the most important aspect of these data is the differences between individuals reflected in the random effects. Here, an overfit MOB longitudinal tree means that splitting is not stopped based on a maximal threshold p-value (0.05 being the default in MOB), but rather only when a minimum node size is reached. Hence the tree is much larger than would typically be the case, being similar to an unpruned CART tree in that regard. For these data, it seems that a single "honest" sized longitudinal tree (the one with 9 terminal nodes) is able to capture most of the covariates' predictive power. Indeed, MOB-based longitudinal random forests, based on either honest sized trees or on large trees, provide only small gains over an honest sized MOB-based longitudinal tree. Moreover, for these data, linear mixed models have similar predictive performance to longitudinal tree-based methods indicating that the population-level structure is not very nonlinear. It is important to note, however, that a random forest based on trees that are constructed to overfit does considerably better than the overfit tree itself, highlighting the ability of random forests to overcome the variability of individual trees in general.

Acknowledgments: This research was partially completed while the authors were visiting the Institute for Mathematical Sciences, National University of Singapore, in 2014.

References

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5 – 32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Bürgin, R. and Ritschard, G. (2015). Tree-based varying coefficient regression for longitudinal ordinal responses. *Computational Statistics and Data Analysis*, **86**, 65 – 80.

De'Ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, **83**, 1105 – 1117.

Eo, S.-H. and Cho, H. (2014). Tree-structured mixed-effects regression modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, **23**, 740 – 760.

Fu, W. and Simonoff, J.S. (2015). Unbiased regression trees for longitudinal and clustered data. *Computational Statistics and Data Analysis*, **88**, 53 – 74.

Hajjem, A., Bellavance, F., and Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics and Probability Letters*, **81**, 451 – 459.

Hajjem, A., Bellavance, F., and Larocque, D. (2014). Mixed effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, **84**, 1313 – 1328.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, **15**, 651 – 674.

Loh, W.-Y. and Zheng, W. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*, **7**, 495 – 522.

Segal, M.R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, **87**, 407 – 418.

Sela, R.J. and Simonoff, J.S. (2012). RE-EM Trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, **86**, 169 – 207.

Singer, J.D. and Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford University Press.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, **17**, 492 – 514.

Zhang, H. (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association*, **93**, 180 – 193.

Functional forms for transition intensities in a progressive three-state model

Robson J.M. Machado[1], Ardo van den Hout[1]

[1] Department of Statistical Science, University College London, United Kingdom.

E-mail for correspondence: robson.machado.14@ucl.ac.uk

Abstract: Continuous-time multi-state models are formulated to investigate transition intensities for discrete states. For the intensities, various functional forms are considered using parametric and semiparametric expressions. Estimation is performed by using maximum likelihood. The models are illustrated with data describing the development of cardiac allograft vasculopathy, a deterioration of the arterial walls, in heart transplantation patients.

Keywords: Continuous-time multi-state models; Smoothing.

1 Introduction

Longitudinal data for transitions between states are common in biostatistics. Multi-state models have been widely applied to this kind of data. Transition probabilities can be derived from transitions intensities. This paper proposes a study of different functional forms for intensities. Estimation is carried out using maximum likelihood.

The models are applied to data for cardiac allograft vasculopathy (CAV) which is a narrowing of the arterial walls. This is the main cause of death in heart transplantation patients. The data are a series of approximately yearly angiographic examinations of heart transplant recipients. The state at each time is a grade of CAV which can be normal, moderate or severe. Dead is the absorbing state and time of death is known within one day. The data contain 2816 rows which are grouped by 614 patients and ordered by years after transplant. Each row represents an examination and contains additional covariates. The process is biologically irreversible and of particular interest is the onset of CAV. Diagnosis of ischaemic heart disease (IHD) and donor age are known to be major risk factors of disease

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

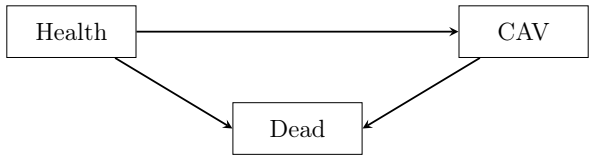

FIGURE 1. Three-state model for disease progression after transplant.

onset (Titman, 2011). In order to investigate this, three-state progressive models are fitted. The states are classified as normal (1) if the patient has not developed the disease, ill (2) if the patient has developed moderate or severe CAV and dead (3) if the patient has died, see Figure 1.

2 Continuous-time multi-state models

Consider a continuous-time Markov processes $\{X(t),\ t > 0\}$ which takes values in the discrete state space $S = \{1, \ldots, m\}$. If $X(t)$ is time-homogeneous transition probabilities are given by

$$p_{rs}(t) = P(X(t) = s | X(0) = r), \tag{1}$$

and the transition probability matrix is defined by $\mathbf{P} = (p_{rs})$ for $r, s \in S$. The transition intensities are given by

$$q_{rs} = \lim_{\Delta t \to 0} \frac{P(X(t + \Delta t) = s | X(t) = r)}{\Delta t}, \tag{2}$$

for $r \neq s$, and can be used to derive the transition probabilities. Let $q_{rr} = -\sum_{s \neq r} q_{rs}(t)$ for all $r \in S$ and define the transition intensity matrix by $\mathbf{Q} = (q_{rs})$. Subject to the initial condition $\mathbf{P}(0) = \mathbf{I}_m$, it is known that $\mathbf{P}(t) = \exp(t\mathbf{Q})$, see Kalbfleisch and Lawless (1985).
Let N denotes the number of individuals. For individual i, n_i denotes the number of observations from the Markov process $\{X(t);\ t > 0\}$ observed at $t_{i1}, \ldots, t_{in_i}$. The dead state is represented by m in the state space S. Let $q_{rs} = q_{rs}(\boldsymbol{\theta})$ depend on a vector of parameters $\boldsymbol{\theta}$. The maximum likelihood inference is conditional on the first observation. The joint distribution of $X_{i2}, \ldots, X_{in_i}$ is

$$P(X_{i2} = x_{i2}, \ldots, X_{in_i} = x_{in_i} | X_{i1} = x_{i1}, \boldsymbol{\theta}, \mathbf{Z}_i), \tag{3}$$

where $\mathbf{Z}_i$ is a $n_i \times p$ covariate matrix. The Markov assumption provides the following expression for (3)

$$\prod_{j=2}^{n_i} P(X_{ij} = x_{ij} | X_{i(j-1)} = x_{i(j-1)}, \boldsymbol{\theta}, \mathbf{Z}_{i(j-1)}), \tag{4}$$

where $\mathbf{Z}_{i(j-1)}$ is the $(j-1)^{th}$ row in $\mathbf{Z}_i$.
In order to build the individual likelihood, it is necessary to verify whether the last observation x_{in_i} is a living state or the dead state. Let $L_i = \prod_{i=1}^{n_i} L_{ij}$ be the ith individual contribution to the likelihood, where

$$L_{ij} = \begin{cases} P(X_{ij} = x_{ij} | X_{i(j-1)} = x_{i(j-1)}) & \text{for } j = 2, \ldots, n_i - 1 \\ C(X_{ij} | X_{i(j-1)}) & \text{for } j = n_i. \end{cases}$$

If a living state is observed,

$$C(X_{in_i} | X_{i(n_i-1)}) = P(X_{in_i} = x_{in_i} | X_{i(n_i-1)} = x_{i(n_i-1)}). \quad (5)$$

If the last observation is a dead state, then known time of death is taken into account by defining

$$C(X_{in_i} | X_{i(n_i-1)}) = \sum_{s=1}^{m-1} P(X_{in_i} = s | X_{i(n_i-1)} = x_{i(n_i-1)}) q_{sm}(t_{n_i-1}). \quad (6)$$

The likelihood function is given by $L(\boldsymbol{\theta}) = \prod_{i=1}^{N} L_i$. In case of no known death times, Kalbfleisch and Lawless (1985) provide an algorithm to estimate $\boldsymbol{\theta}$ based on a quasi-Newton procedure that uses first derivatives of $\log L(\boldsymbol{\theta})$. However, for this investigation the general-purpose optimisation optim in R is used for maximising the log-likelihood function.

3 Time-dependency

A general time-dependent intensity regression model is given by

$$q_{rs}(t) = q_{rs.0}(t) \exp(\boldsymbol{\beta}_{rs}^{\mathrm{T}} \mathbf{z}), \quad (7)$$

where $\mathbf{z}$ is a covariate vector, $\boldsymbol{\beta}_{rs}$ is a parameter vector and $q_{rs.0}(t)$ is the baseline hazard function. Examples of functional forms are

Exponential:	$q_{rs.0}(t) = \exp(\alpha_{rs}), \quad \alpha_{rs} \in \mathbb{R},$
Step function:	$q_{rs.0}(t) = \exp(\alpha_{rs.j}), \quad \alpha_{rs.j} \in \mathbb{R}$ and $k_{j-1} < t \leq k_j,$
Gompertz:	$q_{rs.0}(t) = \exp(\alpha_{rs.1} + \alpha_{rs.2} t), \quad \alpha_{rs.1}, \alpha_{rs.2} \in \mathbb{R},$
Weibull:	$q_{rs.0}(t) = \alpha_{rs} \tau_{rs} t^{\tau_{rs}-1}, \quad \alpha_{rs}, \tau_{rs} > 0,$
Splines:	$q_{rs.0}(t) = \exp(\sum_{k=1}^{K} \alpha_{rs.k} B_k(t)), \quad \alpha_{rs.k} \in \mathbb{R},$

where M is the number of knots for the step function and $B_k(t)$ are B-splines bases with suitable degree where K represents the number of knots (Eilers and Marxs 1996; Kneib 2008).
Time-dependent models are estimated using a piecewise-constant intensities approximation. In this case, the matrix $\mathbf{Q}(t)$ changes at pre-specified time points, but is constant in-between. Let $\mathbf{P}(t_1, t_2)$ with $t_1 < t_2$ represents the probability matrix with entries $p_{rs}(t_1, t_2) = P(X(t_2) = s | X(t_1) = r)$, for $r, s \in S$. The transition probabilities for the likelihood can be derived by using individual follow-up times in the data. For instance, adopting the notation of (4), the probability $P(X_{ij}(t_{ij}) = x_{ij} | X_{i(j-1)} = x_{i(j-1)})$ is obtained by using $\mathbf{Q}(t_{i(j-1)})$ to compute $\mathbf{P}(t_{ij}, t_{i(j-1)})$.

4 Application

For the CAV data, consider the three-state model illustrated in Figure 1. The covariates donor age (*dage*) and *IHD* are added for the transition intensity from 1 to 2. Specifically, the model for this case is

$$q_{12}(t) = q_{12.0}(t)\exp(\beta_1 dage + \beta_2 IHD), \tag{8}$$

where $q_{12.0}(t)$ can assume the functional forms presented in Section 3. For the transitions 1 to 3 and 2 to 3, models are fitted without covariates.

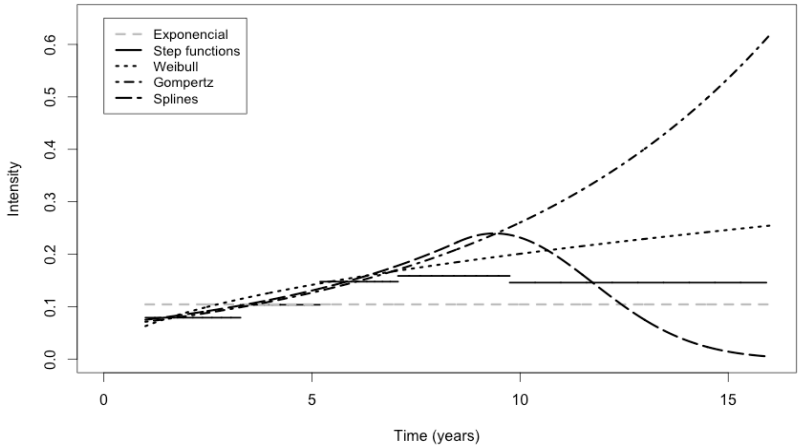

FIGURE 2. Fitted transition intensity from state 1 to 2 for individuals with donors aged 31 and history of IHD.

Only data until 15 years are considered, since after this time data are scarce which may cause identifiability problems. To allow for the Weibull models, time is shifted by plus one year for all analyses. To deal with numerical problems the covariate donor age is centred and re-scaled. The exponential model with five parameters has AIC = 2962. The step function model is fitted with six knots defined by the quantiles of the years. It has 17 parameters and AIC = 2964. The Weibull and Gompertz models with eight parameters each have AIC = 2928 and AIC = 2927, respectively. The last model fitted is the nonparametric with quadratic B-splines and four knots for transition from 1 to 2, and three knots for other transitions. It has twelve parameters and AIC = 2960. The step functions model performs slightly worse than the exponential model, due to its large number of parameters. Both parametric models perform better than the model with splines. Future work will consider a larger number of splines bases and penalties to improve

the modeling. For all models fitted, the estimates $\hat{\beta}_1, \hat{\beta}_2 > 0$, indicating that the donor age and historical of IHD increase the risk of developing CAV. The fitted intensities for 1 to 2 are presented in Figure 2. The shapes of all models are very similar up to 10 years and quite different afterwards. It is due to the fact that there are less data available after that time, see Figure 3.

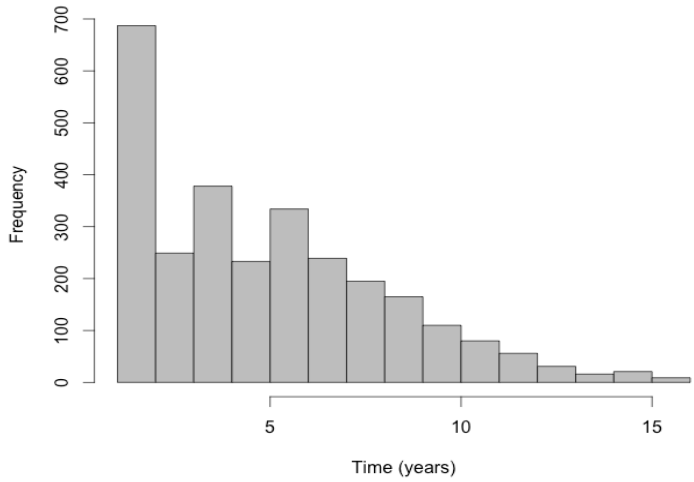

FIGURE 3. Histogram of the observation times in the CAV data.

Acknowledgments: Special Thanks to Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq.

References

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

Kalbfleisch, J.D. and Lawless, J.F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, **80**, 863 – 871.

Kneib, T. and Hennerfeind, A. (2008). Bayesian semiparametric multistate models. *Statistical Modelling*, **8**, 169 – 198.

Titman, A.C. (2011). Flexible nonhomogeneous Markov models for panel observed data. *Biometrics*, **67**, 780 – 787.

Varying single-index signal regression

Brian D. Marx[1]

[1] Experimental Statistics, Louisiana State University, Baton Rouge, LA, USA

E-mail for correspondence: bmarx@lsu.edu

Abstract: The Penalized Signal Regression (PSR) approach to Multivariate Calibration (MVC) assumes a smooth vector of coefficients for weighting a signal to predict the unknown chemical concentration. We allow the PSR coefficient vector to vary smoothly along a covariate (e.g. temperature), which results in a smooth surface on wavelength temperature domain. A slice of this surface effectively estimates the vector of coefficients at any arbitrary temperature. As an added generalization, we relax the implicit assumption of an identity link function by allowing an unknown, but explicit, link function between the linear predictor and the response. Again, we allow the signal's link function to vary smoothly along a covariate, which produces a two-dimensional *link surface*, which is sliced at the same temperature to bend prediction. We term our method: varying-single-index signal regression (VSISR). Estimation is performed using two-dimensional tensor product P-splines. External prediction performance comparisons are made to both the identity link varying-PSR (VPSR) and partial least squares (PLS).

Keywords: Multivariate calibration; Tensor P-spline; Signal regression; Single-index; Varying-coefficient models.

1 Introduction

We take yet another approach to the multivariate calibration problem, in particular where the signal (spectra) regressors appear to have two-dimensional structure. Our application considers NIR spectra (taken over several temperatures) (Marx, 2015). Through simultaneous estimation, we identify and estimate two separate modeling components, both of which are surfaces: (a) a single *smooth regression coefficient vector*, which effectively ensembles a smooth surface while varying along the temperature covariate, and (b) an unknown and nonlinear link function, which too varies along the temperature covariate, yielding a *link surface* and thus extends the work of Eilers, Li and Marx (2009 and 2011). Although the first component 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

linear, the second component explicitly models the nonlinearity, enhancing insight into the process. The combination of these components can lead to a systematic and tractable modeling approach, that is statistical in nature, while in some cases having improved external prediction performance when compared to identity link model variants and partial least squares.

2 Motivating example

We revisit data with permission from Zhenyu Wang and Age Smilde, where the response y comes from the composition (mole fraction) of a mixture of three components (water, 1,2-ethanediol, 3-amino-1-propanol). There are $m = 34$ mixtures: 3 pure, 12 edge, and 19 interior. Corresponding to each ternary mixture, there exists extremely rich spectroscopy regressor information, taken under $\breve{p} = 12$ temperature conditions: (30, 35, 37.5, 40, 45, 47.5, 50, 55, 60, 62.5, 65, 70^o C). Each "signal" actually consists of numerous digitizations ($p = 401$) along the wavelength axis (700 to 1100, by 1 nm). The primary goal is reliable future (external) prediction. The data set brings some unique structure and several challenges: (a) The response is measured *exactly* at the molar level, and only at several dozen concentrations. (b) The rich covariate information has dimension far greater than the number of observations. (c) Internal prediction is *not* of interest, as it could be perfectly done in infinitely many ways. (d) It is the signal regressors themselves, and *not* the responses, that change with changes in the covariate t. We view that the additional covariate t as simply being recorded, while the signal regressors are generated. In this way the classical MVC approach could be more appropriately modified to allow interactive effects of t with the signal regressors, which is related to the work of Eilers and Marx (2003). This structure plays right into the hands of an interesting varying-coefficient model: As temperature affects the signal regressor information, it stands to reason that their corresponding coefficients should also vary with t. With a smoothness constraints within the estimation of both the signal coefficients and the varying-index t, a coefficient surface gels into shape. Original generalizations are made on two fronts: (i) The unknown, but explicit, link function is estimated within the context of a varying-coefficient signal model, and (ii) The link function is further allowed to vary itself, producing a link surface. Details of estimation and regularization of this *doubly*-varying model will follow below.

3 First component: varying-coefficients for signal

The specifics of the data structure are as follows: each observation consists of the data triplet: (y_i, x_i, t_i), where $i = 1, \ldots, N$ and $N = m\breve{p}$, since each y has spectra at each temperature. The response y_i is again a scalar, and independence among the responses is assumed, with common variance

$\text{var}(y) = \sigma^2$. The signal x_i is one-dimensional, consisting of a $p \times 1$ vector of ordered regressors. Completing the triplet is the covariate t_i, e.g. temperature. We first consider modeling rich signal regressors, with coefficients that are allowed to vary with another covariate. We refer to this model as *varying-coefficient penalized signal regression* or VPSR. Given the signal matrix $X = [x_{ij}]$ $(i = 1, \ldots, N; j = 1, \ldots, p)$ and the (unknown) coefficient surface $\alpha(v, t)$, the mean of the ith response can be denoted as $\mu_i = \sum_{j=1}^{p} x_{ij}\alpha(v_j, t_i)$, with v indexing wavelength. This expression explicitly shows the varying-coefficient structure that depends on t. Approximating the coefficient surface by $(n \times \breve{n})$ tensor product B-splines, it can be shown that $\mu = \mathbf{U}\gamma$, where

$$\mathbf{U} = \mathbf{U} \square \breve{\mathbf{B}} = (\mathbf{U} \otimes \mathbf{1}_{\breve{\mathbf{n}}}^{\mathrm{T}}) \odot (\mathbf{1}_{\mathbf{n}}^{\mathrm{T}} \otimes \breve{\mathbf{B}}) \tag{1}$$

is of dimension $N \times (n\breve{n})$, with $U = XB$. The basis B $(\breve{B})$ corresponds to wavelength (temperature). It is useful to view $\mathbf{U}$ as *effective* regressors. The penalized objective function to be minimized is defined as

$$\begin{aligned} Q_P(\gamma) &= \sum_{i=1}^{m}(y_i - \mathbf{u}_i^{\mathrm{T}}\gamma)^2 + \lambda \sum_{r=1}^{n} \gamma_{r\bullet} D_d^{\mathrm{T}} D_d \gamma_{r\bullet}^{\mathrm{T}} + \breve{\lambda} \sum_{s=1}^{\breve{n}} \gamma_{\bullet s}^{\mathrm{T}} D_{\breve{d}}^{\mathrm{T}} D_{\breve{d}} \gamma_{\bullet s} \\ &= ||y - \mathbf{U}\gamma||^2 + \lambda ||P\gamma||^2 + \breve{\lambda} ||\breve{P}\gamma||^2, \end{aligned} \tag{2}$$

where $\mathbf{u}_i^{\mathrm{T}}$ is the ith row of $\mathbf{U}$ and $\gamma_{r\bullet}$ $(\gamma_{\bullet s})$ denotes the rth row (the sth column) of tensor Γ. Define $P = (D_d^{\mathrm{T}} D_d) \otimes I_{\breve{n}}$ and $\breve{P} = I_n \otimes (D_{\breve{d}}^{\mathrm{T}} D_{\breve{d}})$, where I denotes the identity and d denotes the order of the difference penalty. The non-negative tuning parameters allow continuous control over smoothness, but are chosen in a greedy way to minimize external prediction error. For given tuning parameters $(\lambda, \breve{\lambda})$, the VPSR solution for objective (2) is $\hat{\gamma} = (\mathbf{U}^{\mathrm{T}}\mathbf{U} + \lambda P^{\mathrm{T}} P + \breve{\lambda} \breve{P}^{\mathrm{T}} \breve{P})^{-1} \mathbf{U}^{\mathrm{T}} y$.

4 Implementing a two-dimensional link surface

To allow model flexibility, we capture the nonlinear features of the response process through a general (unknown) link function, that can vary across the covariate, yielding a *varying-link surface*. The extended model is of the form $\mu = f(\mathbf{U}\gamma, t)$; f is a two-dimensional function of both the linear predictor $(\eta = \mathbf{U}\gamma)$ and the covariate t. We refer to this model as *varying-coefficient single-index signal regression* or VSISR. We have a two-dimensional surface imbedded within another two-dimensional surface. We have the the varying-link surface, which is estimated with its own tensor P-spline coefficients θ. The objective in (2) can be extended as

$$Q_P^{\star}(\gamma, \theta) = ||y - f(\mathbf{U}\gamma, t)||^2 + \lambda ||P\gamma||^2 + \breve{\lambda} ||\breve{P}\gamma||^2 + \lambda_f ||P_f \theta||^2 + \breve{\lambda}_f ||\breve{P}_f \theta||^2. \tag{3}$$

P-splines are a natural choice: (a) They are regression based and are easy to use and optimize. (b) Heavy smoothing along one variable, using a dth order penalty, leads to polynomial structure of order $d-1$. (c) The *partial* first derivative of f with respect to one variable (denoted as $\dot{f}_\partial(\eta,t)=\frac{\partial f(\eta,t)}{\partial \eta}$) can be easily computed (and is needed in our algorithm). For simplicity in notation, denote $T\{\mathbf{V},y,(\lambda_f,\breve{\lambda}_f),(d_f,\breve{d}_f),(n_f,\breve{n}_f)\}$ as the operation of fitting a tensor product cubic P-spline scatter smoother on $\mathbf{V}=\{v,t\}$ (the input variables) and y (the response) using the penalty tuning parameters $(\lambda_f,\breve{\lambda}_f)$ and difference orders $(d_f,\breve{d}_f)$ on the $n_f\times\breve{n}_f$ knot grid.
For fixed estimate of the surface f (i.e. given θ), and given covariate t, the coefficient vector γ can be estimated using a (first-order) Taylor series approximation of the function f (about the current estimate, γ_0). Specifically, the current $\mu=f(\mathbf{U}\gamma,t)$ can be approximated by

$$f(\mathbf{U}\gamma,t)\approx f(\mathbf{U}\gamma_0,t)+\text{diag}\{\dot{f}_\partial(\mathbf{U}\gamma_0,t)\}\mathbf{U}(\gamma-\gamma_0). \tag{4}$$

Using (4), with fixed f and given t, we have an approximation of $Q_P^\star$

$$\begin{aligned} Q_P^\star &\approx ||y-f(\mathbf{U}\gamma_0,t)-\text{diag}\{\dot{f}_\partial(\mathbf{U}\gamma_0,t)\}\mathbf{U}(\gamma-\gamma_0)||^2+\lambda||P\gamma||^2+\breve{\lambda}||\breve{P}\gamma||^2 \\ &= ||y^\star-\mathbf{U}^\star\gamma||^2+\lambda||P\gamma||^2+\breve{\lambda}||\breve{P}\gamma||^2, \end{aligned} \tag{5}$$

where $y^\star=y-f(\mathbf{U}\gamma_0,t)+\text{diag}\{\dot{f}_\partial(\mathbf{U}\gamma_0,t)\}\mathbf{U}\gamma_0$ and $\mathbf{U}^\star=\text{diag}\{\dot{f}_\partial(\mathbf{U}\gamma_0,t)\}\mathbf{U}$. Note that (5) implies that given f (or θ), the optimal γ that minimizes the right-hand side of (5) can be obtained through the varying-coefficient penalized signal regression solution, $\text{VPSR}\{\mathbf{U}^\star,y^\star,(\lambda,\breve{\lambda}),(d,\breve{d}),(n,\breve{n})\}$.

5 Optimization of the penalty tuning parameters

Reliable external prediction is the primary goal, and the data are split into three disjoint sets: training, validation, and test sets. To start, apply VSISR to the training set and choose "optimal" $(\lambda,\breve{\lambda},\lambda_f,\breve{\lambda}_f)$ to minimize error on the validation set,

$$\text{RMSEV}=\sqrt{\frac{1}{N^{valid}}\sum_{i=1}^{N^{valid}}(y_i-\hat{y}_{vi})^2},$$

where $\hat{y}_{vi}$ is the predicted response for the ith subject in the validation set, using the parameter estimates from the training model. External predictive performance RMSEP can be calculated in a similar fashion, now using the optimal tuning with a model fit that combines the training and validation data,

$$\text{RMSEP}=\sqrt{\frac{1}{N^{test}}\sum_{i=1}^{N^{test}}(y_i-\hat{y}_i)^2}.$$

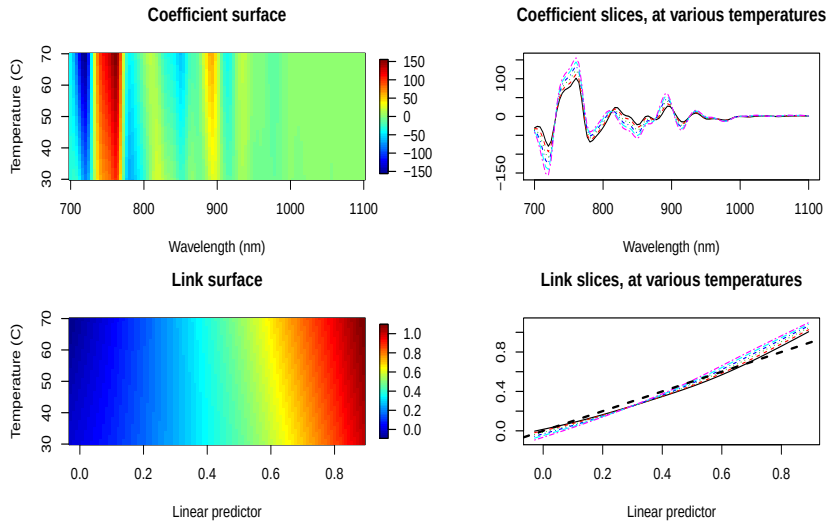

FIGURE 1. 1,2-ethanediol component: Optimal varying signal coefficient surface (top panels); optimal varying link surface (bottom panels).

TABLE 1. External prediction RMSEP using optimal models.

Response	VSISR	VPSR	PLS
Water	**0.0087**	0.0129	0.0367
1,2-ethanediol	**0.0094**	0.0104	0.0134
3-amino-1-propanol	0.0146	**0.0063**	0.0099

Table 1 presents the root mean square error of prediction (RMSEP) for the external prediction set, using optimal VSISR, VPSR, and PLS models. For responses water and 1,2-ethanediol, we find an impressive improvement in external prediction for VSISR when compared to either VPSR or PLS, e.g. reductions as large as 77%. Figure 1 provides image plots for the varying-coefficient surface (upper), the varying-link surface (lower). The right panels display various slices of each surface.

6 Discussion

The notion of the "missing link" is related to the work of Muggeo and Ferrara (2008). The basic appeal of VSISR includes the following: Its simplicity, with doubly-varying coefficient and link surfaces. Unlike many competitors, the indexing associated with the VSISR surfaces can identify potentially "important" regions. The nonlinear structure is targeted, provid-

ing potential process insight. Each smooth surface can have a very general (non-additive) structure. We have highly competitive external prediction ability. There is no "black box" algorithm. No data preprocessing: the entire signals are used with companion covariates. The regularization manages the severely ill-conditioned model. As signal precision increases, the system of equations remains $n\breve{n}$. We note that we modeled one mixture component at a time, and future work could joint modeling, where the sum of the concentrations is constrained to be one.

`Algorithm VSISR`

1. Initializations:
   - Fix the tuning parameter values $(\lambda, \breve{\lambda}, \lambda_f, \breve{\lambda}_f)$ for Steps 1 and 2
   - Fix number of knots $(n, \breve{n}, n_f, \breve{n}_f)$
   - Fix penalty order $(d, \breve{d}, d_f, \breve{d}_f)$
   - Create $\mathbf{U} = U \square \breve{B}$
   - Initialize $\hat{\gamma} = \text{VPSR}\{\mathbf{U}, y, (\lambda, \breve{\lambda}), (d, \breve{d}), (n, \breve{n})\}$ and $\hat{\eta} = \mathbf{U}\hat{\gamma}$
2. Cycle until convergence of $\hat{\gamma}$
   - Estimate $\hat{f}(\hat{\eta}, t)$ and the estimate of the derivative $\dot{f}_\partial(\hat{\eta}, t)$ from $T\{\{\eta, t\}, y, (\lambda_f, \breve{\lambda}_f), (d_f, \breve{d}_f), (n_f, \breve{n}_f)\}$
   - Obtain $y^\star$ and $\mathbf{U}^\star$
   - Update $\hat{\gamma} = \text{VPSR}(\mathbf{U}^\star, y^\star, (\lambda, \breve{\lambda}), (d, \breve{d}), (n, \breve{n}))$
   - Constrain $\hat{\gamma}/||\hat{\gamma}||$ and update $\hat{\eta} = \mathbf{U}\hat{\gamma}$
3. Prediction: $\hat{y}_{new} = \hat{f}(\mathbf{u}_{new}^{\mathrm{T}}\hat{\gamma}, t_{new})$

`end algorithm`

References

Eilers, P.H.C., Li, B., and Marx, B.D. (2009). Multivariate calibration with single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, **96**, 196 – 202.

Eilers, P.H.C. and Marx, B.D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, **66**, 159 – 174.

Marx, B.D. (2015). Varying-coefficient single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, **143**, 111 – 121.

Marx, B.D., Eilers, P.H.C., and Li, B. (2011). Multidimensional single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, **109**, 120 – 130.

Muggeo V. and Ferrara G. (2008). Fitting generalized linear models with unspecified link function: a P-spline approach. *Computational Statistics and Data Analysis*, **52**, 2529 – 2537.

Boosting the discriminatory power of survival models via the C-index and stability selection

Andreas Mayr[1], Benjamin Hofner[1], Matthias Schmid[2]

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] Department of Medical Biometry, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany

E-mail for correspondence: andreas.mayr@fau.de

Abstract: We present a statistical modelling approach for potentially high-dimensional survival data based on boosting a smooth version of the concordance index (C-index). Due to this objective function, the resulting statistical models are optimal with respect to their discriminatory power. We combine the gradient boosting algorithm with stability selection to enhance its variable selection properties. The performance of the approach is demonstrated in a simulation study and an application on biomarkers for breast cancer patients, comparing it to more common penalized Cox-regression models.

Keywords: Time-to-event data; Boosting; Stability selection; Concordance index.

1 Introduction

In the evaluation of biomarkers or gene-signatures for survival data the focus is often on the ability of the new marker to discriminate between patients with larger and smaller survival times ("discriminatory power"). In practice, however, these predictors are often estimated based on Cox-regression, which does not necessarily lead to optimal models for this task. We present a new statistical modelling approach based on boosting the concordance index (C-index) which leads to additive predictors η

$$\eta := \beta_0 + \sum_{l=1}^{p} \beta_l x_l = X^{\mathrm{T}} \boldsymbol{\beta}$$

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

that are optimal with respect to their discriminatory power (Mayr and Schmid, 2014). In other words, the ranking of the linear combinations η show the best agreement with the rankings of the survival times.
Our approach is feasible for high-dimensional data and incorporates variable selection. In combination with stability selection it results in particularly sparse models and is able to identify only the most influential *and* stable predictors.

