Proceedings of the

32nd International Workshop on Statistical Modelling

Volume I

Groningen, Netherlands
3-7 July, 2017

Editors: Marco Grzegorczyk, Giacomo Ceoldo

Proceedings of the 32nd International Workshop on Statistical Modelling
Volume I,
Groningen, July 3-7, 2017,
Marco Grzegorczyk, Giacomo Ceoldo (editors),
Groningen, 2017.

Editors:

Marco Grzegorczyk, m.a.grzegorczyk@rug.nl
University of Groningen
Nijenborgh 9
9747 AG Groningen
The Netherlands

Giacomo Ceoldo, g.ceoldo@student.rug.nl
University of Groningen
Nijenborgh 9
9747 AG Groningen
The Netherlands

Preface

Dear Participants,

First of all, I would like to welcome you to the 32nd International Workshop on Statistical Modelling (IWSM 2017) in the Netherlands, and I wish you a very pleasant stay in Groningen. With around 200,000 inhabitants Groningen is the 7th largest town in the Netherlands, and Groningen has two large universities: The Hanzehogeschool for Applied Sciences (about 20,000 students) and the Rijksuniversiteit Groningen (RUG) with about 30,000 students. The venue of the IWSM 2017 is the Academy Building of RUG and located in the town center of Groningen.

Before you lies one of the two proceedings volumes of IWSM 2017. It is a unique feature within the statistical community that all speakers at this workshop also provide an extended abstract of their talk. This not only provides the participants with a compact written account of interesting contributions, but it also improves the quality of the talks.

Like every year, there was a huge amount of excellent paper submissions, and it was a really challenging task to select from 138 abstracts 56 (41%) for oral presentations. Each paper had to be reviewed and scored by three members of the scientific programme committee. This was a very time-consuming task for the reviewers, and for their valuable efforts I thank all members of the scientific committee: Ernst Wit (RUG), Marijtje van Duijn (RUG), Kenan Matawie (IWSM 2005), Arnošt Komárek (IWSM 2012), Vito Muggeo (IWSM 2013), Thomas Kneib (IWSM 2014), Helga Wagner (IWSM 2015), Jean-François Dupuy (IWSM 2016), Simon Wood (IWSM 2017), Paul Eilers, John Hinde, Dirk Husmeier, Sonja Greven, Edwin van den Heuvel, Jörg Rahnenführer and Korbinian Strimmer.

Some of the abstracts that could not be selected for oral presentation have been given the opportunity to be presented on a poster. On Tuesday (17:15-19:30h) there will be a poster presentation where everybody is invited to meet the researchers and to discuss their ongoing work one-on-one. It will be taken care of drinks and some 'fingerfood' in order not to be distracted by a thirsty throat or a hungry belly. Although it would have been possible to extend the number of presentations, it is an important feature of the IWSM workshops that there are no parallel sessions. This means that each presentation, whether by a PhD student or by a famous statistician, are awarded the same amount of attention. This means that the IWSM is a very coherent meeting, whereby the emphasis lies precisely on the word: 'meeting'. It is a place where junior and senior researchers mix and mingle.

Not coincidently, also this year you will get ample opportunities to meet your fellow participants. On Sunday evening, after the excellent short course by Tom Snijders and before the official start of the conference on Monday, there was already the informal welcome drink gathering in the Pool Restaurant of the Student Hotel. On Monday evening, there will be the official welcome reception in the Spiegelzaal of the Academy Building. On Tuesday, the Pizza & Beer Poster

session in the Academia Restaurant of the Academy Building will encourage lively scientific interactions. On Wednesday, after the excursions, we will reconvene altogether in the Ni Hao restaurant (Gedempte Kattendiep 122). On Thursday evening, if you are hungry again, there will be the official conference dinner in 't Feithhuis (Martinikerkhof 10). Even the conference dinner will be kept quite informal, as the emphasis should be on meeting your fellow participants. It would be great if, long after this conference is over, you could look back on IWSM 2017 and say: '*Groningen was the place where I met 'em all!*'.

I thank the Statistical Modelling Society for trusting in my proposal and for giving me this great opportunity to chair IWSM 2017. Many colleagues from RUG helped me planing and realizing this workshop. I would like to thank all members of my Local Organizing Committee, in particular, Ernst Wit, Ineke Schelhaas, Martijn Wieling, Casper Albers, Wendy Post, Marijtje van Duijn and Hans Burgerhof for their contributions. Also Mariska Pater and Sharon de Puijselaar from the Groningen Congres Bureau have helped me tremendously.

My special acknowledgements go to all sponsors of the IWSM 2017. Without those sponsorships certain things could not have been realized and the programme would certainly have been much sparser. A list of the sponsors of IWSM 2017 can be found on the last pages of this volume.

Last but not least, I would like to thank all authors for the excellent scientific contributions, and I hope that every participant of IWSM 2017 will have a great and especially research-stimulating week in Groningen,

Marco Grzegorczyk
Groningen, 16 June 2017

Scientific committee

- Chair: **Marco Grzegorczyk** (Groningen University, Netherlands)
- **Ernst Wit** (Groningen University, Netherlands)
- **Marijtje van Duijn** (Groningen University, Netherlands)
- **Arnošt Komárek** (Charles University Prague, Czech Republic)
- **Dirk Husmeier** (Glasgow University, Scotland)
- **Edwin van den Heuvel** (Eindhoven University, Netherlands)
- **Helga Wagner** (Linz University, Austria)
- **Jean-François Dupuy** (INSA of Rennes, France)
- **John Hinde** (NUI Galway, Ireland)
- **Jörg Rahnenfüehrer** (TU Dortmund University, Germany)
- **Kenan Matawie** (University of Western Sydney, Australia)
- **Korbinian Strimmer** (Imperial College London, England)
- **Paul Eilers** (Erasmus Rotterdam, Netherlands)
- **Simon Wood** (Bath University, England)
- **Sonja Greven** (LMU Munich, Germany)
- **Thomas Kneib** (Goettingen University, Germany)
- **Vito Muggeo** (Palermo University, Italy)

Local Organizing Committee

- Chair: **Marco Grzegorczyk**
- **Ernst Wit**
- **Ineke Schelhaas**
- **Martijn Wieling**
- **Casper Albers**
- **Wendy Post**
- **Marijtje van Duijn**
- **Hans Burgerhof**
- **Sacha la Bastide**
- **Christine zu Eulenburg**
- **Mark Huisman**
- **Ruud Koning**
- **Mariska Pater**
- **Laura Spierdijk**
- **Christian Steglich**
- **Marieke Timmerman**

Contents

Part I – Invited Papers

Part II – Contributed Papers

Part I – Invited Papers

Nonparametric inference in hidden Markov and related models

Roland Langrock

[1] Department of Business Administration and Economics, Bielefeld University, Germany

E-mail for correspondence: roland.langrock@uni-bielefeld.de

Abstract: Hidden Markov models (HMMs) have been successfully applied in various disciplines, including biology, speech recognition, economics/finance, climatology, psychology and medicine. They combine immense flexibility with relative mathematical simplicity and computational tractability, and as a consequence have become increasingly popular as general-purpose models for time series data. In this talk, I will first introduce the basic HMM machinery and showcase the practical application of HMMs using intuitive examples. I will then demonstrate how the HMM machinery can be combined with penalized splines to allow for flexible nonparametric inference in general-purpose HMM-type classes of models. The focus of the presentation will lie on practical aspects of nonparametric modelling in these frameworks, with the methods being illustrated in economic and ecological real data examples, featuring, inter alia, the famous wild haggis animal, blue whales and the well-known Lydia Pinkham sales data.

Keywords: Animal behaviour; Markov-switching regression; P-splines

---

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

Functional Data Analysis, Spatial Data Analysis and Partial Differential Equations: A fruitful union

Laura M. Sangalli

[1] MOX - Dipartimento di Matematica, Politecnico di Milano, Italy

E-mail for correspondence: laura.sangalli@polimi.it

Abstract: I will discuss an innovative class of regularized regression models for the analysis of spatially distributed data, that merges advanced statistical and numerical analysis techniques. Based on these regression models, I will then present a principal component analysis method that can handle functional signals distributed over complex domains.

Keywords: Penalized regression; functional principal component analysis; data distributed over two-dimensional manifold domains; finite elements.

1 Spatial regression with differential regularization

I will present a novel class of models for the analysis of spatially (or space-time) distributed data, based on the idea of regression with differential regularizations. The models merge statistical methodology, specifically from functional data analysis, and advanced numerical analysis techniques. Thanks to the combination of potentialities from these different scientific areas, the proposed method has important advantages with respect to classical spatial data analysis techniques. Spatial regression with differential regularizations is able to efficiently deal with data distributed over irregularly shaped domains, with complex boundaries, strong concavities and interior holes [Sangalli et al. (2013)]. Moreover, it can comply with specific conditions at the boundaries of the problem domain [Sangalli et al. (2013), Azzimonti et al. (2014, 2015)], which is fundamental in many applications to obtain meaningful estimates. The proposed models have the capacity to incorporate problem-specific priori information about the spatial structure of the phenomenon under study, formalized in terms of a governing partial differential equation [Azzimonti et al. (2014, 2015)]; this very flexible modeling of space-variation allows

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

to naturally account for anisotropy and non-stationarity. Space-varying covariate information is accounted for via a semiparametric framework. The models can also be extended to space-time data [Bernardi et al. (2017)]. Furthermore, spatial regression with differential regularizations can deal with data scattered over non-planar domains, specifically over two-dimensional Riemannian manifold domains, including surface domains with non-trivial geometries [Ettinger et al. (2016), Dassi et al. (2015), Wilhelm et al. (2016)]. This has fascinating applications in the earth-sciences, life-sciences and engineering. The use of advanced numerical analysis techniques, and in particular of the finite element method or of isogeometric analysis, makes the models computationally very efficient. The models are implemented in the R package fdaPDE [Lila et al. (2016)].

2 Smooth principal component analysis for functional signals over complex domains

Based on the regularized regression models outlined above, I will present a regularized method for principal component analysis of functional signals observed over two-dimensional Riemannian manifold domains [Lila et al. (2016)]. This will be illustrated with an application in the neurosciences, studying neuronal connectivity on the cerebral cortex, starting from functional magnetic resonance imaging scans on about 500 healthy volunteers.

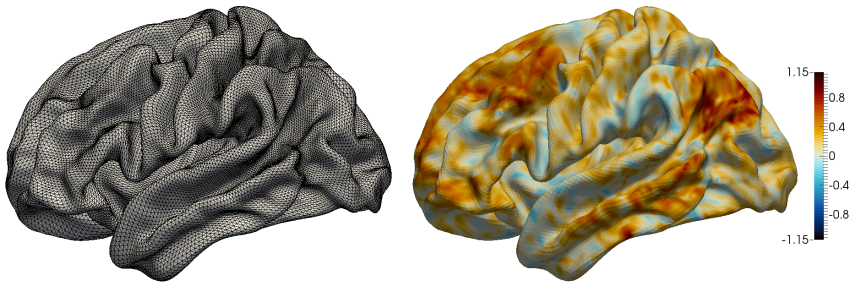


FIGURE 1. Study of high-dimensional neuroimaging signals (data available from The Human Connectome Project Consortium, www.humanconnectomeproject.org). Left: Triangulated surface approximating the left hemisphere of a cerebral cortex. Right: functional connectivity map obtained from fMRI signal. Figure adapted from Lila et al. (2016).

Acknowledgments: This talk is based on joint works with John A.D. Aston, Laura Azzimonti, Mara Bernardi, Bree Ettinger, Michelle Carey, Eardi Lila, Fabio Nobile, Simona Perotto, Jim Ramsay, Piercesare Secchi.

References

Azzimonti, L., Nobile, F., Sangalli, L.M., Secchi, P. (2014). Mixed Finite Elements for spatial regression with PDE penalization. *SIAM/ASA Journal on Uncertainty Quantification*, **2**, 1, pp. 305 – 335.

Azzimonti, L., Sangalli, L.M., Secchi, P., Domanin, M., Nobile, F. (2015). Blood flow velocity field estimation via spatial regression with PDE penalization. *Journal of the American Statistical Association, Theory and Methods*, **110**, 511, pp. 1057 – 1071.

Bernardi, M.S., Sangalli, L.M., Mazza, G., Ramsay, J.O. (2017). A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stochastic Environmental Research and Risk Assessment*, **31**, 1, pp. 23 – 38.

Dassi, F., Ettinger, B., Perotto, S., Sangalli, L.M. (2015). A mesh simplification strategy for a spatial regression analysis over the cortical surface of the brain. *Applied Numerical Mathematics*, **90**, 1, pp. 111 – 131.

Ettinger, B., Perotto, S., Sangalli, L.M. (2015). Spatial regression models over two-dimensional manifolds. *Biometrika*, **103**, 1, pp. 71 – 88.

Lila, E., Aston, J.A.D., Sangalli, L.M. (2016). Smooth Principal Component Analysis over two-dimensional manifolds with an application to Neuroimaging. *Annals of Applied Statistics*, **10**, 4, pp. 1854 – 1879.

Lila, E., Sangalli, L.M., Ramsay, J.O., Formaggia, L. (2016). fdaPDE: functional data analysis and Partial Differential Equations; statistical analysis of functional and spatial data, based on regression with partial differential regularizations, R package version 0.1-4,, http://CRAN.R-project.org/package=fdaPDE.

Sangalli, L.M., Ramsay, J.O., Ramsay, T.O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society Ser. B, Statistical Methodology*, **75**, 4, pp. 681 – 703.

Wilhelm, M., Dede', L., Sangalli, L.M., Wilhelm, P. (2016). IGS: an IsoGeometric approach for Smoothing on surfaces. *Computer Methods in Applied Mechanics and Engineering*, **302**, pp. 70 – 89.

Wilhelm, M., Sangalli, L.M. (2016). Generalized Spatial Regression with Differential Regularization. *Journal of Statistical Computation and Simulation*, **86**, 13, pp. 2497 – 2518.

Minimal Adequate Models: Assessing Uncertainty in Variable Selection

Jelle Goeman[1], Aldo Solari[2]

[1] Leiden University Medical Center, The Netherlands
[2] University of Milano-Bicocca, Italy

E-mail for correspondence: j.j.goeman@lumc.nl

Abstract: We present an alternative approach to variable selection that does not select a single "best" model but attempts to find a collection of models that is "good enough". We call models adequate if they are not significantly worse than the true model. The collection of all adequate models is spanned by a smaller collection of minimal adequate models: the smallest of the adequate models. These minimal adequate models give great insight in model selection uncertainty as well as in collinearity, and are therefore a very practical model building tool. We illustrate the approach with several classical data sets.

Keywords: Model selection; Closed testing; Model misspecification.

1 Variable selection

The goal of variable selection methods in regression is to discard a subset of the covariates without reducing the predictive potential of the remaining variables. Typical variable selection methods find a single "best" model according to a chosen criterion, e.g. AIC or BIC. Variable selection is done to reduce overfit but also for reasons of interpretation. Selected variables are interpreted as important, and discarded variables as irrelevant. Different criteria and different methods, however, can yield very different selected models. Especially when collinearity is present, variable selection methods tend to differ greatly both in which variables are selected and how many.

Clearly, there is uncertainty about the selected model. If we see the single selected model as a point estimate of the true model, then we can say that typical variable selection methods neglect to give standard errors or confidence intervals around the statements they make. Interpreting the selected model in terms of important

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

and irrelevant variables is like interpreting point estimates without an associated measure of uncertainty.

We present a different approach to variable selection that, in contrast, emphasizes the uncertainty in the variable selection process. We adopt a hypothesis testing framework to construct a confidence interval around the true model. Thus, we select not a single model, but a range of models. A model is in the confidence interval if its likelihood is not significantly worse than that of the true model. By construction, the confidence interval contains the true model with probability at least $1 - \alpha$.

Our work builds upon work laid out by Mallows (1973), Aitkon (1974) and Spjotvoll (1977). We show that their work can be seen as a special case of closed testing, which allows their results to be extended outside the scope of linear models e.g. to generalized linear models.

2 Minimal adequate models

The construction of the confidence intervals will be such that if a model is in the confidence set all supersets of the model are also in the confidence set. The confidence set is therefore spanned by its smallest members, the *minimal adequate models*.

The minimal adequate models give great insight in the reliability of inferential statements made as a result of variable selection methods. They can be used to distinguish between variables that must always be selected by a variable selection method and variables that can take each other's roles in the model because they contain the same information, e.g. because of collinearity.

For example, two minimal adequate models can be $\{A, B\}$ and $\{A, C, D\}$. In this case covariate A is necessary for any adequate model, but the role of B can be taken over by the combination of C and D, which together contain the same information. A user may have a preference for model $\{A, B\}$ because it is more sparse or for $\{A, C, D\}$ because it may have a better fit or be more interpretable. In either case the presence of the other minimal adequate model functions as a protection against overinterpretation of the selected model.

3 Model misspecification

A confidence interval for the true model supposes the existence of the true model, which in turn implies that the full model is true. Since this is quite a strong assumption, we will investigate how to relax it. We do this by refining the null hypothesis to be tested for each model: instead of testing whether the reduced model is as good as the true model, we test whether it is as good as the full model. We show that this hypothesis can be conservatively tested even when the full model is not the true model.

4 Application

The use of minimal adequate models will be illustrated with several classical regression data sets, such as Hald's cement data and the famous prostate cancer data set (Hastie et al. 2001).

References

Aitkin, M.A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics*, **16**, 221 – 227.

Hastie, T., Tibshirani, R., and Fiedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

Mallows, E. (1973). Some comments on C_P. *Technometrics*, **15**, 661 – 675.

Spjotvoll, E. (1977). Alternatives to plotting C_P in multiple regression. *Biometrika*, **64**, 1 – 8.

Why bothering about Interval Censoring?

Emmanuel Lesaffre[1], Kris Bogaerts[1], Arnošt Komárek[2]

[1] I-BioStat, KU Leuven, Leuven and UHasselt, Hasselt, Belgium
[2] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

E-mail for correspondence: emmanuel.lesaffre@kuleuven.be

Abstract: We consider here methods to analyze interval-censored survival times. Interval censoring occurs when it is only known that the event happened in-between two examinations. Well-known examples of an interval-censored time are the time until HIV, AIDS, the emergence of a tooth, etc. Most often interval censoring is not appropriately addressed in a statistical analysis and dealt with by methods that handle right censoring of data, e.g. by replacing the interval by the mid-point. Despite several published results it is still too often believed that ignoring the interval-censored character of the data has a minimal impact on the results and conclusions of the statistical analysis. In this contribution we summarize the literature on interval censoring largely from a practical point of view under the frequentist and a Bayesian paradigm. It will be also discussed when it is important to take interval censoring into account.

Keywords: Bayesian inference; Interval censoring; Survival analysis.