2 The *C*-index for survival data

The C-index is a general discrimination measure for the evaluation of prediction models. For time-to-event outcomes, it is defined as

$$C := \Pr(\eta_j > \eta_i | T_j < T_i),$$

where T_j, T_i are the survival times and η_j and η_i the predictors of two observations in an i.i.d. test sample. The C-index measures whether large values of η are associated with short survival times T and vice versa. The interpretation is similar to the wide known AUC: A C-index of 1 represents a perfect discrimination while a C-index of 0.5 will achieved by a completely non-informative marker.
Although following a relatively simple and straight-forward definition, in practice the estimation of the C-index becomes problematic in samples with censoring. Some estimators in the literature simply ignore observation pairs where the smaller survival time was censored, however this can lead to biased results (cf., Schmid and Potapov, 2012). To overcome this issue, Uno et al. (2011) proposed an asymptotically unbiased estimator

$$\hat{C}_{\text{Uno}}(T, \eta) := \frac{\sum_{j,i} \Delta_j (\hat{G}(T_j))^{-2} \text{I}(T_j < T_i) \text{I}(\hat{\eta}_j > \hat{\eta}_i)}{\sum_{j,i} \Delta_j (\hat{G}(T_j))^{-2} \text{I}(T_j < T_i)},$$

where $\hat{G}(t)$ denotes the Kaplan-Meier estimator of the unconditional survival function of T_{cens} (estimated from the learning data).

3 Boosting the *C*-index with stability selection

We apply a gradient boosting algorithm (Bühlmann and Hothorn, 2007) with simple linear models as base-learners to find the optimal predictor η with respect to the C-index. Boosting originally emerged from machine learning, but during the last 15 years has evolved into a powerful tool to estimate the unknown quantities in general regression settings (Mayr et al., 2014). The basic idea is to apply simple regression functions as base-learners (in the easiest case simple linear models) and iteratively fit them one-by-one to the negative gradient of the loss function. In every boosting

iteration only the best-performing base-learner is included in the model, effectively leading to variable selection.
Simply using the C-index as loss function, however, is unfeasible because $\hat{C}_{\text{Uno}}(T, \eta)$ is not differentiable with respect to η_i. To solve this problem, we approximate the indicator function $\text{I}(\hat{\eta}_j > \hat{\eta}_i)$ by the sigmoid function

$$K(\hat{\eta}_j - \hat{\eta}_i) = 1 \Big/ \left(1 + \exp\left(-\frac{(\hat{\eta}_j - \hat{\eta}_i)}{\sigma}\right)\right),$$

leading to a smooth estimator $\hat{C}_{\text{smooth}}(T, \eta)$ that will serve as loss function for the boosting algorithm.
The boosting algorithm then descents the empirical risk via iteratively fitting p separate linear models (one for each predictor) to the negative gradient of $\hat{C}_{\text{smooth}}(T, \eta)$.
To ensure particularly sparse models, we additionally incorporate the stability selection approach proposed by Meinshausen and Bühlmann (2010). The principle idea is to iteratively apply subsampling and carry out boosting on each of the subsamples until a certain number of variables q was selected. Finally, the average selection probabilities are computed for all predictors and only variables that pass a certain threshold π_{thr} are included in the final model. An important advantage of the stability selection approach is that it controls the per-family error rate (PFER) and provides error bounds for the number of false positives (depending on p, q and π_{thr}). For an in-depth overview on stability selection in the context of boosting, see Hofner et al. (2015).

4 Simulation study

We first carried out a simulation study to check the performance of our algorithm under known conditions. We simulated high-dimensional ($p = 1000, n = 500$) survival data based on log-logistic distribution. Only 4 of the 1000 predictors had an actual effect on the outcome, 50% of the outcome observations were censored.
For stability selection we specified $q = 15$ and $\pi_{\text{thr}} = 0.6$, leading to a PFER of 0.172. The resulting coefficients from $B = 100$ simulation runs are presented in Figure 1. On average, boosting with stability selection selected 4.12 (range 3 to 5) predictors. The truly informative predictors (X_1 – X_4) had a selection probability ranging from 96% to 100% (Figure 1 refers to one simulation run) compared to only 0.02% for the non-informative variables.
The resulting models were evaluated on independent test data: The resulting median $\hat{C}_{\text{Uno}} = 0.778$ (range = 0.670 – 0.818) was close to theoretically optimal discriminatory power from the true coefficients ($\hat{C}_{\text{Uno}} = 0.781$) and also outperformed penalized Cox approaches like the Cox lasso ($\hat{C}_{\text{Uno}} = 0.759$) and Cox ridge ($\hat{C}_{\text{Uno}} = 0.760$).

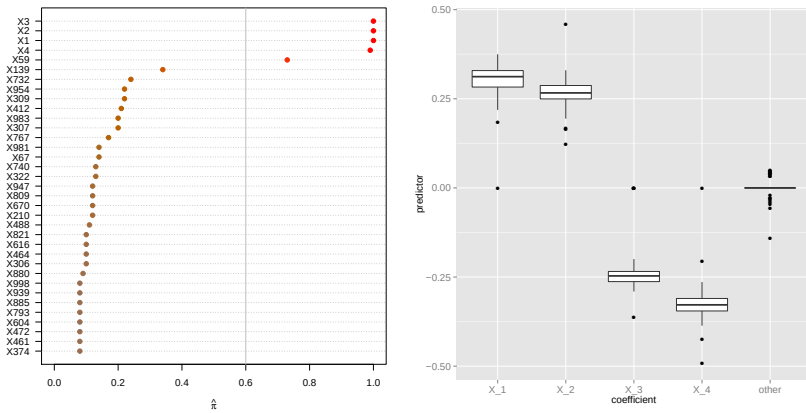

FIGURE 1. Selection probabilities of stability selection from one simulation run (left) and the final coefficients (right) from 100 simulation runs: Only variables $X_1 - X_4$ had a true effect on the survival outcome.

5 Breast cancer data

We finally apply our approach on a data set to build a gene signature for the prediction of the development of distant metastases in breast cancer patients. The data set ($n = 196$) was collected by Desmedt et al. to validate a 76-gene expression signature from the literature. In addition to the expression levels of the 76 genes, four clinical predictor variables were considered (tumor size, estrogen receptor (ER) status, tumor grade and age). Observed metastasis-free survival ranged from 125 days to 3652 days, with 79.08% of the survival times being censored.

To generate separate data sets to evaluate results, we constructed 100 training and test data sets via stratified subsampling. On each of the training samples, we fitted the boosting algorithm for the C-index with and without stability selection as well as lasso and ridge penalized Cox models as benchmarks.

The main results of our analysis are presented in Figure 2. As expected, boosting with the C-index as loss functions leads to a higher discriminatory power (median $\hat{C}_{\text{Uno}} = 0.736$) than the Cox-based approaches. Additionally incorporating stability selection does not further improve the performance on test data ($\hat{C}_{\text{Uno}} = 0.735$), however leads to sparser models, containing on average only 23 (range = 17 – 29) predictors.

6 Discussion

We presented a gradient boosting algorithm to optimize additive predictors with respect to the rank-based C-index, which is a performance criterion

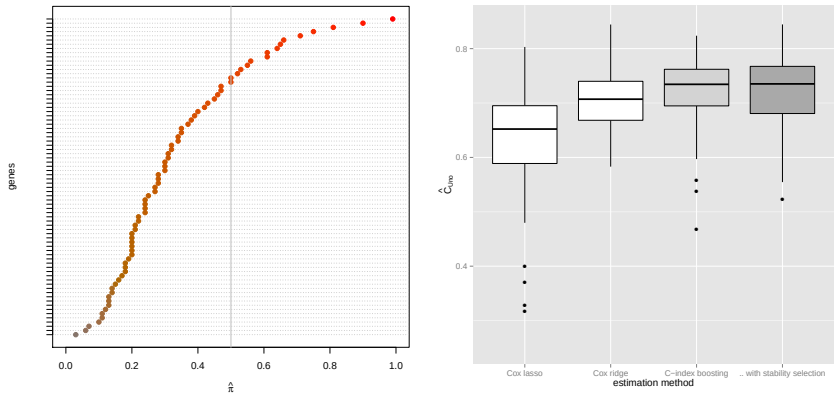

FIGURE 2. Selection probabilities of stability selection on one training set (left) and the resulting C-index on 100 test samples (right) from the breast cancer application comparing the new boosting with and without stability selection with competing penalized Cox approaches.

regarding the discriminatory power of prediction rules. In the context of survival data it measures the association between the prediction model and the observed survival times. Our boosting approach hence results in survival models that are optimal with respect to this C-index.

We combined the approach with stability selection, which is a recently proposed general procedure to enhance variable selection tools via subsampling, effectively controlling the per-family wise error rate. Boosting the C-index in combination with stability selection leads to particularly sparse models: The numerical results from the simulation study and the breast cancer data suggest that the combined algorithm is able to correctly identify the most influential predictors in potentially high-dimensional settings. When it comes to the resulting discriminatory power, boosting the C-index clearly outperformed common Cox-based penalization approaches. This finding should not be further surprising, as our approach in contrast to Cox regression is particularly tailored to optimize the ability of the model to differentiate between observations with smaller and larger survival times. The additional sparsity resulting from stability selection, however, does not necessarily lead to more accurate predictions. It has to be highlighted, that particular sparse and interpretable models and the highest possible discriminatory power are two different goals that may not be achievable at the same time (cf., Hothorn, 2010).

A promising extension of our approach could be to consider also non-linear effects in η. Due to the modular nature of boosting, this could be done relatively easily by incorporating P-spline base-learners instead of the currently used simple linear models.

Acknowledgments: The work on this article was supported by the German Research Foundation (DFG), grant SCHM 2966/1-1 and the Interdisciplinary Center for Clinical Research (IZKF) of the Friedrich-Alexander University Erlangen-Nürnberg (Project J49).

References

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477 – 522.

Desmedt, C., Piette, F., Loi, S., Wang, Y., et al. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter series. *Clinical Cancer Research*, **13**, 3207 – 3214.

Hofner, B., Boccuto, L., and Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*.

Hothorn, T. (2010). Discussion: Stability Selection. *Journal of the Royal Statistical Society, Series B*, **72**, 463 – 464.

Mayr, A. and Schmid, M. (2014). Boosting the concordance index for survival data – a unified framework to derive and evaluate biomarker combinations. *PLoS ONE*, **9**, e84483.

Mayr, A., Binder, H. Gefeller O., and Schmid, M. (2014). The evolution of boosting algorithms – from machine learning to statistical modelling. *Methods of Information in Medicine*, **53**, 419 – 427.

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

Schmid, M. and Potapov, S. (2012). A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in Medicine*, **31**, 2588 – 2609.

Uno, H., Cai, T., Pencina, M., D'Agostino, R., and Wei, L. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, **30**, 1105 – 1117.

Clustering mixed data via latent variable models

Damien McParland[1], Isobel Claire Gormley[1]

[1] School of Mathematical Sciences, University College Dublin, Ireland

E-mail for correspondence: damien.mcparland@ucd.ie

Abstract: A model based clustering procedure for data of mixed type, termed clustMD, is developed using a latent variable model. It is proposed that a latent variable, following a mixture of Gaussian distributions, generates the observed data of mixed type. The observed data may be any combination of continuous, binary, ordinal or nominal variables. The model employs a parsimonious covariance structure for the latent variables, leading to a suite of six clustering models that vary in complexity and provide an elegant and unified approach to clustering mixed data. An expectation maximisation (EM) algorithm is used to estimate the model; in the presence of nominal data a Monte Carlo EM algorithm is required. The clustMD model is illustrated by clustering prostate cancer patients, on whom measurements of mixed type have been recorded.

Keywords: Latent variables; Mixture model; Mixed data; Monte Carlo EM.

1 Introduction

Clustering mixed data has received increasing attention in the literature for some time. Previous approaches include the location mixture model (Hunt and Jorgensen, 1999) while latent variable approaches have been proposed by Browne and McNicholas (2012) and McParland et al. (2014).

The proposed clustMD model is a mixture of latent Gaussian distributions, and provides a parsimonious and computationally efficient approach to clustering mixed data. The mixture of Gaussian models has become a traditional approach to clustering continuous data and parsimonious versions of these models were developed by Fraley and Raftery (1998) among others. A similar parsimony ethos underpins the proposed clustMD model and a suite of models of varying levels of parsimony are developed.

An expectation maximisation (EM) algorithm (Dempster et al., 1977) is used for inference. In the presence of nominal data however the expectation

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

step is intractable so a Monte Carlo EM algorithm is required in such cases. The performance of clustMD is demonstrated by clustering a group of prostrate cancer patients, on whom measurements of mixed type have been recorded.

2 Model

Mixture models are a very useful clustering tool and much research has been devoted to their development. Finite mixture models assume the data arise from a finite number of homogeneous clusters. A detailed explanation of mixture models may be found in McLachlan and Peel (2000). The clustMD model employs a mixture of latent variable models to cluster mixed type data. In brief, the clustMD model assumes the observed J mixed type variables in each observation vector $\underline{y}_i$ are a manifestation of an underlying latent continuous vector, $\underline{z}_i$ (for $i = 1, \ldots, N$), which follows a Gaussian mixture distribution.

Under the clustMD model, continuous variables follow a multivariate Gaussian distribution i.e. if variable j is continuous, $y_{ij} = z_{ij} \sim \mathrm{N}(\mu_j, \sigma_j^2)$. In the case of an ordinal variable, it is supposed that the observed response, y_{ij} is a categorical manifestation of the latent continuous variable, z_{ij}.

For ordinal variable j let $\underline{\gamma}_j$ denote a vector of thresholds which partition the real line. The value of the latent Gaussian variable z_{ij} in relation to $\underline{\gamma}_j$ determines the observed ordinal response y_{ij}. The threshold parameters are not of primary interest in clustMD. Thus, for reasons of identifiability and efficiency, $\gamma_{j,k}$ is fixed such that $\gamma_{j,k} = \Phi^{-1}(\delta_k)$, where δ_k is the proportion of the observed values of variable j that are less than or equal to level k and Φ is the standard Gaussian cumulative distribution function.

Nominal variables are more difficult to model since the set of possible responses is unordered. In this case, a multivariate latent vector is assumed to underlie the observed nominal variable. For nominal variable j with K_j possible responses, the underlying continuous vector has $K_j - 1$ dimensions i.e. $\underline{z}_{ij} = (z_{ij}^1, \ldots, z_{ij}^{K_j-1}) \sim \mathrm{MVN}_{K_j-1}(\underline{\mu}_j, \Sigma_j)$. The observed nominal response y_{ij} is a manifestation of the values of the elements of $\underline{z}_{ij}$ relative to each other and to a threshold, assumed to be 0. That is,

$$y_{ij} = \begin{cases} 1 & \text{if } \max\limits_k \{z_{ij}^k\} < 0 \\ k & \text{if } z_{ij}^{k-1} = \max\limits_k \{z_{ij}^k\} \text{ and } z_{ij}^{k-1} > 0 \text{ for } k = 2, \ldots, K_j. \end{cases}$$

A similar latent variable approach to modelling nominal data is used by the multinomial probit model (Geweke et al., 1994).

The latent continuous data underlying both the ordinal and nominal data are assumed to be Gaussian, as are any observed continuous data. Thus the joint vector of observed and latent continuous data is assumed to follow a multivariate Gaussian distribution $\underline{z}_i \sim \mathrm{MVN}_P(\underline{\mu}, \Sigma)$. Since more than one

latent dimension is required to model each nominal variable P depends on the number of latent dimensions required. This model provides a unified way to simultaneously model continuous, ordinal and nominal data.

This joint model is now embedded in a finite mixture model, facilitating the clustering of mixed data. It is assumed that $\underline{z}_i$ follows a mixture of G Gaussian distributions i.e. $\underline{z}_i \sim \sum_{g=1}^{G} \pi_g \text{MVN}_P(\underline{\mu}_g, \Sigma_g)$ where π_g is the marginal probability of belonging to cluster g and $\underline{\mu}_g$ and Σ_g denote the mean and covariance for cluster g respectively. Gaussian parsimonious mixture models utilise an eigenvalue decomposition of the cluster covariance matrix $\Sigma_g = \lambda_g D_g A_g D_g$. The scaler λ_g controls the cluster volume, D_g is a matrix of eigenvectors of Σ_g and A_g is a diagonal matrix of eigenvalues of Σ_g. The decomposed covariance is constrained in various ways to produce parsimonious models.

The covariance matrix for this model is assumed to be diagonal, meaning that $D_g = I$, the identity matrix. Thus under clustMD $\Sigma_g = \lambda_g A_g$. These parameters can then be constrained to be different or equal across groups and A can also be constrained to be the identity matrix. This gives rise to a suite of 6 clustMD models with varying levels of parsimony. The 6 clustMD models and corresponding constraints are detailed in Table 1.

TABLE 1. Covariance matrix structures of varying degrees of parsimony. Parameters are unconstrained (U), constrained (C) to be equal across groups or equal to the identity (I).

				# Covariance parameters	
Model	λ	A	D	No nominal variables	Nominal variables
EII	C	I	I	1	1
VII	U	I	I	G	$2G-1$
EEI	C	C	I	$1+P$	$C+O$
VEI	U	C	I	$G+P$	$2G+C+O-2$
EVI	C	U	I	$1+GP$	$G(P-2)+C+O-P+2$
VVI	U	U	I	$G(1+P)$	$P(G-1)+O$

If no nominal variables are present in the data set then the clustMD model is identified because the threshold parameters are fixed. However, in the presence of nominal variables, the model as it stands is not identified. Infinitely many combinations of the model parameters will give rise to the same likelihood. A separate volume parameter $\tilde{\lambda}_g$ which applies only to the latent dimensions corresponding to nominal variables is introduced. Thus the diagonal elements of Σ_g corresponding to these dimensions are $\tilde{\lambda}_g a_{gp}$, where a_{gp} is the p^{th} diagonal element of A_g. Various sum to one constraints are applied to each model in order to obtain unique parameter estimates.

3 Model fitting and selection

The clustMD model is fitted using an expectation maximisation (EM) algorithm (Dempster et al., 1977). However, if nominal data are present, the expectation step is not tractable and a Monte Carlo approximation is required. In these cases, the model fitting algorithm is a Monte Carlo EM (MCEM) algorithm. Convergence is guaranteed even though a Monte Carlo approximation is used. However, the monotone increase in the likelihood at each iteration, which a standard EM algorithm guarantees, does not apply here. For more detail on convergence and the Monte Carlo EM algorithm see McLachlan and Krishnan (2008).
The best fitting covariance structure and number of components, is selected using an approximation of the Bayesian Information Criterion (BIC) (Schwarz (1978)). The BIC cannot be evaluated for clustMD models since the observed likelihood relies on the calculation of intractable integrals. However, the observed likelihood may be estimated by assuming that the continuous and categorical variables are independent and utilising the same Monte Carlo approximation that is required for the model fitting algorithm. This approximation has been found to perform well in a simulation study.

4 Application: prostate cancer patients

Twelve mixed type measurements are available for 475 prostate cancer patients who were diagnosed as having either stage 3 or 4 prostate cancer. Eight of the variables are continuous, three are ordinal and one is nominal. The variables analysed are presented in Table 2. The post trial survival status of the patients and the cause of death of those patients who died during the study are available but are not used in the analysis. This data set was analysed by Hunt and Jorgensen (1999) using their location mixture model. The data may be found in Andrews and Herzberg (1985).

TABLE 2. Variables analysed in the prostate cancer data set. The type of variable is denoted by a letter: continuous (C), ordinal (O) and nominal (N). The number in parentheses after the categorical variables indicates the number of possible responses for that variable.

Variable		Variable	
Age	C	Index of tumour stage	C
Weight	C	Serum prostatic acid phosphatase	C
Systolic blood pressure	C	Performance rating	O(4)
Diastolic blood pressure	C	Cardiovascular disease history	O(2)
Serum haemoglobin	C	Bone metastasis	O(2)
Size of primary tumour	C	Electrocardiogram code	N(3)

The suite of 6 clustMD models were fitted to the set of prostate cancer patients with the number of clusters ranging from 1 to 4. The model that maximises the approximate BIC is a 3 cluster model, with the EVI covariance structure. The patients are diagnosed as having either stage 3 or 4 prostate cancer so it is interesting that a 3 cluster model is chosen here. In their paper, Hunt and Jorgensen (1999) sought to identify the cancer stage and only considered 2 cluster models. It is therefore interesting that a 3 cluster model is chosen here. A cross tabulation of the clusters identified by the clustMD model versus the cancer stage diagnosis is given in Table 3. It seems reasonable to expect that clusters 1 and 3 would be similar given that both clusters consist primarily of stage 3 patients. However, comparing the mean vectors for these clusters, it can be seen that patients in cluster 3 are on average heavier and have higher levels of systolic and diastolic blood pressure. They are more likely to have a history of cardiovascular disease and their electrocardiogram score is more likely to indicate a serious anomaly. These differences suggest that a cardiovascular health issue differentiates patients in cluster 3 from those in cluster 1. Indeed, by examining the post trial survival status it can be seen that 21% of cluster 3 patients are alive at the end of the trial, only 7% died of prostatic cancer but 51% died from heart or vascular disease or a stroke during the trial. The remaining 21% died from other causes.

TABLE 3. Cross tabulation of estimated cluster labels versus the diagnosed prostate cancer stage. The Rand index is 0.72 (adjusted Rand is 0.49).

	Stage 3	Stage 4
Cluster 1	207	14
Cluster 2	21	175
Cluster 3	45	13

By comparing the mean vectors across clusters it is clear that patients in cluster 2 have, on average, larger tumours and higher levels of serum prostatic acid phosphatase than patients in clusters 1 and 3. They are also more likely to have bone metastases. Analysing the survival status of clusters 1 and 2 it can be seen that 49% of patients in cluster 2 died from prostatic cancer (compared to 10% in cluster 1), 21% survived until the end of the trial (37% in cluster 1) and 30% died from other causes (57% in cluster 1).

5 Discussion

The clustMD model presented here provides a suite of parsimonious mixture models for clustering mixed type data. The latent variable framework provides an elegant unifying structure for clustering this type of data.

Unless a very large number of Monte Carlo simulations are required for the E-step, due to a sparsely observed response on a nominal variable, the model fitting algorithm is very computationally efficient. Indeed, if no nominal variables are present each model is fitted in a matter of seconds.
Future research directions are plentiful. The most obvious deficiency of the proposed clustMD models is the independence assumption between variables. It would be very beneficial to allow for full covariance matrices which can model the dependencies between variables of mixed type.
The Monte Carlo approximation used in the E-step of the model fitting algorithm is a simple and effective solution but it is not without issues. If the probability of observing a particular response on a nominal variable is small for a particular cluster then a large number of Monte Carlo samples may be required to observe a response in this category. This can slow the model fitting algorithm or even cause instability. A more efficient way to approximate the intractable integrals would greatly improve model fitting.

References

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

Browne, R.P. and McNicholas, P.D. (2012). Model-based clustering and classification of data with mixed type. *Journal of Statistical Planning and Inference*, **142**, 2976 – 2984.

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.

Geweke, J., Keane, M., and Runkle, D. (1994) Alternative computational approaches to inference in the multinomial probit model. *The Review of Economics and Statistics*, **76**, 609 – 632.

Hunt, L. and Jorgensen, M. (1999) Mixture model clustering using the MULTIMIX program. *Australia and New Zealand Journal of Statistics*, **41**, 153 – 171.

McLachlan, G.J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

McParland, D., Gormley, I.C., McCormick, T.H., Clark, S.J., Kabudula, C.W., and Collinson, M.A. (2014). Clustering South African households based on their asset status using latent variable models. *Annals of Applied Statistics*, **8**, 747 – 776.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461 – 464.

Assessing goodness-of-fit for accelerated failure rate models: An insect ecology case-study

Rafael A. Moral[1], John Hinde[2], Clarice G.B. Demétrio[1], Edwin M.M. Ortega[1]

[1] Universidade de São Paulo, Piracicaba, Brazil
[2] National University of Ireland Galway, Galway, Ireland

E-mail for correspondence: `rafael.moral@usp.br`

Abstract: Survival analysis models have been extensively used to analyse time-until-event data. There is a range of extended models that incorporate different aspects, such as overdispersion/frailty, mixtures, and flexible response functions through semi-parametric models. It is important to assess goodness-of-fit of a fitted model so that statistical inference is accurate and reliable. A relatively simple way of doing so is through half-normal plots with simulation envelopes. In this work, we propose an accelerated failure rate model to analyse time-until-event data in a biological control context and present useful tools to assess goodness-of-fit, which are implemented as the `hnp` package in R.

Keywords: Biological control; Exponentiated-Weibull model; Half-normal plots with simulation envelopes; Location-scale modelling.

1 Introduction

In insect studies outcomes of interest vary widely and hence a wide range of data types are obtained, such as count and proportion data, continuous data, and multivariate data. When researching the ecology of insect species, the interest lies in describing several ecological processes, such as predation, competition, prey preference, amongst others. These ecological data can be used to inform decision making regarding many applied fields, especially pest management. Integrated pest management (IPM) comprises of a series of activities with the purpose of reducing damage from insect pests in agroecosystems.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

One common type of ecological data is the time until the occurrence of an event, e.g. death of a predator, time taken for a parasitoid to parasitise its prey, time until a competitor attacked another competitor, time until a predator attacked its prey. Because these variables are continuous and strictly positive, they are usually analysed through survival analysis models or non-parametric tests. Many parametric models have been proposed in the literature and a common objective is to use a well-fitted but not overly complex model to make inference from (Aitkin et al., 2009). It is very important to assess goodness-of-fit of survival models so that inference is not misleading. However, this may be a difficult task. A good alternative is to use diagnostic plots, such as half-normal plots with simulation envelopes (Demétrio et al., 2014).
The main goals of this work are to build and assess goodness-of-fit for location-scale models fitted to data from an experiment on prey preference. Predators were given a choice between parasitised and non-parasitised prey and the time until a predator attacked a prey was observed.

2 Material and methods

2.1 Case-study

A major pest of maize worldwide is the larva of *Spodoptera frugiperda*. Three important natural enemies of this pest are the stinkbug *Podisus nigrispinus* and the earwig *Euborellia annulipes*, both predators, and the parasitoid wasp *Campoletis flavicincta*. The latter lays eggs inside *S. frugiperda* larvae that hatch and develop inside them, killing the larva upon reaching the adult stage. This process causes substantial metabolic changes in the larvae which may be perceived by potential predators.
In a biological control context, it is important that predators act synergistically to promote pest suppression. In this sense, a desirable outcome when controlling *S. frugiperda* with multiple natural enemies, such as the three species mentioned above, would be weak competition between predators and preference for non-parasitised larvae so that the parasitoid's population is also maintained in the system. Information on the time taken by different predator species to attack prey and on prey preference are particularly relevant in this context.

2.2 Experimental design

To study the predatory behaviour of the stinkbug and the earwig when given the choice between parasitised and non-parasitised prey, an experiment was conducted in a completely randomized block design with 50 blocks and four treatments in a 2×2 (species and gender) factorial design, i.e. male and female stinkbugs and earwigs which were fasted for 24 hours prior to experiment set up. Each block consisted of four Petri dishes with

one predator and two *S. frugiperda* larvae, one of which was previously parasitised by *C. flavicincta* and another which was not. Each experimental unit was observed for one hour and the time (in seconds) taken until the predator firstly attacked a larva was recorded, as well as which larva the predator chose to effectively consume first (predator preference).

2.3 Modelling

Let $W_i, i = 1, \ldots, n$, be independent random variables representing recorded times until a predator attacked a larva, following a distribution from the family of exponentiated distributions. Then, the probability density function (pdf) of W_i may be written as

$$f_{W_i}(w_i; \boldsymbol{\theta}_i, a) = a[G(w_i; \boldsymbol{\theta}_i)]^{a-1} g(w_i; \boldsymbol{\theta}_i),$$

with $a > 0$ a shape parameter, $g(\cdot)$ a pdf and $G(\cdot)$ its respective distribution function and $\boldsymbol{\theta}_i$ the vector of parameters of the distribution. Taking the base distribution as Weibull with $g(u; \alpha_i, \gamma_i) = \gamma_i u^{\gamma_i - 1} e^{-\left(\frac{u}{\alpha_i}\right)^{\gamma_i}} / \alpha_i^{\gamma_i}$, $\alpha_i > 0$, $\gamma_i > 0$, and $G(u; \alpha_i, \gamma_i) = 1 - e^{-\left(\frac{u}{\alpha_i}\right)^{\gamma_i}}$, then

$$f_{W_i}(w_i; \alpha_i, \gamma_i, a) = a \left[1 - e^{-\left(\frac{w_i}{\alpha_i}\right)^{\gamma_i}}\right]^{a-1} \frac{\gamma_i}{\alpha_i^{\gamma_i}} w_i^{\gamma_i - 1} e^{-\left(\frac{w_i}{\alpha_i}\right)^{\gamma_i}}, \qquad (1)$$

is the pdf of the exponentiated-Weibull distribution (Nadarajah et al., 2013). Constructing a location-scale model from (1) is straightforward and can be done by finding the distribution of $Y_i = \log W_i$ and reparameterising $f_{Y_i}(y_i; \alpha_i, \gamma_i, a)$ in terms of $\mu_i = \log \alpha_i$, a location parameter, and $\sigma_i = \frac{1}{\gamma_i}$, a scale parameter, which yields the following result:

$$f_{Y_i}(y_i; \mu_i, \sigma_i, a) = \frac{a}{\sigma_i} \left[1 - \exp\left(-e^{\frac{y_i - \mu_i}{\sigma_i}}\right)\right]^{a-1} e^{\frac{y_i - \mu_i}{\sigma_i}} \exp\left(-e^{\frac{y_i - \mu_i}{\sigma_i}}\right). \quad (2)$$

The survival function for model (2) is

$$S_{Y_i}(y_i; \mu_i, \sigma_i, a) = 1 - \left[1 - \exp\left(-e^{\frac{y_i - \mu_i}{\sigma_i}}\right)\right]^a.$$

Assuming right-censoring, with censoring times C_i, we observe response times $T_i = \min(Y_i, C_i)$, together with a censoring indicator $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ if $T_i > C_i$. To incorporate explanatory variables, let $\mu_i = \mathbf{x}_{1i}^{\mathrm{T}} \boldsymbol{\beta}_1$ and $\log \sigma_i = \mathbf{x}_{2i}^{\mathrm{T}} \boldsymbol{\beta}_2$, with $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ $1 \times p_1$ and $1 \times p_2$ covariate vectors and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ $1 \times p_1$ and $1 \times p_2$ parameter vectors. Then, the log-likelihood function under non-informative censoring can be written as

$$l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a) = \sum_{i=1}^{n} \left\{\delta_i \log f_{Y_i}(y_i; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a) + (1 - \delta_i) \log S_{Y_i}(y_i; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, a)\right\}.$$

We did not include block effects in the linear predictors and began by using a species×gender interaction linear predictor for both μ_i and σ_i. We then performed backward selection by removing the interaction effects and subsequent main effects using likelihood-ratio tests. We also studied the inclusion of a prey preference effect in both linear predictors.

2.4 Assessing goodness-of-fit

A useful tool to assess model goodness-of-fit is the half-normal plot of deviance residuals with a model-based simulation envelope, see Demétrio et al. (2014). The envelope is such that under the correct model most of the deviance residuals for the fitted model should lie within it. This tool is implemented as the `hnp` package (Moral et al., 2014) in R (R Core Team, 2014). It is possible to use the `hnp` function to produce these plots for models that are not already coded within the function by providing three helper functions, one to extract the deviance residuals, another to simulate response variables, and finally a function to refit the model to the simulated samples.

3 Results and discussion

The effects of the species×gender interaction were removed from both linear predictors, as well as the effects of gender. The species effect was also removed from the linear predictor for μ_i. Then, the inclusion of prey preference (three levels: non-parasitised larva, parasitised larva and no consumption in one hour) effect was significant on the linear predictor for σ_i. However, it was found that the preference effect could be summarised in two levels, consumption or no consumption (consumption effect). Therefore, after the selection process using likelihood-ratio tests, the final model included the main effects of species and consumption in the linear predictor for σ_i and only an intercept in the linear predictor for μ_i, see Table 1. This suggests that the failure rate function is accelerated for earwigs, which are faster than stinkbugs to attack prey and also for specimens which have effectively consumed prey in one hour of observation.

TABLE 1. Parameter estimates for the selected model.

Parameter	Estimate	Std. error	t value	p-value
a	2.17	0.35	6.15	< 0.0001
σ – Intercept	1.57	0.11	13.87	< 0.0001
σ – Species: Earwig	−0.47	0.10	−4.64	< 0.0001
σ – Consumption: Yes	−0.51	0.14	−3.58	0.0003
μ – Intercept	2.78	0.57	4.87	< 0.0001

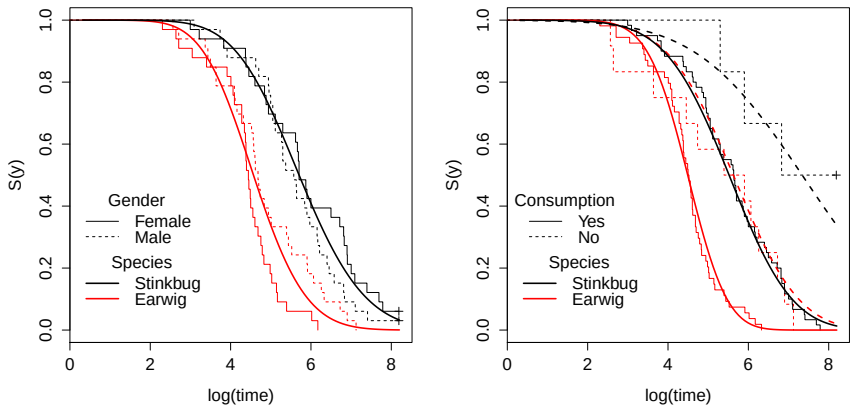

FIGURE 1. Fitted survival curves for the exponentiated-Weibull model for (a) each treatment combination and (b) combination between species and consumption levels.

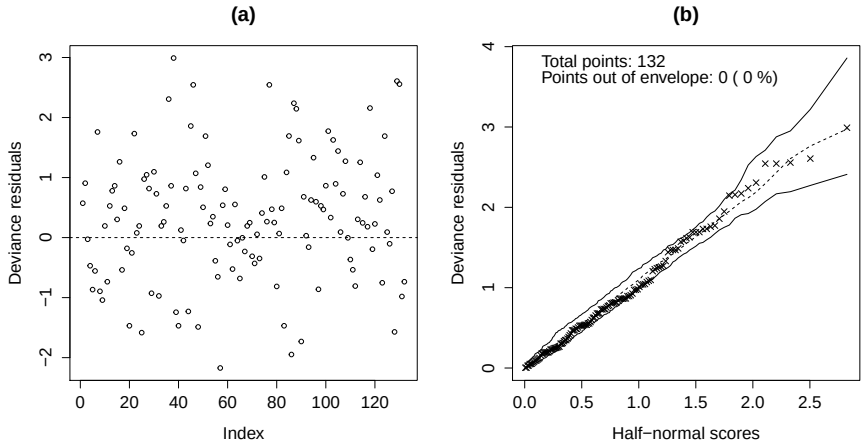

FIGURE 2. (a) Deviance residuals and (b) half-normal plot with simulation envelope for the deviance residuals for the exponentiated-Weibull model fit to the time until attack data.

It was clear that the earwig usually attacks earlier than the stinkbug, and this does not depend on gender, see Figure 1(a). The residual plot and the half-normal plot with simulation envelope also showed that this model fitted the data well, see Figure 2. This is important information from a biological control viewpoint. Also, the fact that prey consumption affected

the survival curves in the same way for both the parasitised and the non-parasitised prey, see Figure 1(b), shows that there is a clear distinction between two groups: specimens that effectively consumed prey in one hour and specimens that did not. This separation between "fast" and "slow" specimens suggests that there is individual variability that must be taken into account when planning biological control strategies.

Acknowledgments: Special Thanks to FAPESP for funding RAM's PhD studies.

References

Aitkin, M., Francis, B., Hinde, J., and Darnell, R. (2009). Statistical Modelling in R. Oxford University Press.

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.

Moral, R.A., Godoy, W.A.C., Demétrio, C.G.B., and Hinde, J. (*unpublished data*). Intraguild predation in overlapping niche predators in agroecosystems.

Moral, R.A., Hinde, J., and Demétrio, C.G.B. (2014). `hnp`: Half-normal plots with simulation envelopes. R package version 1.0. Available at http://CRAN.R-project.org/package=hnp.

Nadarajah, S., Cordeiro, G.M., and Ortega, E.M.M. (2013). The exponentiated Weibull distribution: A survey. *Statistical Papers*, **54**, 839 – 877.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Quantifying the genetic contribution to the variability of count traits

Izabela R.C. Oliveira[12], Geert Molenberghs[23], Clarice G.B. Demétrio[1], Carlos T.S. Dias[1], Cláudio L. Souza[1]

[1] ESALQ, University of São Paulo, Brazil
[2] I-BioStat, Universiteit Hasselt, Belgium
[3] I-BioStat, Katholieke Universiteit Leuven, Belgium

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

Abstract: Heritability and repeatability are important concepts in animal and plant breeding and are quantified based on fitting a model to hierarchical data. When linear models can be used to fit to the data, these attributes are defined as ratios of variance components. Matters are less simple for non-Gaussian outcomes. The focus here is on count outcomes where extensions of the Poisson model are used to describe the data. Expressions for heritability of count traits are derived using the Poisson combined model, which combines a Poisson outcome distribution with normal as well as gamma random effects, to capture both correlation among repeated observations as well as overdispersion, and admits closed-form expressions for the mean, variances and, hence, ratio of variances. The proposed methodology is illustrated using data from plant breeding programs.

Keywords: Combined model; Gamma distribution; Generalized linear mixed model; Overdispersion; Poisson distribution; Random effect.

1 Introduction

Heritability, defined as the proportion of the genetic contribution over the total variability in a phenotype, is useful to quantify the magnitude of improvement in the population and it is used when predicting the outcome of selection practiced among clones, inbred lines, or varieties. Repeatability is also an important concept in quantitative genetics and describes the proportion of phenotypic variance stemming from differences in repeated measures taken on the same individual. A measurement may be said to be repeatable when this proportion is relatively small.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

When the outcomes are normally distributed, linear mixed models are frequently used to estimate the genetic and environmental effects by considering these factors as random terms in the model. In this case, heritability in the broad sense can be quantified as the ratio of the genotypic variance, σ_g^2 say, to the total phenotypic variance, σ_p^2. However, when the trait of interest does not follow a linear model, the genetic and environmental random terms are no longer easily separable from the other model terms. This difficulty arises in particular when one deals with count outcomes. One often models such data using Poisson log-linear models. However, it has been observed recurrently that the mean-variance relationship for the Poisson model may not be met. In this paper, we use the combined models proposed by Molenberghs et al. (2010) for handling overdispersion (Hinde and Demétrio, 1998) and correlated data (Molenberghs and Verbeke, 2005) simultaneously, while obtaining heritability in the broad sense and repeatability based on count traits.

The paper is organized as follows. In Section 2, the motivating cases are described with analyses reported in Section 5. A review of the Poisson combined model for hierarchical and overdispersed count data is the subject of Section 3. The expressions to obtain heritability and repeatability coefficients for count traits are presented in Section 4.

2 Case studies

2.1 The potato breeding study

A total of 31 clones of potato were evaluated, using an augmented blocks design, and several production traits were measured, including the number of large tubers, which are commercially interesting. The clones were replicated from two to six times to the experimental plots, each one consisting of about 10 plants. At 120 days after planting, all produced tubers were harvested and the number of large tubers per plot was counted.

2.2 Inheritance study of trichomes density in tomato

The epidermal outgrowths trichomes in tomato plants are related to resistance to whitefly, a plague of this crop. A completely randomized experiment, with one plant per experimental unit, was performed to study the inheritance of some types of trichomes in tomato, using plants from P_1, P_2, F_1, F_2, $BC_{1(1)}$ and $BC_{1(2)}$ populations. In each plant, three cuts were made and at each of them an area of 1 mm^2 was defined and the numbers of several different trichomes were counted, both in the abaxial and adaxial faces of the leaves. A main interest of this study lies in the repeatability calculation. In this work, we will consider only the glandular trichomes of types IV, VI, and VII counted in the adaxial face.

3 An extended Poisson model to handle hierarchical and overdispersed data

Combining ideas from the overdispersion models and the Poisson-normal model led Molenberghs, Verbeke and Demétrio (2007) and Molenberghs et al. (2010) to a model for repeated count data with overdispersion, assuming that $Y_{ij} \sim \text{Poi}(\lambda_{ij})$ with $\lambda_{ij} = \theta_{ij} \exp\left(\mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\text{T}}\mathbf{b}_i\right)$, $\mathbf{b}_i \sim \text{N}(\mathbf{0}, \mathbf{D})$, $\text{E}(\boldsymbol{\theta}_i) = \text{E}[(\theta_{i1}, \ldots, \theta_{in_i})^{\text{T}}] = \Phi_i$ and $\text{Var}(\boldsymbol{\theta}_i) = \Sigma_i$. The $\boldsymbol{\mu}_i = \text{E}(\mathbf{Y}_i)$ has components: $\boldsymbol{\mu}_{ij} = \phi_{ij} \exp\left(\mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta} + \frac{1}{2}\mathbf{z}_{ij}^{\text{T}}\mathbf{D}\mathbf{z}_{ij}\right)$, and the variance-covariance matrix $\text{Var}(\mathbf{Y}_i) = \mathbf{M}_i + \mathbf{M}_i\left(\mathbf{P}_i - \mathbf{J}_{n_i}\right)\mathbf{M}_i$, where $\mathbf{M}_i$ is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the diagonal and the (j,k)th element of $\mathbf{P}_i$ equals $p_{i,jk} = \exp\left(\frac{1}{2}\mathbf{z}_{ij}^{\text{T}}\mathbf{D}\mathbf{z}_{ik}\right) \frac{\sigma_{i,jk} + \phi_{ij}\phi_{ik}}{\phi_{ij}\phi_{ik}} \exp\left(\frac{1}{2}\mathbf{z}_{ik}^{\text{T}}\mathbf{D}\mathbf{z}_{ij}\right)$. As special cases of this combined model we have the Poisson, Poisson-normal and negative-binomial models.

4 Derivation of heritability for count data

Consider the Poisson-Gamma-Normal model and its variance. Also, without loss of generality, we set $\text{E}(\boldsymbol{\theta}_i) = 1$. Then $\text{Var}(Y_{ij}) = \mu_{ij} + \mu_{ij}(P_{i,jj} - 1)\mu_{ij}$, where $\mu_{ij} = \exp\left(x_{ij}^{\text{T}}\boldsymbol{\beta} + \frac{1}{2}\mathbf{z}_{ij}^{\text{T}}\mathbf{D}\mathbf{z}_{ij}\right) = \mu_{0ij}\mu_{1ij}$, and $P_{i,jj} = \mu_{1ij}(\sigma_{i,jj} + 1)\mu_{1ij} = \mu_{1ij}^2(\sigma_{i,jj} + 1)$. The non-genetic contribution (the variance of the Poisson combined model, for $D = 0$) over the total variability is:

$$\xi_{ij} = \frac{1 + \mu_{0ij}[(\sigma_{i,jj} + 1) - 1]}{\mu_{1ij}\{1 + \mu_{0ij}\mu_{1ij}[\mu_{1ij}^2(\sigma_{i,jj} + 1) - 1]\}}. \tag{1}$$

The heritability, that is, the proportion of the total variability related to the genetic effect is:

$$H_{ij}^2 = 1 - \xi_{ij}.$$

The repeatability is the contribution from overdispersion and between-individual variability over the total phenotypic variability in a population:

$$r_{ij} = \frac{1 + \mu_{0ij}[(\sigma_{i,jj} + 1) - 1]}{\mu_{1ij}\{1 + \mu_{0ij}\mu_{1ij}[\mu_{1ij}^2(\sigma_{i,jj} + 1) - 1]\}}. \tag{2}$$

As before, this ratio places the variance of the Poisson combined model, for $D = 0$, in the numerator and the full variance in the denominator. However, as the interest lies on repeatability, the normal random effect captures the correlation between measures within the same individual. Then, the denominator, which refers to the total phenotypic variance, includes the within-individual component as well. From (2) we can also estimate the within-individual contribution to the total variability, that is, $e_{wij} = 1 - r_{ij}$. The special case of no overdispersion follows easily for heritability and repeatability, by setting $\sigma_{i,jj} = 0$.

5 Analysis of case studies

5.1 The potato breeding study

We considered the combined model and its special cases with linear predictor $\log(\lambda_{ij}) = \beta_0 + b_i$, where β_0 is the effect common to all observations, b_i is the genetic random effect of the ith clone, assumed to be normally distributed with mean 0 and variance σ_g^2. In a second approach, we also evaluated the effect of the total weight of the plot as a covariate added to the linear predictor while modeling the number of large tubers.
When there are no covariates effects, the combined model fits better. The non-genetic contribution for the number of large tubers is $\xi_{ij} \cong 0.47$ and the heritability for this trait is $H_{ij}^2 \cong 0.53$, that is, about 53% of the total phenotypic variation is attributed to genetic variation among clones. When the covariate effect is considered, it captures some amount of variability but still the combined model fits better. The non-genetic contribution over the total variability varied in the interval [0.50; 0.18] and the heritability values varied within the range $H^2 = [0.50; 0.82]$, depending on the covariate value.

5.2 Inheritance study of trichomes density in tomato

We considered the combined model and its special cases with linear predictor $\log(\lambda_{ij}) = \beta_0 + b_i$, where β_0 is the overall effect and b_i is the random effect that captures the variability within the ith plant, assumed to be normally distributed with mean 0 and variance σ_w^2. The combined model is an improvement in fit relative to the other models. Thus, there are correlation and overdispersion effects to be modeled simultaneously. The repeatability is $r_{ij} \cong 9.39 \times 10^{-6}$ for the trichome type IV, $r_{ij} \cong 0.02$ for the trichome type VI and $r_{ij} \cong 0.36$ for the trichome type VII. In all cases, the repeatability values are very low, which indicate that the variability within individuals with respect to phenotypic variability is high.

6 Conclusions

In this paper, we have derived an expression for heritability, based on hierarchical count data, using Poisson-based mixed models. The focus was on the so-called combined model, which brings together a generalized linear model for count data with both normal and gamma random effects, thus accommodating correlation between repeated measures and overdispersion. Importantly, as shown in Molenberghs et al. (2010), special cases of this combined model are the Poisson, Poisson-normal and negative-binomial models.
The combined model and its GLMM sub-model admit closed-form expressions for means, variances, and higher-order moments. As a result, variance

ratios have explicit expressions too. The heritability and repeatability coefficients are sufficiently simple and appealing, in particular in special cases. In the potato study, we have used our methodology to calculate the heritability for the count trait number of large tubers per plot. We considered two scenarios, with and without covariates. In the latter one, we showed how the combined model plays an important role in accommodating extra variability besides the intra-cluster correlation. In the tomato study, we have extended our approach to calculate the repeatability on trichomes count.

In these models, heritability is a function rather than a constant. Practically, heritability and repeatability change with the effects present in the predictor functions. Evidently, one can summarize the functions in a variety of ways, using averages, medians, quartiles, ranges, etc.

Finally, a common interpretation of heritability is in terms of the coefficient of a regression parents-offspring (e.g., "regression of the value of a character measured for the son on the character measured for the father"). This property of the heritability is clearly lost because of the non-normal and hence non-linear nature of our models. This is the price to pay for the use of a model that is more faithful to the data type being recorded.

Acknowledgments: Special Thanks to CAPES and CNPq.

References

Hinde, J. and Demétrio, C.G.B. (1998). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, **27**, 151 – 170.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G., Verbeke, G., and Demétrio, C.G.B. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis* **13**, 513 – 531.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325 – 347.

New qualitative choice models incorporating individual and choice characteristics

Jean Peyhardi[1]

[1] Virtual Plants, CIRAD, AGAP and Inria, 860 rue de St Priest, 34095 Montpellier, France

E-mail for correspondence: jean.peyhardi@gmail.com

Abstract: In the econometric framework, multinomial and conditional logit models are the most usual regression models for qualitative choices. They differ by their parametrization while sharing the canonical link function. This link function can be decomposed into the reference ratio of probabilities and the logistic cumulative distribution function (cdf). We propose to conserve the reference ratio, appropriate for qualitative choices, but to select the cdf among an enlarged family containing the Student cdf for instance. These new qualitative choice models often outperform logit models in terms of likelihood and error rate of classification and stay easily interpretable. This is illustrated with a benchmark dataset of travel demand between Sydney and Melbourn.

Keywords: Qualitative choices; Conditional logit model; Link function; Design matrix.

1 Multinomial and conditional logit models for qualitative choices

Let Y_i be the response variable corresponding to the choice of individual i (with alternatives $j = 1, \dots, J$) and x_i be the vector of individual characteristics (e.g. sex, age). In the context of travel demand, some choice characteristics $\omega_{i,j}$ are also used, such as the cost of alternative j for individual i. In the following we will suppress the individual subscript i without loss of generality.

Luce's choice axiom (Luce, 1959) and the principle of random utility maximisation lead to the logit model defined by

$$\pi_j = \frac{\exp(\eta_j)}{1 + \sum_{k=1}^{J-1} \exp(\eta_k)}$$

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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

for $j = 1, \ldots, J-1$, where $\pi_j = P(Y = j)$. Depending on the form of the linear predictors η_j, we obtain different logit models:

- $\eta_j = \alpha_j + x^{\mathrm{T}}\delta_j$. Individual characteristics x are used with $J-1$ different slopes δ_j. This is the classical multinomial logit model.
- $\eta_j = \alpha_j + \tilde{\omega}_j^{\mathrm{T}}\gamma$ where $\tilde{\omega}_j = \omega_j - \omega_J$. Choice characteristics ω_j are used with common slope δ. This is the conditional logit model introduced by McFadden (1974).
- $\eta_j = \alpha_j + x^{\mathrm{T}}\delta_j + \tilde{\omega}_j^{\mathrm{T}}\gamma$. Individual and choice characteristics are used with respectively different slopes and common slope. This is a combination of the two previous parametrizations.

2 Generalisation of multinomial and conditional logit models

All the classical regression models for categorical data (Tutz, 2012) share the generic equations (Peyhardi et al., 2014)

$$r_j(\pi) = F(\eta_j)$$

for $j = 1, \ldots, J-1$, where r is a C^1-diffeomorphism from the simplex $\Delta = \{\pi \in (0,1)^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\}$ (corner of hypercube) to an open subset of the hypercube $(0,1)^{J-1}$, π is the vector of probabilities $(\pi_1, \ldots, \pi_{J-1})^{\mathrm{T}}$ and F is a continuous and strictly increasing cdf.
Let us remark that the three logit models (defined in Section 1) share the canonical link function, specified by the reference ratio

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \pi_J}$$

and the logistic cdf

$$F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

Therefore, these three logit models are specified by the (reference, logistic, Z) triplet with different design matrices Z respectively equal to

$$Z_1 = \begin{pmatrix} 1 & & & x^{\mathrm{T}} & & \\ & \ddots & & & \ddots & \\ & & 1 & & & x^{\mathrm{T}} \end{pmatrix}, \quad Z_2 = \begin{pmatrix} 1 & & & \tilde{\omega}_1^{\mathrm{T}} \\ & \ddots & & \vdots \\ & & 1 & \tilde{\omega}_{J-1}^{\mathrm{T}} \end{pmatrix},$$

$$Z_3 = \begin{pmatrix} 1 & & & x^{\mathrm{T}} & & \tilde{\omega}_1^{\mathrm{T}} \\ & \ddots & & & \ddots & \vdots \\ & & 1 & & x^{\mathrm{T}} & \tilde{\omega}_{J-1}^{\mathrm{T}} \end{pmatrix}.$$

The reference ratio is mandatory for non-ordered choices whereas the logistic cdf is not (Peyhardi et al., 2014). We thus propose a new class of regression models appropriate for qualitative choices defined by (reference, F, Z_i) models ($i = 1, 2, 3$) where the cdf F can be selected among e.g. the logistic, Gaussian, Laplace, Gumbel, Gompertz, and Student cdfs (with different degrees of freedom $\nu \in \mathbb{R}_+^*$). The heavy tails of Student distributions may markedly improve the model fit and reduce the classification error rates (Peyhardi et al., 2014). Parameter estimates stay interpretable since

$$\frac{\pi_j}{\pi_J} = \frac{F}{1-F}(\eta_j)$$

is strictly increasing with η_j (we have $\pi_j/\pi_J = \exp(\eta_j)$ in the case of the logistic cdf). Finally, this family of reference models for qualitative choices is easily estimated using the standard Fisher's scoring algorithm.

2.1 Fisher's scoring algorithm for reference models

Let us remark that the Fisher's scoring algorithm is simplified in the particular case of the reference ratio (compared to the adjacent, cumulative and sequential ratios) which is a part of the canonical link function. Using the chain rule we obtain the score

$$\frac{\partial l}{\partial \beta} = Z^{\mathrm{T}} D(y - \pi),$$

and the Fisher's information matrix

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial \beta^{\mathrm{T}} \partial \beta}\right) = -Z^{\mathrm{T}} D \mathrm{Cov}(Y) D Z,$$

where

$$D = \mathrm{diag}_{1 \leq j \leq J-1}\left[\frac{f(\eta_j)}{F(\eta_j)\{1 - F(\eta_j)\}}\right],$$

and f is the density function. Remarking that $f = F(1 - F)$ for the logistic distribution, the Fisher's scoring algorithm turns out to be, in this particular case, the algorithm for the canonical link function.

3 Application to travel mode demand

The dataset, used by Greene (2003), contains 210 observations of choice among $J = 4$ travel modes between Sydney and Melbourn (Australia): *air* (1), *bus* (2), *train* (3), and *car* (4). The two individual characteristics are the household income x^1 and the number of people travelling x^2. The three choice characteristics are the terminal time ω_j^1 ($\omega_4^1 = 0$ for car), the amount of time spent traveling ω_j^2 and the in-vehicle cost ω_j^3. The sample is choices

based so as to balance it among the four choices knowing that the true population is dominated by drivers.
The three logit models and other reference models (i.e. $F \neq$ logistic) were estimated. Best results were obtained with (reference, Student$_{\nu=1}$, Z_i) models that markedly outperformed logit models; see Table 1 (we have for instance $l = -192.89$ for logistic versus $l = -169.79$ for Student with the same parametrization Z_2). The (reference, Student$_{\nu=1}$, Z_2) is the best model according to BIC. The proportions between parameters when significant are approximatively conserved comparing logistic and Student models ($\alpha_1/\alpha_2 \simeq 1.43$ for logistic cdf and $\alpha_1/\alpha_2 \simeq 1.72$ for Student cdf for instance). The interesting difference concerns estimate of the slope γ^1 since $\gamma^1/\gamma^2 \simeq 24$ for logistic cdf and $\gamma^1/\gamma^2 \simeq 60$ for Student cdf. In the Student case, the terminal time has a stronger impact on the travel mode choice.
Finally, the selection of F does not a priori change the sign of parameters but may change the proportion between them. Moreover it may reduce the classification error rate and thus increase the precision of predictions.

TABLE 1. Parameters estimates for six reference regression models.

	$F =$ logistic			$F =$ Student$_{\nu=1}$		
	Z_1	Z_2	Z_3	Z_1	Z_2	Z_3
α_1	0.9435	4.7399	6.0351	0.6438	13.7305	15.2387
α_2	1.978	3.3062	4.5045	1.9446	7.955	7.3668
α_3	2.4938	3.9532	5.5735	2.354	8.6827	9.5013
δ_1^1	0.003544		0.007481	0.00496		0.02897
δ_2^1	-0.03033		-0.0209	-0.02581		-0.004521
δ_3^1	-0.05731		-0.05923	-0.06026		-0.0571
δ_1^2	-0.6006		-0.9224	-0.4946		-1.0745
δ_2^2	-0.9404		-0.1478	-1.0836		0.7765
δ_3^2	-0.3098		0.2163	-0.2489		0.7777
γ^1		-0.09689	-0.1012		-0.2548	-0.2597
γ^2		-0.003995	-0.004131		-0.00426	-0.003878
γ^3		-0.01391	-0.008667		-0.01849	-0.01746
$\mathcal{L}$	-253.34	-192.89	-172.47	-253.36	-169.79	-159.02
BIC	554.8	417.86	409.1	554.84	**371.66**	382.21
Error	53.33%	26.19 %	27.14 %	54.29 %	22.38%	21.9%

References

Luce, R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons.

Greene, W. (2003). *Econometric Analysis*. New Jersey: Prentice Hall. 729 – 735.

McFadden, D. (1974). Conditional logit analysis of qualitative choice analysis. *Frontiers in Econometrics*. 105 – 142.

Peyhardi, J., Trottier, C., and Guédon, Y. (2014). A new specification of generalized linear models. *arXiv preprint arXiv:1404.7331*.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press.

Estimating marginal likelihoods from the posterior draws through a geometric identity

Johannes Reichl[1]

[1] Energy Institute at the Johannes Kepler University Linz

E-mail for correspondence: reichl@energieinstitut-linz.at

Abstract: This article develops a new estimator of the marginal likelihood that requires a sample of the posterior distribution as the only input from the analyst. This sample may come from any sampling scheme, such as Gibbs sampling or Metropolis-Hastings sampling. The presented approach can be implemented generically in almost any application of Bayesian modeling and significantly decreases the computational burdens associated with marginal likelihood estimation compared to existing techniques. The functionality of this method is demonstrated in the context of a high-dimensional random intercept probit. Simulation results show that the simple approach presented here achieves excellent stability in low dimensional models, and also clearly outperforms existing methods when the number of coefficients of the model increases.

Keywords: Bayesian statistics; Model evidence; Integrated likelihood; Model selection; Estimation of normalizing constant.

1 Motivation

Bayesian model selection relies on the posterior probabilities of the H candidate models $\mathcal{M}_1, \ldots, \mathcal{M}_H$ conditional on the data (see e.g. Kass and Raftery, 1995). In this article we discuss the estimation of the posterior probabilities $p(\mathcal{M}_h|\boldsymbol{y})$ of the $h = 1, \ldots, H$ candidate models by estimating their marginal likelihoods. Calculating the marginal likelihood is a non trivial integration problem, and as such it is still associated with significant effort on the part of the analyst and potential imprecision in the case of high-dimensional or multi-level models. Comparative studies of existing estimation techniques for the marginal likelihood only provide clear evidence of precision for candidate models of lesser dimensions, while Bayesian analysis frequently requires more complex models (see e.g. Frühwirth-Schnatter and Wagner, 2008).

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

This article presents a technique for estimating the marginal likelihood requiring only a sample of the posterior distribution as an input and is thus implementable as a generic function allowing for it to be used in a variety of applications. As a potentially even more important advantage, the approach shows significantly less sensitivity to an increase in the number of model coefficients compared to existing approaches.

2 The approach

We start by defining the marginal likelihood of model $\mathcal{M}_h$ as

$$p(\mathcal{M}_h|\boldsymbol{y}) = \int_{\boldsymbol{\Theta}_h} p(\boldsymbol{y}|\boldsymbol{\theta}_h)\, p(\boldsymbol{\theta}_h)\, d\boldsymbol{\theta}_h, \tag{1}$$

where $\boldsymbol{\theta}_h$ is a $1 \times K$ vector containing the K coefficients of model $\mathcal{M}_h$. $p(\boldsymbol{y}|\boldsymbol{\theta}_h)$ refers to the likelihood of model $\mathcal{M}_h$, and $p(\boldsymbol{\theta}_h)$ is the prior distribution with domain $\boldsymbol{\Theta}_h$. Suppressing the model index h henceforth, and considering the marginal likelihood of a model $\mathcal{M}$ is the normalizing constant of its posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, we can rewrite Bayes theorem as:

$$\frac{1}{p(\mathcal{M}|\boldsymbol{y})} = \frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}. \tag{2}$$

Let $\mathcal{A}$ be a bounded subset of prior domain $\boldsymbol{\Theta}$, then integrating both sides in (2) over $\mathcal{A}$ gives

$$\frac{1}{p(\mathcal{M}|\boldsymbol{y})} \int_{\mathcal{A}} \mathbf{1}\, d\boldsymbol{\theta} = \int_{\mathcal{A}} \frac{1}{p^*(\boldsymbol{\theta}|\boldsymbol{y})}\, p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}, \tag{3}$$

where $p^*(\boldsymbol{\theta}|\boldsymbol{y})$ is used as an abbreviation for the non-normalized posterior $p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ henceforth. A representation of the marginal likelihood is then found by:

$$p(\mathcal{M}|\boldsymbol{y}) = \int_{\mathcal{A}} \mathbf{1}\, d\boldsymbol{\theta} \Big/ \int_{\mathcal{A}} \frac{1}{p^*(\boldsymbol{\theta}|\boldsymbol{y})}\, p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}. \tag{4}$$

We refer to the right integral in (4) as the non-normalized posterior integral over $\mathcal{A}$ and abbreviate it by $\kappa_{\mathcal{A}}$. Integrating 1 over a $K-$dimensional bounded set has the geometric interpretation of a generalized volume, or hypervolume, and we will refer to the left integral in (4) as the volume of $\mathcal{A}$. This article exploits (4) and presents a new estimator for the marginal likelihood by separately estimating the volume of $\mathcal{A}$ and the corresponding non-normalized posterior integral. We also present a method for choosing $\mathcal{A}$ in such a way that the quotient of these estimators yields a stable estimate of the marginal likelihood.

First, we turn to the non-normalized posterior integral. A common technique for numerical integration is importance sampling, and since the posterior distribution is part of the numerator in the non-normalized posterior integral this suggests the posterior as the importance density. Since draws from the posterior distribution are usually available as a natural output of Bayesian analysis, the importance sampling estimator of the posterior integral $\hat{\kappa}_{\mathcal{A}}$ is available almost ad hoc once $\mathcal{A}$ has been defined.

$$\hat{\kappa}_{\mathcal{A}} = \frac{1}{L}\sum_{l=1}^{L} f(\boldsymbol{\theta}^{(l)}),$$

with

$$f(\boldsymbol{\theta}^{(l)}) = \begin{cases} \frac{1}{p^*(\boldsymbol{\theta}^{(l)}|\boldsymbol{y})}, & \text{if } \boldsymbol{\theta}^{(l)} \in \mathcal{A}, \\ 0, & \text{else}, \end{cases} \tag{5}$$

where $\boldsymbol{\theta}^{(l)}$ with $l = 1, \ldots, L$ refers to the posterior draws after the burn in. Subsequent steps outline a definition of $\mathcal{A}$ allowing a stable estimation of the marginal likelihood from identity (4). Firstly, to ensure the posterior distribution is a proper choice for the importance density, as required in our approach, the region of integration $\mathcal{A}$ must have full support of the posterior distribution. Secondly, to avoid instability, the region of integration $\mathcal{A}$ may only contain points for which the sum in (5) is stable independently of any specific run of the MCMC sampler. To address both of these requirements and at the same time allow for a simple estimation of the volume and the non-normalized posterior integral we define $\mathcal{A}$ as the intersection of two sets $\mathcal{A}_1$ and $\mathcal{A}_2$.

Set $\mathcal{A}_1$ is defined by a threshold value ρ and only those points $\boldsymbol{\theta}$ lie in $\mathcal{A}_1$ whose non-normalized posterior $p^*(\boldsymbol{\theta}|\boldsymbol{y})$ exceeds this threshold, such that:

$$\boldsymbol{\theta} \in \mathcal{A}_1 \; \textit{if} \; p^*(\boldsymbol{\theta}|\boldsymbol{y}) > \rho.$$

Considering the series $p^{*,(1,\ldots,L)} = p^*(\boldsymbol{\theta}^{(1)}|\boldsymbol{y}), \ldots, p^*(\boldsymbol{\theta}^{(L)}|\boldsymbol{y})$, a natural way of determining the threshold ρ is to ensure that the lowest values of $p^{*,(1,\ldots,L)}$ are not destabilizing the sum in (5) by setting ρ as a quantile of $p^{*,(1,\ldots,L)}$, and we define ρ as the median of $p^{*,(1,\ldots,L)}$. Thereby, an almost perfectly stable estimation of (5) is ensured by excluding all values of $p^{*,(1,\ldots,L)}$ from the estimation of $\hat{\kappa}_{\mathcal{A}}$ stemming from the tails of the posterior distribution. To facilitate easy estimation of the volume of $\mathcal{A}$ a second set $\mathcal{A}_2$ is defined, in such a way that the volume of the intersection of $\mathcal{A}_1$ and $\mathcal{A}_2$ can be estimated by means of statistical standard techniques only. As long as $\mathcal{A}_1$ and $\mathcal{A}_2$ have the same dimension K and an intersection $\mathcal{A}$, the volume of this intersection $V_{\mathcal{A}}$ can be written as $V_{\mathcal{A}} = \pi V_{\mathcal{A}_2}$, where π refers to the ratio of points lying in $\mathcal{A}_2$ that also lie in $\mathcal{A}_1$, and $V_{\mathcal{A}_2}$ is the volume of $\mathcal{A}_2$. Hence, for an easy estimation of $V_{\mathcal{A}}$ we define $\mathcal{A}_2$ in such a way that its volume $V_{\mathcal{A}_2}$ can be calculated analytically and that drawing

uniformly from within it is efficient and feasible. Then, π can simply be estimated by drawing K-dimensional vectors $\boldsymbol{\theta}^{(r)}$ uniformly from $\mathcal{A}_2$ such that

$$\hat{\pi} = \frac{1}{R} \sum_{r=1}^{R} I(\boldsymbol{\theta}^{(r)} \in \mathcal{A}_1),$$

where $I(\cdot)$ refers to the indicator function, and R is the number of random draws from within $\mathcal{A}_2$.

Even though other definitions of $\mathcal{A}_2$ are possible, we choose a K-dimensional ellipsoid as $\mathcal{A}_2$, as this choice shows outstanding efficiency of the resulting estimator. The set of points $\boldsymbol{\theta}$ lying in $\mathcal{A}_2$ is thus defined by

$$\boldsymbol{\theta} \in \mathcal{A}_2 \quad if \quad (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\, \boldsymbol{C}\, (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\mathrm{T}} < 1, \tag{6}$$

where $\boldsymbol{C}$ is a positive definite matrix of dimension $K \times K$ with its eigenvectors defining the principal axes of the ellipsoid. $\boldsymbol{\theta}^*$ is a point with support of the posterior distribution, and we define $\boldsymbol{\theta}^*$ as its posterior mode to ensure substantial overlap between $\mathcal{A}_2$ and $\mathcal{A}_1$.

The last step in defining $\mathcal{A}_2$ is thus choosing $\boldsymbol{C}$. Consider matrix $\boldsymbol{R} = \left(\boldsymbol{\theta}^{(1)^{\mathrm{T}}}, \ldots, \boldsymbol{\theta}^{(L)^{\mathrm{T}}}\right)^{\mathrm{T}}$, and its covariance matrix $\boldsymbol{D} = \mathrm{cov}(\boldsymbol{R})$, we define

$$\boldsymbol{C} = (\alpha \boldsymbol{D})^{-1},$$

where α is a scalar with domain $\mathbb{R}^+$, and is employed as a tuning parameter in the presented approach. We recommend to set α in such a way that the resulting intersection of $\mathcal{A}_1$ and $\mathcal{A}_2$ contains about 49% of the L posterior draws of $\boldsymbol{\theta}^{(l)}$, where a theoretic underpinning of this recommendation can be provided by the author upon request.

Algorithm I for the estimation of the marginal likelihood is thus given by:

1. Run an MCMC sampler to obtain L posterior draws $\boldsymbol{\theta}^{(l)}$ after the burn in, calculate the series of non-normalized posterior density values $p^{*,(1,\ldots,L)}$, and set ρ to its median, $\boldsymbol{\theta}^*$ to the posterior mode, and D to the covariance matrix of the posterior draws.