1 Introduction

In survival studies, right censoring is most prevalent and generally dealt with appropriately. Occasionally also left censoring occurs, but in randomized controlled trials and epidemiological studies interval censoring occurs frequently. Left censoring occurs, e.g., in a dental study on emergence of permanent teeth when a tooth emerged prior to the start of the study. An emergence time is then interval-censored when it is only known that the tooth emerged in-between two examinations. Interval censoring also occurs often in HIV/AIDS studies, where time to HIV seroconversion and AIDS are usually determined at planned visits to the clinical researcher. In fact many developments on interval censoring find their origin in HIV/AIDS research. Finally, in cancer trials progression free survival can only be established in the hospital at planned visits. Despite the

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

frequent occurrence of interval censoring, this interval censoring is often treated inappropriately in practice. Note that left and right censoring can be considered as special cases of interval censoring. Also, in practice most often a mix of the censoring types is encountered.

Often interval censoring is bypassed with a single imputation technique, with mid-point imputation being most popular. That is, while the interval-censored survival time is replaced by the mid-point of the interval, the data are analyzed using methods for right-censored data. The effect of inappropriately dealing with interval censoring depends on the size of the intervals, on whether covariates impact the size of the interval and the type of statistical analysis. In the past, absence of statistical software was the main reason for avoiding interval censoring. This is not the case anymore nowadays. This is clearly illustrated in the forthcoming book on interval censoring (Bogaerts, Komárek and Lesaffre, 2017), hereafter referred to as BKL.

One distinguishes case I interval-censored data, also called current status data. This occurs in practice when it is only known whether the event has happend or not at the time of examination. We concentrate in this contribution on case II interval censoring. Namely, we assume that an independent sample of survival times $T_1, \ldots, T_n$ is only observed to lie in intervals $\lfloor l_i, u_i \rfloor$ $(i = 1, \ldots, n)$, where $\lfloor l_i$ means that either l_i is included or not in the interval and the same for u_i. By allowing l_i to be zero, interval censoring reduces to left censoring. On the other hand right censoring is a special case of interval censoring when $u_i = \infty$ (in practice this implies a large value). Further we assume that the censoring mechanism is independent of the true survival times. We will return to this assumption in Section 5.

A popular data set in the statistical literature on interval-censored data comes from a breast cancer study. It consists of the subset of 96 patients who were treated at the Joint Center for Radiation Therapy in Boston between 1976 and 1980. Forty-six patients were randomized to radiation therapy only regimen, while 48 patients to the radiation therapy and adjuvant chemotherapy regimen. The intervals represent the time period during which breast retraction occurred. A graphical representation of the data is shown in Figure 1. Illustrations will also be taken from the Signal Tandmobiel® study, which is a longitudinal dental study that examined, e.g., the emergence distributions of several permanent teeth.

2 Univariate models

2.1 Frequentist approaches

Let the true survival times $T_1, \ldots, T_n$ be i.i.d. with survival distribution $S(\cdot)$. When the survival times are interval-censored with intervals $\lfloor l_i, u_i \rfloor$ $(i = 1, \ldots, n)$, the likelihood to maximize is:

$$L = \prod_{i=1}^{n} \{S(l_i) - S(u_i)\}, \tag{1}$$

with $S(t)$ the unknown but true survival distribution. Peto (1973) was the first to note that the nonparametric maximum likelihood solution of S results in a set

of intervals $\{[p_j, q_j]\}_{j=1}^{m}$ with the following properties: outside these intervals, the estimated survival function is constant. Further, the mass assigned to each of the intervals is well determined but within each interval there is no information as to how that mass is assigned. The intervals are called *regions of possible mass or support* because the maximum likelihood procedure can only tell in which regions there is probability of events to occur. Peto (1973) and Turnbull (1976) suggested a simple reduction algorithm to identify the intervals of possible mass from the data. Further, Turnbull (1976) suggested the *self-consistency algorithm*, a version of the EM algorithm, to determine the *nonparametric maximum likelihood estimator (NPMLE)* of S. Thus, in contrast to the Kaplan-Meier estimator, the NPMLE of the survival function for interval-censored data has no closed solution and must be obtained by an iterative algorithm.

Two versions of the NPMLE are given in Figure 2 obtained from the patients treated with radiotherapy alone in the breast cancer study. The left panel is the NPMLE of the cumulative distribution function of the time to cosmetic deterioration of the breast. Fourteen regions of possible support were found but only to eight regions mass > 0 has been attributed. In Table 1, these intervals are shown. The gray areas indicate that the distribution of probability within the regions of support is not determined. In the right panel the corresponding estimated survival distribution is given but assuming a linear behavior of $\widehat{S}$ in the intervals.

Since the seminal papers of Peto and Turnbull, the classical significance tests in survival analysis to compare two or more groups (logrank test, Gehan-Wilcoxon test, Peto-Prentice-Wilcoxon test, etc.) have been extended to interval-censored observations.

Because for a long time Turnbull's algorithm was not available in statistical software, it was standard to show the Kaplan-Meier estimate based on singly imputed survival times. Alternatively and depending on the application area, also a para-

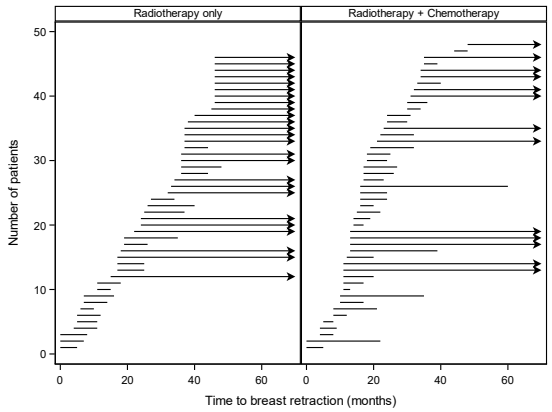


FIGURE 1. Breast cancer study. Observed intervals in months for time to breast retraction of early breast cancer patients per treatment group.

TABLE 1. Breast cancer study. Regions of possible support and NPMLE equivalence classes for the radiotherapy-only group.

$(p_j, q_j]$	(4, 5]	(6, 7]	(7, 8]	(11, 12]	(15, 16]	(17, 18]	(24, 25]
mass	0.046	0.033	0.089	0.071	0	0	0.093

$(p_j, q_j]$	(25, 26]	(33, 34]	(34, 35]	(36, 37]	(38, 40]	(40, 44]	(46, 48]
mass	0	0.082	0	0	0.121	0	0.466

metric estimate was computed. In medical applications the most popular choices are the Weibull and the log-normal distribution. Computations and inference are simpler in the parametric case relying often on Newton-Raphson type of algorithms and standard asymptotic likelihood theory. In-between nonparametric and parametric approaches are flexible estimation methods. Numerous techniques have been proposed that either smooth the hazard, the cumulative hazard or the survival distribution. Popular in this sense is spline smoothing based on cubic splines, B-splines or penalized B-splines adapted to interval-censored data. A smooth solution can also be obtained from, say, a mixture of Gaussian densities for the survival density. Examples of these approaches with software applications in R and SAS software can be found in BKL.

2.2 Bayesian approaches

Parametric analysis of interval-censored observations is fairly standard in classical Bayesian statistical software, such as Win/OpenBUGS or SAS, as long as

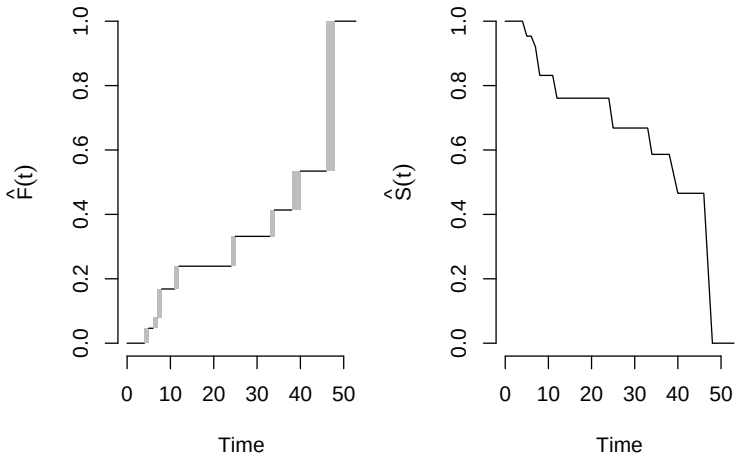


FIGURE 2. NPMLE of the cumulative distribution function (left panel) and NPMLE of the survival function with the additional assumption of a piecewise linear survival curve (right panel).

the chosen survival distribution is supported by the package. More complicated is to perform a Bayesian nonparametric (BNP) analysis. BNP estimation of a cumulative distribution function (and thus of the survival distribution) started with the seminal paper of Ferguson (1973), who introduced the Dirichlet process (DP) prior. The DP prior $D(cS^*)$ is a prior on the survival distributions S defined around a guess survival distribution S^* with variability ruled by a scalar c. Based on a DP prior, Susarla and Van Ryzin (1976) proposed a nonparametric Bayesian approach to estimate the survival function for right-censored survival times. Calle and Gómez (2001) further extended the procedure to interval-censored data. Limiting cases of the posteriors are the Kaplan-Meier for right-censored observations and Turnbull's estimate for interval-censored observations. Until recently no generally available software was available for the BNP approach, this changed with the R package **DPpackage** (Jara, 2007). In the supplementary materials of BKL, some self-written R programs can be found for fitting survival distributions in a nonparametric way as well as illustrations of the use of **DPpackage**.

3 Regression models

Of more interest are survival models that allow for covariates, say $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. In that case the likelihood becomes:

$$L = \prod_{i=1}^{n} \{S(l_i \mid \boldsymbol{X}_i) - S(u_i \mid \boldsymbol{X}_i)\}. \tag{2}$$

While for right-censored data the Cox proportional hazards (PH) model takes a central position because the partial likelihood approach renders the estimation of the baseline hazard obsolete, with interval-censored survival times the baseline hazard/distribution needs to be estimated together with the regression coefficients. We consider here the PH model and the accelerated failure time (AFT) model for interval-censored survival times. Again a variety of approaches were suggested from semiparametric to parametric.

3.1 Frequentist approaches

The likelihood to maximize for the PH model is given by

$$L(\boldsymbol{\beta}, S_0) = \prod_{i=1}^{n} \left\{ S_0(l_i)^{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})} - S_0(u_i)^{\exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})} \right\}, \tag{3}$$

with $\boldsymbol{\beta}$ a p-vector of regression parameters and $S_0(t)$ the baseline survival function.

Finkelstein (1986) extended the nonparametric approach of Turnbull to the proportional hazards model with interval-censored data. For this she assumed the model expressed in (3). Note that likelihood (3) depends only on the baseline hazard through its values at the different observation time points. Let $s_0 = 0 < s_1 < \ldots < s_{K+1} = \infty$ denote the ordered distinct time points of all observed time intervals $\lfloor l_i, u_i \rfloor$ $(i = 1, \ldots, n)$. Further, let $\alpha_{ij} = I\{s_j \in \lfloor l_i, u_i \rfloor\}$ $(j = 1, \ldots, K+1,$

$i = 1, \ldots, n$). To remove the range restrictions on the parameters for S_0, the likelihood is parameterized by $\gamma_k = \log[-\log S_0(s_k)]$ $(k = 1, \ldots, K+1)$. Note that because $S_0(s_0) = 1$ and $S_0(s_{K+1}) = 0$, $\gamma_0 = -\infty$ and $\gamma_{K+1} = \infty$. In terms of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^\top$ the log-likelihood function $\ell(\boldsymbol{\beta}, S_0)$ can be written as

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K+1} \alpha_{ik} \left[\mathrm{e}^{-\zeta_{k-1} \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})} - \mathrm{e}^{-\zeta_k \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})} \right] \right\}, \tag{4}$$

where $\zeta_k = \sum_{m=0}^{k} \exp(\gamma_m)$.

To estimate the mass of the regions of possible support and the regression parameters, Finkelstein proposed the Newton-Raphson algorithm. It turned out that the score equations provide a generalization of the self-consistency algorithm suggested by Turnbull when $\boldsymbol{\beta} = \mathbf{0}$. For the appropriateness of the asymptotic χ^2-distribution for testing $\boldsymbol{\beta} = \mathbf{0}$, it is assumed that K does not increase with the sample size. Farrington's approach (1996) allows to fit the approach of Finkelstein with generalized linear model software. He also provided a technique to select a subset of $L < K$ significant time points s_l $(l = 1, \ldots, L)$. Other approaches suggested for the PH model are: the piecewise exponential model (available in SAS procedure **ICPHREG**) and a variety of smooth approaches some based on spline smoothing.

Another popular semiparametric approach is to apply the partial likelihood approach on multiple imputed (MI) data sets. In the MI approach finite interval-censored survival times are regarded as missing and replaced by a possible survival time given an assumed model. Standard methodology can then be used to analyze the (often between 3 and 10) imputed data sets of right-censored survival times. The results of the multiple analyses are then combined. Pan (2000) proposed two such multiple imputation schemes assuming a particular distribution within the regions of support, but no other assumptions are made.

Finally, here again a parametric approach is easiest to handle, but could be too restrictive in practice.

For the AFT model, the true survival times $T_1, \ldots, T_n$ are assumed to satisfy

$$Y_i = \log(T_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i \qquad (i = 1, \ldots, n), \tag{5}$$

where ε_i are independent and identically distributed with density $g(e)$. In case of interval-censored survival times the likelihood is given by (2) but now with $S(t \mid \boldsymbol{X}) = S_0\{\exp(-\boldsymbol{X}^\top \boldsymbol{\beta})\, t\}$.

Several distributions have been suggested for g, but there exists no semiparametric version of the AFT model. While the parametric AFT model is the simplest to handle it is often too restrictive. More flexible approaches are based on a smooth error density. One option is the Penalized Gaussian Mixture (PGM) model, which assumes that the error density is a mixture of a (large) number of Gaussian densities with fixed means (knots) and with weights that are constrained by a penalty term to produce a smooth density. This approach has been implemented in the R package **smoothSurv**. In Figure 3 the solution from **smoothSurv** is compared to the NPMLE for the emergence distribution of tooth 44 from boys.

3.2 Bayesian approaches

For a long time, the Bayesian PH model could only be fit parametrically to interval-censored observations making use of the statistical packages Win/OpenBUGS and later with the SAS procedure **MCMC**. Unfortunately, a pure semiparametric approach does not seem to be possible here, but recently at least two flexible modelling approaches have been proposed and implemented in software. Wang et al. (2013) proposed a Bayesian PH model with a piecewise constant baseline hazard via a reversible jump MCMC procedure in combination with data augmentation. Their approach fits a dynamic survival model to the data, thereby providing a check for the PH assumption. The method is implemented in the R package **dynsurv**. The recently developed R package **ICBayes** is based on fitting the baseline hazard in a smooth manner using integrated I-splines, see Lin et al. (2015). To this end the relationship of the PH model with a latent nonhomogeneous Poisson process was used in combination with data augmentation.

The parametric Bayesian AFT model can be fitted with BUGS-like and SAS software in very much the same manner as the PH model. We are only aware of the R package **bayesSurv** to fit a smooth AFT model in a Bayesian way. The package is based on the reversible jump MCMC technique, but also a PGM as an error distribution can be fitted. With the package **DPpackage** practitioners have several programs at their disposal for fitting Bayesian AFT models in a semiparametric manner. In the package a Mixture of Dirichlet process prior is used to fit an AFT model for interval-censored observations, which is the basis for the R function **DPsurvint**. This function was applied to examine the dependence of the emergence distribution of a permanent tooth (tooth 44) on gender of a child and the history of caries status of the predecessor deciduous tooth 84 expressed by its dichotomized DMF score (DMF=1, caries on the deciduous tooth, 0 otherwise). Figure 4 shows the posterior predictive survival function for the different gender and DMF combinations.

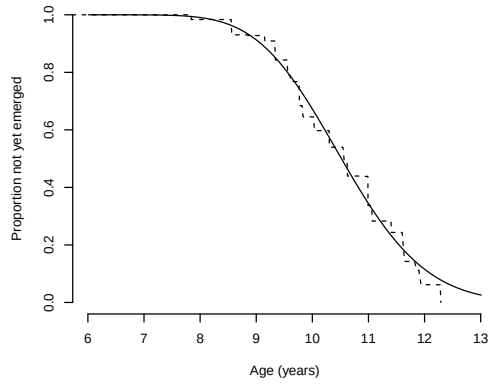


FIGURE 3. Signal Tandmobiel® study (boys). Estimated survival function compared to NPMLE (dashed line) for the time to emergence of tooth 44 estimated using the penalized Gaussian mixture with R package **smoothSurv**.

4 Multivariate models

When several outcomes are measured on a subject who is examined at regular time intervals, we obtain multivariate interval-censored observations. With multivariate outcomes, it is natural to ask for the association between the outcomes. Most of developments have been done for the bivariate case, i.e., when there are two related survival times T_1 and T_2 measured in an interval censored manner. A special case of bivariate interval-censored data are doubly interval-censored times. In that case the T_1 measures the onset of the time-at-risk and $T_2 \geq T_1$ measures the time of the event, and again both T_1 and T_2 are interval-censored.

4.1 Frequentist approaches

Betensky and Finkelstein (1999a) generalized Peto's and Turnbull's argument to bivariate interval censored data. That is, information on the bivariate nonparametric survival function is limited to a number of rectangles bearing (possibly) non-zero mass, again called the regions of possible support. The trigger to develop the bivariate NPMLE of S for interval-censored observations, was the computation of the association between the two true survival times. However, it turned out that the bivariate NPMLE is not a good basis for this because too dependent on the amount right censoring in the data, see Betensky and Finkelstein (1999b).

The absence of statistical software for fitting a rich class of (multi/bi)variate models (for interval-censored data), restricts the use of parametric modelling for multivariate responses. Instead one could use copula models, which disentangle the specification of the association structure and the marginal distributions. The three popular copulas: the Clayton copula, the Gaussian copula and the Plackett copula have been extended to bivariate interval-censored survival times and are implemented in the function **fit.copula** from the R package **icensBKL**

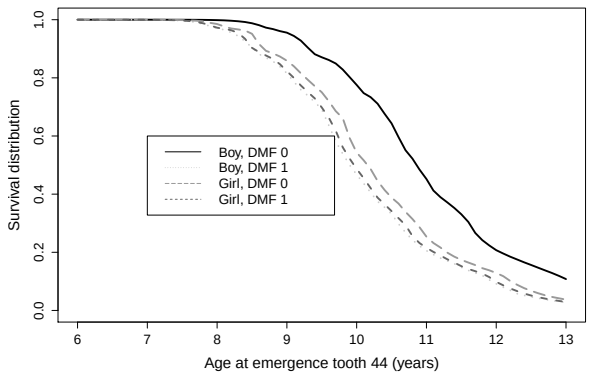


FIGURE 4. Signal Tandmobiel® study. Survival functions for emergence of permanent tooth 44 in gender by DMF groups based on a semiparametric Bayesian model with MDP priors, obtained with R function **DPsurvint**.

(will accompany the BKL book). Even more flexible are the bivariate smoothing techniques, such as the bivariate PGM model implemented in the SAS macro **%smooth**. This macro produced Figure 5 that shows a smooth approximation of the distribution of the true emergence times of the contralateral (left and right) maxillary first premolars (teeth 14 and 24) for boys, collected in the Signal Tandmobiel® study. One can observe that the two emergence times are highly correlated.