2. Define α in such a way that $0.49 \times L$ draws are in $\mathcal{A}$, where , $\boldsymbol{\theta}^{(l)} \in \mathcal{A}$ *if* $\left(p^*(\boldsymbol{\theta}^{(l)}|\boldsymbol{y}) > \rho \;\&\; \left(\boldsymbol{\theta}^{(l)} - \boldsymbol{\theta}^*\right)(\alpha \boldsymbol{D})^{-1} \left(\boldsymbol{\theta}^{(l)} - \boldsymbol{\theta}^*\right)^{\mathrm{T}} < 1\right)$.

3. Draw R points $\boldsymbol{\theta}^{(r)}$ from $\mathcal{A}_2$, count the number $\bar{r}$ of draws for which $p^*(\boldsymbol{\theta}^{(r)}|\boldsymbol{y}) > \rho$, and set $\hat{\pi} = \bar{r}/R$.

4. Estimate the volume of $\mathcal{A}$ as $\hat{V}_{\mathcal{A}} = \hat{\pi} V_{\mathcal{A}_2}$, and obtain the estimator for the non-normalized posterior integral $\hat{\kappa}_{\mathcal{A}}$ from (5).

5. Calculate the final estimator of the marginal likelihood as

$$\hat{p}_{\mathcal{A}}(\mathcal{M}|\boldsymbol{y}) = \frac{\hat{V}_{\mathcal{A}}}{\hat{\kappa}_{\mathcal{A}}}.$$

TABLE 1. PISA Data; logarithm of different marginal likelihood estimators and for five different data sets. Importance sampling and bridge sampling using a mixture importance density constructed as e.g. in Frühwirth-Schnatter and Wagner (2008) are referenced by $\hat{p}_{IS}$ and $\hat{p}_{BS}$; $\hat{p}_{CH}$ refers to Chib's method; and the estimator proposed in this paper is referenced as $\hat{p}_{\mathcal{A}}$; relevant standard errors in parentheses; results from three independent MCMC runs per data set are reported.

US region	K	$\log(\hat{p}_{IS})$	$\log(\hat{p}_{BS})$	$\log(\hat{p}_{CH})$	$\log(\hat{p}_{\mathcal{A}})$
Northeast	28	−363.203	−363.139	−363.086	−363.136
		(0.076)	(0.012)	(0.054)	(0.014)
		−363.028	−363.155	−363.183	−363.167
		(0.185)	(0.012)	(0.049)	(0.010)
		−363.204	−363.169	−363.168	−363.145
		(0.176)	(0.012)	(0.051)	(0.009)
West	35	−503.501	−503.159	−503.219	−503.199
		(0.219)	(0.024)	(0.090)	(0.017)
		−503.781	−503.180	−503.091	−503.174
		(0.262)	(0.024)	(0.097)	(0.012)
		−503.828	−503.125	−503.003	−503.189
		(0.150)	(0.023)	(0.100)	(0.010)
Midwest	43	−667.515	−667.026	−667.117	−667.033
		(0.234)	(0.028)	(0.127)	(0.013)
		−667.364	−667.012	−667.014	−667.031
		(0.243)	(0.029)	(0.123)	(0.012)
		−667.241	−666.962	−667.127	−667.036
		(0.471)	(0.027)	(0.095)	(0.016)
South	57	−891.596	−892.433	−892.246	−892.427
		(0.779)	(0.054)	(0.158)	(0.012)
		−893.633	−892.432	−892.628	−892.432
		(0.536)	(0.058)	(0.255)	(0.013)
		−893.341	−892.568	−892.370	−892.421
		(0.475)	(0.053)	(0.184)	(0.016)
all	142		−2352.958	−2350.232	−2351.877
			(1.001)	(0.447)	(0.017)
			−2354.952	−2351.547	−2351.920
			(1.815)	(1.264)	(0.016)
			−2350.454	−2352.491	−2351.913
			(0.539)	(1.338)	(0.021)

Algorithms for efficiently drawing uniformly from within a hyperellipsoid, and a recursive algorithm returning $\log(V_{\mathcal{A}_2})$ with minimal computing time even for very high K, can be requested from the author. Thus, the calculation of $\hat{\pi}$ and $V_{\mathcal{A}_2}$, and consequently $\hat{V}_{\mathcal{A}}$, is achieved with low computational effort and high precision.

3 Application

In this section the proposed estimation method is applied to a random intercept probit model and a comparison to existing methods is presented, these are Chib's method (1995), importance sampling, and bridge sampling (Meng and Wong, 1996). This paper provides for the first time a comparison of the discussed estimation techniques for a high-dimensional unit-level model, and discloses the shortcomings of existing approaches. As one instance of a comparative study exploring the existing techniques with respect to a unit-level model, Frühwirth-Schnatter and Wagner (2008) estimate a random intercept logit model with up to $K = 25$ coefficients, where in the application shown in this article we increase the number of model dimensions in five applications up to $K = 142$ to demonstrate the extraordinary stability of the presented estimator in comparison to the existing approaches. Data is about reading proficiency in US schools and stems from the 'Program for International Student Assessment' (PISA) as provided in Snijders and Bosker (2012). The data is estimated by a random intercept probit model, and for 5 different partitions of the data with respect to their geographical origin. Table 1 displays the results of the comparative study. While the magnitudes of the estimates can not be compared between the different values of K as these are related to different data, the standard errors allow conclusions about the sensitivity of the respective estimators to an increase of the number of coefficients of the underlying model.

References

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313 – 1321.

Frühwirth-Schnatter, S. and Wagner, H. (2008). Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling. *Computational Statistics and Data Analysis*, **52**, 4608 – 4624.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773 – 795.

Meng, X.-L. and Wong, W.H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, **6**, 831 – 860.

Snijders, T.A. and Bosker, R.J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, 2nd ed.* London: Sage Publishers Ltd.

Fast estimation of multidimensional adaptive P-spline models

María Xosé Rodríguez-Álvarez[1], María Durbán[2], Dae-Jin Lee[3], Paul H.C. Eilers[24]

[1] Department of Statistics and Operations Research, University of Vigo, Vigo, Spain
[2] Department of Statistics, Universidad Carlos III de Madrid, Leganés, Spain
[3] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[4] Erasmus Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: mxrodriguez@uvigo.es

Abstract: A fast and stable algorithm for estimating multidimensional adaptive P-spline models is presented. We call it as *Separation of Overlapping Penalties* (SOP) as it is an extension of the *Separation of Anisotropic Penalties* (SAP) algorithm. SAP was originally derived for the estimation of the smoothing parameters of a multidimensional tensor product P-spline model with anisotropic penalties.

Keywords: Adaptive smoothing; P-splines; Mixed models; SAP algorithm.

1 Introduction

Standard P-splines assume one smoothing parameter for modelling the effect of a covariate across its whole domain. In many applications it is desirable and needed to adapt smoothness locally to the data, and adaptive P-splines have been suggested.
The literature contains several proposals for adaptive P-splines. See, for instance, Krivobokova et al. (2008) and Wood (2011). However, the estimation procedures used by all these approaches can be very slow or even unstable. Based on the proposal by Wood (2011), we generalize the SAP algorithm given in Rodríguez-Álvarez et al. (2014) to deal with the adaptive penalty that is obtained.
For brevity, we mainly focus here on the univariate adaptive approach for Gaussian data. However, the proposal can be extended to the multidimen-

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

sional case, as well as to non-Gaussian responses along the lines of the Generalized Linear Model.

2 Adaptive penalized splines

Consider a regression problem

$$y_i = f(x_i) + \epsilon_i \qquad i = 1, \ldots n,$$

where f is a smooth and unknown function and $\epsilon_i \sim N\left(0, \sigma^2\right)$. Within the P-spline framework (Eilers and Marx, 1996), the unknown function $f(x)$ can be approximated by a linear combination of B-splines basis functions, i.e., $f(x) = \sum_{j=1}^{c} \theta_j B_j(x)$, and smoothness is achieved by imposing a penalty on the regression coefficients $\boldsymbol{\theta}$ in the form

$$P = \lambda \sum_{k=q+1}^{c} \left(\Delta^q \theta_k\right)^2 = \lambda \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{D}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{\theta}, \tag{1}$$

where λ is the smoothing parameter, and Δ^q forms differences of order q on adjacent coefficients, i.e., $\Delta\theta_k = \theta_k - \theta_{k-1}$, $\Delta^2\theta_k = \Delta\left(\Delta\theta_k\right) = \theta_k - \theta_{k-1} - \left(\theta_{k-1} - \theta_{k-2}\right) = \theta_k - 2\theta_{k-1} + \theta_{k-2}$, and so on for higher q. Finally, $\boldsymbol{D}$ is simply the matrix representation of Δ^q.

As can be observed in (1), the same smoothing parameter λ applies to all coefficient differences, irrespective of their location. For more flexibility, we may think of assuming a different smoothing parameter for each difference

$$\sum_{k=q+1}^{c} \lambda_{k-q} \left(\Delta^q \theta_k\right)^2 = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{D}^{\mathrm{T}} \mathrm{diag}(\boldsymbol{\lambda}) \boldsymbol{D} \boldsymbol{\theta}, \tag{2}$$

where $\boldsymbol{\lambda} = \left(\lambda_1, \ldots, \lambda_{c-q}\right)^{\mathrm{T}}$. Note that this approach would imply as many smoothing parameters as coefficients (minus q), which could lead to undersmoothing and unstable computations. Given the local and ordered nature of the coefficient differences, we may model the smoothing parameters λ_k as a smooth function of k (its position) and use B-splines for this purpose (here no penalty is assumed)

$$\boldsymbol{\lambda} = \boldsymbol{C} \boldsymbol{\phi},$$

where $\boldsymbol{C}$ is a B-spline regression matrix of dimension $(c-q) \times p$ with $p < (c-q)$, and $\boldsymbol{\phi} = \left(\phi_1, \ldots, \phi_p\right)^{\mathrm{T}}$ is the new vector of smoothing parameters. Performing some simple algebraic operations, it can thus be shown that the *adaptive* penalty (2) is

$$\boldsymbol{\theta}^{\mathrm{T}} \left(\sum_{l=1}^{p} \phi_l \boldsymbol{D}^{\mathrm{T}} \mathrm{diag}\left(\boldsymbol{c}_l\right) \boldsymbol{D} \right) \boldsymbol{\theta}, \tag{3}$$

where $\boldsymbol{c}_l$ denotes the column l of $\boldsymbol{C}$.

3 Estimation algorithm

Estimation of the P-spline model subject to the penalty defined in (1) can be carried out based on the equivalence between P-splines and mixed models (Currie and Durban, 2002)

$$\boldsymbol{f} = \boldsymbol{B\theta} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha},$$

where $\boldsymbol{X} = \left[\mathbf{1}_n | \boldsymbol{x} | \cdots | \boldsymbol{x}^{(q-1)}\right]$ and $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{D}^{\mathrm{T}} \left(\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}\right)^{-1}$. It can be shown that the inverse of the variance-covariance matrix of the random effects $\boldsymbol{\alpha}$ is

$$\boldsymbol{G}^{-1} = \frac{1}{\tau^2}\boldsymbol{I}_{c-q},$$

where $\tau^2 = \sigma^2/\lambda$. Here, only one variance component, τ, is present, shrinking or penalizing the $\boldsymbol{\alpha}$ towards zero. The variance component can then be estimated on the basis of the iterative Schall (1991) algorithm. When applying the same mixed model parameterization to the adaptive P-spline with the penalty defined in (3), $\boldsymbol{G}^{-1}$ becomes

$$\boldsymbol{G}^{-1} = \sum_{l=1}^{p} \frac{1}{\tau_l^2} \mathrm{diag}\left(\boldsymbol{c}_l\right), \tag{4}$$

where $\tau_l^2 = \sigma^2/\phi_l$. In this case each random effect is shrunk by several variance components, making the application of Schall unfeasible. In the paper by Rodríguez-Álvarez et al. (2014) the Schall algorithm was extended to deal with multiple penalties on the same coefficients, with the penalties coming in that case from two (or more) spatial dimensions. However, multiple penalties can arise in a broader class of situations, as in our adaptive approach. Given that $\boldsymbol{G}^{-1}$ in (4) is expressed as a linear combination defined over the variance components, it can be shown that the SAP algorithm can be generalized to the estimation of τ_l. Specifically, in each iteration the variance components estimates are updated, until convergence, according to

$$\widehat{\tau}_l^2 = \frac{\boldsymbol{\alpha}^{\mathrm{T}} \mathrm{diag}\left(\boldsymbol{c}_l\right) \boldsymbol{\alpha}}{\mathrm{ed}_l},$$

where

$$\mathrm{ed}_l = \mathrm{trace}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}\frac{\mathrm{diag}\left(\boldsymbol{c}_l\right)}{\tau_l^2}\boldsymbol{G}\right),$$

with $\mathbf{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}$ and $\mathbf{V} = \sigma^2\mathbf{I}_n + \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathrm{T}}$. In principle, these traces involve several $n \times n$ matrices. However, an efficient computation can be achieved since: (a) $\boldsymbol{G}\mathrm{diag}\left(\boldsymbol{c}_l\right)\boldsymbol{G}$ is a diagonal matrix; and, (b) $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Z}$ can be easily computed (see equation (8) in Rodríguez-Álvarez et al., 2014). Finally, note that $\sum_{l=1}^{p} \mathrm{ed}_l = \mathrm{trace}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Z}\boldsymbol{G}\right) = \mathrm{trace}\left(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{P}\right)$, where $\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{P}$ corresponds to the hat matrix of the random effects.

4 Adaptive penalty in two dimensions

Extension of the univariate P-spline model given in Section 2 above to the modeling of two-dimensional (2D) surfaces is usually based on the tensor product of univariate B-spline basis, with the penalty matrix being defined as (see Eilers and Marx, 2003, for further details)

$$\gamma_1 \left(\mathbf{I}_{c_2} \otimes \boldsymbol{D}_1\right)^{\mathrm{T}} \left(\mathbf{I}_{c_2} \otimes \boldsymbol{D}_1\right) + \gamma_2 \left(\boldsymbol{D}_2 \otimes \mathbf{I}_{c_1}\right)^{\mathrm{T}} \left(\boldsymbol{D}_2 \otimes \mathbf{I}_{c_1}\right),$$

where γ_1 and γ_2 are the smoothing parameters (we assume anisotropy) and $\otimes$ denotes the Kronecker product. Following the same reasoning used for the univariate case, in the adaptive case each smoothing parameter γ_d $(d = 1, 2)$ is replaced by a vector of smoothing parameters $\boldsymbol{\gamma}_d$, where each component is associated with one coefficient difference. To reduce the dimension, $\boldsymbol{\gamma}_d$ is then modeled by means of B-splines. However, since we still are in the two dimensional case, the tensor product of B-spline basis is used. Specifically,

$$\begin{aligned}\boldsymbol{\gamma}_1 &= \left(\boldsymbol{C}_{11} \otimes \boldsymbol{C}_{12}\right) \boldsymbol{\phi}_1 = \boldsymbol{C}_1 \boldsymbol{\phi}_1,\\ \boldsymbol{\gamma}_2 &= \left(\boldsymbol{C}_{21} \otimes \boldsymbol{C}_{22}\right) \boldsymbol{\phi}_2 = \boldsymbol{C}_2 \boldsymbol{\phi}_2,\end{aligned}$$

where $\boldsymbol{C}_{11}$, $\boldsymbol{C}_{12}$, $\boldsymbol{C}_{21}$ and $\boldsymbol{C}_{22}$ are B-spline regression matrices of dimension $(c_1 - q_1) \times p_{11}$, $c_2 \times p_{12}$, $c_1 \times p_{21}$ and $(c_2 - q_2) \times p_{22}$, respectively, and $\boldsymbol{\phi}_1 = (\phi_{11}, \ldots, \phi_{1p_{11}p_{12}})$ and $\boldsymbol{\phi}_2 = (\phi_{12}, \ldots, \phi_{1p_{21}p_{22}})$. Accordingly, the *adaptive* penalty matrix in two dimensions can be then expressed as

$$\begin{aligned}&\sum_{m=1}^{p_{11}p_{12}} \phi_{1m} \left(\mathbf{I}_{c_2} \otimes \boldsymbol{D}_1\right)^{\mathrm{T}} diag\left(\boldsymbol{c}_{1,m}\right) \left(\mathbf{I}_{c_2} \otimes \boldsymbol{D}_1\right) +\\ &\sum_{s=1}^{p_{21}p_{22}} \phi_{2s} \left(\boldsymbol{D}_2 \otimes \mathbf{I}_{c_1}\right)^{\mathrm{T}} diag\left(\boldsymbol{c}_{2,s}\right) \left(\boldsymbol{D}_2 \otimes \mathbf{I}_{c_1}\right).\end{aligned}$$

$\boldsymbol{c}_{d,l}$ denoted the column l of $\boldsymbol{C}_d$.

5 An application

To illustrate our proposal, we use data consisting of photon counts of diffracted x-ray radiation as a function of the angle of diffraction. The dataset can be found in the `R`-package `diffractometry`. Given that the outcome variable represents count data, a Poisson model was adopted. We compared the performance of the SOP algorithm with the method given in Wood (2011), as it is implemented in the `R`-package `mgcv`. In both cases, we used second-order differences and 200 B-splines for the curve and 80 for the adaptive penalty. Results are shown in Figure 1. The result of `mgcv` was almost identical to our proposal, so it is not depicted. Our algorithm

took less that 3 seconds, whereas `mgcv` was around 1000 times slower. We also applied the proposed algorithm to the analysis of simulated 2D data. A sample size of 2000 Gaussian data with $\sigma = 0.1$ was simulated, second-order differences were used, and we chose 15 marginal B-splines for the surface and 8 for the adaptive penalty. Note that this configuration yields to 128 ($8 \times 8 \times 2$) variance components. It should be noted, however, that the implementation of the 2D adaptive in the `mgcv` package uses a different adaptive penalty that the one presented on Section 4. In that approach, and for the same configuration, the number of variance components is 64. Here our algorithm was around 30 times faster than Wood (2011)'s approach, providing a computing time of about 22 seconds. Figure 2 depicts the graphical results. As for the univariate case, the result of `mgcv` was very similar to SOP and it is not shown.

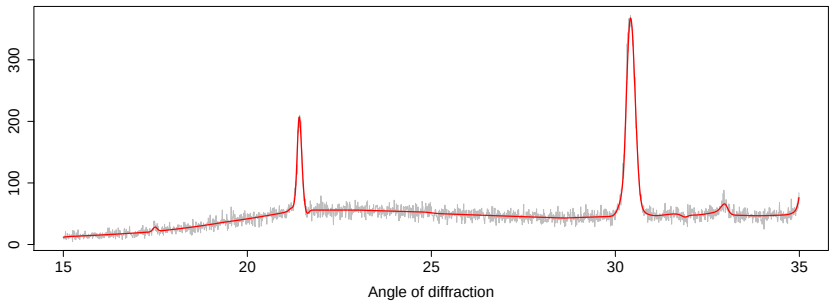

FIGURE 1. Smooth effect of the angle of diffraction on the x-ray radiation. Grey: Raw data. Red: SOP algorithm.

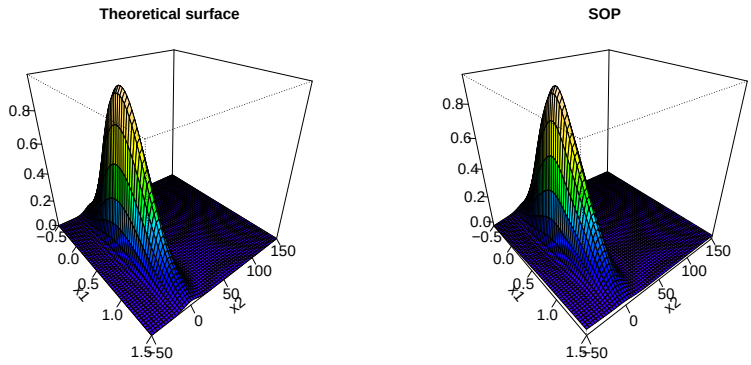

FIGURE 2. Left: simulated theoretical surface. Right: fitted surface given by the SOP algorithm.

Acknowledgments: The authors would like to express their gratitude for the support received in the form of the Spanish Ministry of Economy and Competitiveness grants MTM2011-28285-C02-01 and MTM2011-28285-C02-02. M.X. Rodríguez-Álvarez thanks the Agrupamento INBIOMED from DXPCTSUG-FEDER unha maneira de facer Europa (2012/273).

References

Currie, I. and Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, **4**, 333 – 349.

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

Eilers, P.H.C. and Marx, B.D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*, **66**, 159 – 174.

Krivobokova, T., Crainiceanu, C.M., and Kauermann, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, **17**, 1 – 20.

Rodríguez-Álvarez, M.X., Lee, D-J., Kneib, T., Durbán, M., and Eilers, P.H.C. (2014). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing* (To appear). DOI: 10.1007/s11222-014-9464-2.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719 – 721.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B*, **73**, 3 – 36.

BTL-Lasso – A penalty approach to heterogeneity in paired comparison data

Gunther Schauberger[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-Universität Munich, Germany

E-mail for correspondence: gunther.schauberger@stat.uni-muenchen.de

Abstract: A new method to handle heterogeneity in paired comparison data is proposed. The preference of an item over another item is modelled depending on covariates of the subjects. The model allows for heterogeneity between subjects as the preference for an item can vary between subjects depending on subject-specific covariates. The model is estimated with a regularized estimation approach penalizing the differences between the subject-specific parameters corresponding to covariates. The specific penalty term allows for variable selection and for clusters of items regarding certain covariates. The method is applied to data from a pre-election study from Germany.

Keywords: BTL-Lasso; Paired comparison; Bradley-Terry; Lasso; Heterogeneity.

1 The Bradley-Terry-Luce model for binary and ordinal response

The Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952) is a popular choice when it comes to the modeling of paired comparison data. If $\{a_1, \ldots, a_m\}$ denote the set of items to be compared, one models the probability that item a_r is preferred over item a_s. In the simple case of preferring either the first or the second item, the BTL model has the form

$$P(Y_{(r,s)} = 1|r, s) = \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)}, \tag{1}$$

where the random variable $Y_{(r,s)}$ indicates which item is preferred. If $Y_{(r,s)} = 1$, item a_r is preferred over a_s, if $Y_{(r,s)} = 0$ item a_s is preferred. The parameter γ_r represents the strength or attractiveness of item a_r. For identifiability, the restriction $\sum_{r=1}^{m} \gamma_r = 0$ is applied. The ordinary BTL model can

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

be seen as a logistic regression model. In a more general setting, also other link functions than the logit link are possible.
In some applications, paired comparison data can or should not be reduced to binary decisions, for example in sport events where also draws are possible. Therefore, paired comparison data can also have an ordinal response (Tutz, 1986). In this case, model (1) is extended to

$$P(Y_{(r,s)} \leq k|r,s) = \frac{\exp(\theta_k + \gamma_r - \gamma_s)}{1 + \exp(\theta_k + \gamma_r - \gamma_s)} \tag{2}$$

and can be seen as a cumulative logit model. The parameters $\theta_1, \ldots, \theta_K$ represent threshold parameters for the different levels of the response $Y_{(r,s)} \in \{1, \ldots, K\}$. Here, $Y_{(r,s)} = K$ corresponds to a strong preference of a_r over a_s and $Y_{(r,s)} = 1$ corresponds to a strong preference of a_s over a_r. In general, model (1) can be seen as a special case of model (2) for $K = 2$.

2 Modeling heterogeneity in paired comparison data

2.1 Including subject-specific covariates in BTL models

Models (1) and (2) implicitly assume, that the attractiveness of an item is the same for all subjects. This rather strict assumption can be relaxed by taking into account subject-specific covariates that can have an influence on the attractiveness of certain items. Then the parameter γ_r for the attractiveness of item a_r is replaced by the term $\gamma_{ir} = \beta_{r0} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_r$, which represents the attractiveness of item a_r for subject i. Here, $\boldsymbol{x}_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$ is a subject-specific covariate vector with p covariates and $\boldsymbol{\beta}_r^{\mathrm{T}} = (\beta_{r1}, \ldots, \beta_{rp})$ represents the effects of the covariates on the attractiveness of item a_r. Incorporating this term into model (2), one obtains

$$\begin{aligned} P(Y_{i(r,s)} \leq k|r,s,\boldsymbol{x}_i) &= \frac{\exp(\theta_k + \gamma_{ir} - \gamma_{is})}{1 + \exp(\theta_k + \gamma_{ir} - \gamma_{is})} \\ &= \frac{\exp(\theta_k + (\beta_{r0} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_r) - (\beta_{s0} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_s))}{1 + \exp(\theta_k + (\beta_{r0} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_r) - (\beta_{s0} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_s))} \\ &= \frac{\exp(\theta_k + \beta_{r0} - \beta_{s0} + \boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s))}{1 + \exp(\theta_k + \beta_{r0} - \beta_{s0} + \boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s))}. \end{aligned} \tag{3}$$

For identifiability, restrictions are necessary, we use the symmetric side constraints $\sum_{r=1}^{m} \beta_{rj} = 0$ for $j = 0, 1, \ldots, p$. Again, Model (3) can be seen as a cumulative logit model or, more precisely, a proportional odds model. The linear predictors for different categories of the response only differ with regard to the threshold parameters θ_k. The remaining parameters are independent from the single response categories. For a response variable consisting of K ordered categories, one models $K - 1$ cumulative probabilities

$P(Y_{(r,s)} \leq 1 \mid r,s), \ldots, P(Y_{(r,s)} \leq K-1 \mid r,s)$. The probability for a single response category is represented by the difference $P(Y_{(r,s)} = k \mid r,s) = P(Y_{(r,s)} \leq k \mid r,s) - P(Y_{(r,s)} \leq k-1 \mid r,s)$. Therefore, $P(Y_{(r,s)} \leq k)$ has to be greater or equal $P(Y_{(r,s)} \leq k-1)$ for $k = 1, \ldots, K$ to have non-negative probabilities for all single categories. As the probabilities only differ with respect to the threshold parameters, this is ensured if $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_K$. The threshold for the last category K is restricted to $\theta_K = \infty$ so that $P(Y_{(r,s)} \leq K) = 1$ holds. Furthermore, it is reasonable to put further restrictions on the threshold parameters to ensure equal probabilities for corresponding categories if the order of the paired comparison is reversed. Therefore, we use the restrictions $\theta_k = -\theta_{K-k}$ and, if K is even, additionally $\theta_{K/2} = 0$. For example, these restrictions ensure that $Y_{(r,s)} = K$ (strong preference of a_r over a_s) has the same probability as $Y_{(s,r)} = 1$. Due to these restrictions, $\lfloor \frac{K-1}{2} \rfloor$ (free) threshold parameters have to be estimated. Again, in the special case of binary response ($K = 2$) all threshold parameters are omitted.

2.2 Regularization

Model (3) contains a rather complex parameter structure. On the one hand, we have the parameters from the regular (ordinal) BTL model, namely the threshold parameters and, for each item r, a parameter β_{r0} for its basic attractiveness. They are essential for a meaningful interpretation of the model and, therefore, will not be penalized. On the other hand, we have an additional parameter for each combination between the items and the covariates. These parameters will be penalized, but not regarding to their absolute values but regarding to the absolute values of differences between such parameters corresponding to the same covariate. The respective penalty term is denoted by

$$J(\alpha) = \sum_{j=1}^{p} \sum_{r<s} w_{rsj} |\beta_{rj} - \beta_{sj}|,$$

where $r, s \in \{1, \ldots, m\}$, $\alpha^{\mathrm{T}} = (\theta_1, \ldots, \theta_{K-1}, \beta_{10}, \ldots, \beta_{mp})$ and w_{rsj} is a weight parameter. A similar type of penalty has, e.g., been proposed by Oelker et al. (2014). The penalty has the effect, that the parameters referring to the same covariate are shrunk towards each other. When the tuning parameter is high enough, the differences can be shrunk to exactly zero so that the effect of a covariate is the same for two (or more items). Therefore, this penalty leads to clusters of items which share the same effect of a certain covariate. With growing tuning parameter, these clusters become bigger until all items form a single cluster. In that case, due to the sum-to-zero constraints all parameters are zero and, therefore, the covariate is irrelevant for the attractiveness of the items.

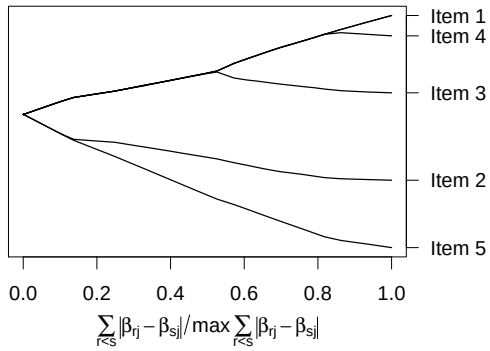

FIGURE 1. Exemplary coefficient paths for a covariate j in a setting with $m = 5$ different items.

For illustration, Figure 1 shows the coefficient paths corresponding to a covariate j for a toy example with $m = 5$ items. The paths are drawn along the (normed) penalty term $\sum_{r<s} |\beta_{rj} - \beta_{sj}|$ for covariate j. It can be seen that the penalty enforces a clustering of the items which is reduced when the penalty is relaxed. The first split creates two clusters, a first cluster consisting of items 1,3 and 4, and a second cluster consisting of items 2 and 5. Next, the second cluster is split again so that items 2 and 5 now form clusters of their own. Finally, also the first cluster is decomposed until every item forms a cluster of its own.

Zou (2006) proposed the so-called adaptive lasso as a derivative of the regular lasso. In contrast to regular lasso, it provides consistency in terms of variable selection. In adaptive lasso, the single penalty terms are weighted with the inverses of the unpenalized ML-estimates. Similar to this concept the weight parameter w_{rsj} are defined by $w_{rsj} = |\beta_{rj}^{\text{ML}} - \beta_{sj}^{\text{ML}}|^{-1}$. Small differences in the ML-estimates are penalized stronger than bigger differences which has the effect that the clustering of the parameters is enforced.

3 Application to pre-election data from Germany

The proposed method is applied to data from the German Longitudinal Election Study (GLES) from 2013, see Rattinger et al. (2014). The participants were asked to rank the five most important parties (CDU/CSU, SPD, Greens, Left Party, FDP) for the upcoming federal election on a scale from -5 to 5. From these responses, paired comparisons were generated as the differences between all parties, ending up with ordered paired comparisons between -10 and 10. The response was narrowed down to an ordered response with five categories.

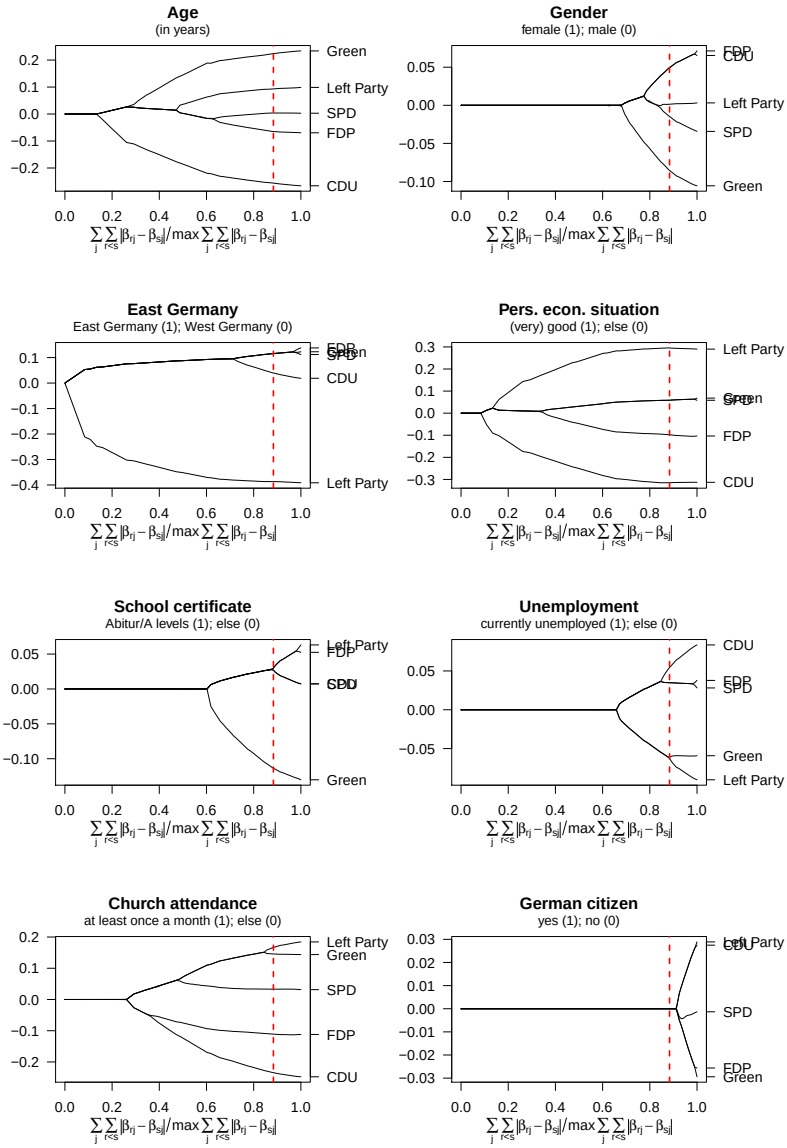

FIGURE 2. Coefficient paths separately for all six covariates. Dashed vertical lines represent optimal model according to 10-fold cross-validation.

Figure 2 shows the corresponding coefficient paths and entails short variable descriptions for the eight covariates we used. The coefficient paths are drawn separately for each covariate. The paths illustrate how the penalty

term enforces clustering of the different parties. The dashed vertical lines represent the optimal model according to the 10-fold cross-validation.
The coefficient paths allow for interesting insights into how the preference of the voters for certain parties depends on characteristics of the voters themselves. Let us first consider the covariate unemployment. With respect to unemployment, the parties can be divided into two main clusters. The Left party and the Greens in one cluster, CDU, SPD and FDP in another cluster. Basically, this could be interpreted as a separation between the older and more established parties (SPD, CDU, FDP) and the younger parties (Greens and Left Party). In the optimal model, the second cluster of parties can be further divided into a cluster of SPD and FDP and a cluster only consisting of CDU. For gender, four different clusters are identified in the final model. The Greens are much more attractive for female than for male voters and form a cluster of their own. The SPD and the Left party seem almost equally attractive for males and females while the CDU and the FDP are more attractive for males. For the variable school leaving certificate, a very sparse solution with only two clusters (Greens vs. all other parties) emerged confirming the reputation of the Greens to be a party for academics. The German citizenship was completely eliminated from the model, naturalized citizens do not systematically prefer other parties. For the variables age and church attendance, all parties are very different and every parties forms a cluster of its own.

References

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, **39**, 324 – 345.

Oelker, M., Gertheiss, J., and Tutz, G. (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling*, **14**, 157 – 177.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., and Wolf, C. (2014). *Pre-election Cross Section (GLES 2013)*. GESIS Data Archive.

Tutz, G. (1986). Bradley-Terry-Luce models with an ordered response. *Journal of Mathematical Psychology*, **30**, 306 – 316.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418 – 1429.

Variable selection in mixture models for ordinal response with an uncertainty component

Micha Schneider[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: Micha.Schneider@stat.uni-muenchen.de

Abstract: Mixture Models provide the opportunity to model discrete human choices as a combination of preference and uncertainty structure. In CUB models the preference is represented by shifted binomial random variables and the uncertainty by a discrete uniform distribution. This concept can be extended to a wider class of models by using classical ordinal response models as the cumulative model for the preference structure. To reduce model complexity variable selection via lasso regularization is proposed. The variable selection leads to stable parameter estimates and easy-to-interpret results in both model components.

Keywords: Mixture models; Variable selection; CUP model; Ordinal response.

1 Mixture models

The probability that an individual i chooses the category r from ordered categories $\{1, \dots, k\}$ given the explanatory variables $\boldsymbol{z}_i, \boldsymbol{x}_i$ is composed of the individual's propensity towards uncertainty and preference structure. The mixture distribution denoted as CUB (combination of uncertainty and binomial) as considered, for example, by Iannario and Piccolo (2012) has been defined by

$$P(R_i = r|\boldsymbol{z}_i, \boldsymbol{x}_i) = \pi_i b_r(\xi_i) + (1 - \pi_i)p_r^U \qquad r \in \{1, \dots, k\}, \qquad (1)$$

where π_i is the propensity or mixture weight. The preference structure is modelled by a shifted binomial distribution $b_r(\cdot)$ with parameter ξ and the uncertainty by a uniform distribution p_r^U with probability $1/k$ for each of the response categories. The uncertainty includes nearly all kinds of indecision related to the nature of human choices like willingness to respond, lack of time, partial understanding etc.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

The parameters π_i and ξ_i are linked to the covariates $(\boldsymbol{z}_i^{\mathrm{T}}, \boldsymbol{x}_i^{\mathrm{T}})$ by the logit models

$$\text{logit}\,(\pi_i) = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}; \quad \text{logit}\,(\xi_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}; \qquad i = 1, \ldots, n. \tag{2}$$

The first model links the explanatory variables $\boldsymbol{z}_i$ to the probability of a structured response while the second model links the explanatory variables $\boldsymbol{x}_i$ to the choice of a specific response category given the individual response is in a structured way. The covariates $\boldsymbol{z}_i^{\mathrm{T}}$ and $\boldsymbol{x}_i^{\mathrm{T}}$ may be identical, completely different or overlap. If the preference structure cannot be captured by the simple shape of a binomial distribution (which requires a single mode, for instance), the use of the CUB model is not appropriate. Then a more flexible approach is necessary.

2 An extended class of models

Although the binomial distribution for the preference structure in the CUB models can be motivated by a counting process of selection among the k categories the main reason for the choice of the binomial distribution in the structured part of the mixture is its simplicity. An extended class of models is obtained by replacing the rather restrictive binomial model by more flexible ordinal models while the uniform distribution as an uninformative distribution is retained. The so called CUP model (combination of uncertainty and preference), described in Tutz et al. (2014), has the general form

$$P(R_i = r|\boldsymbol{x}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1 - \pi_i)P_U(U_i = r), \tag{3}$$

where R_i represents the observed response and Y_i, U_i are the unobserved random variables taking values from $\{1, \ldots, k\}$. $P_U(U_i = r) = 1/k$ represents the uniform distribution, while the distribution of Y_i is determined by $P_M(Y_i = r|\boldsymbol{x}_i)$, which can be any ordinal model. Traditional models are the cumulative logit model

$$\log\left(\frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1,$$

or the adjacent categories model

$$\log\left(\frac{P(Y_i = r+1|\boldsymbol{x}_i)}{P(Y_i = r|\boldsymbol{x}_i)}\right) = \gamma_{0r} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}, \quad r = 1, \ldots, k-1$$

(see Agresti, 2013; or Tutz, 2012). These models are more flexible and can handle complex ordinal data structures. The CUB model is a special case in which the binomial distribution is used to model the preference structure. As in CUB the omission of the uncertainty component in CUP models typically yields biased parameter estimates.

3 Variable selection

Since two independent sets of parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, are involved and it is typically not known which variables are relevant for the mixture and which for the preference structure variable selection is quite important in mixture models. It is proposed to use an adopted version of the lasso (Tibshirani, 1996) to the mixture models. The penalized log-likelihood that is to be maximized is given by

$$l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{\gamma}) - J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the un-penalized log-likelihood and $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is a specific penalty term. Let the vectors $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ be partitioned into $\boldsymbol{z}_i^{\mathrm{T}} = (\boldsymbol{z}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{z}_{ig}^{\mathrm{T}})$ and $\boldsymbol{x}_i^{\mathrm{T}} = (\boldsymbol{x}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{x}_{ih}^{\mathrm{T}})$ such that each components refer to a single variable. For example, the vector $\boldsymbol{z}_{ij}$ can represent all the dummy variables that are linked to the jth variable, or represent the power functions of the jth variable if one includes polynomial terms. The corresponding predictors are $\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}$ and $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}$ with corresponding partitioning of the parameter vectors, $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_g^{\mathrm{T}})$ and $\boldsymbol{\gamma}^{\mathrm{T}} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_h^{\mathrm{T}})$, respectively. The proposed penalty then has the form

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_1 \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta},j}} \|\boldsymbol{\beta}_j\| + \lambda_2 \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma},j}} \|\boldsymbol{\gamma}_j\| \tag{4}$$

where λ_1 and λ_2 are the tuning parameters for the selection of $\boldsymbol{z}$ and $\boldsymbol{x}$ variables, respectively. The weights $df_{\boldsymbol{\beta},j}$ are defined as the number of parameters collected in the corresponding parameter vector $\boldsymbol{\beta}_j$, the weights $df_{\boldsymbol{\gamma},j}$ are defined in the same way. The penalty enforces the selection of variables in the spirit of the group lasso (Yuan and Lin, 2006) rather than selection of single parameters. If each parameter group consists of only one element the approach is equivalent to the lasso adapted to mixture models.

4 Computational aspects

The mixture models considered in the previous sections can be estimated by an adapted version of the EM-Algorithm proposed by Dempster et al. (1977). The information about the membership to the uncertainty or structure component is treated as missing data. Then the log-likelihood contribution of observation i is given by

$$l_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = z_i^* \left(\log(\pi_i \cdot f_M(r_i|\boldsymbol{x}_i, \boldsymbol{\gamma}))\right) + (1 - z_i^*) \left(\log((1 - \pi_i) \cdot 1/k)\right),$$

where z_i^* takes the value 1 if observation i belongs to the structure component and zero if observation i belongs to the uncertainty component. The probability π_i depends on the individual characteristics by

$$\pi_i = 1/(1 + e^{-\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}}).$$

During the EM-Algorithm $\hat{z}_i^*$ is estimated to obtain $\hat{\boldsymbol{\beta}}$, $\hat{\pi}_i$ and $\hat{\boldsymbol{\gamma}}$. Dampster et al. (1977) showed that under weak conditions the EM algorithm finds a local maximum of the likelihood function. Hence it is sensible to use different start values to find the solution of the maximization problem. A detailed description of the use of the EM-Algorithm in CUP mixture models can be found in Tutz et al. (2014).
If a penalty term is included the algorithm has to be modified. First, one replaces the usual log-likelihood by the corresponding penalized log-likelihood. Specific algorithms are needed to account for the penalty term for which the derivative does not exist at critical values. We used a version of the FISTA algorithm by Beck and Teboulle (2009) that was provided by Wolfgang Pößnecker.

5 Application

In the following, the methods are applied to the data from the Survey on Household Income and Wealth (SHIW) by the Bank of Italy. The data set consists of 3816 respondents from the wave of 2010. The response is the happiness index indicating the overall life well-being measured on a Likert Scale from 1 (very unhappy) to 10 (very happy). Several covariates as, for example, age, marital status, area of living and educational degree are available.
In the beginning of the selection process all available covariates are included. For simplicity we chose the same λ-values for λ_1 and λ_2. The lowest BIC was found for $\lambda \approx 20.4$. Figure 1 shows the coefficient paths for both parameter sets $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Each line type stands for one parameter group. Because of the penalty term there are some parameters which are selected in both parameter set as for example marital status or area of living and others which are only selected in one of the two sets.

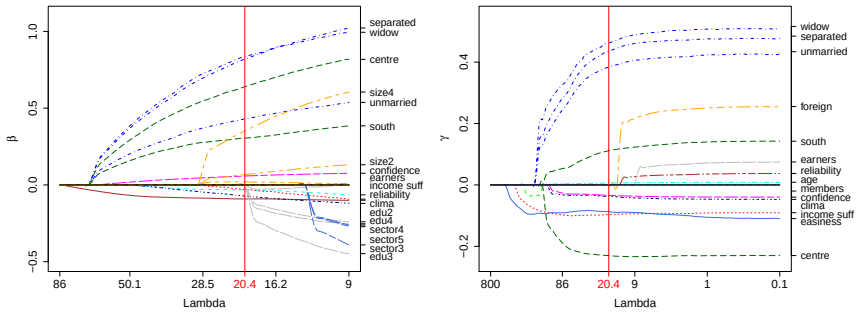

FIGURE 1. Coefficients paths of β (left) and γ (right).

Table 1 shows the results of the penalized model. Figure 2 illustrates the effects of the selected covariates. It is seen that the marital status "widow"

corresponds to high values of unhappiness and high certainty (small $1-\pi$). In contrast, the status "married" indicates happiness but a large amount of uncertainty in the response. From the plot for the variable area it is seen that the people living in the north have large uncertainty and medium happiness whereas people from the south tend to categories that indicate unhappiness with a middle level of uncertainty.

TABLE 1. Coefficients of the chosen model.

Covariates	Concomitant(β)	Structure(γ)
Constant	0.779	0.647
Marital status: Unmarried	0.429	0.384
Marital status: Separated	0.822	0.438
Marital status: Widow	0.840	0.465
Area: Centre of Italy	0.645	-0.230
Area: South of Italy	0.307	0.112
Confidence in people	0.058	-0.034
Interview atmosphere	-0.069	-0.038
Income sufficient	-0.028	-0.097
Age (centered)		0.006
Income earners	-0.032	
Reliability of information	-0.093	
Easiness to answer		-0.087
District size Cat2	0.067	
District size Cat3	0.019	
District size Cat4	0.350	

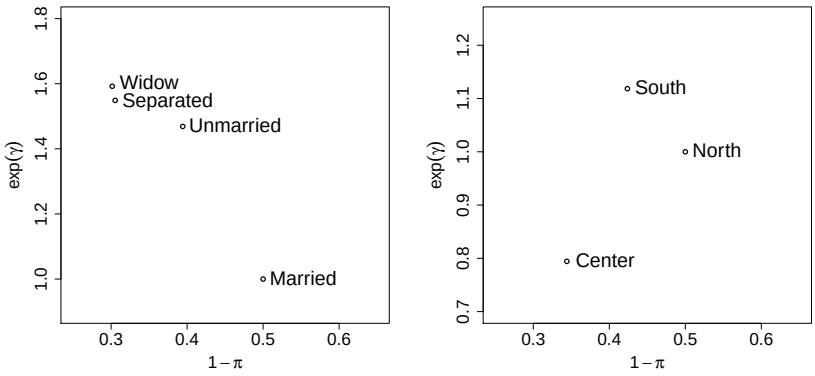

FIGURE 2. Effects of the categorical covariates marital status (left) and area of living (right) in the structure and uncertainty component.

Acknowledgments: Thanks to Wolfgang Pößnecker who provided the R-Code for the EM-Algorithm and Maria Iannario and Domenico Piccolo who provided the SHIW data.

References

Agresti, A. (2013). *Categorical Data Analysis*, (3rd ed.), New York: Wiley.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **2**, 183 – 202.

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.

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

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

Piccolo, D. (2006). Observed information matrix in mub models. *Quaderni di Statistica*, **8**, 33 – 78.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267 – 288.

Tutz, G. (2012). *Regression for Categorical Data.* Cambridge: Cambridge University Press.

Tutz, G., Schneider, M., Innario, M., and Piccolo, D. (2014). Mixture models for ordinal responses to account for uncertainty of choice, *Department of Statistics, LMU München, Technical Report*, **175**.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49 – 67.

Penalized logistic regression for small or sparse data: interval estimators revisited

Marianna Siino[1], Salvatore Fasola[1], Vito M.R. Muggeo[1]

[1] Dip. Scienze Economiche, Aziendali e Statistiche, University of Palermo, ITALY

E-mail for correspondence: marianna.siino01@unipa.it

Abstract: This paper focuses on interval estimation in logistic regression models fitted through the Firth penalized log-likelihood. In this context, many authors have claimed superiority of the likelihood ratio statistic with respect to the (wrong) Wald statistic via simulation evidence. We re-assess such findings by detailing the inferential tools also including in the comparisons the (right) Wald statistic and other statistics neglected in previous literature. In particular, we assess performances of the CIs estimators by simulation and compare them in a real data set. Differently from previous findings, the likelihood ratio statistic does not appear to be the best inferential device in Firth penalized logistic regression.

Keywords: Penalized likelihood; Logistic regression; Sandwich formula; Score-based CIs; Gradient-based CIs.

1 Introduction

The logistic regression equation reads as $\text{logit}(\pi_i) = \sum_{j=1}^{K} x_{ij}\beta_j$ where $\pi_i = E[Y|x_i]$, Y is the dichotomic response variable and $\boldsymbol{x}_i$ a K-dimensional covariate vector. To estimate the regression parameters β_j, Firth (1993) suggested to modify the classical score functions $U_j(\boldsymbol{\beta})$ through

$$U_j^*(\boldsymbol{\beta}) = U_j(\boldsymbol{\beta}) + \frac{1}{2}\text{tr}\left\{I(\boldsymbol{\beta})^{-1}\partial I(\boldsymbol{\beta})/\partial\beta_j\right\}, \qquad j = 1, \ldots, K \tag{1}$$

corresponding to the penalized log-likelihood

$$\ell^*(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \log|I(\boldsymbol{\beta})|^{\frac{1}{2}}, \tag{2}$$

where the penalty $|I(\boldsymbol{\beta})|^{\frac{1}{2}}$ is known as the Jeffrey's invariant prior and $I(\boldsymbol{\beta})$ is the Fisher Information matrix.

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Since the Firth approach allows to remove the first order $O(n^{-1})$ bias of the MLEs, it is widespread in practice, especially in medical statistics involving small samples, rare events, unbalanced data or highly predictive covariates. The Firth approach has been discussed by Heinze and Schemper (2002) to address the issue of monotone likelihood, also known as phenomenon of separation between the response and any covariate. In fact, it always produces finite estimates even when some classical estimate diverges to $\pm\infty$. In the literature, some suggestions to bypass this problem have been discussed, including the omission of the variable that causes the problem of separation, *ad hoc* adjustments of the data or setting the relevant parameter to a high value. However, these alternatives are not appropriate when the interest is to evaluate the main and joint effects with other covariates. Exact logistic regression represents a possible solution, but conditional maximum likelihood estimates are not always obtainable and the distribution of the sufficient statistics is degenerate in the presence of continuous covariates (Mehta and Patel, 1995). Therefore, penalized maximum likelihood inference is more convenient when exact results are unavailable or too conservative (see for instance Heinze, 2006).

We focus on the construction of confidence intervals for the β_js, based on test statistics computed using penalized likelihood quantities. Existing approaches consider CIs based on the penalized likelihood ratio and the Wald statistics (Heinze and Schemper, 2002; Bull et al., 2007). However, we note two possible shortcomings in these papers: first, they consider an inappropriate Wald statistic as discussed below; second, other statistics, such as the score and the more recent gradient statistics (Terrell, 2002), are ignored. See Muggeo and Lovison (2014) for a graphical characterization of the four likelihood-based statistics.

In the next section, we describe the different test statistics used to construct CIs in the penalized logistic regression. In Section 3, they are compared via simulation in terms of coverage levels and average width, while in Section 4 we apply them in a real example. A discussion follows in the last section.

2 Methods

Let $T(\beta_{0j})$ be a pivot statistic. A $(1-\alpha)100\%$ confidence interval for the parameter of interest β_j is defined as

$$CI = \{\beta_{0j} \in \mathbb{R} : z_{\frac{\alpha}{2}} \leq T(\beta_{0j}) \leq z_{1-\frac{\alpha}{2}}\}, \tag{3}$$

where $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ are the quantiles of the standard normal distribution for a given α.

The likelihood ratio statistic currently discussed in the literature is

$$L = \text{sign}(\hat{\beta}_j^* - \beta_{0j})\sqrt{-2\{\ell^*(\hat{\boldsymbol{\beta}}^*) - \ell^*(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^*)\}},$$

where $\hat{\boldsymbol{\beta}}^*$ is the full (unrestricted) penalized ML estimate, and $\hat{\boldsymbol{\beta}}^*_{\mathbf{0}}$ is the restricted penalized ML estimate obtained by fixing β_j at β_{0j}.
The Wald statistic is

$$W = \frac{\hat{\beta}^*_j - \beta_{0j}}{\sqrt{\mathrm{I}^{-1}(\hat{\boldsymbol{\beta}}^*)_{jj}}},$$

which uses the jth element of the main diagonal of the Information to estimate the variance of $\hat{\beta}^*_j$.
However, it should be stressed that $\mathrm{I}^{-1}(\cdot)$ is not an appropriate formula to compute the variance matrix of the estimator. In fact, from basics of inference, the variance of the ML estimator comes from the sandwich formula, reducing to the inverse Information only if the model is correctly specified and the second Bartlett identity holds. Clearly, the second Bartlett identity does not hold in the Firth approach, namely $E[-H^*(\boldsymbol{\beta})] \neq \mathrm{I}(\boldsymbol{\beta})$, where H^* is the Hessian depending on the penalty. Thus, a more reliable variance estimate for $\hat{\boldsymbol{\beta}}^*$ is given by the sandwich formula

$$V(\hat{\boldsymbol{\beta}}^*) \approx H^*(\hat{\boldsymbol{\beta}}^*)^{-1}\mathrm{I}(\hat{\boldsymbol{\beta}}^*)H^*(\hat{\boldsymbol{\beta}}^*)^{-1}. \tag{4}$$

For large samples the penalty effect in (2) vanishes and the simple approximation $V(\hat{\boldsymbol{\beta}}^*) \approx \mathrm{I}^{-1}(\hat{\boldsymbol{\beta}}^*)$ holds; however in small to moderate samples $\mathrm{I}^{-1}(\hat{\boldsymbol{\beta}}^*)$ typically overestimates the variance of $\hat{\boldsymbol{\beta}}^*$. This crucial issue appears to have been overlooked in literature, causing a (pointless) 'bad reputation' of the Wald statistic.
We denote the Wald statistic based on the sandwich variance by

$$W_S = \frac{\hat{\beta}^*_j - \beta_{0j}}{\sqrt{\mathrm{v}_{jj}(\hat{\boldsymbol{\beta}}^*)}},$$

where $\mathrm{v}_{jj}(\hat{\boldsymbol{\beta}}^*)$ is computed using the formula (4) having a numerical approximation of $H^*(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^*$.
Likelihood ratio and Wald represent the most famous likelihood based statistics used for interval estimators in penalized logistic regression, but other asymptotically equivalent statistics are available.
We also consider the score statistic

$$S = U^*_j(\hat{\boldsymbol{\beta}}^*_{\mathbf{0}})\sqrt{\mathrm{I}^{-1}(\hat{\boldsymbol{\beta}}^*_{\mathbf{0}})_{jj}},$$

where $\mathrm{I}^{-1}(\hat{\boldsymbol{\beta}}^*_{\mathbf{0}})_{jj}$ is the jth element on the main diagonal of the inverse of the variance of the conditional score U^*_j. The gradient statistic (Terrel, 2002) is

$$G = \mathrm{sign}(\hat{\beta}^*_j - \beta_{0j})\sqrt{(\hat{\beta}^*_j - \beta_{0j})U^*_j(\hat{\boldsymbol{\beta}}^*_{\mathbf{0}})}.$$

Curiously, the score statistic is well known in the mainstream inference background, but its use appears to be quite limited in practical applications. All the aforementioned statistics have a standard Gaussian null limit distribution, thus the confidence interval (3) can be computed.

3 Simulation study

In this section, we show the results of a small simulation study which compares CIs based on W, W_S, L, S, and G. We generate Bernoulli data $Y_i \sim Ber(\pi_i)$ where $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$ with $\beta_0 = 1$, $\beta_1 \in \{0.5, 1.5\}$, and three different sample sizes $n \in \{20, 50, 100\}$. For each scenario we consider a balanced binary variable $x_i = I(i > n/2)$ and a continuous covariate with equally spaced values, i.e. $x_i = i/n$.

TABLE 1. Simulation results based on 1000 runs: empirical coverage level (CL) and average width (AW) of the CIs based on the Wald (W), the Wald with the sandwich variance (W_S), likelihood ratio (L), score (S) and gradient (G) statistics.

		$x_i = I(i > n/2)$				$x_i = i/n$			
		$\beta_1 = 0.5$		$\beta_1 = 1.5$		$\beta_1 = 0.5$		$\beta_1 = 1.5$	
n	Test	CL	AW	CL	AW	CL	AW	CL	AW
20	W	0.991	4.69	0.984	5.52	0.994	7.25	0.991	8.00
	W_S	0.970	4.09	0.934	4.43	0.982	6.70	0.981	7.16
	L	0.957	4.87	0.973	6.05	0.968	7.40	0.981	8.40
	S	0.951	3.83	0.934	4.31	0.968	6.40	0.972	7.20
	G	0.957	5.59	0.972	7.55	0.962	7.97	0.977	9.29
50	W	0.968	2.81	0.970	3.65	0.968	4.73	0.971	5.63
	W_S	0.958	2.66	0.957	3.15	0.962	4.55	0.943	5.22
	L	0.959	2.83	0.965	3.87	0.959	4.75	0.950	5.80
	S	0.961	2.58	0.947	3.19	0.955	4.44	0.953	5.38
	G	0.950	2.93	0.967	4.41	0.955	4.86	0.947	6.05
100	W	0.955	1.94	0.959	2.59	0.965	3.31	0.966	3.96
	W_S	0.946	1.90	0.938	2.36	0.959	3.25	0.951	3.80
	L	0.946	1.94	0.940	2.67	0.954	3.31	0.954	4.01
	S	0.948	1.86	0.945	2.40	0.951	3.19	0.957	3.85
	G	0.943	1.96	0.937	2.84	0.953	3.34	0.950	4.08

Table 1 reports the empirical coverage level (CL) and the relevant average width (AW) of the 95% CIs for β_1. Broadly speaking, the likelihood ratio statistic does not perform the best. In general, the CIs performances, in terms of coverage level, are similar using L, S, W_s and G statistics; therefore, score, modified Wald and gradient-based CIs result to be good alternatives to the likelihood-based CIs. The coverage level of the CIs based on W_S is closer to the nominal level than the CIs based on W, due to the better approximation of the asymptotic variance. Moreover, CIs based on W_S have a smaller average width than the confidence intervals based on W. In fact, also for $n = 100$, the CIs based on W are the worst both in terms of CL and AW. Overall, score-based CIs appear to perform slightly

better than the competitors in terms of CL and AW, also in the most difficult scenarios with small samples ($n = 20$) and strong predictors causing sparsity and sampling zeros ($\beta_1 = 1.5$). Note that, even if the expected value of the penalized score (1) is not null, it does not affect the performance of the score-based CIs, according to the simulation results. On the other hand, its variance is correctly estimated by the classical Fisher information matrix, since the penalty term added to the score in (1) is not random. Moreover, since the core statistic is a sum of random variables it is expected to converge more rapidly to the Normal distribution. The simple gradient statistic, which does not require computation of the second derivatives, provides acceptable CIs in terms of CL but too much larger AW. It is worth pointing out that it is derived as the geometric mean between S and W, so that its distribution may be affected by the same inaccuracies of W. As expected, differences attenuate in large samples.

4 Example: osteogenic sarcoma data

We consider data from Metha and Patel (1995) concerning $n = 46$ patients with osteogenic sarcoma. The three year disease-free interval (DFI3) is the response, while the explanatory categorical variables are gender (SEX), the presence of any osteoid pathology (AOP) and lymphocytic infiltration (LI). The finite classical MLE does not exist since there is separation caused by the variable LI. We estimate a penalized logistic regression model with additive linear effects and compute the 95% confidence intervals for β_{LI} and results are reported in Table 2.

TABLE 2. 95% confidence intervals for β_{LI} based on the five pivot statistics.

95%CI	W	W_S	L	S	G
Inf	-5.504	-4.637	-7.363	-4.805	-10.147
Sup	0.582	-0.286	-0.188	-0.104	-0.356

It is worth emphasizing how, according to W, the variable LI would not be significant, leading to misleading conclusions. Moreover, the 95% CI based on S is the narrowest, while the one based on the gradient statistic is the widest. The results are coherent with the previous simulation findings.

5 Conclusions

We have introduced three 'new' interval estimators for the coefficients in penalized logistic regression, based on the score, the gradient and a modified version of the Wald statistic where the variance is computed using the sandwich formula. The score-based CIs have the best performances,

especially in terms of average width. The gradient-based CI is also a good alternative and, moreover, its ease of computation represents a noteworthy advantage since it just needs estimates and first derivatives. According to the simulation study, the modified Wald CIs perform better than the usual Wald CIs in terms of coverage level and average width, since the sandwich variance is a more appropriate estimator of the asymptotic variance of the penalized maximum likelihood estimators. The likelihood ratio statistics does not necessarily represent therefore, the best approach to construct confidence intervals in penalized logistic regression.

References

Bull, S.B., Lewinger, J.P., and Lee, S.S.F. (2007). Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine*, **26**, 903-918.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27 – 38.

Heinze, G. and Schemper, M. (2002). A solution for the problem of separation in logistic regression. *Statistics in Medicine*, **21**, 2409 – 2419.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, **25**, 4216 – 4226

Mehta, C.R. and Patel, N.R. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine*, **14**, 2143 – 2160.

Muggeo, V. and Lovison G. (2014). The 'three plus one' likelihood-based test statistics: unified geometrical and graphical interpretations. *The American Statistician*, **68**, 302–306.

Terrell, G.R. (2002). The gradient statistic. *Computing Science and Statistics*, **34**, 206-215.

Doubly robust-based GEE for the analysis of longitudinal ordinal missing data

Jose Luiz P. da Silva[1], Enrico A. Colosimo[1], Fabio N. Demarqui[1]

[1] Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

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

Abstract: GEE is a well-known method for the analysis of non-Gaussian longitudinal data. However, it is only valid under the strong assumption of completely at random data (MCAR). In this work, it is proposed a doubly robust method for the analysis of longitudinal ordinal data with intermittently missing response and covariate that are missing at random (MAR). The approach combines ideas of imputation and inverse probability methods. Simulation results revealed better performance of the proposed method compared to single robust competitors. The method is applied to a data set related to Analgesia Pain in Childbirth study.

Keywords: Doubly robust GEE; Missing at random; Multiple imputation; Proportional odds model; Weighted GEE.

1 Introduction

Generalized Estimating Equations (GEE) (Liang and Zeger, 1986) have computational simplicity and it allows populational interpretation of the parameters of interest. However, in its basic form, it is valid only under MCAR (Liang and Zeger, 1986).

In the current paper, it is proposed a doubly robust approach for the analysis of longitudinal ordinal data with intermittently missing response and covariate that are MAR. The proposed approach (DRGEE) combines ideas from Multiple Imputation GEE (MIGEE) (Little and Rubin, 1987), and Weighted GEE (WGEE). For consistency, it requires only the weight or the imputation model to be correct.

This work was motivated by the Analgesia in Childbirth study which was conducted in Minas Gerais state, Brazil. The main objective of that 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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

was to compare two techniques of analgesia for labor pain in 49 patients.

2 Outline of GEE and WGEE

Let $O_{it} \in \{1, \ldots, J\}$ be the ordinal response for subject i $(i = 1, \ldots, n)$ at time t $(t = 1, \ldots, T_i,\ T_i \leq T)$. It can be defined $Y_{itj} = I(O_{it} = j)$ for $j = 1, \ldots, J$, where $I(A)$ denotes the indicator function. Y_{itj} is converted into the vector $\boldsymbol{Y}_{it} = (Y_{it1}, \ldots, Y_{it(J-1)})^{\mathrm{T}}$ and let $\boldsymbol{Y}_i = (Y_{i1}^{\mathrm{T}}, \ldots, Y_{iT_i}^{\mathrm{T}})^{\mathrm{T}}$ the stacked response vector. Let $\boldsymbol{X}_i = (X_{i1}^{\mathrm{T}}, \ldots, X_{iT_i}^{\mathrm{T}})^{\mathrm{T}}$ denotes the covariate vector that may be missing for the i-th subject, and $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{Z}_{iT_i}^{\mathrm{T}})^{\mathrm{T}}$ the matrix of explanatory variables that are always observed. Let $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i | \boldsymbol{X}_i, \boldsymbol{Z}_i)$. In this work it is assumed a cumulative logit link, that is,

$$\text{logit}\left[\Pr(O_{it} \leq j | X_{it}, Z_{it})\right] = \beta_{0j} + X_{it}\beta_x + \boldsymbol{Z}_{it}^{\mathrm{T}}\boldsymbol{\beta}_z, \qquad j = 1, \ldots, J-1. \quad (1)$$

Formulation in (1) implies a proportional odds model (McCullagh, 1980). Main interest is in regression parameters $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0,J-1}, \beta_x, \boldsymbol{\beta}_z^{\mathrm{T}})^{\mathrm{T}}$ estimated by using GEE (Liang and Zeger, 1986; Lipsitz et al., 1994):

$$\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}, \quad (2)$$

where $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^{\mathrm{T}}}$ and $\boldsymbol{V}_i$ is a "working covariance" matrix usually decomposed into the form $\boldsymbol{V}_i = \boldsymbol{F}_i^{1/2} \boldsymbol{C}_i \boldsymbol{F}_i^{1/2}$, where $\boldsymbol{F}_i$ is a matrix containing the marginal variances and $\boldsymbol{C}_i$ is the marginal correlation matrix.

For WGEE, some further notation is needed. It can be defined $R_{it} = 0$ if O_{it} and X_{it} are missing, $R_{it} = 1$ if only X_{it} is observed, $R_{it} = 2$ if only O_{it} is observed, and $R_{it} = 3$ if O_{it} and X_{it} are both observed. Let $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{iT_i})^{\mathrm{T}}$, and $\bar{\boldsymbol{R}}_{it} = (R_{i1}, \ldots, R_{i,t-1})$. The marginal probability $\Pr(\boldsymbol{R}_i = \boldsymbol{r}_i | \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ can be obtained through conditional models $\lambda_{itk} = \Pr(R_{it} = k | \bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, $k = 0, 1, 2, 3$. It is assumed MAR and additionally

$$\Pr(R_{it} = r_{it} | \bar{\boldsymbol{R}}_{it}, \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) = \Pr(R_{it} = r_{it} | \bar{\boldsymbol{R}}_{it}, \bar{\boldsymbol{O}}_{it}^{o}, \bar{\boldsymbol{X}}_{it}^{o}, \bar{\boldsymbol{Z}}_{it}), \quad (3)$$

for each time t, where $\bar{\boldsymbol{O}}_{it}^{o}$ and $\bar{\boldsymbol{X}}_{it}^{o}$ are the histories of observed responses and covariates up to time $t-1$. To model λ_{itk} it is adopted a politomic logistic regression. Let $\pi_{it} = \Pr(R_{it} = 3 | \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $\pi_{itt'} = \Pr(R_{it} = 1, R_{it'} = 3 | \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) + \Pr(R_{it} = 3, R_{it'} = 3 | \boldsymbol{O}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$.