For survival outcomes the association measures Spearman's rank correlation, Kendall's tau and the global and local cross-ratio are in use. These measures can be estimated by plugging in sample values in the population versions of the associations. This can be done for parametric models, but when based on the PGM approach a goodness-of-fit check for the parametric models is obtained.

To graphically explore the association structure of multivariate observations one can use a biplot. On a biplot the original p-dimensional outcome is projected onto 2 (or 3) dimensions displaying individuals as points and variables as vectors. If the 2-dimensional plot captures most of the original variability, then the projections of the points on the vectors provide useful visual information on the characteristics of (groups of) individuals. The biplot has been extended to multivariate interval-censored observations (Cecere et al., 2013) and implemented in the function **IC-Biplot** of the package **icensBKL**.

Hierarchical models, called frailty models in the survival context, provide yet another way to model multivariate interval-censored outcomes. Conditional on a random intercept, the outcomes are then assumed independent. This class of models has been also extended to the interval-censored case. Again various illustrations of methodologies and software can be found in BKL.

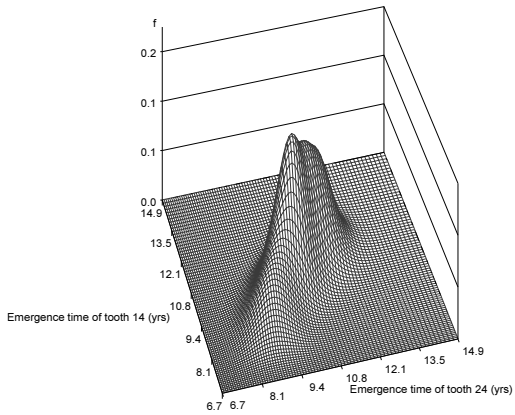


FIGURE 5. Signal Tandmobiel® study. Density of penalized normal mixture model for emergence of permanent teeth 14 and 24 obtained from SAS using macro **%smooth**.

4.2 Bayesian approaches

Parametric frailty models can be fit with standard Bayesian software such Win/OpenBUGS and the SAS procedure **MCMC**. More challenging is to fit multivariate models for interval-censored data in a semiparametric manner. A few approaches have been suggested to fit the frailty distribution in a flexible manner. The approach of Komárek and Lesaffre (2007) builds on the penalized Gaussian mixture idea. Let $(T_{i1}, \ldots, T_{in_i})^\top$ be independent random vectors representing times-to-event of the i-th cluster which are observed as intervals $\lfloor l_{il},\, u_{il} \rfloor$ and $\boldsymbol{X}_{il}$ be the covariate vector for the lth observation in the ith cluster ($i = 1, \ldots, n;, \, l = 1, \ldots, n_i$). In the random-effects AFT model, the (i, l)-th event time is expressed as

$$\log(T_{il}) = \boldsymbol{X}_{il}^\top \boldsymbol{\beta} + b_i + \varepsilon_{il} \qquad (i = 1, \ldots, n;\, l = 1, \ldots, n_i), \tag{6}$$

where ε_{il} are (univariately) i.i.d. random errors with a density g_ε and $b_1, \ldots, b_n$ are cluster-specific i.i.d. random-effects with a density g_b. The approach then consists in expressing either of densities g_ε and g_b as a univariate PGM. The approach was implemented in the R package **bayesSurv** and illustrated with data from the Signal Tandmobiel® study. More specifically the software was used to examine the impact of caries (now or in the past) of deciduous teeth and their successors. Another option is to use the package **DPpackage**, which provides functions that allow for a multivariate semiparametric approach.

5 Discussion

The list of statistical approaches extended to deal with interval-censored data is endless. In fact, each statistical approach developed for fully observed or right-censored data can be extended to interval-censored data. Additional topics that have been investigated with interval-censored data: competing risks, multi-state models, interval-censored covariates, etc. We also omitted here the discussion of doubly interval-censored observations, important for HIV/AIDS research.

Finally, the majority of the developments (if not all) have been done under the assumption of non-informative independent censoring. This assumption is violated when the censoring intervals are associated with the actual and unobserved time-to-event. This may happen more often in practice than assumed, and may affect the conclusions considerably. Developments that deal with informative censoring are therefore desirable.

To conclude, there is no reason anymore to bypass interval censoring since there is ample software available for a great variety of problems.

References

Betensky, R. A. and Finkelstein, D. M. (1999a). A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, **18**, 3089 – 3100.

Betensky, R. A. and Finkelstein, D. M. (1999b). An extension of Kendall's coefficient of concordance to bivariate interval censored data. *Statistics in Medicine*, **18**, 3101 – 3109.

Bogaerts, K., Komárek, A. Lesaffre, E. (2017). *Survival Analysis with Interval-Censored Data: A Practical Approach with examples in R, SAS and BUGS*. Boca Raton: CRC/Chapman and Hall

Calle, M. L. and Gómez, G. (2001). Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods. *Journal of Statistical Planning and Inference*, **98**, 73 – 87.

Cecere, S. and Groenen, P. J. F. and Lesaffre, E. (2013). The interval-censored biplot. *Journal of Computational and Graphical Statistics*, **22**, 123 – 134.

Jara, A. (2007). Applied Bayesian non- and semiparametric inference using DP-package. *R News*, **7**, 17 – 26.

Farrington, C. P. (1996). Interval censored survival data: A generalized linear modelling approach. *Statistics in Medicine*, **15**, 283 – 292.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209 – 230.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**, 845 – 854.

Komárek, A. and Lesaffre, E. (2007). Bayesian accelerated failure time model for correlated censored data with a normal mixture as an error distribution. *Statistica Sinica*, **17**, 549 – 569.

Lin, X. and Cai, B. and Wang, L. and Zhang, Z. (2015). A proportional hazards model for interval-censored failure time data. *Lifetime Data Analysis*, **21**, 470 – 490.

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, **56**, 199 – 203.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, **22**, 86 – 91.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, **71**, 897 – 902.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290 – 295.

Wang, X. and Chen, M.-H. and Yan, J. (2013). Bayesian dynamic regression models for interval censored survival data with application to children dental health. *Lifetime Data Analysis*, **19**, 297 – 316.

Language comprehension as a multiple label classification problem

R. Harald Baayen[1], Tino Sering[1], Cyrus Shaoul[2], Petar Milin[3]

[1] Eberhard Karls University of Tübingen, Germany
[2] Landmark College, USA
[3] University of Sheffield, UK

E-mail for correspondence: harald.baayen@uni-tuebingen.de

Abstract: The initial stage of language comprehension is a multi-label classification problem. Listeners or readers, presented with an utterance, need to discriminate between the intended words and the tens of thousands of other words they know. We propose to address this problem by pairing a network trained with the learning rule of Rescorla and Wagner (1972) with a second network trained independently with the learning rule of Widrow and Hoff (1960). The first network has to recover from sublexical input features the meanings encoded in the language signal, resulting in a vector of activations over all meanings. The second network takes this vector as input and further reduces uncertainty about the intended meanings. Classification performance for a lexicon with 52,000 entries is good. The model also correctly predicts several aspects of human language comprehension. By rejecting the traditional linguistic assumption that language is a (de)compositional system, and by instead espousing a discriminative approach (Ramscar, 2013), a more parsimonious yet highly effective functional characterization of the initial stage of language comprehension is obtained.

Keywords: multi-label classification, language comprehension, error-driven learning, Rescorla-Wagner, Widrow-Hoff

Table 1 presents 10 simple sentences. When reading these sentences, the letters and their combinations succeed in bringing to the fore a small number meanings while dismissing thousands of others as irrelevant. Sentences present the reader with a multi-label classification problem.

We address this problem as follows. First, we represent the orthographic input by means of letter trigrams. For the first sentence, these are `#Ma Mar ary ry# y#p #pa pas ass sse sed ed# d#a #aw awa way ay#` (the `#` symbol represents the space character). Letter trigrams provide a much richer

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

representation of the visual input than do orthographic words. For the data in Table 1, there are $n = 104$ distinct letter trigrams, to which we refer as cues.

The second column lists the lexical meanings (lexomes) that are the targets of classification. Lexomes are pointers to locations in a high-dimensional semantic vector space (defined below). Note that past-tense word forms such as *passed* (regular) and *ate* (irregular) are coupled with the lexomes PASS and EAT as well as with past tense (PAST). Likewise, the two word forms *apple* and *pie* are coupled with one lexome APPLEPIE, and the three expressions with the word forms *kicked the bucket*, *passed away*, and *died*, are all linked with the same lexome DIE.

TABLE 1. Sentences, lexomes in the message, and frequency of occurrence (F). The total number of learning events is $k = 771$.

	Sentence	Lexomes in the message	F
1	Mary passed away	MARY DIE PAST	40
2	Bill kicked the ball	BILL KICK PAST DEF BALL	100
3	John kicked the ball away	JOHN KICK PAST DEF BALL AWAY	120
4	Mary died	MARY DIE PAST	300
5	Mary bought clothes for the ball	MARY BUY PAST CLOTHES FOR DANCEPARTY	20
6	Ann bought a ball	ANN BUY PAST INDEF BALL	45
7	John filled the bucket	JOHN FILL PAST DEF BUCKET	100
8	John kicked the bucket	JOHN DIE PAST	10
9	Bill ate the apple pie	BILL EAT DEF APPLEPIE	3
10	Ann tasted an apple	ANN TASTE PAST INDEF APPLE	33

Is it possible to discriminate between the targeted lexomes given the letter trigrams in the sentences? We will show that considerable headway can be made by an error-driven incremental multi-label classifier that comprises two simple networks, each with only an input layer and an output layer. In what follows, we first provide a formal definition of the algorithm, and illustrate it for the sentences in Table 1. We then turn to a more realistic example in which lexomes targeted in around a million of utterances have to be discriminated from some 52,000 other lexomes.

1 An algorithm for multiple label classification

The problem of incremental learning of multi-label classification is defined by a sequence of events at which a set of features (henceforth cues) are present and generate predictions about classes (henceforth outcomes), only some of which are actually present in the learning event. The mismatch between predicted outcomes and the outcomes actually present in a learning event provides the error driving learning.

From a total of n distinct cues and m possible outcomes, only small subsets will be present in a given learning event. Let k denote the number of unique learning events (learning events may repeat, cf. *good morning* and *tickets please*). We index a specific learning event in the sequence $\mathbf{t}$ (of length $K \geq k$) of learning events by t. The classification problem is defined by $\mathbf{t}$, a sparse $n \times k$ cue matrix $\mathbf{C}$ which

is 1 whenever a given cue is present in a specific event and zero otherwise, and a sparse $m \times k$ target matrix $\mathbf{T}$ that is 1 whenever an outcome is present and zero otherwise.

Classification proceeds in two steps, using two networks. The first network has cues as inputs and outcomes as outputs. It is defined by an $m \times n$ matrix $\mathbf{W}$ of connection weights from cues (columns) to outcomes (rows). Given $\mathbf{W}$, the predicted support (henceforth activation) for a specific outcome given the cues in the learning event is obtained by summation of the weights on the connections from these cues to that outcome. The $m \times k$ activation matrix $\mathbf{A}$ specifies these activations for all outcomes across all unique learning events:

$$\mathbf{A} = \mathbf{WC}.$$

The classification performance of this first network is assessed by checking whether the outcomes with the highest activations are those of the targeted lexomes.

As shown by Danks (2003), if over a sequence of learning events no further changes in the weight matrix take place other than the tiny increments and decrements that come with individual updates, i.e., when the weight matrix has entered a state of equilibrium, then, given the incremental learning rule of Rescorla and Wagner (1972) (see below), $\mathbf{W}$ can be estimated straight from conditional probabilities characterizing the input. Let $\mathbf{E}$ specify pairwise conditional probabilities of cues given cues,

$$\mathbf{E} = \begin{pmatrix} \Pr(c_0|c_0) & \Pr(c_1|c_0) & \dots & \Pr(c_n|c_0) \\ \Pr(c_0|c_1) & \Pr(c_1|c_1) & \dots & \Pr(c_n|c_1) \\ \dots & \dots & \dots & \dots \\ \Pr(c_0|c_n) & \Pr(c_1|c_n) & \dots & \Pr(c_n|c_n) \end{pmatrix},$$

and let $\mathbf{F}$ denote a matrix specifying conditional probabilities of outcomes given cues,

$$\mathbf{F} = \begin{pmatrix} \Pr(o_0|c_0) & \Pr(o_1|c_0) & \dots & \Pr(o_n|c_0) \\ \Pr(o_0|c_1) & \Pr(o_1|c_1) & \dots & \Pr(o_n|c_1) \\ \dots & \dots & \dots & \dots \\ \Pr(o_0|c_n) & \Pr(o_1|c_n) & \dots & \Pr(o_n|c_n) \end{pmatrix}.$$

Danks' equilibrium equations state that

$$\mathbf{F} = \mathbf{EW}^T,$$

which can be solved using the generalized inverse.

When a weight matrix is calculated in this way, the effect of the exact order of learning events is lost. Furthermore, a Danks weight matrix dampens the consequences of the frequencies of occurrence of cues and outcomes in the input space, while highlighting the contrasts that allow cues to discriminate between outcomes. Thus, the Danks weight matrix is useful when there is no information on the sequence of learning events (e.g., when only the frequency of learning events is available but not their order) and when interest is directed specifically to an idealised endstate of learning.

Preferably, the weight matrix $\mathbf{W}$ is estimated by repeated application of the learning rule of Rescorla & Wagner (1972) to the learning events $\mathbf{t}$. The update at learning event t,

$$\mathbf{W}^t = \mathbf{W}^{t-1} + \mathbf{\Delta}_{rw}$$

depends on the learning rate η (typically set at 0.001) regulating the magnitude of the changes to the weight matrix, on the predictions for the outcomes as gauged by the activations of these outcomes given the cues, and on whether the outcomes are actually present in the learning event. Specifically, let $\mathbf{c}$ denote the transpose of that column vector of $\mathbf{C}$ specifying which cues are present at the current learning event t, and let $\mathbf{o}$ denote the transpose of that column vector of $\mathbf{T}$ detailing which outcomes are present at t, and let $\mathbf{J}$ denote an $m \times n$ all-ones matrix. Let the (row) vector $\mathbf{a}_1$ to specify the activations of those outcomes that are present in the learning event while setting to zero the activations for all other outcomes:

$$\mathbf{a}_1 = (((\mathbf{J} \cdot \mathbf{o})^T \cdot \mathbf{c})^T \cdot \mathbf{W})\mathbf{i}.$$

Here, $\mathbf{i}$ is a row unit vector of length n. Note that $((\mathbf{J} \cdot \mathbf{o})^T \cdot \mathbf{c})^T$ is 1 for all cue-outcome combinations that are present in the learning event, and zero elsewhere. Next, let the (row) vector $\mathbf{a}_0$ represent the activations of those outcomes not present in the learning event, again given the cues in that learning event, and let it be zero for all other outcomes:

$$\mathbf{a}_0 = (((\mathbf{J} \cdot [\mathbf{1} - \mathbf{o}])^T \cdot \mathbf{c})^T \cdot \mathbf{W})\mathbf{i}.$$

$((\mathbf{J} \cdot [\mathbf{1} - \mathbf{o}])^T \cdot \mathbf{c})^T$ is 1 for all cue-outcome pairs where the cue is present but the outcome not, and zero elsewhere. The update to the weight matrix, $\mathbf{\Delta}_{rw}$, can now be defined as follows:

$$\mathbf{\Delta}_{rw} = \eta\{((\mathbf{J} \cdot \mathbf{o})^T \cdot \mathbf{c})^T \cdot (\mathbf{1} - \mathbf{a}_1) - ((\mathbf{J} \cdot [\mathbf{1} - \mathbf{o}])^T \cdot \mathbf{c})^T \cdot \mathbf{a}_0\}.$$

For cue-outcome pairs that are both in the learning event, the update of their weight is given by the difference from the maximal activation, 1 by definition. As the summed activations $\mathbf{a}_1$ tend to be less than 1, weights will be strengthened. For cue-outcome pairs where the cue is present but the outcome is not, the corresponding connection weight is decreased by the summed activations $\mathbf{a}_0$. Estimation of $\mathbf{W}$ using incremental updating over the sequence of learning events is fast, first because only parts of the weight matrix require updating (efferent weights from cues not present in the learning event are left untouched), and also because the updates to individual outcomes are independent and hence allow for parallelization.

The activation matrix $\mathbf{A} = \mathbf{WC}$ specifies, for each unique learning event and for each outcome, the joint support provided by the cues in that learning event for that outcome.

Although class predictions based on $\mathbf{A}$ can do well for small constructed data sets, they lack precision for large real data sets. Prediction accuracy can be further improved by a second network that is given the task to predict the target $\mathbf{T}$ from the activation matrix $\mathbf{A}$:

$$\mathbf{T} = \mathbf{DA}.$$

The prediction matrix

$$\mathbf{P} = \mathbf{DA}$$

is the resulting approximation of $\mathbf{T}$. Although $\mathbf{D}$ can be calculated using the generalized inverse of $\mathbf{A}$, computation costs can be prohibitive for large numbers

of learning events. It is therefore preferable to estimate $\mathbf{D}$ as follows:

$$\begin{aligned} \mathbf{T} &= \mathbf{DA} \\ \mathbf{TA}^T &= \mathbf{DAA}^T \\ \mathbf{Y} &= \mathbf{DX}, \end{aligned}$$

which leads to $\mathbf{D} = \mathbf{YX}^{-1}$. Since $\mathbf{X}$ is $m \times m$, and since generally $m \ll k$, computational costs are much lower when calculating $\mathbf{X}^+$ as compared to calculating $\mathbf{A}^+$.

The prediction matrix can also be estimated iteratively by means of the update rule of Widrow and Hoff (1960). This update rule, which specifies the update $\mathbf{\Delta}_{wh}$ to the $m \times m$ second weight matrix $\mathbf{D}$, is important, first, as it allows us to assess the consequences of how the order of learning events affects classification, and second, because for large numbers of training events (in the order of hundreds of millions), it is not feasible to actually calculate $\mathbf{A}$ (and $\mathbf{P}$).

Let $\mathbf{Z}$ denote an $m \times m$ matrix initialized with zeroes, let $\mathbf{a}$ denote the column vector of the activation matrix $\mathbf{A}$ giving the predicted activations for the current learning event, and let $\mathbf{o}$ denote the transpose of the corresponding column vector of the target matrix $\mathbf{T}$. The Widrow-Hoff update to $\mathbf{Z}$ is:

$$\mathbf{\Delta}_{wh} = \eta\{\mathbf{a}(\mathbf{o} - \mathbf{a}^T\mathbf{Z})\}.$$

We take the transpose to obtain $\mathbf{D} = \mathbf{Z}^T$.

The weights for the two networks ($m \times n$ for the Rescorla-Wagner network, and $m \times m$ for the Widrow-Hoff network) can be estimated in two ways. One possibility is to first estimate $\mathbf{W}$ and then estimate $\mathbf{D}$. Alternatively, one can update both networks in tandem for each successive learning event. In this case, it is not necessary to calculate $\mathbf{A}$. Note that when estimating

$$\mathbf{P} = (\mathbf{WC})^+\mathbf{TWC}$$

we 'inject' error twice: once during the estimation of $\mathbf{W}$ and again during the estimation of $\mathbf{P}$.

The equilibrium equations are implemented in the `ndl` package for `R` on CRAN. An efficient Python implementation for incremental learning of $\mathbf{W}$ is available at `github.com/quantling/pyndl`. An implementation of incremental learning for `R` is available (for `linux` only) upon request from the authors. Software for efficient updating of $\mathbf{D}$ by Widrow-Hoff is currently under development.

Returning to the example of Table 1, first consider classification performance when $\mathbf{W}$ and $\mathbf{D}$ are estimated independently, using incremental updating for the former, and the generalized inverse for the latter. In this case, for each of the 10 sentences, the lexomes in that sentence have the highest prediction values in $\mathbf{P}$.

When the two networks are updated in tandem, with at each learning event first an update of $\mathbf{W}$ and then an update of $\mathbf{D}$, accuracy varies with the (random) order in which the 771 learning events are made available to the model. For one such random order, the proper lexomes had the highest ranks in $\mathbf{A}$ for 9 out of 10 sentences. The one sentence with an error is *John kicked the bucket*, where DEF (the lexome for the definite article) intrudes with a higher activation before DIE, which is found at the next rank (4).

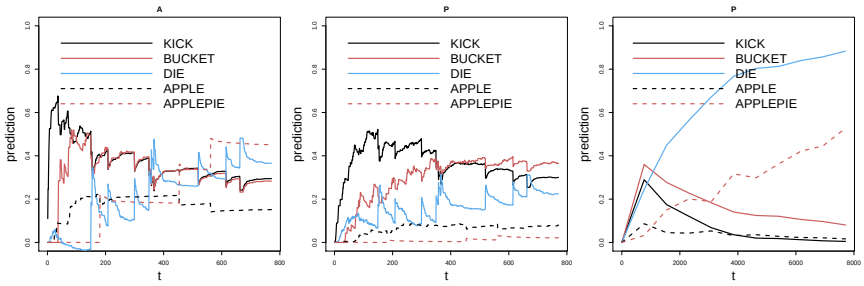


FIGURE 1. Prediction strengths for selected lexomes in the learning events of sentences 8 and 9 in Table 1, using incremented coupled Rescorla-Wagner and Widrow-Hoff. Left and center panels: frequencies as in the table; right panel: frequencies increased tenfold.

Figure 1 illustrates this incremental training regime. The left and center panels show the predictions based on **A** and **P** when training proceeds on a random order of 771 learning events, and the right panel when training proceeds on 7710 learning events. Solid lines represent key lexomes from sentence 8 in Table 1: KICK and BUCKET for the unintended literal reading and DIE for the intended idiomatic reading. Dashed lines represent the competitors APPLE and APPLEPIE in sentence 9. The spiky behavior in the left and center panels reflects the learning and unlearning that unfolds as outcomes competing for the same cues are encountered. Comparison of the left and center panels shows that the Rescorla-Wagner network learns much faster than the Widrow-Hoff network. By the end of the learning sequence, the former, but not the latter network succeeds in giving the intended lexomes higher prediction scores. The rightmost panel shows that with sufficient experience, the model learns that *kick the bucket* means DIE, and that an *apple pie* is not an apple but a particular kind of pie.

An important property of this approach to language comprehension is that the correct lexomes are selected without any worries about regular or irregular verbs, literal versus idiomatic expressions, finding boundaries between words, decomposing words into parts, or disambiguating homographs. Given the assumption that understanding drives the recalibration of weights, the rich information available in the combinatorics of sublexical cues and lexomes is sufficient for multiple label classification to be effective.

2 Multiple label classification with 52,000 classes

To clarify whether this approach scales up, we applied our algorithm to the TASA corpus (Zeno, 1995), a collection of texts comprising a total of 10,807,146 words representing 52,401 word types. Lemmatization was carried out with `TreeTagger` (`www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`), which distinguished 90,339 lemmata, of which 37,938 occurred once. To keep computations tractable, the model was trained on all words occurring at least twice and 351 hapax legomena that occurred in a precompiled list of words. Hapax legomena that were not included were replaced by the dummy word `HAPAX`, resulting in a total of 52,401

lexomes. Learning events were sentences in the TASA corpus. Sequences of more than 8 words were split at the next available occurrence of *and* or *or*. This resulted in a total of 992,752 learning events. The multi-label classification challenge is to predict the appropriate lexomes (out of 52,401) given the letter trigrams of the (possibly inflected) words in the learning events.

Using the `ndl2` package for R, **W** (52,401 lexomes × 11,724 letter trigrams) was estimated using all learning events. To keep computations tractable for the second network, two learning events were selected randomly from a precompiled list of 8866 targeted lexomes, resulting in a total of 17,455 learning events (in 276 cases there was overlap with two or more lexomes in the same event, and for one word, there was only 1 learning event available). The total number of outcomes in this subset of learning events was 19,020. With these restrictions, the matrices **A** (19,020 lexomes × 17,455 learning events), **D** (19,020 × 19,020 lexomes) and **P** (19,020 lexomes × 17,455 events) could be estimated straightforwardly.

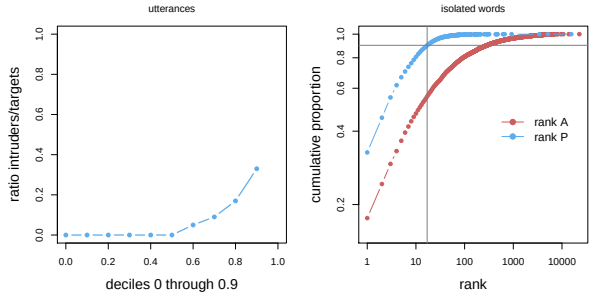


FIGURE 2. Left: Quantiles of the ratio of intruders (false positives) to targets (correct identifications), full utterances. Right: Rank and corresponding cumulative proportion based on **A** (red) and **P** (blue), isolated words.

The left panel of Figure 2 presents the ratio of intruders (lexomes with an activation exceeding that of the least activated target lexome) to the number of targeted lexomes. The median number of intruders is zero, at the 8th decile the ratio is 0.17, and at the 9th, it is 0.33. At the 10th decile, we find cases with vast numbers of intruders, leading to a maximal ratio of 1208.9. Examples of intruders are *down* for the sentence *The aleuts were housed in abandoned rundown gold mines or fish canneries*, and *field* and *success* for the sentence *He is an ecologist who studied succession in abandoned cornfields*.

We also tested identification performance when target lexomes were presented in isolation. The right panel of Figure 2 plots in blue cumulative proportion (out of a total of 7179) against rank based on **P**: 34% of lexomes had the highest prediction value, 88% of the targeted lexomes had at most a rank of 16 (indicating 15 intruders with higher activations). As show by the red curve, performance based on **A** instead of **P** is substantially worse. Human lexical decision performance, as gauged using the British Lexicon Project (BLP, Keuleers et al. 2012) was for the present data at 90% correct. As the lexical decision task does not require actual identification, but only sufficient evidence for lexicality, it appears that human subjects tolerate around 16 intruders.

As shown in Figure 3, the model also predicts power-transformed lexical decision

response times ($t' = -1000t^{-1}$). For all but the first decile, log activation $a_i = \mathbf{W}\mathbf{c}_i$ (with $\mathbf{c}$ the vector specifying the present and absent cues in the input, and i indexing a specific lexome) shows a nearly linear effect with negative slope. Log rank prediction (the log rank of $p_i = \mathbf{D}\mathbf{W}\mathbf{c}_i$) has a smaller effect that is again negative and nearly linear, but now for the first nine deciles. The 90% decile of the rank is at 18, which is close to the cut-off at rank 17 for lexicality decisions in the right panel of Figure 2. Apparently, the same range of ranks influences both decisions and reaction times.

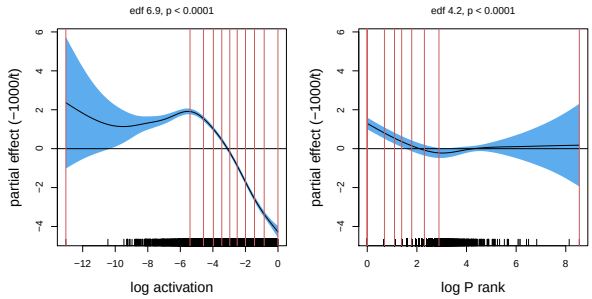


FIGURE 3. Partial effects in a GAM fitted to power-transformed ($-1000t-1$) reaction times. Left: log activation; Right: log prediction rank. Vertical lines denote deciles. The 90% decile of log prediction rank is at rank 18 (red lines indicate deciles). Regression analyses were carried out with GAMs (Wood, 2006).

$\mathbf{P}^T$ defines a semantic vector space (cf. Landauer & Dumais, 1997), and lexomes are indices or pointers for locations in this space. By way of illustration of the semantic nature of $\mathbf{P}^T$, the left panel of Figure 4 presents partial effects for human semantic similarity ratings for word pairs (Bruni et al., 2014) as predicted from correlations of the corresponding column vectors of $\mathbf{P}^T$ (left). For 90% of the data points, a nearly linear relation is observed. Clearly, extreme values are unreliable as predictors. Similarity in $\mathbf{P}^T$-space, i.e., similar prediction values across events and thus greater similarity of experiences communicated, correctly predicts greater perceived semantic similarity.

The column vectors of $\mathbf{D}$ also define a lexomic space, but similarities in this space turn out to be positively correlated with the Levenshtein distance between the orthographic forms of the two words. As shown in the right panel of Figure 4, the more different two word forms are, the lower their perceived semantic similarity.

3 Concluding remarks

Multi-label classification is a hard problem, not only for statistics, but also for humans. For instance, in auditory word recognition, isolated words taken from conversational speech have recognition rates between 20% and 40% (Arnold et al., 2017). In the visual lexical decision task, undergraduate students perform near chance on the lower-frequency words (Baayen et al., 2017). From this perspective, the model's performance, with training on a mere 10 million words, is too good to be true. This is, of course, due to the model being given perfect feedback,

whereas human learning tends to proceed under uncertainty and lack of full understanding.

Given that the model presents a simplified perspective on the first stage of comprehension — understanding the words — several of its features are remarkable. First, the traditional linguistic assumption that language is a (de)compositional system is replaced by a perspective in which the language signal is a code that discriminates between possible messages (Ramscar 2013, Shannon, 1956).

Second, the model is parsimonious with only one free parameter, the learning rate η. And although **W** and **D** can be very large, most of the weights are close to zero. E.g., for **W**, only 5,885 weights exceed 0.1 (0.00058% of the total number of weights), and only 195 weights are greater than 0.5. Arnold et al. (2017) show for auditory comprehension that **W** can be pruned down to a fraction of the original weights without noticeable loss of accuracy.

Third, the classifier implements a three-layer network that differs from back-propagation networks in that there is direct error injection twice, once for **W** using the Rescorla-Wagner equations, and once for **D**, using Widrow-Hoff (or the generalized inverse). Importantly, the power of the first network should not be underestimated. Although ever since the criticism of the perceptron by Minsky & Papert (1972), two-layer networks have been regarded as far too restricted for any classification tasks requiring more than the simplest linear separation, it turns out that actually, with an appropriate choice of cues, Rescorla-Wagner networks can solve much more interesting problems. Figure 5 illustrates this for a simple example with two classes (represented by gray and red points) that in $R \times R$ are not linearly separable (left panel). When the data are re-represented by identifiers for rows and columns (right panel), a Rescorla-Wagner network correctly predicts the highest activations for around 210 of the 260 elements of the red class (see Baayen and Hendrix, 2017, for detailed comparison with other machine learning classifiers, and also Ghirlanda, 2005).

Fourth, more sophisticated features than letter trigrams can be used as cues, such as the frequency band summary features used by Arnold et al. (2017) for modeling auditory word recognition, and for reading the histogram of oriented

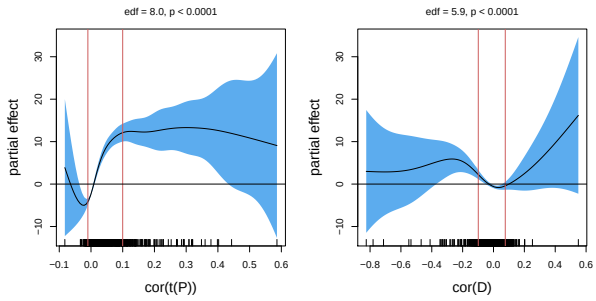


FIGURE 4. Partial effects of the correlations of row vectors of **P** (left) and column vectors of **D** as predictors of human similarity ratings for 2,369 word pairs. Red vertical lines indicate 5% and 95% percentiles. Regression analyses were carried out with GAMs (Wood, 2006).

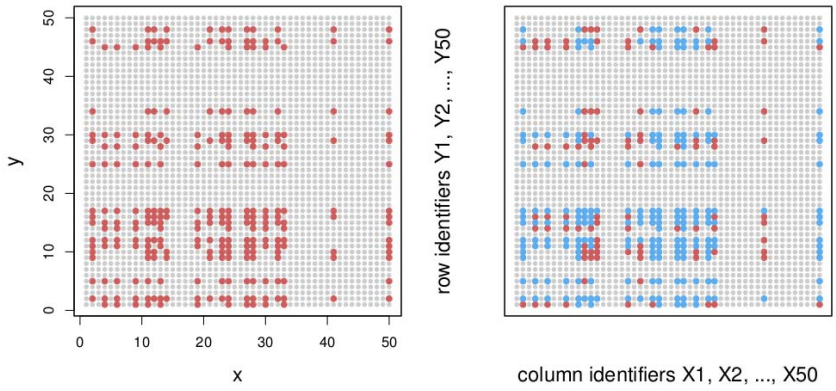


FIGURE 5. A non-linearly separable classification problem with a majority class in gray (2240) and a minority class in red (260). Left: data points in a Cartesian grid ($x = 1, 2, \ldots, 50; y = 1, 2, \ldots, 50$). Right: rerepresentation with row and column identifiers as cues for a Rescorla-Wagner network: hits in blue, misses and false alarms in red.

gradients feature descriptor proposed by Dalal and Triggs (2005).

Finally, the model is transparent to interpretation. $\mathbf{W}$ specifies the support provided by sublexical features for lexomes. $\mathbf{D}$ transforms activation vectors that are still strongly influenced by form similarity into vectors closer to the targeted lexomes, which in turn results in a semantic vector space, $\mathbf{P}^T$.

References

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., and Baayen, R.H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS-ONE*, **12**, e0174623.

Baayen, R. H., and Hendrix, P. (2017). Two-layer networks, non-linear separation, and human learning. In Wieling, M., Kroon, M., van Noord, G., and Bouma, G. (Eds.) From Semantics to Dialectometry. Festschrift in honor of John Nerbonne. London, College Publications, 13–22.

Baayen, R. H., Tomaschek, F., Gahl, S., and Ramscar, M. (2017). The Ecclesiastes principle in language change. In Hundt, M., Mollin, S., and Pfenninger, S. (Eds.) *The changing English language: Psycholinguistic perspectives*. Cambridge, Cambridge University Press.

Bruni, E. and Tran, N.K. and Baroni, M. (2014). Multimodal distributional semantics, *Journal of Artificial Intelligence Research*, **49**, 1–47.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR'05*, volume 1, 886–893.

Danks, D. (2003). Equilibria of the RescorlaWagner model. *Journal of Mathematical Psychology*, **47**, 109–121.

Ghirlanda, S. (2005). Retrospective revaluation as simple associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, **31**, 107–111.

Keuleers, E. and Lacey, P. and Rastle, K. and Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words, *Behavior Research Methods*, **44**, 287–304.

Landauer, T.K. and Dumais, S.T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, **104**, 211–240.

Minsky, M. and Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: The MIT Press.

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, **46**, 377–396.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: A. H. Black and Prokasy, W. F. (Eds.), *Classical conditioning II: Current research and theory*. New York: Appleton Century Crofts.

Shannon, C. E. (1956). The bandwagon, *IRE Transactions on Information Theory*, **2**, 3.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. *1960 WESCON Convention Record* Part IV, 96–104.

Wood, S. (2016). *Generalized additive models*. New York: Chapman & Hall.

Zeno, S.M. and Ivens, S.H. and Millard, R.T. and Duvvuri, R. (1995). *The educator's word frequency guide*. New York: Touchstone Applied Science.

Part II – Contributed Papers

The Multivariate Method of Simulated Quantiles for Portfolio Optimisation

Paola Stolfi[1], Mauro Bernardi[2], Lea Petrella[3]

[1] Department of Economics, Roma Tre University, Italy
[2] Department of Statistical Sciences, University of Padova, Italy
[3] MEMOTEF Department, Sapienza University of Rome, Italy

E-mail for correspondence: paola.stolfi@uniroma3.it

Abstract: The sparse multivariate method of simulated quantiles is proposed as a likelihood–free alternative to indirect inference procedures that does not rely on an auxiliary model specification. We extend the asymptotic theory and we show that the sparse–MMSQ estimator enjoys the oracle properties under mild regularity conditions. The method is applied to estimate the parameters of the Skew Elliptical Stable distribution with the aim of finding optimal portfolios within a Mean–Value–at–Risk objective function.

Keywords: Projectional quantiles; Sparse regularisation; Skew Elliptical Stable distribution; Portfolio optimisation.

1 Introduction

In this paper we extend the method of simulated quantiles (MSQ) of Dominicy and Veredas (2013) to a multivariate framework (MMSQ). The MMSQ is then applied to estimate the parameters of the distribution of a portfolio of asset returns with the purpose of allocating a fixed amount of preexisting wealth among alternative risky assets. The asset allocation problem requires the prior selection of an appropriate distribution for modelling the multivariate structure of financial returns being characterised by the presence of skewness, heavy–tails and positive tail–dependence which may prevent the existence of the moments. The Skew–Elliptical Stable distribution (SESD) introduced by Branco and Dey (2001) extends the elliptical Stable distribution using the skewing mechanism of Azzalini and Dalla Valle (1996) and it is particularly useful to model skewed and heavy–tailed continuous data. Despite their prominent role in the financial literature, multivariate Stable distributions have not been massively employed because of

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

the inferential issues that can arise when fitting these distributions to data. The lack of a closed form expression for the density function as well as moments finiteness prevent either the use of valid likelihood–based inferential procedures and the moment–matching estimators. Furthermore, the inferential problem becomes even more relevant as the dimension of the distribution increases. The method of simulated quantiles (MSQ), like alternative likelihood–free procedures, is based on the minimisation of the distance between appropriate quantile–based statistics evaluated on the true and simulated data. Specifically, the MSQ estimates the vector of unknown parameters ϑ by solving the following minimisation problem

$$\hat{\vartheta} = \arg\min_{\vartheta} \left(\hat{\mathbf{\Phi}}_{\mathbf{y}} - \tilde{\mathbf{\Phi}}^{R}_{\vartheta} \right)' \mathbf{W}_{\vartheta} \left(\hat{\mathbf{\Phi}}_{\mathbf{y}} - \tilde{\mathbf{\Phi}}^{R}_{\vartheta} \right), \tag{1}$$

where $\hat{\mathbf{\Phi}}_{\mathbf{y}}$ is the vector of quantile–based statistics evaluated on the available sample of observations $\mathbf{y} = (y_1, y_2, \ldots, y_T)$, while $\tilde{\mathbf{\Phi}}^{R}_{\vartheta} = \frac{1}{R} \sum_{t=1}^{R} \tilde{\mathbf{\Phi}}^{r}_{\vartheta}$ is the average of the vector of quantile–based statistics $\tilde{\mathbf{\Phi}}^{r}_{\vartheta}$ based on the r–th vector of simulations from the postulated model $\mathbf{y}^r = (y_1^r, y_2^r, \ldots, y_T^r)$ with $r = 1, 2, \ldots, R$, and $\mathbf{W}_{\vartheta}$ is a positive definite weighting matrix. The success of the method strongly depends on a careful selection of the vector of quantile–based measures $\mathbf{\Phi}$ even in the univariate setting where quantiles are well defined. The lack of a natural ordering in the multivariate setting instead requires an accurate definition of the concept of quantile. Here, we rely on the notion of projectional quantile recently introduced by Hallin et al. (2010) and Kong and Mizera (2012). This notion of multivariate quantile makes the estimator flexible and it allows us to deal with non–elliptically contoured distributions.