Following Chen and Zhou (2011) define a weight matrix $\boldsymbol{\Delta}_i = [\delta_{itt'}]$, with elements $\delta_{itt'} = \{I(R_{it} = 1, R_{it'} = 3) + I(R_{it} = 3, R_{it'} = 3)\} / \pi_{itt'}$ for $t \neq t'$, $\delta_{itt} = I(R_{it} = 3)/\pi_{it}$. Let $\boldsymbol{M}_i = \boldsymbol{F}_i^{-1/2} (\boldsymbol{C}_i^{-1} \cdot \boldsymbol{\Delta}_i) \boldsymbol{F}_i^{-1/2}$ where "$\cdot$" denotes the Hadamard product. The weighted GEE (WGEE) for $\boldsymbol{\beta}$ are given by

$$\sum_{i=1}^{n} \boldsymbol{D}_i \boldsymbol{M}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}. \quad (4)$$

3 Doubly robust GEE for longitudinal ordinal data

Adding a term of expectation zero, say $\phi(\cdot)$, to (4) gives rise to the so-called *doubly robust* estimators. Chen and Zhou (2011) showed that the optimal choice is given by $\phi_{opt} = E_{(\boldsymbol{Y}_i^m, \boldsymbol{X}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\}$, with $\boldsymbol{N}_i = \boldsymbol{F}_i^{-1/2} \left\{\boldsymbol{C}_i^{-1} \cdot (\mathbf{1}\mathbf{1}^T - \boldsymbol{\Delta}_i)\right\} \boldsymbol{F}_i^{-1/2}$, where $\mathbf{1}$ is a vector of 1's, and $\boldsymbol{Y}_i^m$ and $\boldsymbol{X}_i^m$ denote the missing components of $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$, respectively. A doubly robust estimate for $\boldsymbol{\beta}$ can be obtained by solving

$$\sum_{i=1}^{n} \left[\boldsymbol{D}_i \boldsymbol{M}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) + E_{(\boldsymbol{Y}_i^m, \boldsymbol{X}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{X}_i^o, \boldsymbol{Z}_i, \boldsymbol{R}_i)} \{\boldsymbol{D}_i \boldsymbol{N}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\}\right] = \mathbf{0}. \tag{5}$$

Estimates for missing data parameters as well as the second term in (5) can be obtained by MLE. Under correct specification of at least one of the predictive models, the estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and covariance matrix as in Chen and Zhou (2011).

4 Simulation study

A small simulation study was conducted in order to quantify bias and precision under misspecification of the predictive models. It was considered $T_i = T = 3$ repeated ordinal measures (with three categories) and two covariates (quantitative and qualitative). The true marginal model is

$$\text{logit}\, \Pr(O_{itj} \leq j | X_{it}, Z_{it}) = \beta_{0j} + \beta_1 X_{it} + \beta_2 Z_{it}, \qquad j = 1, 2, \tag{6}$$

where Z_{it} is normal with unity variance and mean $(0, 0.5, 1)$ for $t = 1, 2, 3$. The binary covariate X_{it} may be missing at some time points and is generated according to

$$\text{logit}\, \Pr(X_{it} = 1 | \bar{X}_{it}, Z_{it}) = \gamma_0 + \gamma_1 X_{i,t-1} + \gamma_2 Z_{it}. \tag{7}$$

It is assumed $\beta_{01} = -0.4$, $\beta_{02} = 1.2$, $\beta_1 = -0.5$, $\beta_2 = 0.5$, $\gamma_0 = \log(1)$, $\gamma_1 = 2$ and $\gamma_2 = 2$. Data were generated assuming a constant $\rho = 0.9$ correlation between the latent vectors.

As independent estimating equations were fitted, R_{it} can be defined as the indicator of observing both O_{it} and X_{it}, and it was taken

$$\log \frac{\Pr(R_{it} = 1)}{\Pr(R_{it} = 0)} = \psi_{0t} + \psi_1 I(R_{i,t-1} = 1) + \psi_2 O^*_{i,t-1} + \psi_3 X^*_{i,t-1} + \psi_4 Z_{it}, \;\; t = 2, 3, \tag{8}$$

where $O^*_{i,t-1} = O_{i,t-1}$, if $O_{i,t-1}$ is observed and 0 otherwise, and $X^*_{i,t-1} = X_{i,t-1}$if $X_{i,t-1}$ is observed and 0 otherwise. The true values are taken as $\psi_{02} = 6.6$, $\psi_{03} = 6$, $\psi_1 = 2$, $\psi_2 = -2$, $\psi_3 = -2$ and $\psi_4 = 2$. It was observed about 24% of missing observations under this setup.

TABLE 1. Simulation results for incomplete covariate and response data.

	Empirical bias				Standard Error				Empirical Coverage			
	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z	β_{01}	β_{02}	X	Z
Complete	1.4	0.2	-0.0	0.9	0.156	0.167	0.187	0.072	0.96	0.94	0.94	0.95
Available	-14.8	7.7	-23.2	-18.1	0.159	0.171	0.194	0.077	0.93	0.91	0.92	0.87
WGEE(r^+)	-0.2	0.8	5.6	1.2	0.182	0.198	0.235	0.095	0.95	0.94	0.94	0.93
WGEE(r^-)	15.6	-4.5	-22.2	-3.9	0.184	0.205	0.222	0.088	0.95	0.94	0.92	0.94
MIGEE(x^+)	6.2	-1.9	-10.0	-1.4	0.162	0.174	0.201	0.076	0.95	0.94	0.94	0.95
MIGEE(x^-)	10.3	-2.7	-18.2	-3.9	0.160	0.172	0.200	0.076	0.94	0.92	0.92	0.94
DRGEE(x^+, r^+)	-0.1	0.8	2.8	1.3	0.185	0.202	0.242	0.089	0.96	0.95	0.95	0.94
DRGEE(x^-, r^+)	-0.7	1.1	3.8	1.4	0.194	0.213	0.251	0.088	0.96	0.95	0.95	0.94
DRGEE(x^+, r^-)	2.5	0.0	-2.1	0.1	0.178	0.198	0.224	0.082	0.96	0.96	0.94	0.94
DRGEE(x^-, r^-)	13.9	-3.7	-24.3	-5.5	0.176	0.194	0.228	0.084	0.95	0.94	0.93	0.94

"$^+$" indicates correctly specified model

"$^-$" indicates misspecified model omitting the X_t predictor

It was considered ordinary GEE for the complete and available data, weighted GEE (WGEE), multiple imputation (MIGEE) by chained equations (with 10 multiple imputations), and the proposed doubly robust version (DRGEE). In order to investigate robustness of these methods, the covariate X_{t-1} was omitted from the covariate model (7) or the missing data model (8).

Results are summarized in Table 1 for $n = 300$. In each of the $S = 1000$ Monte Carlo replications it was obtained the relative percentage bias, defined as $100 \times (\hat{\beta} - \beta)/\beta$, the standard error obtained through the sandwich estimator, and the empirical coverage probability probability for 95% confidence intervals.

The MAR missingness impact is observed for all the regression parameters, the largest relative bias occur in binary covariate X.

Doubly robust method requires the simultaneous specification of two predictive model. When at least one of them is correctly specified the resulting estimator is still consistent. Estimates are, on average, closer to those obtained with fully observed data compared to single robust WGEE or MIGEE. This behavior can be observed to all parameters, specially for that associated with the incomplete binary variable X. Regarding the uncertainty of parameter estimates, it is noted that the standard error from DRGEE is greater than multiple imputation, but of the same order as of the weighted method. Further, the efficiency of the doubly robust estimates appears relatively more sensitive to misspecification of the weight model than the covariate model. Empirical coverage rates were acceptable for correctly specified WGEE and MIGEE as well as for DRGEE when at least one of predictive models are correctly specified.

Simulation results also indicated that the bias of doubly robust estimators when both the covariate model and the missing data model are incorrect was the same magnitude of misspecified WGEE or MIGEE.
Monte Carlo simulation for samples sizes 50, 150 and 600 (not shown) reached similar conclusions.

5 Data analysis: analgesia in childbirth

This study was conducted in Minas Gerais state, Brazil, in order to compare two techniques of analgesia (standard epidural analgesia and continuous intravenous infusion of remifentanil) for labor pain of 49 patients. The response of interest is the intensity of pain (1: mild, 2: moderate, and 3: intense), assessed at 0, 60, and 90 minutes. Predictor variables were GROUP (0: peridural; 1: remifentanil), AGE (in years), DU (uterine dilatation), and OXYT (consumption of oxytocin, ordinal with three levels). PAIN and OXIT were missing for 9 patients at time 60 and for 18 patients at time 90.
For the ordinal response it was used the following proportional odds model

$$\operatorname{logit}\Pr(PAIN_{itj} \leq j|\boldsymbol{u_{it}}) = \beta_{0j} + \boldsymbol{u}_{it}^{\mathrm{T}}\boldsymbol{\beta}, \qquad j = 1, 2, \quad t = 1, 2, 3, \tag{9}$$

where $\boldsymbol{u}_{it}$ is formed by TIME, GROUP, AGE, DU and OXIT.
For the missing data process R_{it} was defined as the indicator of observing both $PAIN_{it}$ and $OXIT_{it}$, and take the following form

$$\log\left(\frac{\Pr(R_{it}=1)}{\Pr(R_{it}=0)}\right) = \psi_{0t} + \boldsymbol{w}_{it}^{\mathrm{T}}\boldsymbol{\psi}, \qquad t = 2, 3, \tag{10}$$

where $\boldsymbol{w}_{it}$ includes GROUP, AGE, DU, histories of OXYT and PAIN.
It was assumed the following model for the missing covariate OXYT

$$\operatorname{logit}\Pr(OXYT_{itj} \leq j|\boldsymbol{v_{it}}) = \gamma_{0j} + \boldsymbol{v}_{it}^{\mathrm{T}}\boldsymbol{\gamma}, \qquad j = 1, 2, \quad t = 2, 3, \tag{11}$$

where $\boldsymbol{v}_{it}$ includes main effects for GROUP, AGE, and DU.
Results are shown in Table 2. The chained equations approach was used for MIGEE with the number of multiple imputation set to 10. WGEE was fitted using model (10) for the weights. DRGEE was applied considering (10) and (11) as the models for the weights and the missing covariate, respectively. It was used an independent working correlation.
TIME effect is non significant for all methods. All methods provide the same conclusion for effects of GROUP, DU and OXYT. However, it can be noticed that p-value for AGE effect goes from a non-significant of 0.06 in the standard GEE to a significant one in DRGEE, as well as for the other two missing data approaches. The conclusion is that older women have lower chance of experiencing mild pain than young women.

TABLE 2. Regression parameters for the analgesia in birth data.

	Available			WGEE			MIGEE			DRGEE		
Parameter	Est	SE	P	Est	SE	P	Est	SE	P	Est	SE	P
INTER1	1.80	1.308	0.17	1.83	1.267	0.15	1.60	1.237	0.20	1.65	1.315	0.21
INTER	3.30	1.314	0.01	3.26	1.310	0.01	3.11	1.242	0.01	3.07	1.351	0.02
TIME	-0.18	0.286	0.52	-0.16	0.315	0.62	-0.11	0.273	0.69	-0.19	0.293	0.51
GROUP	-1.22	0.445	0.01	-1.24	0.439	0.01	-1.06	0.409	0.01	-1.01	0.390	0.01
AGE	-0.07	0.035	0.06	-0.07	0.033	0.03	-0.06	0.031	0.04	-0.07	0.035	0.04
DU	-0.36	0.158	0.02	-0.35	0.168	0.04	-0.37	0.156	0.02	-0.34	0.166	0.04
OXYT(=2)	0.73	0.554	0.19	0.71	0.554	0.20	0.95	0.542	0.08	0.82	0.531	0.12
OXYT(=3)	1.26	0.510	0.01	1.43	0.503	0.00	1.41	0.488	0.00	1.47	0.475	0.00

6 Conclusion

When longitudinal ordinal data are of interest, the DR estimator is a nice alternative. It is attractive in the sense that it needs only the correct specification of at least one of the models, but not necessarily both. Simulation results have indicated that, when at least the covariate model or missing data model is correct, the doubly robust estimators are consistent and present small-sample bias comparable to single robust alternatives MIGEE or WGEE. The proposed method presented good coverage probabilities, as well as its competitors but with a slight larger variance than multiple imputation. We hope that, in practical applications, the proposed estimator has a great potential of reducing the bias if the MAR assumption is correct.

References

Chen, B. and Zhou, X.H. (2011). Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates. *Biometrics*, **67**, 830 – 842.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13 – 22.

Lipsitz, S.R., Kim, K., and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, **13**, 1149 – 1163.

Little, R.J. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley New York.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, **42**, 109 – 142.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106 – 121.

Statistical models for dynamics in extreme value processes

Bernhard Spangl[1], Sascha Desmettre[2], Peter Ruckdeschel[3]

[1] Institute of Applied Statistics and Computing, BOKU, Vienna, Austria
[2] Department of Mathematics, TU Kaiserslautern, Germany
[3] Institute for Mathematics, University Oldenburg, Germany

E-mail for correspondence: bernhard.spangl@boku.ac.at

Abstract: We study four different approaches to model time-dependent extremal behavior: dynamics introduced by (a) a state-space model (SSM), (b) a shot-noise-type process with GPD marginals, (c) a copula-based autoregressive model with GPD marginals, and (d) a GLM with GPD marginals (and previous extremal events as regressors). Each of the models is fit against data, and from the fitted data, we simulate corresponding paths according to the respective fitted models. At this simulated data, the respective dependence structure is analyzed in copula plots and judged against its capacity to fit the corresponding inter-arrival distribution.

Keywords: Time-dependence; State-space model; GLM; Shot-noise-type process; Extreme values.

1 Motivation and issues

A challenge in dealing with extreme events in river discharge data is to capture well time dynamics of these extremes, in particular in the presence of seasonal effects and trends.
We will provide models which are able to capture the extreme behaviour and provide simple and parsimonious, but yet flexible dynamics.
The goal is to be able to assess the magnitude of extreme events as well as the inter-arrival time distribution of those extremes by simulation (see also Khaliq et al., 2006).

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 Data generating processes

To address these issues, we discuss four different approaches to incorporate dynamics into extreme value processes. For different choices of parameters, each of these approaches is illustrated by typical realizations to assess the induced dynamics and by lagged PP plots to grasp the respective dependence structures.

2.1 SSM approach

As a first approach we propose to use the following model:

$$X_t = \mu_t + \sigma_t v_t,$$

with

$$\begin{aligned} \mu_t &= \mu_0 + \varphi X_{t-1}, \\ \log \sigma_t &= \log \sigma_0 + \gamma \log(1 + (X_{t-1} - \mu_{t-1})^2). \end{aligned}$$

This model is comparable to an AR-EARCH one and can be easily extended to an ARMA-EGARCH model or other even more complex ones.
The iid innovations v_t are generated according to the following two-step procedure. First, we choose $v_t' \sim F_0$, where, e.g., $F_0 = \mathcal{N}(0,1)$. Then, given a certain threshold τ the innovations are enriched by $v_t'' \sim$ GPD, i.e.,

$$v_t = \begin{cases} v_t' & : \ |v_t'| < \tau \ , \\ v_t'' \sim \mathrm{GPD}(\mu = 0, \sigma = 1, \xi) & : \ \text{else.} \end{cases}$$

ξ is chosen according to the maximum domain of attraction of F_0.
This state-space model approach allows for separate estimation of time-dependency (via filtering) and of GPD-parameters, μ, σ, and ξ, of the marginal distribution of the innovations. Typical realizations are shown in Figure 1, together with the corresponding lagged PP plots.

2.2 Shot-noise-type approach

As a second approach we suggest to use the following shot-noise-type process:

$$X_t = U_t Y_t + (1 - U_t) v_t,$$

where $U_t \sim \mathrm{Bin}(0, \beta)$ and $v_t \sim G = \mathrm{GPD}(\mu, \sigma, \xi)$. Let G and G^{-1} also denote the corresponding cdf and quantile function, respectively. Moreover, $\{U_t\}_t$ and $\{v_t\}_t$ are stochastically independent, and (U_t, v_t) is independent of $\{X_s\}_{s<t}$. Then $Y_t = f_\beta(X_{t-1})$ with

$$\begin{aligned} f_\beta &= G^{-1} \circ f_\beta^0 \circ G, \\ f_\beta^0 &= u/[(1-u)\beta + u]. \end{aligned}$$

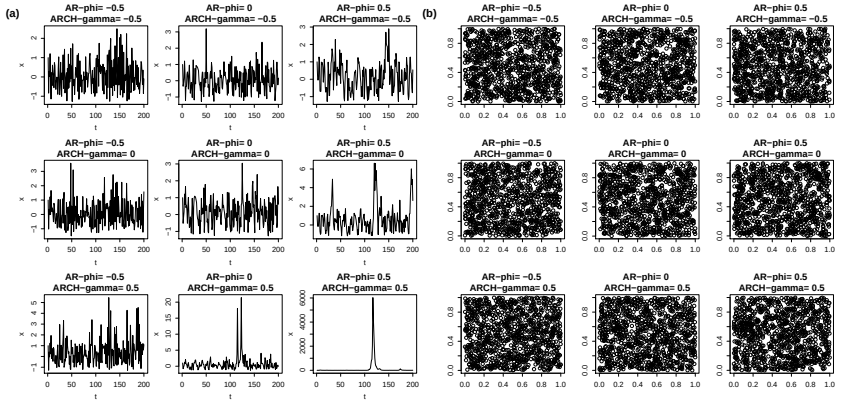

FIGURE 1. SSM approach: (a) typical realizations, and (b) lagged PP plots.

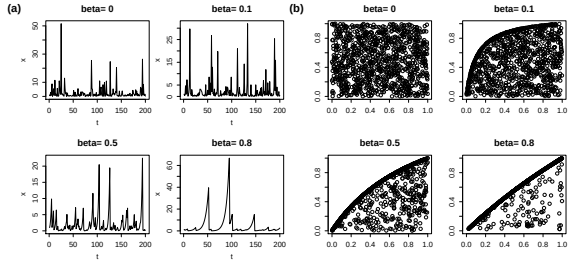

FIGURE 2. Shot-noise-type approach: (a) typical realizations, and (b) lagged PP plots.

Hence, $X_t \sim G$. The parameter $\beta \in [0, 1]$ controls the dependency. Again, this approach allows for separate estimation of time-dependency (number of dropdowns) and of GPD-parameters, μ, σ, ξ, of the marginal distribution. Figure 2 shows typical realizations of the shot-noise-type (SNT) approach, together with the corresponding lagged PP plots. We refer to Desmettre et al. (2015) for the definition and theoretical foundations of this process as well as its application in liquidity risk management.

2.3 Gaussian-copula approach

Next, we propose to use a copula-based autoregressive model defined by

$$X_t = G^{-1} \circ F \circ Y_t,$$

with

$$Y_t = \rho Y_{t-1} + \bar{\rho} v_t,$$

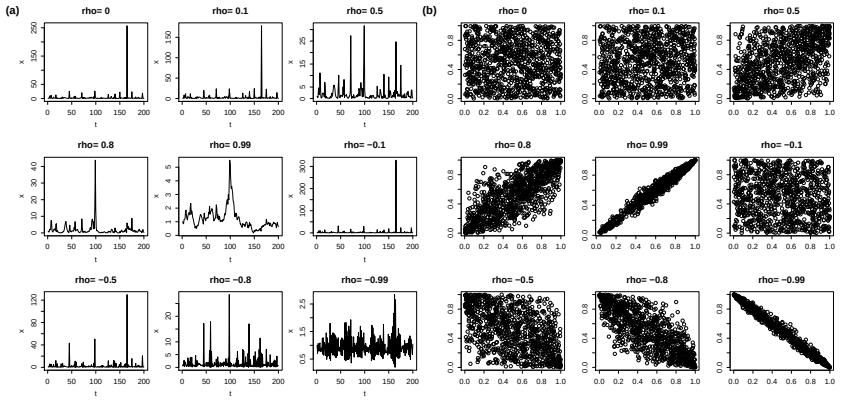

FIGURE 3. Gaussian-copula approach: (a) typical realizations, and (b) lagged PP plots.

where $\bar{\rho}^2 + \rho^2 = 1$, and $v_t \overset{\text{iid}}{\sim} F = \mathcal{N}(0,1)$. Again, let G and G^{-1} denote the cdf and quantile function of the GPD(μ, σ, ξ). Hence, $X_t \sim G$, and the parameter $\rho \in [-1, 1]$ controls the dependency. In the same way we may transform a more complex model, e.g., a Gaussian ARMA-model to a GPD-process. This again separates the estimation of time-dependency (Gaussian ARMA) and of GPD-parameters, μ, σ, ξ, of the marginal distribution.
In Figure 3 we see typical realizations of the Gaussian-copula (GC) approach, together with the corresponding lagged PP plots.

2.4 GLM approach

Last, we suggest a parameter driven approach. We define our GPD-process by

$$X_t \sim \text{GPD}(\mu, \sigma_t, \xi_t).$$

The scale parameter σ_t is given by

$$\sigma_t = \ell_1(\beta_1 + \beta_2 h_1(X_{t-1})),$$

with ℓ_1 and h_1 properly chosen, e.g., $\ell_1 \colon \mathbb{R} \to \mathbb{R}_+$, and $h_1(x) = x^2$.
The shape parameter ξ_t is given by

$$\xi_t = \ell_2(\gamma_1 + \gamma_2 h_2(X_{t-1})),$$

with ℓ_2 and h_2 again properly chosen, e.g., $\ell_2 \colon \mathbb{R} \to (-0.5, 2.5)$, and $h_2(x) = \log(1 + |x|)$. h_2 controls the tails. Moreover, we note that it is essentially to chose the codomain of ℓ_2 equal to $(-0.5, 2.5)$.
Typical realizations of the GLM approach are plotted in Figure 4. Here, lagged PP plots do not make sense as the dependence structure is modeled by fitting the GPD-parameters by GLMs where previous observations enter as regressors. See Pupashenko et al. (2014) for theoretical foundations.

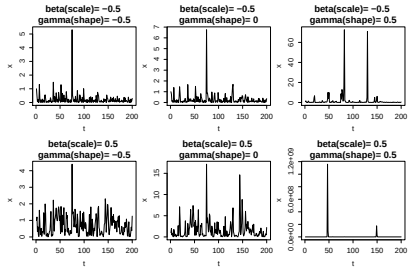

FIGURE 4. GLM approach: typical realizations.

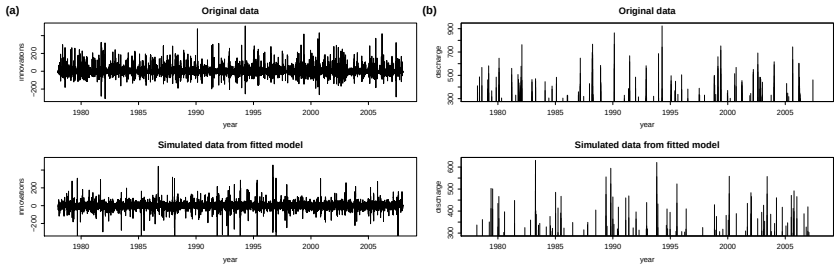

FIGURE 5. Real data vs. simulation: (a) SSM approach, and (b) shot-noise-type approach.

3 Evaluation and real world data set

To evaluate our four approaches we use daily average discharge data of the Danube river at Donauwörth from 1978 to 2008.
We obtain corresponding model fits and simulated data as visible in the following pictures.
In Figures 5 and 6 the upper panels always show the original data and the lower panels one simulated path.
Figure 5a shows the innovations of the original series after filtering and the simulated innovations using the SSM approach. Here, instead of using an EARCH model as proposed in Section 2.1, we use a GARCH(1,1) model with t-distributed innovations. Comparing the two series in Figure 5a we see that in the lower panel the volatility clusters are reproduced quite well. However, negative innovations are over-represented in the simulated series.
In Figure 5b only the extreme events above a threshold of 300 of the original data as well as the simulated ones using the shot-noise-type approach are plotted. Comparing the two plots in Figure 5b we note that the distribution of extreme events, i.e., the inter-arrival time distribution of extremes, coincide well, whereas the level of discharge is slightly under-estimated by the simulated process.

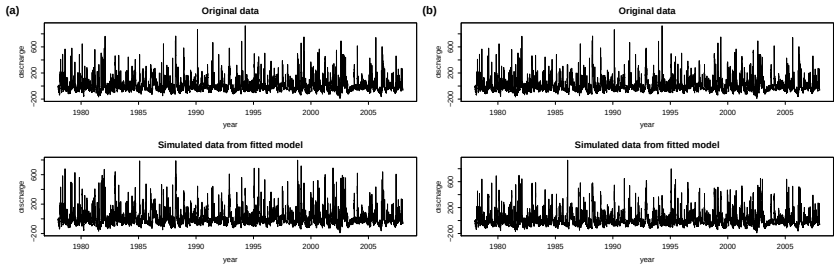

FIGURE 6. Real data vs. simulation: (a) Gaussian-copula approach, and (b) GLM approach.

In the upper panels of Figure 6 the original detrended and deseasonalized series is displayed. The lower panel of Figure 6a shows one simulated path using the Gaussian-copula approach whereas in Figure 6b we see a simulated path using the GLM approach. Comparing the lower panels of Figures 6a and 6b with the upper ones we see that the level of discharge is estimated very well by the Gaussian-copula and the GLM approach. However, we note that using these two approaches we are not able to assess the inter-arrival times.

4 Summary and conclusion

We presented four flexible and parametric approaches to model dynamics as well as extremes. We remark that all models proposed are parsimonious ones. Moreover, we are able to get grip on the inter-arrival time distribution of extremes. The SSM and GLM approaches are able to model complex dynamics. The SNT as well as the SSM approach are able to assess the inter-arrival time distribution of the extreme events well. The GC approach can easily be extended to capture multivariate dependency.

References

Desmettre, S., de Kock, J., Ruckdeschel, P., and Seifried, F.T. (2015). Generalized Pareto Processes and Liquidity. *Working Paper*.

Khaliq, M.N., Ouarda, T.B.M.J., Ondo, J.-C., Gachon, P., and Bobee, B. (2006). Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review. *Journal of Hydrology*, **329**, 534 – 552.

Pupashenko, D., Ruckdeschel, P., and Kohl, M. (2015). L_2 differentiability of generalized linear models. *Statistics and Probability Letters*, **97**: 155 – 164.

Spatio-temporal censored model of precipitation climatology

Reto Stauffer[1], Nikolaus Umlauf[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: reto.stauffer@uibk.ac.at

Abstract: Flexible spatial-statistical models are widely used to create climatological estimates. Although most models assume a normally distributed response this assumption can lead to inaccurate estimates for certain variables. Precipitation, for instance, is physically limited to values ≥ 0 so that it might be seen as left-censored. We develop a novel spatial-statistical additive model for location, scale, and shape which can handle censored normal distributed responses. This article presents a precipitation climatology over complex terrain with a daily temporal resolution on a 0.5 x 0.5 km grid. The results demonstrate that the new method outperforms existing methods and is able to resolve local effects quite well compared to single-station estimates. Our model enables the creation of climatologies with a high spatio-temporal resolution, yet does not need extensive tuning. Overall, the model is easily adaptable to new applications.

Keywords: Spatial climatology; Spatial modelling; Precipitation; Censoring.

1 Introduction

The northern part of Tyrol in the Eastern Alps is reaching from $465m$ up to $3798m$, the highest peak in Austria. Due to the complex topography and the location within the Alps, Tyrol is strongly related to precipitation. During the winter the local economy, tourists, transportation, and the local population is strongly related to snow conditions while during the summer season extreme precipitation amounts can lead to floods or droughts.
Using spatial-statistical models to create climatologies is not new. However, existing methods are often based on monthly or even yearly means or sums. Plausible reason: there is no need to handle zero-observation events and the

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

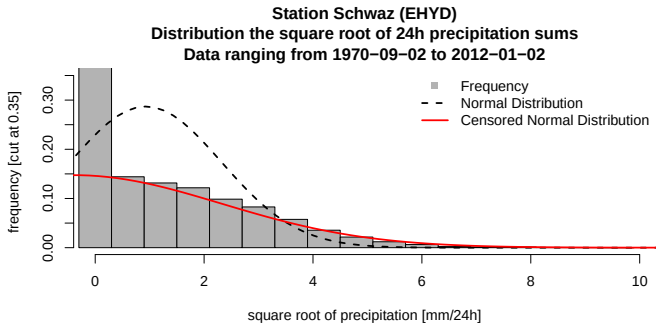

FIGURE 1. Sample distribution of 24h precipitation sums at station Schwaz (535 *m.a.s.l*, Tyrol; observations square-root transformed) containing 14975 observations. The lines show the PDFs for a normal distribution (black, dashed), and for a normal distribution left-censored at 0 (red, solid). Ordinate cut at 0.35. Class $[-0.3, 0.3]$ has a frequency of 0.95.

assumption of a normally distributed responses is often apposite. On a daily base zero-observations occur very frequently. We developed an extended additive model for location, scale, and shape for censored normally distributed response. The field of possible applications is reaching from climatological trend analysis, torrent control, or local spatial planning, to the use as priori information for extreme-value modelling or weather forecasting.

2 Skewness of observation distribution

Precipitation observations are typically skewed distributed. A common approach to remove major parts of the skewness is to transform the observed values as proposed by Box and Cox (1964). For precipitation, a square root transformation is commonly used (Hutchinson, 1998b). Furthermore the data show a strong point mass on 0 due to the fact that about two third of all days are "dry" without precipitation. As precipitation can only be ≥ 0, the distribution can be seen as left-censored at 0 (Cohen, 1959).
Figure 1 shows the distribution of observed square root transformed daily precipitation sums for station Schwaz. Additionally, two probability distribution functions (PDFs) are shown. The first one shows a normal distribution with in-sample mean and standard deviation, the second shows a left-censored normal distribution with its censoring point at 0. The latter one is obviously more appropriate for the given sample distribution.

3 Censored spatial model

For the spatial modelling we are using a new R package called `bamlss` (Umlauf, 2015), a package for Bayesian Additive Models for Location, Shape

and Scale. The model can be written as:

$$y^* \sim N(\mu, \sigma^2),$$

where y^* is the latent response, η is the location, and σ the scale parameter of the latent unobservable process. Location and log-scale are expressed by an additive combination of explanatory variables. In the "simple" example shown in Figure 2 & 3 these are: `yday` (day of the year), `alt` (altitude), and `lon/lat` (longitude/latitude).

$$\eta_\bullet = \varphi_0 + g(\text{yday}) + g(\text{alt}) + g(\text{lon}, \text{lat}),$$

where $\varphi_0 = \beta_0$ for η_{mu}, $\varphi_0 = \gamma_0$ for η_σ, and g are non-linear functions determined by coefficients β and γ respectively

$$\mu = \eta_\mu, \qquad \log(\sigma) = \eta_\sigma, \qquad \mathbf{y} = \max(\mathbf{0}, \mathbf{y}^*).$$

Given the response y, and the explanatory variables x, the unknown parameters for η_μ (β) and for η_σ (γ) can be found by maximum likelihood:

$$L(\beta, \gamma | y, x) = \prod_{i=1}^{N} f(y_i | x_i, \beta, \gamma)^{I(y_i > 0)} \cdot F(0 | x_i, \beta, \gamma)^{I(y_i = 0)},$$

where f and F are the probability density function and the cumulative distribution function of the normal distribution respectively.
Preliminary results can be found in Figure 2 & 3 showing the effects included in the model formulas $\eta_\bullet$. The 110 stations used for the model are reaching up to $2290m$ but, about 90 % are located below $1600m$ which leads to an increasing log-scale with higher altitude as shown in Figure 2.

4 Model comparison and intermediate findings

Figure 4 shows a comparison of the seasonal effects between single-station estimates (colorized) and the mean seasonal effect on the full spatial model (black). The seasonal effect strongly varies between the south and the north side of the Alps. The spatio-temporal model captures the characteristic of the major proportions, however, the south side stations (reddish colors) should have a different effect. To take this into account an additional effect (e.g. spatially variable seasonal effects) should be added but has not been done yet.
To check the improvement of the new censored spatio-temporal estimates, we compared the results of the censored models (using bamlss) against some generalized additive models based on monthly mean observations (using mgcv). As the temporal resolution of these models differ (daily vs. monthly) two questions arise: (i) What is the skill of the new daily-based climatology

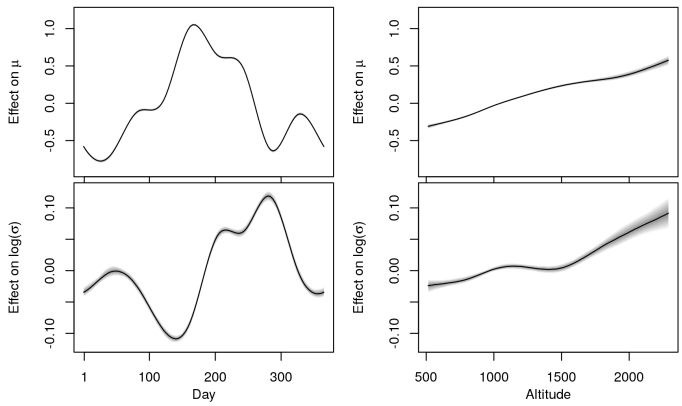

FIGURE 2. Centred effects for location (top row) and log-scale (bottom row) on square root transformed observations. Left/Right: effect on day of the year/altitude. Effects estimated on 30 years of data (1982-2012), 110 different stations. Highest precipitation amounts expected during summer period (μ). Log-scale seems to be highest in late summer/autumn (day 200–300; late July to September) which is related to the convective season.

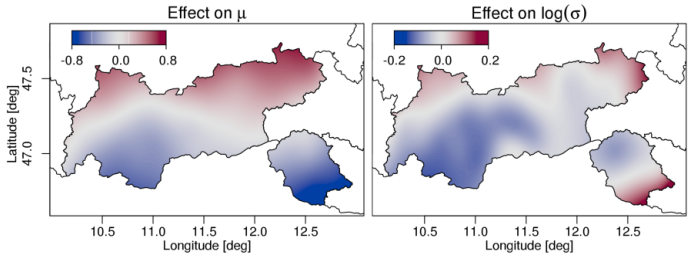

FIGURE 3. Centred spatial longitude/latitude effect over North Tyrol for location (left), and log-scale (right). Same data basis as Figure 2. The location effect shows the increasing amounts of precipitation towards the northern parts (Alpine ridge). Lowest in dry inner Alpine valleys. The log-scale effect seems to be strongly related to the station density (boundary-effect).

when aggregated to monthly means vs. a model directly trained on monthly means? (ii) How is the skill on a daily basis?