An issue frequently observed in high dimensions is the curse of dimensionality, i.e., the situation where the number of parameters grows quadratically or exponentially with the dimension of the problem. In those circumstances, the right identification of the sparsity pattern becomes crucial since it reduces the number of parameters to be estimated. Those reasonings motivate the use of sparse estimators that automatically shrink to zero some parameters, such as, for example, the off diagonal elements of the variance–covariance matrix. In this paper, we penalise the quadratic objective function of the MMSQ by adding a smoothly clipped absolute deviation (SCAD) ℓ_1–penalisation term that shrinks to zero the off–diagonal elements of the scale matrix of the postulated distribution. We extend the asymptotic theory in order to accommodate sparse estimators, and we prove that the resulting sparse–MMSQ estimator enjoys the oracle properties of Fan and Li (2001) under mild regularity conditions.

The remainder of the paper is organised as follows. Section 2 briefly describes the MMSQ and its sparse extension. Section 3 introduces the multivariate Skew–Elliptical distribution and discusses some of its properties which are useful for the application to the portfolio optimisation problem. We apply the MMSQ to the portfolio optimisation problem in Section 4. Portfolio optimisation has a long tradition in finance initiated by the Markowitz's (1952) seminal paper that introduced the mean–variance (MV) approach. The MV approach relies on quite restrictive conditions about the underlying distribution of asset returns that are relaxed here by assuming that financial returns follow a multivariate SESD. Moreover, since these distributions do not admit finite second moment, we consider a portfolio allocation problem where the expected return is traded–off against higher Value–at–Risk (VaR) profiles that make investment less attractive.

2 The multivariate method of simulated quantiles

In this section we briefly describe the MMSQ. To this aim, we first introduce the following definition of projectional quantiles. Let $\mathbf{Y} \in \mathbb{R}^d$ be random vector, $\mathbf{u} \in \mathbb{S}^{d-1}$ be a vector in the unit sphere and $\tau \in (0,1)$, then the $\tau\mathbf{u}$ projectional quantile of $\mathbf{Y}$ is defined as

$$q^{\tau\mathbf{u}} \in \left\{ \arg\min_{q \in \mathbb{R}} \Psi_{\tau\mathbf{u}}(q) \right\}, \tag{2}$$

where

$$\Psi_{\tau\mathbf{u}}(q) = \mathbb{E}\Big[\rho_\tau \left(\mathbf{u}'\mathbf{Y} - q \right) \Big], \tag{3}$$

and $\rho_\tau(z) = z\left(\tau - \mathbf{I}_{(-\infty,0)}(z)\right)$. Clearly the $\tau\mathbf{u}$–projectional quantile is the τ–quantile of the univariate random variable $\mathbf{u}'\mathbf{Y}$. This feature makes the definition of projectional quantile particularly appealing in order to extend the MSQ to the multivariate setting because, once the direction is properly chosen, it reduces to the univariate quantile. The multivariate method of simulated quantiles is now introduced. Let $F_{\mathbf{Y}}(\cdot, \vartheta)$ be the distribution function of $\mathbf{Y}$ which depend on a vector of unknown parameters $\vartheta \subset \Theta \in \mathbb{R}^k$ and let:

- *(i)* $\mathbf{q}_\vartheta^{\boldsymbol{\tau}\mathbf{u}} = (q_\vartheta^{\tau_1\mathbf{u}}, q_\vartheta^{\tau_2\mathbf{u}}, \ldots, q_\vartheta^{\tau_s\mathbf{u}})$ be a $s \times 1$ vector of projectional quantiles at given confidence levels $\tau_i \in (0,1)$ with $i = 1, 2, \ldots, s$, and $\mathbf{u} \in \mathbb{S}^{d-1}$;
- *(ii)* $\boldsymbol{\Phi}_{\mathbf{u},\vartheta} = \boldsymbol{\Phi}\left(\mathbf{q}_\vartheta^{\boldsymbol{\tau}\mathbf{u}}\right)$ be a $b \times 1$ vector of quantile functions assumed to be continuously differentiable with respect to ϑ for all $\mathbf{Y}$ and measurable for $\mathbf{Y}$ and for all $\vartheta \subset \Theta$;
- *(iii)* $\hat{\mathbf{q}}^{\boldsymbol{\tau}\mathbf{u}} = (\hat{q}^{\tau_1\mathbf{u}}, \hat{q}^{\tau_2\mathbf{u}}, \ldots, \hat{q}^{\tau_s\mathbf{u}})$ and $\hat{\boldsymbol{\Phi}}_{\mathbf{u}} = \boldsymbol{\Phi}\left(\hat{\mathbf{q}}^{\boldsymbol{\tau}\mathbf{u}}\right)$ be the corresponding sample counterparts;

and assume that $\boldsymbol{\Phi}_{\mathbf{u},\vartheta}$ cannot be computed analytically but it can be empirically estimated on simulated data. At each iteration $j = 1, 2, \ldots$ the MMSQ compute $\tilde{\boldsymbol{\Phi}}^R_{\mathbf{u},\vartheta_j} = \frac{1}{R}\sum_{r=1}^R \tilde{\boldsymbol{\Phi}}^r_{\mathbf{u},\vartheta_j}$, where $\tilde{\boldsymbol{\Phi}}^r_{\mathbf{u},\vartheta_j}$ is the function $\boldsymbol{\Phi}_{\mathbf{u},\vartheta}$ computed at the r–th simulation path from $F_{\mathbf{Y}}\left(\cdot, \vartheta^{(j)}\right)$. The parameters are subsequently updated by minimising the distance between the vector of quantile measures calculated on the true observations $\hat{\boldsymbol{\Phi}}_{\mathbf{u}}$ and that calculated on simulated realisations $\tilde{\boldsymbol{\Phi}}^R_{\mathbf{u},\vartheta_j}$. The subscript $\mathbf{u}$ denotes that those quantities depend on a set of directions, that should be selected in such a way that they fully characterise the feature the corresponding parameter of the distribution identifies. We establish consistency and asymptotic normality of the proposed estimator. The MMSQ estimator is then extended in order to achieve sparse estimation of the scaling matrix. Specifically, the SCAD ℓ_1–penalty of Fan end Li (2001) is introduced into the MMSQ objective function as follows

$$\hat{\vartheta} = \arg\min_{\vartheta} \left(\hat{\boldsymbol{\Phi}}_{\mathbf{u}} - \tilde{\boldsymbol{\Phi}}^R_{\mathbf{u},\vartheta}\right)' \mathbf{W}_\vartheta \left(\hat{\boldsymbol{\Phi}}_{\mathbf{u}} - \tilde{\boldsymbol{\Phi}}^R_{\mathbf{u},\vartheta}\right) + n\sum_{i<j} p_\lambda\left(|\sigma_{ij}|\right), \tag{4}$$

where $\mathbf{W}_\vartheta$ is a $b \times b$ symmetric positive definite weighting matrix, $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=1}^d$ is the scale matrix and $p_\lambda(\cdot)$ is the SCAD ℓ_1–penalty:

$$p'_\lambda(\gamma) = \lambda\left\{ \mathbf{I}_{(\gamma \le \lambda)} + \frac{(a\lambda - \gamma)}{(a-1)\lambda}\mathbf{I}_{(\gamma > \lambda)} \right\} \tag{5}$$

3 Multivariate Skew Elliptical Stable distribution

Here, we introduce the definition of Skew–Elliptical Stable distribution introduced by Branco and Dey (2001). Specifically, we consider a slightly different parameterisation from Branco and Dey (2001), having the interesting property that the diagonal elements of the scale matrix do not affect the overall skewness of the distribution. Let $(\mathbf{X}', Y)'$ be a Normal random vector of dimension $(d+1)$ conditional on the latent factor $\zeta \sim \mathcal{S}_{\frac{\alpha}{2}}(\bar{\omega}_\alpha, 1, 0)$ where $\mathcal{S}_{\frac{\alpha}{2}}(\bar{\omega}_\alpha, 1, 0)$ denotes the univariate totally right–skewed α–Stable distribution with scale $\bar{\omega}_\alpha = \left(\cos\frac{\pi\alpha}{2}\right)^{\frac{2}{\alpha}}$ and shape parameter $\alpha \in (0, 2)$, i.e.

$$\begin{bmatrix} \mathbf{X} \\ Y \mid \zeta \end{bmatrix} \sim \mathcal{N}_{d+1}(\mathbf{0}, \zeta\boldsymbol{\Omega}_\delta), \tag{6}$$

where $\boldsymbol{\Omega}_\delta = \begin{bmatrix} \bar{\boldsymbol{\Omega}} & \boldsymbol{\delta} \\ \boldsymbol{\delta}' & 1 \end{bmatrix}$ and $\bar{\boldsymbol{\Omega}}$ is a proper correlation matrix, symmetric and positive definite with $|\sigma_{ij}| < 1$, for $i, j = 1, 2, \ldots, d$ and $i \neq j$. Then the random variable $\mathbf{Z} = (\mathbf{X} \mid Y > 0)$ is Skew Elliptical Stable distributed, i.e., $\mathbf{Z} \sim \mathcal{SESD}_d(\alpha, \mathbf{0}, \bar{\boldsymbol{\Omega}}, \boldsymbol{\delta})$ with density

$$f_{\mathbf{Z}}(\mathbf{z}, \alpha, \mathbf{0}, \bar{\boldsymbol{\Omega}}, \boldsymbol{\delta}) = 2\int_0^{+\infty} \phi_d(\mathbf{z}, \mathbf{0}, \zeta\bar{\boldsymbol{\Omega}})\, \Phi_1\left(\frac{\boldsymbol{\lambda}'\mathbf{z}}{\sqrt{\zeta}}\right) h(\zeta)\, d\zeta, \tag{7}$$

where $\phi_d(\cdot)$ and $\Phi_1(\cdot)$ denote the density of the multivariate Normal distribution and the cumulative density function of the univariate Normal distribution, respectively, $h(\zeta)$ is the density of the mixing variable and $\boldsymbol{\lambda} = \left(1 - \boldsymbol{\delta}'\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}\right)^{-\frac{1}{2}} \bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta} \in \mathbb{R}^d$ denotes the vector of skewness parameters. Furthermore, the transformation $\mathbf{Y} = \boldsymbol{\xi} + \boldsymbol{\omega}\mathbf{X}$ where $\boldsymbol{\omega} = \text{diag}\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_d\}$ is $\mathbf{Y} \sim \mathcal{SESD}_d(\alpha, \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\delta})$. Among the attractive properties of the SESD, the closure under linear combinations is of relevance for the portfolio optimisation problem discussed in the next Section.

4 Portfolio application

We consider a portfolio allocation problem, where, at each time $t = 1, 2, \ldots, T$, the investor's wealth allocation is based on the choice of the vector of optimal portfolio weights $\mathbf{w}_t > 0$ by minimising the following objective function

$$\arg\min_{\mathbf{w}_t} -\mathsf{E}_t(\mathbf{w}_t'\mathbf{Y}_{t+1}) - \kappa\mathsf{VaR}_t^\lambda(\mathbf{w}_t'\mathbf{Y}_{t+1}), \text{ s.t. } \mathbf{w}_t'\mathbf{1} = 1, \tag{8}$$

where $\mathbf{Y}_t \sim \mathcal{SESD}(\alpha, \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\delta})$, $\mathsf{E}_t(\mathbf{w}_t'\mathbf{Y}_{t+1})$ and $\mathsf{VaR}_t^\lambda(\mathbf{w}_t'\mathbf{Y}_{t+1})$ denote the portfolio's expected return and the portfolio Value–at–Risk at level $\lambda \in (0, 1)$ evaluated at time t for the period $(t, t+1]$, respectively. Here, $\kappa \geq 0$ denotes the investor's risk aversion parameter: the larger κ, the higher is the penalisation for the risk profile of the selected portfolio. The empirical application is structured as follows. We consider a basket of weekly returns of seventeen MSCI European indexes, covering the period from January 6th, 1995 to November 25th, 2016. Then, for each week, from April 23, 2010 to the end of the sample period, we estimate the SESD parameters using a rolling windows of $n = 800$ observations

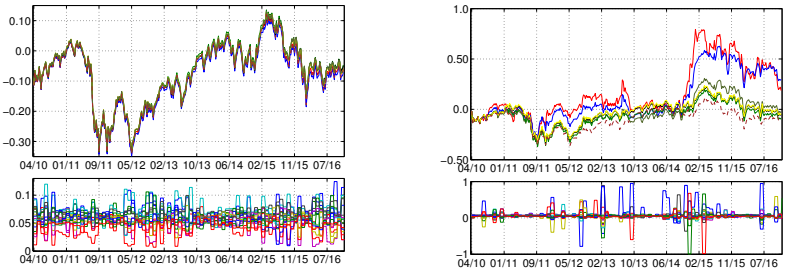


FIGURE 1. Mean–VaR$_{0.95}$ optimal portfolio results over the period from April 23th, 2010 to the November 25, 2016. Figures plot the optimal portfolios cumulative returns for different values of the risk aversion parameter κ in the top panel and the optimal portfolio weights for $\kappa = 10$ in the bottom panel. The dotted thinned brown line represents the equally weighted portfolio cumulative returns which has been added for comparison.

for the MMSQ and of $n = 200$ for the Sparse–MMSQ. The optimal tuning parameters of the SCAD penalty are selected by K–fold cross validation, with $K = 5$. As regards the portfolio allocation exercise, for each window, we solve equation (8) for the vector of optimal allocations $\mathbf{w}_t$, where the portfolio expected returns and VaR are calculated exploiting the closure property with respect of the linear combination of the SESD: $Z_{t+1} = \mathbf{w}_t'\mathbf{Y}_{t+1} \sim \mathcal{SESD}\left(\alpha, \mathbf{w}_t'\boldsymbol{\xi}, \mathbf{w}_t'\boldsymbol{\Omega}\mathbf{w}_t, \boldsymbol{\omega}_Z^{-1}\mathbf{w}_t'\boldsymbol{\omega}\boldsymbol{\delta}\right)$, where $\boldsymbol{\omega}_Z = (\boldsymbol{\Omega}_Z \odot I_h)^{\frac{1}{2}}$ and $\odot$ denotes the Hadamart multiplication. Moreover, the VaR confidence level is fixed at $\lambda = 0.99$ and several levels of investors' risk aversion are considered $\kappa = \{0.10, 0.5, 1.0, 2.0, 5.0, 10.0\}$. We present the results of the empirical portfolio performance evaluation. To this end, we forecast the one–step ahead conditional returns' distribution over the whole sample period. The sequence of predictive distributions delivered by the competing models, are then used to build the mean–VaR optimal portfolios with and without the short selling constraint in equation (8). Figure 1, reporting the cumulative returns for the MMSQ (on the left) and the Sparse–MMSQ (on the right), evidence that, as the risk aversion coefficient κ increases from $\kappa = 0.1$ to $\kappa = 10$, cumulative returns increase as well. This evidence is stronger for the Sparse–MMSQ meaning that the shrinkage effect induced by the estimation method have a positive impact on the estimation of the scale matrix and, as a consequence, the portfolio results greatly benefited from a better estimate of the dependence structure among assets. The bottom panels of Figure 1 plot the evolution over time of the optimal weights. Optimal weighs for the Sparse–MMSQ are characterised by a marked heterogenous behaviour while those implied by the MMSQ are flat and display lower levels of diversification.

References

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.

Branco, M.D. and Dey, D.K. (2001). A general class of multivariate skew-elliptical distributions. *J. Multivariate Anal.*, 79(1):99–113.

Dominicy, Y. and Veredas, D. (2013). The method of simulated quantiles. *J. Econometrics*, 172(2):235–247.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Hallin, M., Paindaveine, D., and Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth. *Ann. Statist.*, 38(2):635–669.

Kong, L. and Mizera, I. (2012). Quantile tomography: using quantiles with multivariate data. *Statist. Sinica*, 22(4):1589–1610.

Markowitz, H.M. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

Gaussian Process model for evolving 3D lip curves

Irene Mariñas[1], Adrian Bowman[1], Vincent Macaulay[1]

[1] School of Mathematics and Statistics, University of Glasgow, UK

E-mail for correspondence: i.marinas.1@research.gla.ac.uk

Abstract: This paper addresses the problem of estimating a ridge curve embedded in a three-dimensional surface that changes over time. The main challenge is to exploit the details of surface shape, while maintaining computational feasibility. A Gaussian Process approach is adopted.

Keywords: Gaussian Processes; Shape analysis; 3D curves; Lip morphology

1 Introduction and motivation

FIGURE 1. Three-dimensional curves representing an upper lip, embedded in a three-dimensional facial image, during the emotion *Disgust*, at time points 3, 13 and 37.

This study is motivated by the shape of the lips in a three-dimensional facial image (surface) and their variation over the expression of different emotions such as disgust, fear, anger, happiness, et cetera (Figure 1). To record the expressions, a large number of pictures of a person producing the emotion are taken with a stereophotogrammetric camera system, which leads to a set of data in four dimensions (the three spatial dimensions plus time). The statistical analysis of information on shape has been a research topic of considerable interest since the earliest part of the twentieth century, but it has developed substantially in the present century especially thanks to advances in computational tools. Interest in shape analysis of the human face began because of its applications in biology, medicine and psychology.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

2 Gaussian Process model for three-dimensional curves evolving over time

A ridge curve embedded in a three-dimensional surface can be expressed in terms of a continuous index, i.e. its arc-length (rescaled to be from 0 to 1), $s \in [0, 1]$, and the discrete label (i.e. coordinate), $c \in \{x, y, z\}$ (Mariñas et al. 2016). If the surface is, moreover, changing over time, a time component $t \in \mathbb{R}$ can be added to the model. A GP can be then specified as:

$$w(t, c, s) \sim GP\big(m(t, c, s), k(t, t', c, c', s, s')\big). \tag{1}$$

Let $\mathbf{s} = (s_1, ..., s_n)^{\mathrm{T}}$ for a choice of n values of s. Each coordinate can be represented as a function of the arc-length and the time: $w(t, x, s) = x(t, s)$, $w(t, y, s) = y(t, s)$ and $w(t, z, s) = z(t, s)$ (Figure 2).