The in-sample validation of four selected models is shown in Figure 5. Skill scores of mean absolute and root mean squared errors are shown. Model "mon-A/cens-A" and "mon-B/cens-B" are comparable as they have identical explanatory effects. The setup of model "A" is shown in Chapter 3, setup "B" contains additional topographic regressors.

As shown our new model is comparable to the one estimated on monthly mean values when aggregated to a monthly resolution (Figure 5). On a

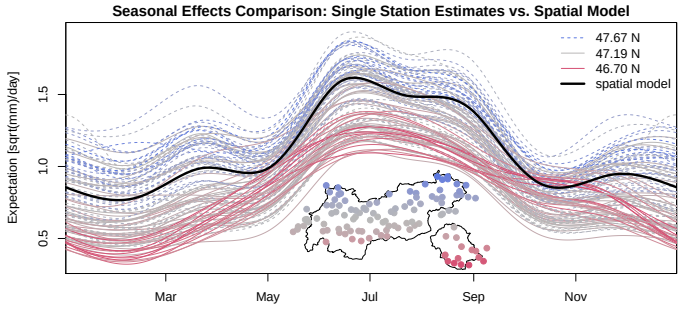

FIGURE 4. Each line plotted shows the climatology for one of the stations estimated via a single-station based censored bamlss model. The colors correspond to the latitudinal location of the station (see subplot; overview over North Tyrol). In black: mean effect from the spatio-temporal model shown in Figure 2 & 3.

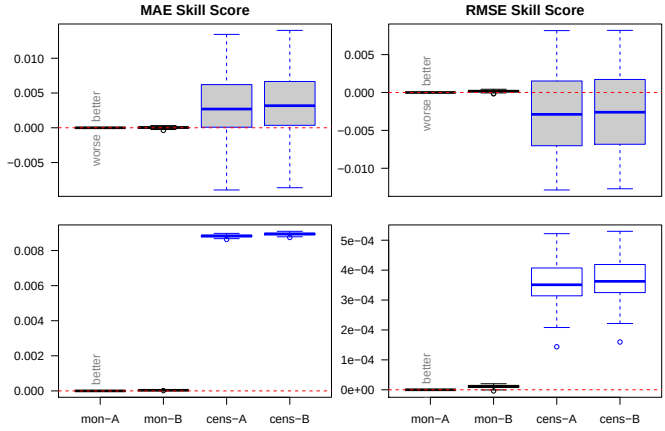

FIGURE 5. Skill scores of bootstrapped mean absolute (MAE) and root mean squared (RMSE) errors validated on monthly-mean base (top row), and on a daily basis (bottom row). Model "mon-A" chosen as reference model. Black: normal distributed models on monthly means (mgcv). Blue: censored normal distributed models on daily observations (bamlss). The unit of all data is in "square root of daily mean" ($\sqrt{millimeter}$ per day).

daily basis the new model has significantly lower errors (positive skill).

5 Conclusion and outlook

The presented method has some major advantages over existing ones. Treating precipitation as left censored makes it unnecessary to reduce the temporal resolution (e.g., take monthly means to avoid zero-observations). Furthermore additional daily-based explanatory variables can be included

to create e.g, wind-direction dependent climatological estimates. In comparison to commonly used tools for precipitation analysis, the new approach needs less priori knowledge and less extensive tuning. Furthermore, the new method outperforms existing ones even when aggregated to a monthly resolution.
As shown, not all effects are captured yet. Searching for additional variables has to be one of the next steps. Due to the size of the data set additional (multi-dimensional) terms are computationally expensive (memory & CPU). Therefore we are working on the code of the bamlss package to optimize the performance/resource management.

Acknowledgments: The ongoing project is supported by the Austrian Science Fund (FWF): TRP-290.

References

Box, G.E. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211 – 252.

Cohen, A.C.J. (1959). Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, **1**, 99 – 113.

Hutchinson, M.F. (1998a). Interpolation of rainfall data with thin plate smoothing splines – Part I. *Journal of Geographic Information and Decision Analysis*, **2**, 168 – 185.

Hutchinson, M.F. (1998b). Interpolation of rainfall data with thin plate smoothing splines – Part II. *Journal of Geographic Information and Decision Analysis*, **2**, 152 – 167.

Umlauf, N. (2015). A conceptional Lego toolbox for Bayesian distributional regression models. *30th International Workshop on Statistical Modelling.*.

P-spline quantile regression with a mixed model algorithm

Federico Torretta[1], Vito M.R. Muggeo[1], Paul H.C. Eilers[2]

[1] University of Palermo, Italy
[2] Wageningen University and Research Centre, The Netherlands

E-mail for correspondence: federico.torretta@unipa.it

Abstract: In this paper we propose an iterative algorithm to select the smoothing parameter in semiparametric quantile regression. The functional form of the covariate effects is unspecified and expressed via B-spline bases, and additional unpenalized linear terms are allowed within the regression equation. We present results from some simulations and real-data analysis.

Keywords: Quantile regression; P-spline; Schall algorithm; Smoothing parameter selection.

1 Introduction

Nonparametric quantile regression (NQR) aims to model covariate effects on the response quantiles without imposing any rigid and parametric relationship with covariates. However current implementations of nonparametric quantile regression, as implemented, for instance, in the R packages quantreg or cobs, are far from being efficient. In fact the traditional methods employed in literature rely on criteria working on a pre-specified grid of the smoothing parameter. Therefore one has to select several values for the smoothing parameter and to fit the model at each candidate value; the final model is selected according to the best value of the criterion, such as the usual cross-validation score, the Schwartz (or Bayesian) Information Criterium or even more recently the so-called L-curve (e.g., Andriyana et al., 2014). The computational burden becomes particularly expensive when the regression equation involves multiple additive components leading to a multidimensional grid of smoothing parameters.
To overcome the aforementioned shortcoming, this paper aims to set up an iterative algorithm for smoothing parameter selection in additive NQR

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

such as $Q_\tau = s_1(x_1)+s_2(x_2)+\cdots+z^{\mathrm{T}}\beta$ where the smooth functions $s(\cdot)$ are expressed via low rank B-splines, and penalties are set on the corresponding coefficients to avoid overfitting. Section 2 describes the proposed algorithm, Section 3 presents results from some simulations. An analysis of dataset concerning growth of the seagrass *P. oceanica* is included in Section 4.

2 Methods: the Schall algorithm for NQR

To begin with, we assume a single smooth term in the linear predictor and suppress dependence on the percentile τ for the sake of simplicity. Hence $s(x_i) = B(x_i)^{\mathrm{T}}\theta$ where $B(x_i) = (B_1(x_i), \ldots, B_J(x_i))^{\mathrm{T}}$ includes the J values of the basis functions at x_i, and $\mathcal{L}(\theta) = \sum_i \rho(y_i - B(x_i)^{\mathrm{T}}\theta)$ is the objective function for the quantile regression of interest. The penalized objective function to be minimized is $\mathcal{L}_\lambda(\theta) = \mathcal{L}(\theta) + \lambda||D_d\theta||_1$, where $\lambda||D_d\theta||_1 = \lambda\sum_{j=1}^{J-d}|\Delta^d\theta|_j$ is the L_1 norm of the dth order differences of coefficients, and λ controls the amount of smoothing of the fitted curve. As $\lambda \to \infty$ the resulting fitted quantile curve tends to a polynomial of degree $d-1$ with all dth order differences equal to zero; for relatively moderate values of λ, only some differences will be zero leading to a piecewise $(d-1)$th order polynomial fitting.
To avoid a grid search on a specified grid of λ values, we exploit the Schall algorithm which was previously proposed for the linear mixed models (Schall, 1991) and it has also been employed for expectile smoothing (Schnabel and Eilers, 2009). However in the NQR framework, or more generally in L_1-problems, it has never been discussed. The proposed approach is based on the following steps:

1. Fix a (small) value for the smoothing parameter $\lambda^{(0)}$;
2. Fit the NQR by minimizing the objective $\mathcal{L}(\theta) + \lambda^{(0)}||D_d\theta||_1$;
3. Compute $\hat{\sigma}^2_\theta = ||D_d\hat{\theta}||^2_2/edf$ and $\hat{\sigma}^2_e$;
4. Compute $\hat{\lambda} = \hat{\sigma}_e/\hat{\sigma}_\theta$;
5. Set $\hat{\lambda} \to \lambda^{(0)}$ and repeat steps 2. to 4. till convergence.

The layout of the algorithm appears straightforward, but there are some remarks to be emphasized, especially in step 3 and 4. First notice we consider the ratio of standard deviations (rather than variances) which is reasonable within a L_1 framework. Also, the model *edf*s are computed via the trace of the hat matrix derived from a parametric smooth approximation of the objective $\mathcal{L}(\theta)$ (Muggeo et al., 2012), that allows to partition the total model *edf* among the multiple smooth terms included in the model. Finally to estimate the error variance σ^2_e, some approaches could be followed, including estimates based on the sum of the asymmetrically weighted squared or

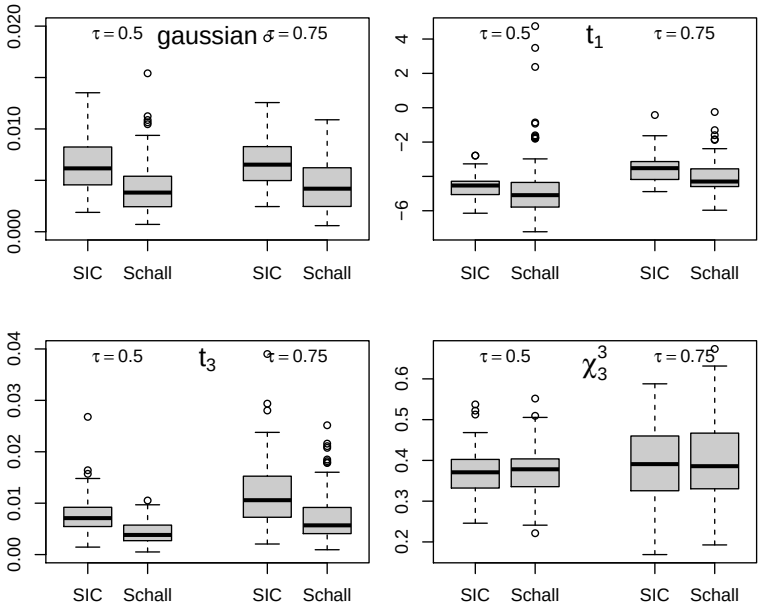

FIGURE 1. Comparing the proposed approach (Schall) versus the smoothing spline via SIC based on rqss() in quantreg package (SIC): Root Mean Integrated Square Error for different scenarios. In the panel of t_1, values are on log scale.

absolute residuals. Yet another approach, which simulations have shown to perform somewhat better, relies on estimating the variance via the asymmetric Laplace likelihood, $\hat{\sigma}_e^2 = \hat{\psi}^2 \frac{(1-2\tau+2\tau^2)}{(1-\tau^2)\tau^2}$, where $\hat{\psi}$ is an estimate of the scale parameter of the asymmetric Laplace distribution.
The aforementioned outline refers to a single smooth term, however the computational cost is unchanged if multiple additive smooth terms are included: the term-specific *edf* are computed by summing the relevant elements of the main diagonal of the pseudo hat matrix appropriately partitioned. Thus the term-specific variances and smoothing parameters can be obtained according to the ratios of step 3 and step 4 respectively. This makes Schall algorithm very attractive in the multidimensional case where the multidimensional grid search gets substantially unfeasible.

3 Simulation results

In this section we present results from a limited simulation experiment comparing the performance of two smoothers within NQR models. More specifically we compare smoothing splines with a total variation penalty and λ selected via SIC (i.e. in R `AIC(rqss(y~qss(x,lmbd),tau),k=-1)`)

versus the P-spline smoother with third order ($d = 3$) difference penalty and λ selected via the Schall algorithm using a Laplace-based error variance estimate. Data have been generated by $y_i = g_0(x_i) + 0.2e_i$, where the signal is $g_0(x_i) = \sin(2\pi x_i)$, $x_i \sim U(0,1)$, and the *iid* e_is come from four different distributions: Gaussian, t_1,t_3 or a centred χ^2_3, namely $\chi^2_3 - 3$. The sample size is $n = 100$, and two percentiles $\tau = 0.50$ and $\tau = 0.75$ have been considered. Figure 1 shows the RMISE (root mean integrated square error) over the replicates. P-splines with selection of λ via the Schall algorithm show lower RMISE than that of smoothing splines along with the SIC, with no substantial difference for asymmetric errors. Hence the lighter computational burden appears to go with a better statistical efficiency, although additional scenarios, for instance different signals or more extreme quantiles, should be considered.

4 Additive NQR in practice

We apply the proposed framework to data of seagrass *P. oceanica* in Mediterranean sea. Interest lies in modelling the 0.9-quantile of annual length as a function of age and depth. Thus the additive quantile regression model is $Q_{0.9}(\texttt{lenght}) = s_1(\texttt{age}) + s_2(\texttt{depth})$. We employ two B-spline bases with a $d = 3$ difference order penalty and the aforementioned algorithm to select the two smoothing parameters. Convergence is attained in less than 10 iterations. Figure 2 portrays the fitted smoothed relationships with estimated degrees of freedom $edf_{\texttt{age}} = 4.85$ and $edf_{\texttt{depth}} = 7.47$ computed via the trace of the hat matrix coming from the parametric smooth approximation of Muggeo et al. (2012).

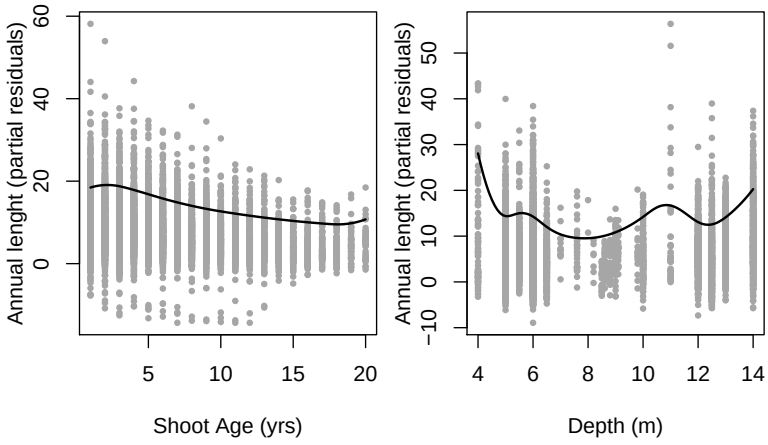

FIGURE 2. The fitted smooth effects of Age and Depth for the 90th quantile curve of Annual length in P. oceanica seagrass.

5 Conclusion

In this paper we have proposed an iterative algorithm to 'estimate' the smoothing parameter in P-spline NQR models. The methodology has been discussed for additive terms, but extension to varying coefficients or even random effects for longitudinal data appears to be straightforward.
In terms of computational burden, the method is clearly less demanding than grid-search, especially in additive NQR having multiple smooth terms. In terms of statistical performance, results from the simulation experiment suggest that the proposed approach performs equal or even better than SIC, especially with Gaussian errors. It also should be emphasized that the computation of the degrees of freedom in our framework represents a valid alternative to the common Koenker's method based on the number of non-zero residuals in the model. However this point deserves further investigation along with some theoretical issues: for instance, unlike mean regressions, it is not clear what is the exact objective being optimized via the Schall algorithm in the NQR framework.

References

Andriyana, Y., Gijbels, I., and Verhasselt, A. (2014). P-spline quantile regression in varying coefficient models *Test*, **23**, 153 – 194.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91**, 74 – 89.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press.

Muggeo, V.M.R., Sciandra, M., and Augugliaro, L. (2012). Quantile regression via iterative least squares computations. *Journal of Statistical Computation and Simulation*, **82**, 1557 – 1569.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719 – 727.

Schnabel, S. and Eilers, P.H.C. (2009). Optimal expectile smoothing. *Computational Statistics and Data Analysis*, **53**, 4168 – 4177.

Modeling creditworthiness using a generalized linear mixed-effects approach

Laura Vana[1], Bettina Grün[2], Kurt Hornik[1]

[1] WU Wirtschaftsuniversität Wien, Department of Finance, Accounting and Statistics, Austria
[2] Johannes Kepler University Linz, Department of Applied Statistics, Austria

E-mail for correspondence: laura.vana@wu.ac.at

Abstract: Based on a multivariate generalized linear mixed-effects approach, we aim at modeling creditworthiness using key financial ratios from an extensive list of potential variables that drive a firm's financial health as covariates and ratings and default information as dependent variables. The model allows to estimate the latent creditworthiness as well as rater-specific biases. While the origin of bias in credit ratings has appeared in the literature under different assumptions, the proposed model allows to empirically quantify the rater-specific deviations from a firm's latent creditworthiness and analyze to which financial ratio the bias might be attributable. To assess if including variable selection based on a spike-and-slab prior and explicitly including financial ratios to model rater bias improves the model, the out-of-sample predictive performance of these models is compared.

Keywords: Creditworthiness; Spike-and-slab prior; Variable selection; Rater bias.

1 Introduction

Despite the criticism for failing to assess risk accurately and for the lack of transparency in their rating methodology, credit rating agencies (CRAs) maintain a powerful role as financial market participants and have a huge impact on the cost of funding. We propose a multivariate generalized linear mixed-effects model that incorporates binary and continuous responses in order to predict creditworthiness based on financial ratios. The model allows to separate latent creditworthiness and rater-specific biases.
The topic of credit rating bias has been incorporated in various theoretical models, i.e., Vasiliki and Veldkamp (2009), and the proposed framework also enables the explicit modeling of rater biases for each CRA. Modeling

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

the bias as a constant permits the identification of a general over-pessimistic or over-optimistic behavior in the rating procedure of the CRAs, while modeling the bias as a linear combination of financial ratios shows which weights the CRAs assign to the different ratios when determining a firm's financial health.
Several different models are estimated where some make use of Bayesian variable selection with a spike-and-slab prior (Mitchell and Beauchamp, 1988) to account for model uncertainty and the potential redundancy of some financial ratios. The out-of-sample predictive performance of the models is compared using 10-fold cross-validation.

2 Methods

2.1 Model

Suppose that for each of N financial statements i we have a corresponding vector of K financial ratios $\boldsymbol{x}_i$, a default indicator D_i and probability of default (PD) estimates $[PD_{ij}]_{j\in J_i}$ from a non-empty subset J_i of all available raters. A characteristic of having a panel of ratings from different sources is that the data is available only in an unbalanced design, i.e., ratings are not available from all raters for each financial statement. Note that the proposed model explicitly accounts for this structure and J_i does not need to contain all raters. Let the random variable PD_i be the true PD corresponding to financial statement i and $S_i = \Phi^{-1}(PD_i)$ and $S_{ij} = \Phi^{-1}(PD_{ij})$ be the true PD scores and the rating scores, respectively, where Φ^{-1} is the inverse of the normal cumulative distribution function.
We assume that:

$$\begin{aligned} p(D_i = 1|S_i) &= \Phi(S_i), & y_{ij} &= S_i + \epsilon_{ij}, \qquad j \in J_i \\ S_i &\sim N(\nu_i, \tau^2), & \epsilon_{ij} &\sim N(\mu_{ij}, \sigma_j^2), \end{aligned}$$

where the observations are the default indicators D_i and the rating scores $\boldsymbol{y}_i = [y_{ij}]_{j\in J_i}$. Moreover, we take ν_i and μ_{ij} as linear combinations of the $\boldsymbol{x}_i$: $\nu_i = \alpha_0 + \boldsymbol{x}_i^\top \boldsymbol{\alpha}$ and $\mu_{ij} = \gamma_{j0} + \boldsymbol{x}_i^\top \boldsymbol{\gamma}_j$. The model with a constant bias, i.e., $\boldsymbol{\gamma}_j = (0, \ldots, 0)$, will be referred to as the reduced model.

2.2 Priors

In order to perform Bayesian variable selection among the ratios that are assumed to affect both the default variable D_i and rating scores $\boldsymbol{y}_i$, as well as ratios that may cause rater biases, we use a spike-and-slab prior on the $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}_j$ coefficients, which consists of a Dirac distribution as the spike and a normal distribution as slab.
Let $\boldsymbol{\delta}^{(\boldsymbol{\alpha})} = \left(\delta_1^{(\boldsymbol{\alpha})}, \ldots, \delta_K^{(\boldsymbol{\alpha})}\right)$ and $\boldsymbol{\delta}^{(\boldsymbol{\gamma}_j)} = \left(\delta_1^{(\boldsymbol{\gamma}_j)}, \ldots, \delta_K^{(\boldsymbol{\gamma}_j)}\right)$ be indicator variables with $\delta_k^{(\cdot)} = 1$ if the corresponding coefficient is allocated to the

slab component. The indicator variables can be drawn in a hierarchical way:

$$p\left(\delta_k^{(\cdot)} = 1 | \omega_k^{(\cdot)}\right) = \omega_k^{(\cdot)}, \qquad \omega_k^{(\cdot)} \sim \mathcal{B}e(a_{\omega^{(\cdot)}}, b_{\omega^{(\cdot)}}).$$

For the slab component we use Zellner's g-prior, with a hyper-prior on the shrinkage factor, $\frac{g}{g+1} \sim \mathcal{B}e(1, 3/2-1)$. We choose the conjugate prior, i.e., a gamma distribution, for the precision parameters σ_j^{-2} and τ^{-2} with hyper-parameters equal to 0.001 and an uninformative normal prior centered at zero for the intercepts α_0 and γ_{j0}. The prior inclusion probabilies $\omega_k^{(\alpha)}$ are drawn hierarchically from $\mathcal{B}e(1,1)$, while no hyper-prior is used for $\boldsymbol{\gamma}_j$ where it is a priori assumed that one covariate is included.

2.3 Predictive performance

The out-of-sample performance in predicting the rating scores is measured by the mean squared prediction error (MSPE) of each rater j:

$$\widehat{\text{MSPE}}_j = \frac{1}{N} \sum_{i=1}^{N} \left(y_{ij}^{\text{new}} - \frac{1}{M} \sum_{m=1}^{M} \hat{y}_{ij}^{(m)} \right)^2$$

where M is the number of Markov Chain Monte Carlo (MCMC) draws and $\hat{y}_{ij}^{(m)} = \boldsymbol{x}_i^\top \boldsymbol{\alpha}^{(m)} + \mu_{ij}^{(m)}$.

3 Data and results

We build a data set containing 70 financial variables from Compustat® for more than 2300 corporates in the US over the period 2003–2013. Issuer credit ratings from Standard & Poor's (S&P), Moody's and Fitch were obtained from Compustat® and from the CRAs themselves, respectively. The binary default indicator is built from the ratings data. The sample contains 14412 observations and 276 defaults. We map the ordinal credit ratings to PD estimates by using the empirical default rates for each rating class published annually by the CRAs.

Model estimates are obtained using MCMC methods with the software JAGS (Just Another Gibbs Sampler; Plummer 2003) with $M = 10000$ iterations, 1000 burn-in iterations and a thinning interval of 20. A visual inspection of the trace plots indicates the convergence of the chains.

We use 10-fold cross-validation and compare the out-of-sample performance in predicting the rating scores of four models: M_1 the reduced model without spike component, M_2 the reduced model with spike component on the $\boldsymbol{\alpha}$ coefficients, M_3 the model with $\mu_{ij} = \gamma_{j0} + \boldsymbol{x}_i^\top \boldsymbol{\gamma}_j$ without spike component, and M_4 the model with the spike-and-slab prior on both the $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}_j$ coefficients.

TABLE 1. Estimated rater MSPEs for the four models.

	M_1		M_2		M_3		M_4	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
S&P	0.295	0.008	0.296	0.009	0.294	0.008	0.294	0.009
Moody's	0.316	0.022	0.316	0.022	0.311	0.021	0.318	0.018
Fitch	0.771	0.064	0.771	0.064	0.770	0.064	0.757	0.063

Preliminary results show that M_4 contains on average 32 non-zero $\boldsymbol{\alpha}$ coefficients and 23 financial ratios have a posterior inclusion probability of 100%. A sensitivity analysis reveals that the choice of $a_{\omega(\boldsymbol{\alpha})}$ and $b_{\omega(\boldsymbol{\alpha})}$ does not have a significant effect on the posterior estimates. Regarding the model for the rater bias we observe that the posterior model size for Moody's is higher than for the other two CRAs with ratio equity/assets being systematically underestimated. This is not surprising as Moody's rating methodology differs slightly from the others as they rate companies according to the recovery rate in case of default, while the other two CRAs' most important metric is the probability of default.
The results for the predictive performance evaluation are given in Table 1. On a rater level, S&P has lowest MSPE while Fitch has the highest. Note that for Fitch the fewest ratings are available. The differences between the models are small. This suggests that the reduced model performs as well as the model where the bias is a linear combination of ratios, suggesting that aiming at explaining the rater bias by the financial ratios does not lead to superior results.

References

Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, **83**, 1023 – 1032.

Plummer, M. (2003). JAGS: A program analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.*

Vasiliki, S. and Veldkamp, L. (2009). Ratings shopping and asset complexity: A theory of ratings inflation. *Journal of Monetary Economics*, **56**, 678 – 695.

Joint modelling of quantile regression and survival time of lung function decline in cystic fibrosis patients

Elisabeth Waldmann[1], David Taylor-Robinson[2]

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg
[2] Department of Public Health and Policy, University of Liverpool

E-mail for correspondence: elisabeth.waldmann@liv.ac.uk

Abstract: When modelling repeated measurement and time to event data simultaneously, analysis is often based on combining mixed models with survival analysis. Some data sets however have a more complex structure than the one underlying normality assumption. In the case of modelling lung function of cystic fibrosis (CF) patients over a long period of time the impact of the covariates differs between quantiles of the dependent variable. An important question in CF research is to understand the impact of the lung function trajectory on the onset of a variety of different infections. This work aims at explaining the setup of a Bayesian joint quantile model and illustrates it using data from the United Kingdom CF registry.

Keywords: Longitudinal data; Survival analysis; Joint modelling, Quantile regression.

1 Cystic fibrosis in the United Kingdom - the registry

Cystic fibrosis (CF) is one of the most common serious genetic diseases in the western world. It has an impact on a number of organs, primarily the lung, pancreas and liver. The UK CF registry collects longitudinal data on all patients with CF. It currently holds approximately 10.000 patients This analysis uses data on 1557 patients seen twice or more between 1995 and 2009. It explores the trajectory of 8806 lung function measurements and the influence of various covariates on risk of infection. To this end the participants of the study attend annual examinations in which, among other factors, their lung function is measured as the percentage of air exhaled

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

within the first second of exhalation. This value is called forced expiratory volume (%FEV1) and is typically decreasing over time in CF patients.

2 Bayesian joint modelling

CF patients are also more susceptible to pulmonary infections such as *Pseudomonas aeuriginosa* (PA). The lung function decreases significantly faster when patients are infected by one or more of those diseases (see Qvist et al., 2015). Instead of interpreting the disease as the trigger, some practitioners believe that lower lung functions lead to higher susceptibility for infections. Joint models tackle this problem by investigating the connection between a mixed model and a survival model. In our case the dependent variable in the mixed model is the lung function measured as the %FEV1 value and the event of the survival model is the point of infection. There are several approaches to joint modelling, the one that will be referred to in the following is of the type in Faucett and Thomas (1996). The Monte Carlo Markov Chain (MCMC) approach that is described in the following is also based on this paper. The likelihoods are hence as follows:

$$f(Y_i|\eta_i,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}^{J_i}} \exp\left(-\sum_{j=1}^{J_i} \frac{(y_{ij}-\eta_i(t_{ij}))^2}{\sigma^2}\right) \quad (1)$$

$$f(s_i,d_i|Y_i) = \{\lambda_0(s_i)\exp(\eta_i(s_i))\}^{d_i} \exp\left(\int_{e_i}^{s_i} \lambda_0(u)\exp(\gamma\eta_i(u))\,du\right), \quad (2)$$

where n is the number of individuals, J_i the number of observations of the ith individual, e_i is the time individual i enters the study and s_i is either the time of infection or last time observation. The dependent variable y_{ij} is the jth observation of individual i. To cope with the baseline hazard $\lambda_0(t)$, the time is split into a grid $t_0,\ldots,t_K$ for which $\lambda_0(t)$ is approximated by a piecewise linear function with values λ_k for $k=1,\ldots,K$. The integral in equation (2) can hence be approximated by a sum of K integrals. The association parameter γ links the mixed model to the survival analysis and quantifies the strength of the influence. In the CF study the predictor $\eta_i(t_{ij})$ was chosen to be simply a random slope and intercept: $\eta_i(t_{ij}) = a_i + b_i t$. The extension to more elaborate effects is straightforward. In the following the full conditional for an arbitrary part $\eta_{ijl}(t_{ij})$ (l indicating that it is just

the lth part of the predictor) of the predictor presented as an example:

$$p(\eta_{ijk}|\cdot) \propto \exp\left(-\sum_{j=1}^{J_i}\frac{\left(y_{ij}-\eta_{i(t_{ij})}\right)^2}{\sigma^2}\right) p(\eta_{ijk}(t_{ij})) \{\lambda_0(s_i)\exp(\gamma\eta_i(s_i))\}^{d_i}\exp\left(-\int_0^{s_i}\lambda_0(t)\exp(\gamma\eta_i(u))\,du\right).$$

As the full conditional has no closed form distribution, the random intercepts were sampled with adaptive rejection sampling (ARS). The same holds for the random slope as well as for any other effects possibly introduced into the model. The full conditional for γ consists of the second term of the likelihood and is also generated by ARS. The model variance σ^2 is not linked to the survival part of the model, the full conditional is hence conform with the result in Bayesian mixed models:

$$\sigma^2|\cdot \sim \text{invGamma}\left(a_0+\frac{n}{2}, b_0+\sum_{i=1}^{n}\sum_{j=1}^{J_i}(y_{ij}-\eta_i t_{ij})^2\right),$$

where a_0 and b_0 are parameters of an inverse gamma prior. The relevant part of the likelihood for the full conditional of the λ_k is (2), which reduces to a gamma distribution:

$$\lambda_k|\cdot \sim \text{Gamma}\left(\sum_{i=1}^{n}d_{ik}+1, \sum_{i=1}^{n}\int_{\max(t_k,e_i)}^{\min(t_{k+1},s_i)}\exp(\gamma\eta_{ij}(t))\,dt\right).$$

3 Bayesian joint quantile regression

The above described data set shows strong differences in the covariates when measuring the impact on different quantiles. The structure of the presented model will hence be extended to a model describing the association between a mixed model quantile regression and the survival model. The ultimate goal of this experiment is to model the survival function based on a set of quantiles mimicking the whole distribution. Bayesian inference on quantile regression can be conducted by using the asymmetric Laplace distribution (ALD) as auxiliary distribution. The ALD contains the check function $\rho_\tau(u) = u(\tau - I(u > 0))$, where τ is the quantile of interest, which when minimized leads to estimators for quantile regression parameters and can be rewritten as a location scale mixture, which allows to conduct MCMC studies similar as in mean regression:

$$y_{ij} \sim \text{ALD}\left(\eta_{ij\tau}+\xi w_{ij}, \sigma^2 w_{ij}\phi\right).$$

Here w_{ij} is a weight variable with $w_{ij} \sim \text{Exp}(1/\sigma^2)$ and $\phi = \frac{2}{\tau(1-\tau)}$ and $\xi = \frac{1-2\tau}{\tau(1-\tau)}$ are quantile specific auxiliary variables. The major change in the setup is in the random and the model variance. Only the likelihood (1) changes:

$$\frac{1}{\sqrt{2\pi\sigma^2 w_{ij}\phi}^{J_i}} \exp\left(-\sum_{j=1}^{J_i} \frac{(y_{ij} - (\eta_{ij}(t_{ij}) + \xi w_{ij}))^2}{2\sigma^2 w_{ij}\phi}\right),$$

and the full conditionals have to be adapted accordingly. The full conditional for the model variance is a gamma distribution, similar to the above described but with slight changes in the parameters: the shape $a_{\sigma^2} = a_0 + \frac{3N}{2}$ and the scale $b_{\sigma^2} = b_0 + \sum_{i=1}^{n}\sum_{j=1}^{J_i} \frac{(y_{ij}-(\eta_{ij\tau}+\xi w_{ij}))^2}{2w_{ij}\phi} + w_{ij}$. Neither the full conditional for the hyper parameters nor for γ and the λ_ks change to the above explained. The full conditional for the weights w_{ij} stays the same as in Bayesian quantile regression: an inverse Gaussian distribution.

4 Results

The Bayesian joint quantile model was applied to the CF data set. Lung function measured in %FEV1 was the dependent variable in the mixed model part and the onset of PA is the event time in the survival model. To make the association comparable, an extra fixed effects term was added, but only the random effect was entered in the association term in the survival model. Figure 1 displays the different in the association parameter for nine different quantiles. The association parameter was considered significant, if 95% of the values in the sample are on one side of 0 and the corresponding boxes hence colored in grey.

To understand those values we have to look at them in the context of Figure 2. This figure shows a selection of the mixed model part for four different quantiles. The black lines display the overall quantile specific regression estimation, while the grey lines stand for the deviation from this overall estimation per patient. We can see that the individual lines scatter rather differently below and above the black lines. The impact of this deviation on the hazard rate is quantified by γ. Negative γ values mean that the patients which deviate negatively from the overall trend result in having a multiplicative constant bigger than one and hence have and increased risk of infection. The size of γ can naturally only be interpreted in the light of the size of the random effects.

Note that besides the conclusions about the connection between lung function decline and onset of infections, it is also worth mentioning that the decline of the lung function is slightly steeper for the lower quantiles, i.e. for patients that already have bad lung functions.