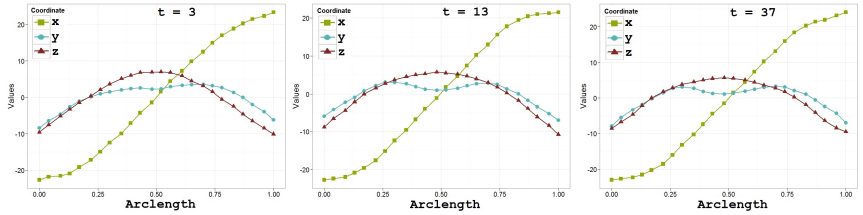


FIGURE 2. 3D upper lip curves, from the emotion *Disgust*. Each coordinate represented as function of the arc-length, at time points 3, 13 and 37.

Then a three dimensional curve at time t can then be notated as:

$$\mathbf{W}(t) = \begin{bmatrix} \mathbf{x}(t) & \mathbf{y}(t) & \mathbf{z}(t) \end{bmatrix}^{\mathrm{T}}, \tag{2}$$

where $\mathbf{x}(t) = (x(t, s_1) \cdots x(t, s_n))^{\mathrm{T}}$, and similarly for $\mathbf{y}(t)$ and $\mathbf{z}(t)$. The sequence for a choice of T values of t, $\mathbf{t} = (t_1, ..., t_T)^{\mathrm{T}}$: $(\mathbf{W}(t_1) \cdots \mathbf{W}(t_T))^{\mathrm{T}} = \mathbf{W} \sim N_{3Tn}(\mathbf{m}, \mathbf{K})$, where $\mathbf{m}$ is the mean, assumed to be zero, and $\mathbf{K}$ is the covariance matrix. Separability is assumed such that: $k(t, t', c, c', s, s') = k_t(t, t')k_c(c, c')k_s(s, s')$ (Rasmussen 2006), i.e. $\mathbf{K} = \mathbf{K}_t \otimes \mathbf{K}_c \otimes \mathbf{K}_s$.

- If the process is assumed Markovian, the Ornstein-Uhlenbeck (OU) covariance function can be used, i.e., $k_t(t, t') = \exp(-|t - t'|/\mu)$, with hyperparameter μ, the time-scale. Hence, $\mathbf{K}_t$ represents the covariance of curves at different time points, with $(i, j)^{th}$ element equal to $k_t(t_i, t_j)$.
- For the 3×3 matrix $\mathbf{K}_c$, two hyperparameters were specified: κ_1, the correlation between x and y or z, and κ_2, between y and z:

$$\mathbf{K}_c = \begin{pmatrix} 1 & \kappa_1 & \kappa_1 \\ \kappa_1 & 1 & \kappa_2 \\ \kappa_1 & \kappa_2 & 1 \end{pmatrix}. \tag{3}$$

- The space-covariance function used is the Squared-Exponential (SE), i.e., $k_s(s, s') = \sigma_f^2 \exp\left(-\frac{1}{2}(s - s')^2/\lambda^2\right)$, with hyperparameters: σ_f^2, the signal variance and λ, the length-scale. Therefore, $\mathbf{K}_s$ represents the covariance matrix for the n arc-length inputs, with $(i, j)^{th}$ element equal to $k_s(s_i, s_j)$.

3 Conditional dependencies, likelihood and predictive distributions

The distribution for the first three-dimensional curve at time $t = 1$, $\mathbf{W}(1) \sim N_{3n}(\mathbf{0}, \mathbf{K}_c \otimes \mathbf{K}_s)$, can be factorised as:

$$\begin{aligned} \mathbf{x}(1) &\sim N_n(\mathbf{0}, \mathbf{K}_s), \\ \mathbf{y}(1) \mid \mathbf{x}(1) &\sim N_n(\kappa_1 \mathbf{x}(1), (1-\kappa_1^2)\mathbf{K}_s), \\ \mathbf{z}(1) \mid \mathbf{x}(1), \mathbf{y}(1) &\sim N_n\big(\left[\{\kappa_1 - \kappa_1\kappa_2\}\,\mathbf{x}(1) + \{\kappa_2 - \kappa_1^2\}\,\mathbf{y}(1)\right] / \left[1-\kappa_1^2\right], \\ &\quad \left[1 - \{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2\} / \{1-\kappa_1^2\}\right]\mathbf{K}_s\big). \end{aligned} \tag{4}$$

Subsequent three-dimensional curves in the sequence can be conditioned on the previous time-point (by the Markov property):

$$\mathbf{W}(t) \mid \mathbf{W}(t-1) \sim N_{3n}\left(\kappa \mathbf{W}(t-1), (1-\kappa^2)\mathbf{K}_c \otimes \mathbf{K}_s\right), \tag{5}$$

where $\kappa = \exp(-1/\mu)$, assuming the time difference between curves is one.

Analogous conditional dependencies between coordinates can be used:

$$\begin{aligned} \mathbf{x}(t) \mid \mathbf{W}(t-1) &\sim N_n\left(\kappa \mathbf{x}(t-1), (1-\kappa^2)\mathbf{K}_s\right), \\ \mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1) &\sim N_n\big(\kappa \mathbf{y}(t-1) + \kappa_1\left[\mathbf{x}(t) - \kappa \mathbf{x}(t-1)\right], \\ &\quad (1-\kappa^2)\mathbf{K}_s(1-\kappa_1^2)\big), \\ \mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{y}(t), \mathbf{W}(t-1) &\sim N_n\Bigg(\kappa \mathbf{z}(t-1) + \frac{1}{1-\kappa_1^2}\left[\{\kappa_1 - \kappa_1\kappa_2\}\{\mathbf{x}(t) - \right. \\ &\quad \kappa \mathbf{x}(t-1)\} + \{\kappa_2 - \kappa_1^2\}\{\mathbf{y}(t) - \kappa \mathbf{y}(t-1)\}\big], \\ &\quad \left.(1-\kappa^2)\left[1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right]\mathbf{K}_s\Bigg). \end{aligned} \tag{6}$$

Given the hyperparameters $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu, \kappa_1, \kappa_2)$, the total log-likelihood of the sequence can be calculated as:

$$\log p(\mathbf{W} \mid \boldsymbol{\theta}) = \log p(\mathbf{W}(1) \mid \boldsymbol{\theta}) + \sum_{i=2}^{T} \log p(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}). \tag{7}$$

At each time point t, $\log p(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) = \log p(\mathbf{x}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) + \log p(\mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1), \boldsymbol{\theta}) + \log p(\mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{y}(t), \mathbf{W}(t-1), \boldsymbol{\theta})$.

Marginal predictions at time $q \in \mathbb{R}$ can be done at a set of test points $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)$ for each coordinate x, y and z, using:

$$\begin{aligned} \mathbf{W}^*(q) \mid \mathbf{W} \sim \big(&[\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s}][\mathbf{M} \otimes \mathbf{K}_c \otimes \mathbf{K}_s]^{-1}\mathbf{W}, \\ &\mathbf{K}_c \otimes \mathbf{K}_{s^*} - [\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s}][\mathbf{M} \otimes \mathbf{K}_c \otimes \mathbf{K}_s]^{-1}[\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s}]^{\mathrm{T}}\big), \end{aligned} \tag{8}$$

where $\mathbf{K}_{s^*s}$ denotes the $n^* \times n$ matrix of the covariances evaluated at all pairs of training and test points, $\mathbf{M}$ has $(i,j)^{th}$ element equal to $\kappa^{|i-j|}$ and

$$\mathbf{L} = \begin{bmatrix} \exp(-|q-1|/\mu) & \exp(-|q-2|/\mu) & \cdots & \exp(-|q-T|/\mu) \end{bmatrix}. \tag{9}$$

The matrix $\mathbf{L}$ will change depending on the value of q. Analogous conditional dependences between coordinates can be calculated.

4 Fitting the model for the emotion *Disgust* and classification of emotions

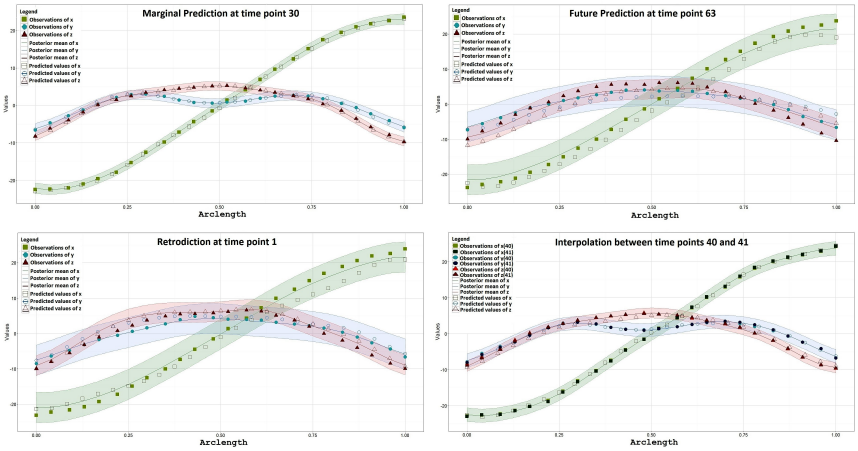


FIGURE 3. Observations, posterior means and predicted values for 3-dimensional lip curves of the emotion *Disgust*.

The model was fitted for the 3D upper lip curves from the emotion *Disgust* (Figure 2), of 61 pictures, i.e. $\mathbf{t} = (1, \ldots, 61)^{\mathrm{T}}$. Optimal hyperparameters were found by maximum likelihood. To ease the optimization process, the number of hyperparameters can be reduced by finding the signal variance, σ_f, that maximises the log-likelihood function analytically. The small difference between time points makes the hyperparameter μ very large. Through experimentation it was observed that the larger the value of μ, the smaller its effect on the likelihood function. Moreover, it was found that μ and κ_2, the correlation between coordinates y and z, were negatively correlated. For this reason, the hyperparameters κ_1 and κ_2 were optimised for a series of individual time points, using a GP model for three-dimensional curves (Mariñas et al. 2016), and then fixed to the mean of these values. Optimization is then carried out for the remaining hyperparameters. An additive normal error of standard deviation 0.5 mm was added to the model of observations to accommodate errors in the observed facial surface. The optimal values found are: $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_f, \hat{\lambda}, \hat{\mu}, \hat{\kappa}_1, \hat{\kappa}_2) = (5.3383, 0.1553, 25.5058, -0.0164, 0.7850)$, with respective SE: $0.0565, 0.0027, 2.5851, 0.0059, 0.0286$. The same time-points ($\mathbf{t} = (1, \ldots, 61)^{\mathrm{T}}$) were considered to make predictions at 25 spatial-points (Figure 3). Retrodiction was done at time $q = -1$, using the data at time point 1. Prediction at time $q = 63$, conditioned on the data from the last curve available, i.e. at time 61, was also made. Marginal prediction was done at time $q = 30$, using the observed values at that time, and interpolation at $q = 40.75$, using the observed values at times 40 and 41. The posterior means are displayed with 2 standard deviations bands (shown dotted). Note how the error bands expand as predictions are further away from the observations.

Methods of classification are being studied to categorise each emotion in terms of the correlation parameters. An initial approach was to perform Principal Com-

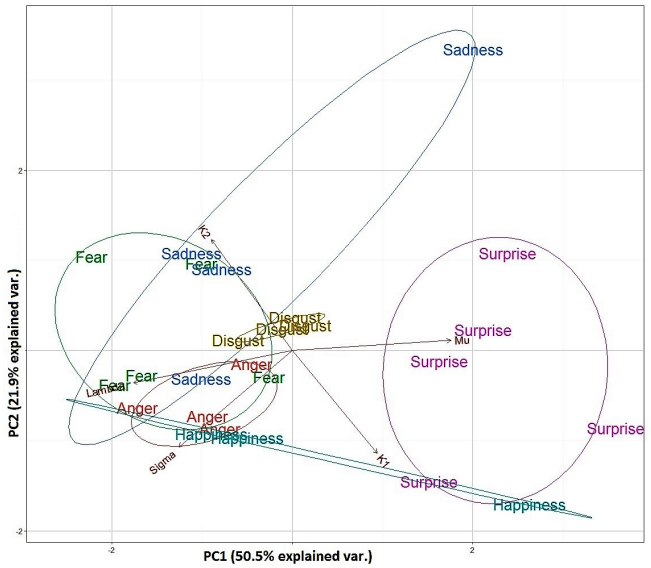


FIGURE 4. Biplot of the (scaled) first two principal components.

ponent Analysis (PCA) (Figure 4). Data consists of six different emotions: anger, disgust, fear, happy, sadness and surprise. There are available sequences from 60 images, in the case of the expression of disgust, to sequences of about 180 images, in the case of happiness. Data were collected by Oliver Garrod and colleagues from the School of Psychology at the University of Glasgow from a 25 year old actress performing each expression between three and five times. There was no stimulation: she based the expressions on the Facial Action Coding System proposed by Paul Ekman (Ekman et al. 1997).

5 Conclusions and further lines of investigation

The use of shape information, expressed in a continuous and multivariate scale raises a number of very interesting issues from a methodological perspective. The lip curve data represent a peculiar scenario due to their high smoothness both spatial and temporally, therefore special adjustments had to be made, which are not needed in other sequences of three-dimensional curves. Nonetheless, the model interpolates the data well and produces accurate predictions. When it comes to the classification of the different emotions using PCA, it can be seen that this is limited by the small number of replicates available. The first two principal components explain together 72.4% of the variability. It is clear from the biplot that *Surprise* is the most different from the rest of the emotions, and *Disgust* the one with less variability between replicates, which can be interpreted as having a series of strong unique characteristics that do not allow much change. For a better understanding of the differences between the emotions, in terms of their correlation parameters, it would be necessary to increase the number of replicates, as well as adding more subjects to the study, to account for the variability across

people. The notion of a shape evolving in time can be extended to a phylogenetic setting, where branching points in the evolution can occur. The aim is to develop statistical methods by which shape information on organisms can be used to reconstruct a phylogenetic tree.

Acknowledgments: IM is grateful to the School of Mathematics and Statistics at the University of Glasgow for her research studentship.

References

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: The MIT Press.

Mariñas, I., Bowman, A. and Macaulay, V. (2016). *Modelling the shape of emotions*. In *Proceedings of the 31st International Workshop on Statistical Modelling*, Dupy, J-F. and Josse, J. (Eds.), pp. 201-206.

Ekman, P. and Rosenberg, E. (1997). *What the face reveals: Basic and aplied studies of spontaneous expressions using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Construction of High-resolution Linkage Maps Using Discrete Graphical Models

Pariya Behrouzi[1], Ernst C. Wit[1]

[1] Johann Bernoulli Institute, University of Groningen, Netherlands

E-mail for correspondence: P.Behrouzi@rug.nl

Abstract: Linkage maps are important for fundamental and applied genetic research. In this article, we introduce an algorithm to construct high-quality and high-density linkage maps for diploid and polyploid species. We employ a sparse Gaussian copula graphical model and the nonparanormal skeptic approach to construct linkage maps. We compare our method with other available method when the data are clean and contain no missing observations and when data are noisy and incomplete. In addition, we implement the method on real genotype data of barley. We have implemented the method in the R package "qploidMap" which is freely available at CRAN.

Keywords: Graphical models; Gaussian Copula; Linkage map; Polyploids.

1 Genetic background on linkage map

A linkage map provides a fundamental resources to understand the order of markers for the vast majority of species whose genome are yet to be sequenced. Furthermore, it is an essential ingredient to identify genes associated with different traits such as, disease resistance in plants or animals. Diploid organisms contain two sets of chromosomes, whereas polyploids contain more than two sets of chromosomes. Here, we refer to diploids and polyploids as q-ploid $q \geq 2$, where in diploids q = 2, triploids q = 3, and so on. The genotype of any q-ploid organism at each single locus on the genome can be either homozygous if all q allele copies of an organism are identical, or heterozygous otherwise.

1.1 Meiosis and Markov dependence.

Linkage mapping is possible because of a biological crossover process, which occurs during meiosis. Assume a sequence of ordered SNP markers $X_1^c, \ldots, X_{p_c}^c$

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

along chromosome c. Due to the *genetic linkage*, neighboring markers across a chromosome are linked. The key biological fact during meiosis is that markers at different chromosomes segregate independently. Given the above mentioned fundamental genetic concepts, X_j^c can be defined as

$$X_j^c = \begin{cases} 1 & \text{parental marker at locus } j, \\ 0 & \text{otherwise.} \end{cases}$$

The sequence $X_1^c, X_2^c, \ldots, X_{p_c}^c$ forms a Markov chain with the state space S where the genotype state of a genetic marker on a genome depends only on the genotype state in the previous genetic marker. Thus, genotype state of marker X_j in following set $\{X_1^c, \ldots, X_{j-1}^c, X_j^c, X_{j+1}^c, \ldots X_{p_c}^c\}$ can be written as follow $P(X_j^c|X_{j-1}^c, X_{j-2}^c, \ldots, X_1^c) = P(X_j^c|X_{j-1}^c)$. This property implies that neighboring markers have conditional dependence relationship. Two complications arise when (i) a population contains heterozygous genotypes, in this case we define $Y_j^c = \sum_{k=1}^{q} X_{jk}^c$, (ii) we do not know what "parental" is, in that case the conditional independence relationships are more complicated. Treating the SNP markers according to the Markovian assumption yields the graphical model. The variables vector $X = \{X_1^c, \ldots, X_p^c\}$, where p is the total number of given markers in a genome, is a discrete graphical model with a joint distribution $P(X)$ which can be factorized as $P(X) = \prod_{c=1}^{C} \prod_{j=1}^{p_c} f_{j,j+1}^{(c)}(x_j^{(c)}, x_{j+1}^{(c)})$. Here, C defines the number of chromosomes in a genome, and p_c stands for the number of markers in chromosome c.

2 Algorithm to construct linkage map

We propose to build a linkage map in two steps. First, reconstruct an undirected conditional independence graph between SNP markers in a genome. Second, determine the correct order of markers in the obtained linkage groups.

2.1 Construct undirected graphical model

To reconstruct the conditional independence graph between SNP markers in a q-ploid species we propose two methods, latent graphical model, and the nonparanormal skeptic. The former method can deals with missing values, whereas the latter is computationally faster.

Latent graphical model

A relatively straightforward approach to discover the conditional independence relationships among markers is to assume underlying continuous variables $Z_1, \ldots, Z_p$ for the markers $Y_1, \ldots, Y_p$ such that $Z_j = \Phi^{-1}(\widehat{F}_j(Y_j))$, where $\Phi^{-1}(.)$ is the standard Gaussian quantile function. We assume $Z \sim N_p(0, \Theta)$ where $\Theta = \Sigma^{-1}$ contains all the conditional independence relationship between SNP markers. Furthermore, we implement the EM algorithm which iteratively finds penalized

maximum likelihood estimation of parameter $\widehat{\Theta}$. Using the extended rank likelihood in the E-step we compute the expected complete penalized log-likelihood as follow

$$Q_\lambda(\widehat{\Theta} \mid \Theta^{(m)}) = \frac{n}{2}[\log|\Theta| - tr(\bar{R}\Theta) - p\log(2\pi)] - \lambda||\Theta||_1$$

where $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} E_{Z^{(i)}}(Z^{(i)}Z^{(i)t}|y^{(i)}, \widehat{\Theta}^{(m)})$, and λ is a nonnegative tuning parameter. To calculate $\bar{R}$ we propose two different approaches, namely Gibbs sampling and approximation method (Behrouzi et al. (2017)). The M-step is the maximization problem which can be solved efficiently using graphical lasso (Fridman et al., 2008).