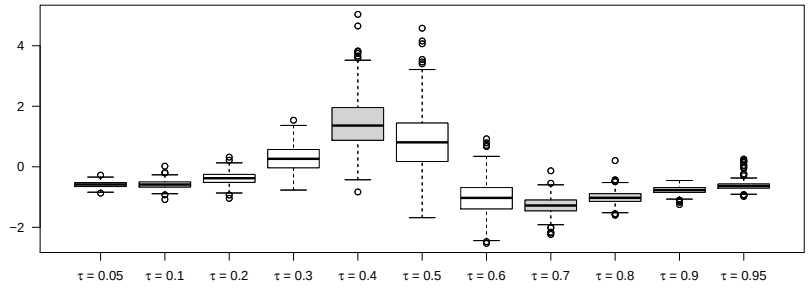

FIGURE 1. MCMC samples of the association parameter γ for different quantiles.

5 Summary and outlook

The model described here shows how joint modelling of longitudinal data and time to events can be extended to including quantile regression into the survival model. Besides including a set of quantiles, as mentioned above it is for example also possible to replace the quantile model by a Bayesian distributional regression or with any other kind of non mean regression in order to get a deeper understanding of the connection between a longitudinal process and time to event observations.

References

Faucett, C. and Thomas, D. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, **15**, 1663 – 1685.

Qvist, T., Taylor-Robinson, D., Waldmann, E., Olesen, H., Rønne Hansen, C., Mathiesen, M.I., Høiby, N., Katzenstein, T.L., Smyth, R.L., Diggle P., and Pressler, T. (2015). Infection and lung function decline; comparing the harmful effects of cystic fibrosis pathogen. *Work in progress*.

Waldmann, E., Kneib, T., Yue, Y., Lang, S., and Flexeder, C. (2013). Bayesian semiparametric additive quantile regression. *Statistical Modelling*, **13**, 223 – 252.

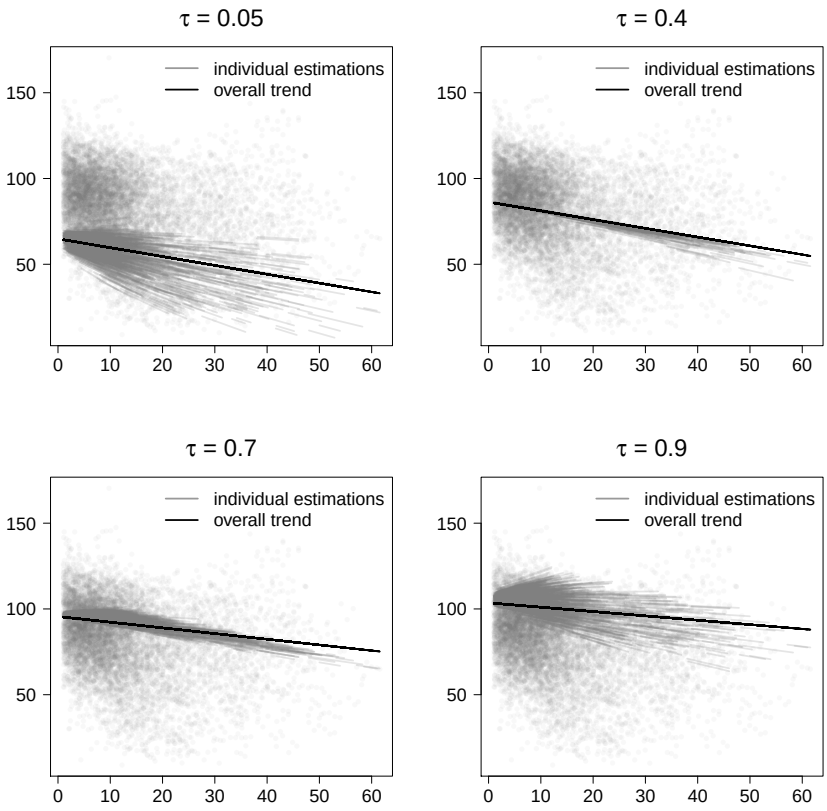

FIGURE 2. Linear effect of age on lung function with: overall effect in black, individual effects grey.

Estimation of two-sided choice models: an application to public school choice

Matthew Weldon[1], Andrew Titman[1]

[1] Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

E-mail for correspondence: m.weldon@lancaster.ac.uk

Abstract: This paper introduces a methodology for estimating choice models on both sides of a many-to-one matching market, such as a school choice setting. We introduce a model that relates an observed matching to the parameters of random utility models for both agents and institutions, via a stability condition on the observed matching. The stability partial likelihood can be fitted using standard optimisation methods. We present simulation results for identifiability and consistency, and the results of a pilot study, estimating parental preferences for school attributes in a UK public school setting.

Keywords: School choice; Stable matchings; Discrete choice.

1 Introduction

Consider the problem of studying the choices that people make over which school to send their children to, or which university to attend, or which employment to take. In these contexts it is often difficult to find useful data about agents' preferences to inform research. The classic "revealed preference" approach would be to use the observed choices themselves as revealing agents' preferences for the various options. However, in the settings described above, choice is usually constrained by scarcity of places at institutions, and the fact that the institutions themselves have preferences (or act as if they have preferences) over whom to admit. These institutional preferences may themselves be of substantive interest.

Simply ignoring constrained choice sets would lead to unacceptable bias in estimates of choice parameters. Indeed, discrete choice researchers go to great lengths to account for the endogeneity of choice sets. However, methods in use, such as inverse probability weighting or constraining choice

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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

sets, are somewhat ad-hoc and do not take account of the uncertainty that these methods add to inferences.

The problem of empirically modelling preferences in two-sided matching markets has received increasing attention recently. It has been shown that, under an assumption of stability, parameters of underlying random utility models for agents on both sides of a matching are identifiable (cf. Agarwal and Diamond, 2013; Echenique et al., 2013; Logan et al., 2008). So far, however, methods to estimate models for unaggregated data have suffered from computational intractability or partial identification that has restricted their use in real-world applications.

Logan et al. (2008) employ a Gibbs MCMC method to model preferences for marriage partners (i.e. a one-to-one matching). However, this requires sampling from the constrained conditional distribution of each of the $M \times N$ unobserved utilities in each iteration, which is computationally intensive for large matchings. Echenique et al. (2013), propose a method-of-moments estimator for repeated 2×2 matchings for the purpose of investigating identification, but do not elaborate on how this estimator could be used in applied work.

In this paper, we present a "stability likelihood" for many-to-one matchings that admits a flexible class of random utility models. We show that, by specifying and estimating a model for preferences on the supply side (that is, the institutions) we can not only account for constrained choices on the demand side, but can also provide estimates of institutional choice parameters. In some settings, such as labour-market research, these supply-side preferences may be interesting in their own right.

2 The stability likelihood

Let $\mathcal{A} = \{a_1, \ldots, a_N\}$ be a set of agents which we will call students, and let $\mathcal{S} = \{s_1, \ldots, s_M\}$ (with $N > M$) be a set of agents we will call schools. Each student has a set of preferences over all of the schools, and each school has an ordering over all of the students.

A matching is defined as a set of school–student pairs $\mathbf{m} = \{(a, s) : \{s\} = \mathrm{m}(a)\ \&\ a \in \mathrm{m}(s)\}$. Let $\mathcal{M}$ denote the set of all possible matchings; we are interested in the set of matchings that satisfy the stability property, $\mathcal{M}^* \subseteq \mathcal{M}$, in the sense first defined by Gale and Shapley (1962), which we will define below.

Students' preferences for schools, and schools' orderings over students, can be modelled as arising from an underlying set of utilities. Let U_{as} be student a's utility for school s, and V_{sa} be school s's utility for student a. We treat utilities on both sides equally. Students' utilities for schools can be modelled as a function of observable attributes and unobservable random variation. For example, $U_{as} = \alpha_s + x_{as}^{\mathrm{T}}\beta + \epsilon_{as}$. Likewise, schools' preferences for students are may be given by $V_{sa} = z_{sa}^{\mathrm{T}}\delta + \eta_{sa}$, where ϵ_{as} and η_{sa} are

match-specific unobserved variation, whose distribution may be specified by the researcher. In our examples we have used $\epsilon_{as}, \eta_{sa} \sim N(0,1)$. Let $\theta = (\alpha, \beta, \delta)$.

There is no need for exclusion restrictions on the utility models for identification. In particular, X and Z can contain some or all of the same covariates. For example, distance as a choice criterion may enter both models, and in many-to-one markets separate parameters will be identified for each.

Although we have a simple model relating parameters θ to utilities, and utilities to preferences, a likelihood of the form $\mathrm{P}(\mathbf{m}|\theta)$ is intractable, as it depends not only on the preferences but also on the unknown matching mechanism used. In the absence of a generative model for the matching itself, we base inference on a model for the stability of the observed matching, conditional upon the matching: $\mathrm{P}(\mathbf{m} \in \mathcal{M}^*|\mathbf{m}, \theta)$.

In order to derive a tractable likelihood for stability, we first define stability. Stability of a matching $\mathbf{m}$ is equivalent to the condition that each school-student pair who are *not* matched would not mutually prefer to match with each other than their current matches.

Let $\overline{U}_a = U_{as}\ \forall s : \{s\} = \mathrm{m}(a)$. This is the student's utility for her allocated school. Similarly, let $\overline{V}_s = \min_{a \in \mathrm{m}(s)}\{V_{sa}\}$ be the utility of the worst student matched with school s. For each unmatched pair (a, s), the pair blocks the matching iff $U_{as} > \overline{U}_a$ and $V_{as} > \overline{V}_s$. If there are no such pairs, the matching is stable.

For a given matching $\mathbf{m}$, the probability of pairwise stability is equivalent to the probability that no unmatched pair (a, s) blocks the matching. The conditional likelihood of stability, given the matching, the parameters and the unobserved $\overline{\mathbf{V}}$ is defined as

$$\begin{aligned}&\mathrm{P}(\mathbf{m} \in \mathcal{M}^*|\mathbf{m}, \theta, \overline{\mathbf{V}}) \\ &= \prod_{a \in \mathcal{A}} \int_{\mathbb{R}} \left\{ \prod_{\{s \notin \mathrm{m}(a)\}} 1 - \mathrm{P}(U_{as} > u)\mathrm{P}(V_{sa} > \overline{V}_s) \right\} p_{\overline{U}_a}(u)\, \mathrm{d}u.\end{aligned}$$

The integrals over $p_{\overline{U}_a}(u)$ are easily approximated using gaussian quadrature rules. The joint likelihood for stability and the unobserved $\overline{\mathbf{V}}$ is

$$\mathrm{P}(\mathbf{m} \in \mathcal{M}^*, \mathbf{v}|\mathbf{m}, \theta) = \mathrm{P}(\mathbf{m} \in \mathcal{M}^*|\mathbf{m}, \theta, \overline{\mathbf{V}}) \prod_{s \in \mathcal{S}} p_{\overline{V}_s}(v|\mathbf{m}, \delta),$$

where

$$\begin{aligned}p_{\overline{V}_s}(v|\mathbf{m}, \delta) &= \frac{\mathrm{d}}{\mathrm{d}v}\left[1 - \prod_{a \in \mathrm{m}(s)} \mathrm{P}(V_{sa} > v|\delta)\right] \\ &= \left[\prod_{a \in \mathrm{m}(s)} \mathrm{P}(V_{sa} > v|\delta)\right]\left[\sum_{a \in \mathrm{m}(s)} \frac{p_{V_{sa}}(v)}{\mathrm{P}(V_{sa} > v|\delta)}\right].\end{aligned}$$

The random variables $\overline{\mathbf{V}}$ are considered nuisance parameters. It is possible to integrate over their distribution using MCMC methods, although in practice we have found that profile likelihood estimates of the substantive parameters, evaluated at the MLE of $\overline{\mathbf{V}}$, are very close to the marginal estimates at a fraction of the computational cost. Simulation studies conducted using several specifications give positive results for identification, consistency and efficiency of estimates computed by this method.

3 School choice study

Of substantive interest are the determinants of parental preferences for UK public schools. We have data on 838 pupils allocated to seven schools in a medium sized town in the North West of England. There are four community schools and three faith schools. For each school-pupil pair, we have the distance between the pupil's home and the school, the pupil's ethnicity, and the ethnic composition of the school. On the pupil side, we wish to model preferences for proximity and possible ethnic in-group preferences (homophily), as well as estimating a fixed effect for each school to account for other aspects of popularity.

Pupils apply to schools by ranking up to three schools, and the allocation is computed using a stable matching algorithm, which guarantees the validity of the stability assumption. We have additional data in the form of the total number of first preference applications received for each school. This information has not been used in estimation, and so is useful for externally validating the fitted model.

Estimates are presented in Table 1. Negative distance effects are present on the supply- and demand-sides. In addition, the effects for ethnic in-group preferences are positive, but are not robust to sampling variation. For example, an asian student has to choose between two schools that are identical apart from the fact that the second school has 10% more asian students. A 95% confidence interval for the odds ratio of preferring the second school is (1.02, 1.70).

The estimated school effects are compared to numbers of first choice applications to these schools in Table 1. The model appears to be able to capture the ranking of the schools as revealed by application numbers.

4 Conclusion

We have presented, in general terms, a flexible, tractable partial-likelihood for estimating the parameters of a two-sided preference model, which uses the information within a stable many-to-one matching. This stability–likelihood can be incorporated into maximum-likelihood or Bayesian approaches, and a number of methods are possible. In this paper we have

TABLE 1. (a) Estimated co-efficients and standard errors for a two-sided model fit to seven secondary schools in England. (b) School effects are compared to known application numbers.

Parameter	MLE	Std err.
Supply-side:		
δ_1 Dist. × Comp.	-0.51	0.092
δ_2 Dist. × Faith	-0.062	0.046
Demand-side:		
β_1 Distance (km)	-0.84	0.070
β_2 Asian × % asian	0.024	0.012
β_3 White Br. × % wbr.	0.013	0.0086

Sch. effects:	MLE	Std err.	Applications
α_1	2.73	0.54	212
α_2	1.77	0.27	201
α_3	0.40	0.64	166
α_4	0.13	0.28	130
α_5	0*	–	101
α_6	-0.06	0.21	117
α_7	-0.61	0.59	100

* Reference school.

focused on a computationally convenient method that profiles-out the unobserved threshold utilities.

Because it is a fully-specified likelihood, albeit only modelling part of the data generating process, the model accounts for the uncertainty that comes from not knowing each decision maker's true choice sets.

In a pilot study we have presented results for proximity and ethnic in-group preferences for schools that are tentatively interesting, and exhibit the potential usefulness of the model for social choice research in challenging data environments.

References

Agarwal, N. and Diamond, W. (2013). Identification and Estimation in two-sided matching markets. **Cowles Foundation discussion paper**, **1905**. Yale University.

Echenique, F., Lee, S., and Shum, M. (2013). Partial identification in two-sided matching models. In: *Advances in Econometrics*, **31**, 117 – 139.

Gale, D. and Shapley, L. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, **69**, 9 – 15.

Logan, J., Hoff, P., and Newton, M. (2008). Two-sided estimation of mate-preferences for similarities in age, education and religion. *Journal of the American Statistical Association*, **103**, 559 – 569.

Menzel, K. (2014). *Large matching markets as two-sided demand systems.* Technical report, New York University.

Estimation of a general heaping model via random-walk Metropolis algorithms

Ariane Würbach[12], Sabine Zinn[1]

[1] Leibniz Institute for Educational Trajectories (LIfBi), Bamberg, Germany
[2] Chair of Statistics and Econometrics, Otto-Friedrich-University Bamberg, Germany

E-mail for correspondence: ariane.wuerbach@lifbi.de

Abstract: Heaping is a general tendency of survey respondents to report income data rounded off to the nearby modulos. There can be differences to which degree heaping occurs, i.e. possible modulos can be 100, 500, or 1000. We developed a general method to account for different heaping patterns. A mixture model describes the underlying Dagum distribution and the supposed heaping mechanism. For the sake of simplicity, we assume that the probabilities of heaping to specific heaping points (modulos) are constant within predefined intervals. The parameters of the mixture model are estimated simultaneously using three different random-walk Metropolis algorithms: a single-block algorithm, a multiple-block algorithm (MB), and a randomized multiple-block algorithm (RMB). The results are compared by inefficiency factors and marginal likelihoods. Results from a simulation study show that estimates are better approximated by either MB or RMB algorithms. The proposed method is applied to income data of the National Educational Panel Study (NEPS). The performance is evaluated by posterior predictive checks and demonstrates a good fit of the general model.

Keywords: Heaping; Random-walk Metropolis algorithm; Multiple-block scheme.

1 A general model for heaping

When people are asked to report their monthly income they tend to round off, or even refuse to answer. Rounding off data causes abnormal concentrations of reported values at certain *heaping points* and might hamper statistical inference. In order to facilitate adequately modeling of heaped data, we introduce a general method that explicitly specifies the true underlying distribution in combination with the heaping pattern present in a data set. The true income is assumed to follow a Dagum 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

and the heaping probabilities are stated to be constant within predefined intervals. The log-likelihood of this mixture model is given by

$$\ell(\theta|z_i) = \left[1 - \sum_{b=1}^{S} \rho_b \int_{s_b}^{t_b} f(z_i|\psi) dz_i \mathbb{I}_{(z_i \in \mathbb{R}^+)} + \sum_{b=1}^{S} \rho_b \left[F(t_b|\psi) - F(s_b|\psi)\right] \mathbb{I}_{(z_i \in H)}\right].$$

The observed values $\mathbf{z} = \{z_i\}_{i=1}^n$ can take values in $\mathbb{R}_0^+$. The vector $\theta = \{\psi, \rho\}$ denotes the parameters for the underlying income distribution and the heaping probabilities for rounding off to heaping points h_b, where $H = \{h_b\}_{b=1}^S$ is the set of heaping points with $h_b \in \mathbb{N}_0$ mod (100, 500, 1000). The intervals within which values might be rounded off to the heaping points h_b are denoted by $I_b = [s_b, t_b]$. $f(z_i|\psi)$ is the Dagum density function of the true values, and $F(z_i|\psi)$ the corresponding distribution function with $\psi = \{a, b, p\}$, where a, p are shape parameters and b is a scale parameter.

2 Bayesian estimation of the global heaping model

We use three random-walk Metropolis (RWM) algorithms in order to estimate the parameters of the model. The posterior means serve as MCMC estimates. Although convergence of MCMC simulations is almost always assured, often very large convergence times may be implied, or even worse, one is lead to belief that convergence is achieved while important aspects of the target distribution are left unexplored (*pseudo convergence*), see Jackman (2009). Consequently, careful selection and thoughtful adjustment of the tuning parameters in MCMC settings is vital to construct an efficient, well-mixing MCMC sampler.

The first RWM algorithm applied is a single-block sampler (S-RWM). Here, the posterior distribution is sampled in one block for all parameters. The second algorithm is a fixed block version (MB-RWM). The parameters of the underlying distribution form one block, and the parameters of the heaping mechanism are separated into clear-cut blocks following a natural blocking strategy, i.e. all probabilities describing a specific modulo (4M-RWM, considered modulos are 0, 100, 500 and 1000), or probabilities covering the same interval (8I-RWM, i.e. the range of z_i is divided into eight intervals). In the third RWM algorithm, all parameters are clustered randomly into blocks (RMB-RWM), i.e. at each iteration the blocks are newly constructed with varying sizes and composition.

3 Simulation study

We simulate a data set assuming the Dagum distribution with $a = 3.6$, $b = 2416$, $p = 0.43$ as true underlying distribution. Furthermore, we assume a heaping mechanism which can be described by 20 heaping probabilities that give the probability of a true income value to be rounded

off to a specific modulo. As suitable prior and proposal distributions the multivariate normal distribution is chosen assuming that the parameters are a priori independent, with $\mu = (\{0.2\}_{b=1}^{S}, 3, 2000, 0.4)^{\mathrm{T}}$, and covariance $\Sigma = \mathrm{diag}(\{0.01\}_{b=1}^{S}, 0.1, 100, 0.01)$. The corresponding definition of the proposal density is $\mathcal{N}(\theta^{(t-1)}, \Sigma)$. For the construction of the MB-RWM and RMB-RWM algorithms the components of θ are split into vector blocks. In a Metropolis step, θ is updated either in one block, or for each vector block conditioned on the most current value of the parameters in the remaining blocks (Chib and Ramamurthy, 2010). We run 10 independent RWM

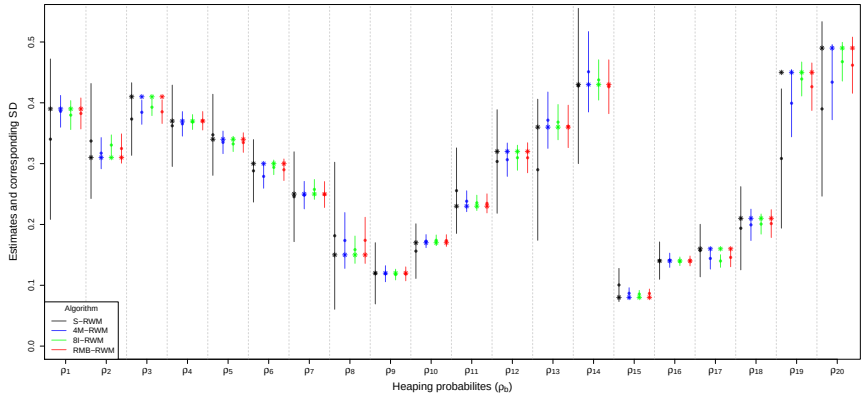

FIGURE 1. Averaged posterior means and total standard deviation of the RWM algorithms over 10 runs with varying starting values. The true parameters are marked by a star.

chains with different starting values for each algorithm, with a chain length of $T = 10{,}000$ iterations ($t = 1, \ldots, T$) following a burn-in of $n_0 = 1000$ iterations. Figure 1 gives the combined results of the applied RWM algorithms displaying the averaged MCMC estimates with their corresponding total standard deviation (SD) and the true values of the data generating process (DGP). We find that almost all estimates are close to the true parameter values or even within one SD. Only ρ_{19} in S-RWM, ρ_3 in 4M-RWM and RMB-RWM as well as ρ_{17} in 8I-RWM are not so close. Though, marginal posterior densities exemplarily given for ρ_3 and ρ_{19} (Figure 2) show that the true parameters are covered by the 95% highest posterior density region. Furthermore, we see that in the blocking algorithms the MCMC estimates are closer to the true parameter values and the SD get remarkably smaller than in the S-RWM. Also the *inefficiency factors* (Ineff) as well as the *marginal likelihoods* (Chib and Ramamurthy, 2010) point to a better performance of the MB-RWM algorithms as compared to the S-RWM algorithm. In summary, we find that the Ineff are highest for S-RWM, within the range (521.0;1048.9) and decline with blocking (for 4M-RWM

the Ineff are in the range (77.95;692.64) and in the range (92.19;491.70) for the RMB-RWM). The 8I-RWM shows the smallest Ineff. Only five parameter values have an Ineff above 100, but below 305, all other parameters have an Ineff of lower than 50 or 100, which indicates highest efficiency. Further, we find the marginal likelihood estimate of the 8I-RWM algorithm to be the greatest (−66,237.12) as compared to S-RWM (−66,363.06), 4M-RWM (−66,285.66), and RMB-RWM (−66,283.61).

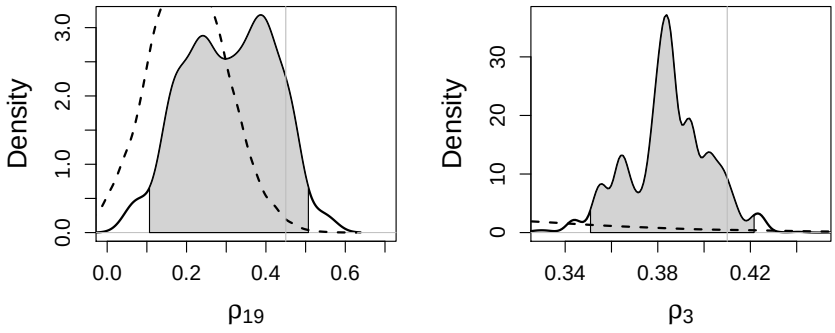

FIGURE 2. Marginal posterior plots for selected parameter values of the S-RWM (left) and 4M-RWM (right) algorithm. Solid lines indicate posterior distributions and dashed lines the prior distributions. The gray colored vertical lines give the position of true value from the DGP. The highest posterior density distribution (95%) is shaded in gray.

4 Application to NEPS income data

The proposed method is applied to income data from the adult cohort of the National Educational Panel Study (NEPS). *Posterior predictive checks* based on 10 replicates constructed with the posterior means from each RWM algorithm demonstrate a good fit of the general model with regard to selected descriptive statistics (mean, quartiles). This finding is also supported by the histogram from observed and replicated data, as exemplarily shown in Figure 3 for one replicate. We refer to Table 1 for a summary of the corresponding two-sided posterior predictive p-values. Small posterior predictive p-values indicate that the real and replicated data do systematically deviate from each other. Considering, however, the absolute differences gives the following ranges: (-62.14;-18.08) for the mean, (-180.00;-82.42) for the median, (-24.19;0.65) for the first quartile, and for the third quartile (-106.84;0). These deviations are rather small and can be

TABLE 1. Two-sided posterior predictive p-values for descriptive statistics.

Descriptive	S-RWM	4M-RWM	8I-RWM	RMB-RWM
mean	0.6000	< 0.0001	< 0.0001	0.5000
median	< 0.0001	< 0.0001	< 0.0001	< 0.0001
1. quartile	0.6000	0.1000	< 0.0001	0.2500
3. quartile	0.6000	0.1000	0.2500	0.2500

attributed to a less satisfactory model fit concerning parameter b, the scale parameter of the Dagum distribution. Similar problems have already been described and discussed in the context of maximum likelihood estimation of the general heaping model, see Zinn and Würbach (2014). A possible solution to counteract this obstacle is using the log-normal as underlying distribution. Besides the replacement of the true underlying distribution, further steps include the specification of varying heaping mechanisms.

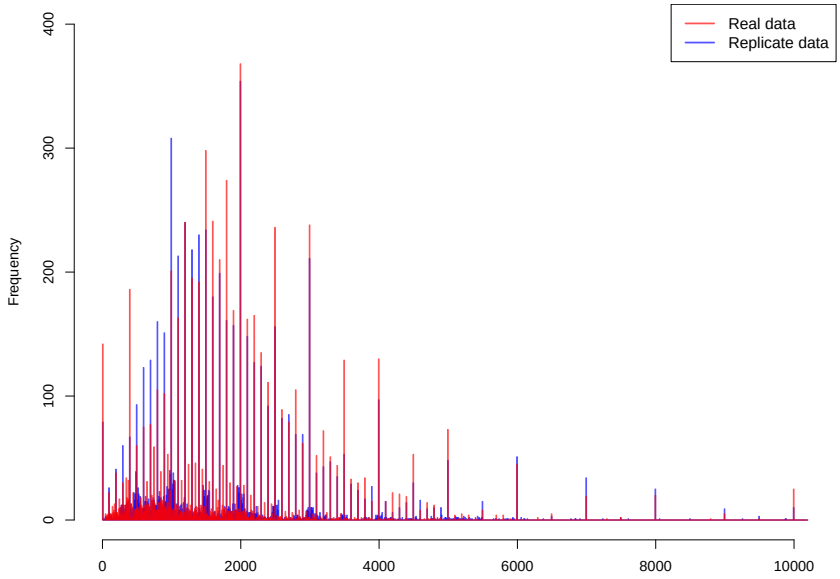

FIGURE 3. Individual net income data from the adult cohort of the German National Educational Panel Study (orange) and replicated data (blue). Overlapping proportions are colored in red.

Notes

We also tested the *uniform-in-each-direction* proposal density ($\theta^* = \theta^{(t-1)} + \epsilon$, $\epsilon = 0.05$), see Jackman (2009). It yields very good approximations to the true parameter values, i.e. for most of the parameters the estimates lie within one standard deviation of the posterior mean. The sampler does quickly converge and the inefficiency factor is always lower than 300 in S-RWM, and 150 in 8I-RWM. However, the uniform proposal density has a strong theoretical drawback. It does not ensure that the whole support of the target distribution is covered at each iteration, see Chib and Greenberg (1995). The consequence of such a restricted support is that the log marginal likelihood becomes incalculable.

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 6 - Adults, doi:10.5157/NEPS:SC6:1.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

References

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, **49**, 327 – 335.

Chib, S. and Jeliazkov, I. (2001). Marginal Likelihood from the Metropolis-Hastings Output. *Journal of the American Statistical Association*, **96**, 270 – 281.

Chib, S. and Ramamurthy, S. (2010). Tailored randomized block MCMC methods with application to DSGE models. *Journal of Econometrics*, **155**, 19 – 38.

Gelman, A., Meng, X.-L., and Stern, H.S.] (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733 – 807.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Chichester, U.K.: Wiley.

Zinn, S. and Würbach, A. (2014). A statistical approach to account for heaping patterns: An application to self-reported income data. *NEPS Working Paper*, **40**, Bamberg, Leibniz Institute for Educational Trajectories (LIfBi).

Index

Aderhold, Andrej, 204
Aerts, Marc, 215, 238
Agresti, Alan, 244
Alba, Anna, 163
Ambrosino, Chiara, 3
Ameraoui, Abdelkader, 89
Armstrong, Ben, 187

Böhning, Dankmar, 238
Banerjee, Sudipto, 23
Bartolucci, Francesco, 94
Berger, Moritz, 100
Bernardi, Mauro, 106
Bonou, Wilfried, 169
Bottone, Marco, 106
Boukhetala, Kamal, 89
Bremhorst, Vincent, 112
Brockhaus, Sarah, 117

Camarda, Carlo G., 123
Chandler, Richard E., 3
Christensen, Nelson, 145
Colombi, Roberto, 129
Colosimo, Enrico A., 354
Cook, Bruce D., 23
Couturier, Dominique-Laurent, 210
Crespo Cuaresma, Jesus, 135

De Mulder, Wim, 139
Demétrio, Clarice G.B., 308, 314
Demarqui, Fabio N., 354
Denuit, Michel, 273
Desmettre, Sascha, 360
Dias, Carlos T.S., 314
Dupuy, Jean-François, 89
Durbán, María, 330

Edwards, Matthew C., 145
Eilers, Paul H.C., 123, 151, 181, 330, 372

Fallah, Lida, 157
Fasola, Salvatore, 348
Fernández-Fontelo, Amanda, 163
Finley, Andrew O., 23
Frühwirth-Schnatter, Sylvia, 35, 175
Francis, Brian, 94
Frasso, Gianluca, 169
Fuest, Andreas, 117

Gampe, Jutta, 181
Gasparrini, Antonio, 187
Giordano, Sabrina, 129
Goerg, Georg M., 192
Gormley, Isobel Claire, 302
Grün, Bettina, 135, 377
Greven, Sonja, 51, 117
Groll, Andreas, 198
Grow, André, 139
Grzegorczyk, Marco, 204

Hastie, Trevor, 198
Heller, Gillian, 210
Heritier, Stephane, 210
Hermans, Lisa, 215
Herwartz, Helmut, 249
Hinde, John, 67, 157, 308
Hofmarcher, Paul, 135
Hofner, Benjamin, 221, 296
Hornik, Kurt, 377
Hsu, Chiu-Hsieh, 227
Hu, Chengcheng, 227
Humer, Stefan, 135
Husmeier, Dirk, 204

Iannario, Maria, 232

Jagan, Kavya, 267
Jaspers, Stijn, 238
Jin, Yuxue, 192

Kateri, Maria, 244
Kenward, Michael G., 187, 215
Klein, Nadja, 249, 255
Klingenberg, Bernhard, 261
Kneib, Thomas, 249, 255
Koehler, Jim, 192
Kosmidis, Ioannis, 267
Kreyenfeld, Michaela, 112

Lambert, Philippe, 112, 169, 273
Lang, Joseph B., 232
Larocque, Denis, 279
Lee, Dae-Jin, 330

Machado, Robson J.M., 285
Marx, Brian D., 290
Mayr, Andreas, 117, 296
Mayr, Georg J., 366
McParland, Damien, 302
Messner, Jakob W., 366
Meyer, Renate, 145
Molenberghs, Geert, 139, 215, 314
Moral, Rafael A., 308
Moser, Mathias, 135
Muggeo, Vito M.R., 348, 372

Nassiri, Vahid, 215

Oliveira, Izabela R.C., 314
Ortega, Edwin M.M., 308

Pandolfi, Silvia, 94
Pennoni, Fulvia, 94
Petrella, Lea, 106
Peyhardi, Jean, 319
Posekany, Alexandra, 175
Puig, Pere, 163
Putter, Hein, 181

Reichl, Johannes, 324
Remy, Nicolas, 192
Rodríguez-Álvarez, María Xosé, 330
Ruckdeschel, Peter, 360

Schauberger, Gunther, 336
Schmid, Matthias, 296
Schneider, Micha, 342
Siino, Marianna, 348
Silva, Jose Luiz P., 354
Simonoff, Jeffrey S., 279
Smith, Adam, 221
Souza, Cláudio L., 314
Spangl, Bernhard, 360
Stauffer, Reto, 366

Taylor, Jeremy, 227
Taylor-Robinson, David, 381
Titman, Andrew, 387
Torretta, Federico, 372
Tutz, Gerhard, 100, 198, 336, 342

Umlauf, Nikolaus, 366

van den Hout, Ardo, 285
Van der Elst, Wim, 215
Vana, Laura, 377
Verbeke, Geert, 139, 215, 238

Waldmann, Elisabeth, 381
Weldon, Matthew, 387
Würbach, Ariane, 392

Zeileis, Achim, 366
Zhou, Yuzhen, 23

Zinn, Sabine, 392
Zou, Anyi, 267

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