TABLE 1. Comparison the two methods over 50 independent run, where $q = 2$

		Grouping Accuracy		Ordering Accuracy	
Missing rate	Error rate	DGMmap	MSTMap	DGMmap	MSTMap
p=500 & n=200					
0	0	**1.00** (0.00)	0.55 (0.34)	**1.00** (0.00)	0.90 (0.09)
0.05	0.05	**1.00** (0.00)	0.10 (0.07)	**0.77** (0.04)	0.61 (0.12)
0.10	0.10	**1.00** (0.00)	0.01 (0.00)	**0.60** (0.03)	0.18 (0.23)
0.15	0.15	**1.00** (0.00)	0.01 (0.00)	**0.56** (0.01)	0.00 (0.00)
p=1000 & n=200					
0	0	**1.00** (0.00)	0.61 (0.36)	**1.00** (0.00)	0.91 (0.06)
0.05	0.05	**1.00** (0.00)	0.04 (0.03)	**0.56** (0.00)	0.51 (0.09)
0.10	0.10	**1.00** (0.00)	0.44 (0.16)	0.52 (0.00)	**0.78**(0.02)
0.15	0.15	**1.00** (0.01)	0.05 (0.00)	**0.52** (0.00)	0.60 (0.13)

Nonparanormal SKEPTIC

Alternatively, we use the nonparanormal skeptic approach (Liu et al., 2012) to compute the correlation matrix. In this approach, a sample correlation matrix Γ can be computed from pairwise rank correlations, namely Kendall's tau $\widehat{\tau}_{jl}$, and Spearman's rho $\widehat{\rho}_{jl}$.

$$\widehat{\Gamma}_{jl} = \begin{cases} \sin(\frac{\pi}{2}\widehat{\tau}_{jl}) & j \neq l \\ 1 & j = l \end{cases}, \quad \widehat{\Gamma}_{jl} = \begin{cases} 2\sin(\frac{\pi}{6}\widehat{\rho}_{jl}) & j \neq l \\ 1 & j = l. \end{cases}$$

To estimate the graph we use the graphical lasso. To determine the number of linkage groups we use the EBIC model selection which picks the penalty term that minimizes the EBIC value over $\lambda > 0$.

2.2 Ordering markers in each linkage group

Assume that a set of d markers have been assigned to the same linkage group. Let $G(d, E_d)$ be a sub-graph. To order d markers, the algorithm uses multi-dimentional scaling (MDS) method to post processing of each linkage group. The goal of using the MDS is to find an one-dimensional map such that the distances between markers fit a given set of measured partial correlation that indicate how far markers are.

Table 2 Estimated number of linkage groups (LGs) for OWB data set

	Estimated # LG	Size of the LGs
qploidMap	**7**	**140, 199, 211, 187, 236, 182, 173**
MSTMap	1	1328

Comparison of ordering accuracy between qploidMap and MSTMap. In this Table assumed MSTMAP has estimated correctly the number of LGs in the OWB data set.

Linkage Group (LG)	Sensitivity Score	
	qploidMap	MSTMap
1	0.86	**0.96**
2	**0.78**	0. 52
3	0.78	**0.92**
4	**0.74**	0.49
5	**0.71**	0.38
6	**0.61**	0.50
7	**0.70**	0.61
Average	**0.74**	0.63

Figure 1: Estimated genetic map in qploidMap

3 Data analysis

3.1 Simulations

We set up simulations to generate q-ploid genotype data. We compare the performance of our proposed method with MSTMAP(Wu et al. (2008)) for different ranges of missing rates and genotyping errors when $q = 2$. The results of comparisons are provided in Table 1. We note that high values of the grouping and ordering accuracy scores indicate good performance. These results suggest that the proposed approach performs well compare with the other method.

3.2 Construct linkage map in *Barley*

A barley genotyping dataset is used in the literature to compare different map construction methods for real-world diploid data. This genotyping dataset is generated from a doubled haploid population which allows to achieve homozygous individuals, $Y = \{0, 1\}$. Barley genotype data is the result of crossing Oregon Wolfe Barley Dominant with Oregon Wolfe Barley Recessive. The Oregon Wolfe Barley (OWB) data includes $p = 1328$ markers that were genotyped on $n = 175$ individuals which 0.02% genotypes are missing. The barley dataset is expected to yield 7 linkage groups, one for each of the 7 barley chromosomes.

Table 2 shows that our method estimated correctly the number of chromosomes for the OWB dataset. Whereas, The MSTMAP grouped all 1328 markers as one chromosome. The ordering accuracy scores are higher in qploidMap compare with the MSTMAP, except in chromosomes one and three. Also, the size of the markers within each linkage group is consistent with the number of markers that has been reported in Cistue et al. (2011).

4 Conclusion

Construction of linkage map is the most fundamental step required for a detailed genetic study in any species. We propose to build a linkage map in two steps: First, reconstructing an undirected conditional independence graph between SNP markers in a genome, Second determining the order of markers in the obtained chromosomes from the first step. Our simulations show that the proposed method outperforms the alternative method in terms of linkage map quality. In the application of our method in Barley, the proposed method construct more accurate linkage map compare with the alternative method.

References

Behrouzi, P. and Wit, E. C (2017). Detecting Epistatic Selection with Partially Observed Genotype Data using Copula Graphical Models. *Submitted*.

Cistu, L., et al. (2011). Comparative mapping of the Oregon Wolfe Barley using doubled haploid lines derived from female and male gametes. *Theoretical and applied genetics*, 1399-1410.

Liu, H., F. Han, M. Yuan, J. Lafferty, and L. Wasserman (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 2293-2326.

Wu, Y., et al. (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet*, 4(10), e1000212.

Inference in complex systems using multi-phase MCMC sampling with gradient matching burn-in

Alan Lazarus[1], Dirk Husmeier[1], Theodore Papamarkou[1]

[1] School of Mathematics and Statistics, University of Glasgow, UK

E-mail for correspondence: a.lazarus.1@research.gla.ac.uk

Abstract: We propose a novel method for parameter inference that builds on the current research in gradient matching surrogate likelihood spaces. Adopting a three phase technique, we demonstrate that it is possible to obtain parameter estimates of limited bias whilst still adopting the paradigm of the computationally cheap surrogate approximation.

Keywords: Parameter inference; Delayed Rejection Adaptive Metropolis; Surrogate likelihood; Markov Chain Monte Carlo; Gradient matching

1 Introduction

Statistical inference in nonlinear differential equations (DE) is challenging. The log-likelihood landscape is typically multimodal and every parameter adaptation, e.g. in an MCMC simulation, requires a computationally expensive numerical integration of the DEs. Using numerical methods to solve the equations results in prohibitive computational cost; particularly when one adopts a Bayesian approach in sampling parameters from a posterior distribution. Alternatively, one can try to reduce this computational complexity by obtaining an interpolant to the data from which one can obtain a comparative objective function that matches the gradients of the interpolant and the DEs. By sampling on this cheap representative likelihood surface, bias is introduced to the modelling problem. Current research focuses on reducing this bias by introducing a regularising feedback mechanism from the DEs back to the interpolation scheme (e.g. Niu et al. 2016). The idea is to make the interpolant maximally consistent with the DEs. Although this paradigm has proved to improve performance over naïve gradient matching, the feedback loop fails to fully eradicate bias in the final estimate.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

For this reason, a natural progression would be to sample from the true likelihood space whilst reducing computational complexity in the discarded burnin steps. Assuming this hypothesis, we postulate the use of a surrogate likelihood in the burnin phase alone. Through an example possessing multimodal likelihood, we will show the ability of the algorithm to avoid any local entrapment whilst obtaining accurate parameter estimates.

2 Method

Consider time-dependent observations $\mathbf{y}(t) = \boldsymbol{x}(t) + \boldsymbol{\epsilon}$—where $\boldsymbol{x}(t)$ denotes the signal and $\boldsymbol{\epsilon}$ independent additive zero mean Gaussian noise with variance parameter σ^2—whose signals are governed by a system of differential equations:

$$\frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}(t), \boldsymbol{\theta}) \tag{1}$$

dependent on some (partially) unknown parameters $\boldsymbol{\theta}$. Assuming Gaussian noise, we place a GP prior on the latent variable $\mathbf{x}$

$$\mathbf{x}(t) \sim \mathcal{GP}(0, k(\mathbf{t}, \mathbf{t}')), \tag{2}$$

leading us to a Gaussian distribution, $p(\mathbf{x}_i|\boldsymbol{\phi}_i) = \mathcal{N}(\mathbf{x}_i|\mathbf{0}, \mathbf{K}_i)$ for an arbitrary set of time points $\mathcal{T} = \{t_1, ..., t_n\}$ with entries of $\mathbf{K}_i$ given by evaluating kernel function k at each element of $\mathcal{T} \times \mathcal{T}$ (Rasmussen and Williams, 2006). Under our assumption of Gaussian noise, we consider the joint distribution, $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\phi}, \sigma)$. Marginalising over latent variables $\mathbf{x}$ provides a zero mean distribution for the observations:

$$\mathbf{y} \sim \mathcal{N}(0, \mathbf{K} + \sigma^2\mathbf{I}) \tag{3}$$

(see Dondelinger et al. (2013) for details). Considering the joint distribution between our signal and observed values, we may implement an elementary transformation of a Gaussian distribution to obtain the posterior distribution for our signal with mean given by:

$$\mu(\tau) = k(\tau, \mathcal{T})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \tag{4}$$

where $k(\tau, \mathcal{T})$ denotes evaluation of the kernel function at τ over $\mathcal{T}$. Subsequently, firstly estimating the hyperparameters via ML, we adopt the mean of the posterior as a representation of our signal $\mathbf{x}$. This allows us to proceed under the supposition that we have a fixed interpolant for the true signal. Given that the derivative of eq. 4,

$$\frac{\partial\mu(\tau)}{\partial\tau} = k'(\tau, \mathcal{T})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \tag{5}$$

is the mean of the Gaussian distributed DE derivative (see section 7.5 of Vanhatalo et al., 2015), we may consider:

$$f(\hat{\mathbf{x}}(t), \boldsymbol{\theta}) \sim \mathcal{N}\Big(\frac{d\hat{\mathbf{x}}(t)}{dt}, \gamma^2\mathbf{I}\Big), \tag{6}$$

where γ^2 is a fictitious noise term (assumed equal for both gradients) and $\hat{\mathbf{x}}$ is given by eq. 4. Contrary to the work done by Dondelinger et al. (2013), fixing the GP hyperparmeters $\boldsymbol{\phi}$ and the interpolant $\hat{\mathbf{x}}$ allows us to abandon the Gibbs sampling routine at this stage of the algorithm, further reducing the overall computational burden. The corresponding negative log-likelihood term provides a gradient matching objective function:

$$\pi(\boldsymbol{\theta}) = n \log\gamma^2 + \frac{1}{2\gamma^2}\left|\left|\frac{d\hat{\mathbf{x}}(t)}{dt} - f(\hat{\mathbf{x}}(t), \boldsymbol{\theta})\right|\right|^2 \tag{7}$$

which gives a representative computationally tractable surface as a surrogate for the log-likelihood. This involves the term γ^2 representing the mismatch between the gradient obtained from the differential equation and that from explicit differentiation of the GP posterior mean. This parameter will be sampled throughout the surrogate burnin phase of the algorithm. The proposed sampling scheme involves three phases. In the initial burnin phase, samples are drawn from the surrogate distribution in eq. 7 using a Delayed Rejection Adaptive Metropolis[1] (DRAM) scheme (Haario et al. (2006)). Assuming a degree of similarity between the surrogate and true likelihood surfaces, this drives the sampler towards the global minimum of the true likelihood function until a PSRF[2] value of 1.1 has been achieved. From here, we initialise a corrective phase in the true likelihood space, correcting for any bias introduced by the inconsistencies between the surrogate and true likelihood spaces. Sampling with DRAM, this phase is concluded upon obtainment of a PSRF equal to 1.1. The proceeding sampling phase replicates this corrective phase with sampling steps recorded until we achieve a PSRF value of 1.05. The stepwise decrease in target PSRF values allows time for the adaptive component of AM to learn the topology and adjust the covariance accordingly. We adopt an uninformative Inv-Gamma(0.001,0.001) prior for σ^2 and γ^2 and a $Ga(4, 0.5)$ prior for the parameters of the DE. All parameters are sampled on the log scale to account for the positivity constraint.

3 Results

We assess performance on the following DE model of circadian oscillation[3]:

$$\frac{dp_1}{dt} = \frac{k_1}{36 + k_2 p_2} - k_3, \quad \frac{dp_2}{dt} = k_4 p_1 - k_5 \tag{8}$$

This is a notoriously challenging problem due to the extreme multimodality of the likelihood. Following Girolami et al. (2010), we focus on the inference of two parameters (k_3 and k_4), setting the other parameters and initial conditions to

[1] Obtained using the adaptive Metropolis component of DRAM with *modMCMC* function in the *FME* package in R.

[2] Obtained at intervals of 20 steps using the *gelman.diag* function from the *coda* package in R.

[3] We used the same differential equations as in Girolami et al 2010. The actual Goodwin oscillator is of a slightly different form, where the terms k_3 and k_5 are replaced by $k_3 p_1$ and $k_5 p_2$, respectively.

the same fixed values as in Girolami et al. (2010). Five sets of initial parameter values for k_3 and k_4 were obtained using a Sobol sequence over the domain $[0,5]^2$. Figure 1 shows the chain moving through the k_3-k_4 parameter domain. Comparing with the traditional method, we observe the ability of the proposed method to evade the various local minima. PSRF values of 17.1, 10.4 and 12.3 were obtained for k_3, k_4 and σ^2 simulations respectively after 10000 steps using the traditional DRAM method in true likelihood space. Comparatively, the proposed method required 1690 steps in surrogate space, 1690 in the corrective phase and 1010 in the sampling phase to achieve a PSRF of 1.05. The number of numerical integration steps required are given in Table 1 for each of the ten DRAM chains. In Figure 2, boxplots are given that provide the distribution of

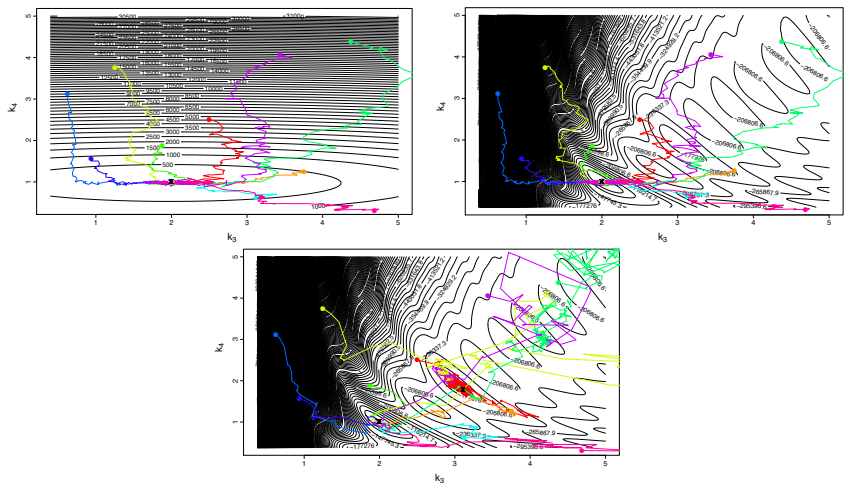


FIGURE 1. Ten chains generated with the proposed method shown in the parameter domain of the negative surrogate log-likelihood space (topleft) and negative log-likelihood space (topright). The bottom plot shows simulations generated using the traditional method. The true parameter value is given by a point at (2,1). The black crosses denote the final point of each chain.

TABLE 1. Number of numerical integration steps (N) for the traditional method. The number required in the proposed scheme is 2700.

Chain Index	1	2	3	4	5	6	7	8	9	10
N	25583	29778	29835	28452	29708	29672	29768	29530	29859	29847

bias in our sampled parameter estimates for each of the five chains. Figure 3 provides RMS deviation in function space obtained using eq. 9,

$$RMS_{function} = \sqrt{\frac{1}{n}\left|\left|\mathbf{x} - \hat{\mathbf{x}}\right|\right|^2} \tag{9}$$

where $\mathbf{x}$ denotes the true signal and $\hat{\mathbf{x}}$ denotes the numerical solution of the DE for one parameter sample from the sample phase of the multiphase approach and the post burnin period of the traditional method. This provides a measure of the predictive accuracy of the MCMC samples.

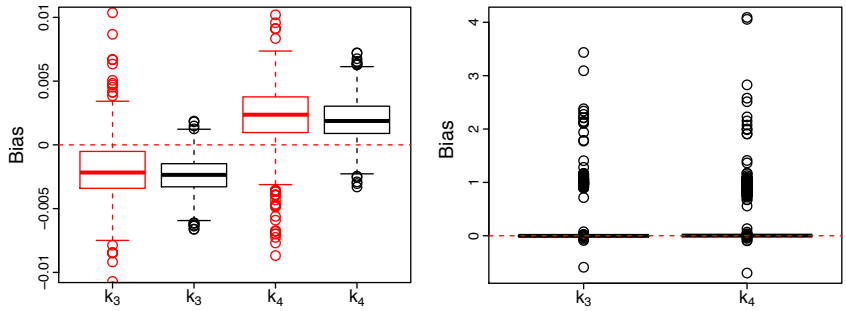


FIGURE 2. Boxplots showing the distribution of bias for both methods (left) where red boxes give the bias of the standard DRAM samples (without outliers) and the black boxes give the bias of the proposed method. The plot on the right gives the bias in both parameters for the DRAM method with outliers included.

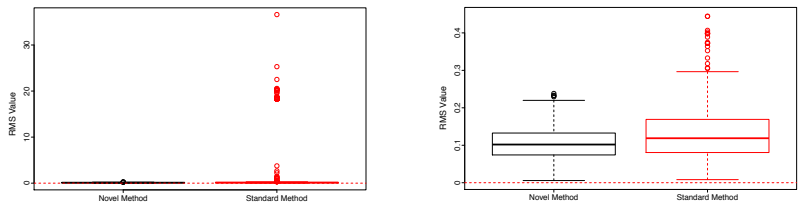


FIGURE 3. Functional RMS comparison between the proposed method (black) and DRAM (red). On the left, we include outliers (from DRAM) and, on the right, these are removed to enable better scalability of the plots for comparison. The red dotted line denotes a functional RMS equal to zero.

4 Conclusion

Our work considers the sampling in DE parameter inference as a computationally efficient three-phase scheme that achieves low levels of bias in sampled parameter estimates (Figure 2). Achieving a PSRF of 1.05, we observe the ability of the algorithm to converge in the parameter space of the circadian oscillator system of equations; a model for which the standard DRAM procedure fails to replicate this success (bottom of Figure 1). Considering the results in function space, we observe the superior performance of the proposed method compared with the traditional DRAM method, showing that the performance improvement is witnessed in both domains of study. These features, along with the vast improvement in computational efficiency, demonstrate improved parameter inference compared with the traditional method.

Acknowledgments: This project was supported by a grant from the Engineering and Physical Sciences Research Council (EPSRC) of the UK, grant reference number EP/L020319/1.

References

Dondelinger, F. et al. (2013). *ODE parameter inference using adaptive gradient matching with Gaussian processes*. Proceedings of Machine Learning Research, Volume 31, pp 216-228.

Girolami, M. et al. (2010). *System identification and model ranking: the Bayesian perspective*. Learning and Inference in Computational Systems Biology. MIT Press, pp. 201-230.

Haario, H. et al. (2006). *DRAM: Efficient adaptive MCMC*. Statistics and Computing, Volume 16, Issue 4, pp 339-354.

Niu, M. et al. (2016). *Fast Parameter Inference in Nonlinear Dynamical Systems using Iterative Gradient Matching*. Proceedings of Machine Learning Research, Volume 48, pp 1699-1707.

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Vanhatalo, J. et al. (2015). *Bayesian Modeling with Gaussian Processes using the GPstuff Toolbox*. MIT Press.

Pragmatic order selection in hidden Markov models

Jennifer Pohle[1], Roland Langrock[1]

[1] University Bielefeld, Germany

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

Abstract: We discuss the notorious problem of order selection in hidden Markov models, i.e. of selecting an adequate number of states, highlighting typical challenges arising when analyzing complex real data. Extensive simulations are used to demonstrate why standard information criteria tend to favor models with undesirably large numbers of states. We also propose a pragmatic step-by-step approach to implement order selection.

Keywords: Information Criteria; Model Selection; Time Series

1 Introduction

Hidden Markov models (HMMs) are time series models for observations that are driven by an underlying finite-state Markov chain. Despite their popularity, identifying the most appropriate number of states has proven to be notoriously difficult in practice (when the focus lies on inference on the data-generating process, i.e. in an unsupervised learning situation). In particular, it has been demonstrated that information criteria, when applied in empirical settings, often lead to the selection of much larger numbers of states than seem desirable (see, e.g., Langrock *et al.*, 2015). This behavior can partly be explained by the complexity of many real data sets. In addition to the features that actually motivate the use of state-switching models (such as multimodality and autocorrelation), real time series data often exhibit additional patterns, e.g. outliers, seasonal fluctuations or non-trivial dependence structures. When neglecting these features in the specification of an HMM, then additional states within the model can "mop up" the neglected structure in the data, therefore providing an improved model fit, however at the price of reduced interpretability.

In this paper, we demonstrate the above points using simulations, and suggest a pragmatic approach to choose the number of hidden states, which takes into

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

account formal criteria for guidance, but also stresses the importance of the study aim, expert knowledge and model checking procedures.

2 Basic formulation of hidden Markov models

An HMM comprises an observed time series $\{X_t\}_{t=1}^T$ which is assumed to be driven by an unobservable state process $\{S_t\}_{t=1}^T$. In the basic model formulation, $\{S_t\}$ is assumed to be a first-order Markov chain, characterized by the state transition probabilities $\gamma_{ij} = \Pr(S_t = j \mid S_{t-1} = i)$, $i, j = 1, \ldots, N$. In addition, it is usually assumed that the observations are conditionally independent of each other, and of past states, given the current state: $p(X_t|X_{t-1}, .., X_1, S_t, ..., S_1) = p(X_t|S_t)$. Thus, the distribution of each observed variable X_t is completely determined by the current state S_t. These two assumptions complete the basic model formulation.

3 Simulation Studies

We use simulations to investigate the performance primarily of the AIC and the BIC when it comes to selecting a suitable number of hidden states. We also investigate the performance of the integrated completed likelihood (ICL) criterion (Biernacki *et al.*, 2001), which has a stronger focus on the model's ability to partition the data. The integrated completed likelihood is obtained using the most probable (Viterbi-decoded) state sequence $\hat{s}$:

$$\text{ICL} = -2 \log \mathcal{L}_c(x, \hat{s}) + p \log(T),$$

where $\mathcal{L}_c(x, \hat{s})$ denotes the (approximate) complete-data likelihood function given the observed time series $x = (x_1, \ldots, x_T)$, and p is the number of model parameters (see Zucchini *et al.* (2016), for details on how to evaluate the likelihood of an HMM).

We showcase six scenarios where there is additional structure in the data that is not accommodated within the basic HMM formulation detailed above. Each type of additional structure considered may be found in real data, where the assumptions made with the basic HMM formulation typically are overly simplistic.

Scenario 1 (benchmark, correct model specification): We simulate data using a two-state gamma HMM without additional structure, primarily as a benchmark for the subsequent scenarios (see Fig. 1).

Scenario 2 (outliers): The data are generated using the benchmark model, adding uniformly distributed errors to 0.5% of the observations generated.

Scenario 3 (inadequate emission distribution): Modified emission distribution within state 2, using a nonparametrically constructed density with a form similar to the gamma distribution, but exhibiting a heavy tail.

Scenario 4 (temporal variation): Transitions probabilities dependent on the time of the day, i.e. state occupancy exhibits within-day variation.

Scenario 5 (semi-Markov state process): Replacing the geometric dwell-time distribution within state 2 of the benchmark model by a Poisson.

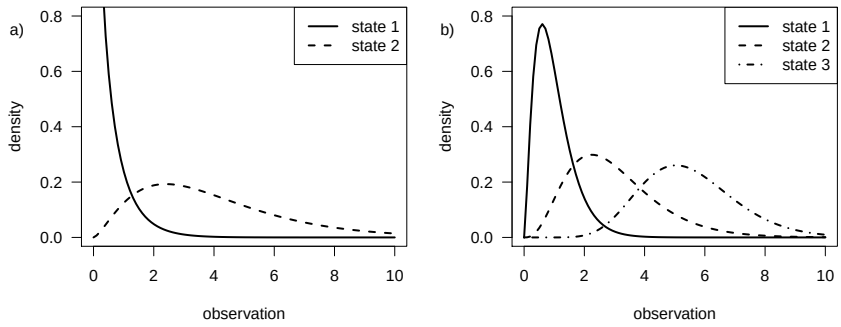


FIGURE 1. a) Densities specified in Scenario 1; b) Densities used in Scenario 7.

Scenario 6 (violation of conditional independence assumption): Time-varying mean parameters of the state-dependent gamma distributions, generated using autoregressive processes of order 1.

For each scenario, 100 data sets were generated ($T = 5000$). In each run, stationary gamma HMMs with 2–5 states were fitted using maximum likelihood estimation (neglecting the additional structure in scenarios 2–6).

Table 1 shows, for each simulation scenario, the percentages of runs in which the models with 2–5 states are chosen by AIC, BIC and ICL, respectively. Both AIC and BIC mostly failed to detect the true number of states, in all five scenarios with model misspecifications, with AIC doing worse than the BIC, due to the higher penalty on model complexity in the latter.

The ICL did fairly well in our simulations, which can be explained by the tendency of the ICL to favor non-overlapping solutions, i.e. HMMs where the state-dependent distributions are clearly distinct. This tendency is also pointed out in Biernacki *et al.* (2001). However, crucially, this behavior will not always be desirable. To demonstrate this point, we carried out an additional simulation using more overlapping states.

Scenario 7 (stronger overlapping emission distributions): We simulate data using a three-state gamma HMM without additional structure, where the states overlap more strongly than in the previous scenarios (see Fig. 1). Furthermore, in this scenario transition probabilities are specified as $\gamma_{11} = \gamma_{22} = 0.8$.

For this scenario, we generated 100 data sets, with $T = 2000$ observations each, and fitted stationary gamma HMMs with 2–4 states to each data set.

The results obtained for Scenario 7 are displayed in Table 2. We find that the ICL here does indeed tend to favor models with too few states. While the AIC performed equally poorly as the ICL — often selecting four and hence too many states — the BIC always chose the correct number of three states in this scenario.

In complex real data, we will usually find several violations of the assumptions involved in the basic HMM formulation. While the deviations may be relatively minor, they may effectively accumulate, such that order selection can in fact even be more problematic than with just a single, yet stronger assumption violation. While conceptually it would seem natural to simply overcome the limitations of HMM formulations that cause criteria-based order selection to fail, this is usually

TABLE 1. Percentages of runs in which the models with 2–5 states are chosen by AIC, BIC and ICL, for all simulation scenarios.

simul.		number of hidden states selected			
scenario	criterion	2 (%)	3 (%)	4 (%)	5 (%)
1	AIC	37	43	20	–
(benchmark)	BIC	100	–	–	–
	ICL	100	–	–	–
2	AIC	–	47	49	4
(outliers)	BIC	30	70	–	–
	ICL	58	42	–	–
3	AIC	–	27	60	13
(inadequate emission	BIC	–	100	–	–
distribution)	ICL	26	71	3	–
4	AIC	–	–	57	43
(temporal variation)	BIC	14	84	2	–
	ICL	100	–	–	–
5	AIC	–	14	74	12
(semi-Markov state	BIC	–	100	–	–
process)	ICL	100	–	–	–
6	AIC	–	–	28	72
(violation of cond.	BIC	5	95	–	–
indep. assumption)	ICL	100	–	–	–

TABLE 2. Percentages of runs in which the models with 2–4 states are chosen by AIC, BIC and ICL, for the additional simulation scenario 7.

simul.		number of hidden states selected		
scenario	criterion	2 (%)	3 (%)	4 (%)
7	AIC	–	54	46
(stronger overlapping	BIC	–	100	–
emission distributions)	ICL	62	36	–

not a useful strategy in practice, both because of computational considerations and as corresponding highly parameterized models may distract from the actual study aim.

4 Pragmatic Oder Selection

Given the difficulties outlined above, we suggest the following pragmatic step-by-step approach to selecting the number of states of an HMM. The proposed strategy applies only to the unsupervised learning case, where inference related to the state process is of primary interest.

Step 1 decide *a priori* on the candidate models, in particular the minimum and the maximum number of states that seem plausible, and fit the corresponding range of models;
Step 2 closely inspect each of the fitted models, in particular plotting their estimated state-dependent distributions and considering their Viterbi-decoded state sequences;
Step 3 use model checking, in particular residuals, to obtain a more detailed picture of the fitted models;
Step 4 use model selection criteria for guidance as to how much improvement, if any, is obtained for each increment in the number of states;
Step 5 make a pragmatic choice of the number of states taking into account findings from Steps 2-4, but also the study aim, expert knowledge and computational considerations;
Step 6 in cases where there seems to be no strong reason to prefer one particular model over another (or several other) candidate model(s), results for each of these models should be reported.

The resulting choice of the number of states will necessarily be somewhat subjective. However, a corresponding analysis nevertheless will be as scientific, if not more scientific, than any allegedly objective choice of the number of states. Furthermore, we have made the experience that a thorough implementation as detailed in **Steps 1-4** will usually make it fairly easy to pick a suitable N.

5 Conclusion

Model selection criteria are problematic with respect to choosing the number of states of an HMM applied to complex real data within an unsupervised learning framework. In particular, any structure in the data that is neglected in the model formulation will, to some extent, be mopped up by additional model states that do not have a clear interpretation anymore. The ICL criterion appears to overcome several of the problems associated with the more established AIC and BIC, yet it does not come without its own limitations, namely a sensitivity to overlapping state-dependent distributions.

We proposed a pragmatic step-by-step approach to order selection which, while lacking objectivity, we believe is the best possible practical solution. As pointed out in Hennig (2015) in the context of cluster analysis, it is crucial that the individual researcher's modeling decisions, and in particular the rationale underlying the selection of the number of states, need to be made transparent.

Overall, the selection of the number of states clearly is an important yet challenging issue, which requires statistical expertise (when applying model selection and model checking tools) and modeling experience, but also a good understanding and intuition of the data and research question at hand (in order to arrive at a sensible choice of the number of states).

References

Biernacki, C., Celeux, G. and Govaert, G. (2001). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on pattern analysis and machine intelligence*, **22**, 719 – 725.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, **64**, 53 – 62.

Langrock, R., Kneib, T., Sohn, A. and DeRuiter, S.L. (2017). Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, **71**, 520–528.

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

Spatial Conditional Overdispersed Models: Application to count area data

Vicente Núñez-Antón[1], Edilberto Cepeda-Cuervo[2]

[1] Department of Econometrics and Statistics, University of the Basque Country UPV/EHU, Bilbao, Spain
[2] Department of Statistics, National University of Colombia, Bogotá, Colombia

E-mail for correspondence: vicente.nunezanton@ehu.eus

Abstract: We propose alternative models for the analysis of count data featuring a given spatial structure. We assume that the overdispersion data structure partially results from the existing and well justified spatial correlation between geographical adjacent regions, so an extension of existing overdispersion models that include spatial neighborhood structures within a Bayesian framework is proposed. Finally, their usefulness is illustrated by fitting them to infant mortality rates and to data including the proportion of mothers who, after giving birth to their last child, underwent a postnatal screening period in Colombia.

Keywords: Count area data; Infant mortality; Overdispersion; Postnatal screening period rates; Spatial statistics.

1 Introduction and motivating examples

Generalized linear models are commonly used in medical data analysis, specially for their flexibility in the modelling of the distribution of the response variable. One of the most relevant contributions in the development of the generalized linear models theory centers on count data, such as infant mortality rates or the proportion of mothers who underwent a postnatal screening period. When the data show some type of overdispersion, this phenomenon can be generated from very different sources, which should be appropriately assessed in order to be able to reduce the impact wrong model formulations not able to capture all of the variability in the data have on the results from the statistical analysis. Several models have been proposed following the theory originally developed for overdispersed binomial and Poisson models. Therefore, taking into account the existing spatial association between observations on the variable under study is

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

very relevant in the statistical analysis of this type of data. One would desire that the data provide spatially ordered information for the variable under study, so that the general idea that observational units closer to a specific observation may have some influence on its distributional behavior may be assessed and included in the model specification. Several techniques have been proposed to model spatial correlation, but interest in the proposals here lies in trying to quantify this correlation with the use of a parametric model that includes a set of parameters able to model this effect. Therefore, we propose alternative overdispersion models that include basic neighborhood structures in spatial models, which are based on a self-constructed covariate, defined by using a series of non stochastic spatial weights that somehow establish the strength of the spatial dependence between the geographical units being considered in the study. The proposed models include spatial neighborhood structures both for the mean and for the overdispersion parameter specification, so that the spatial association can be quantified both in the mean and overdispersion behavior, by using neighborhood structure assumed by the researcher performing the analysis. Thus, our models are able to quantify the spatial association related to the neighborhood structures proposed or assumed by the researcher. In addition, they include regression structures associated to the overdispersion parameter, to the different factors that may be related to the existing heteroscedasticity, and to the spatial association not explained by the assumed neighborhood structures that could be related to other covariates or to model structures specified in the proposed dispersion model. In order to illustrate the usefulness of the proposed models we apply them to two real data sets. The first one corresponds to the analysis of the number of children under 5 years who died in the 5-year period 2000-2005 in Colombia. The second one corresponds to the study of the proportion of mothers who had given birth to their last child between 1999 and 2005 and who underwent a postnatal screening period. In both data sets, we have variables available at the departmental level, such as the percentage of the population that has basic services not being satisfactorily attended to, the percentage of young people who had access to a higher academic achievement level, and the resources provided for academic achievement or education and integral attention for young children per household for families coming from the general participation system. All of these variables will be used as covariates for the proposed models.

2 (Generalized) spatial conditional overdispersion models

Generalized overdispersion models for count data, in which regression structures are assumed for both the mean and overdispersion parameters, may be an appropriate alternative to analyze count area data, where spatial association between observations is present. However, given that these models do not provide information related to the strength of the spatial association between observations of the variable under study, some more general models should be specified. In the spatial overdispersion models proposed here, this effect can be easily estimated, an issue that is addressed by proposing the use of a weight matrix in the model, where its parameter estimates are associated to the corresponding lag variable modelling this neighborhood association. The specific choice of the elements of

the weight matrix is a very relevant issue, mainly because it should be the result of both interpreting and understanding the spatial structure the variable under study has. In the proposed models we assume that observational units correspond to each of the areas in the study, where we control for or model the variation generated by the existing spatial correlation between experimental units, we use the weights w_{ij}, which are the elements in the model reflecting the level of dependence between the spatial units in the study, indexed by i and j. These values are the corresponding elements of the weight matrix $\mathbf{W}$. One of the most commonly used specifications for the weight matrix $\mathbf{W}$ assumes that $w_{ij} = 1/n_i$, if the observational unit or region j belongs to the neighborhood of an observational unit i, or if there is geographical first or second order contiguity between regions i and j, where n_i is the number of first or second order adjacent regions for region i; and $w_{ij} = 0$ otherwise. The generalized spatial conditional models proposed here are structurally different from the Poisson or binomial CAR and SAR models. More specifically, the proposed spatial conditional models are formulated in terms of generalized overdispersed models, which include spatial lag structures in both the mean and dispersion regression structures. In addition, nonstructural random effects are also included in both of the aforementioned regression structures, so that the models are able to capture the overdispersion not explained by the spatial structures included in the proposed spatial models. Our proposed spatial conditional overdispersion regression model assumes that the spatial variable under study, $Y_i,\ i = 1, \ldots, n$, conditioned on the values in all of the neighborhoods of the i-th region, but not including the i-th region itself (i.e., $Y_{\sim i}$), has a overdispersed conditional distribution denoted by $f(y_i|y_{\sim i}),\ i = 1, \ldots, n$, where the conditional mean and the conditional dispersion parameters follow given regression structures that, besides some covariates affecting the response variable, also include spatial lags of the variable under study. These models assume that the conditional overdispersion density functions follow either a Poisson or a binomial distribution, leading to the proposal of the (generalized) spatial conditional Poisson, negative binomial, normal Poisson, binomial, beta binomial and binomial normal regression models, respectively. As an illustrative example of these model proposals, we consider the Poisson normal model for overdispersed count data. In this model, the overdispersion is included in the model with the use of a normally distributed random effect term in the mean model. In this way,

$$g(\lambda_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \nu_i, \tag{1}$$

where $g(.)$ is usually the logarithm function, $\mathbf{x}_i$ is the $q \times 1$ vector of explanatory variables for the i-th observation, $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown regression parameters, and $\nu_i \sim N(0, \tau)$. In this model, $(Y_i|\nu_i),\ i = 1, \ldots, n$, follows a Poisson distribution with mean $\lambda_i = E(Y_i|\nu_i)$. In the spatial conditional normal Poisson model, if $Y_i,\ i = 1, \ldots, n$, represent area count data from different regions or areas, such as departments or states, a portion of the existing overdispersion can be explained by the neighborhood spatial structured assumed by the researcher, which is given by the spatial conditional normal Poisson model where we assume that $(Y_i|Y_{\sim i}, \nu_i)$ follows a Poisson distribution with mean $\lambda_i = \mathrm{E}(Y_i|Y_{\sim i}, \nu_i)$, with

$$g(\lambda_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \rho\mathbf{W}_i\mathbf{y} + \nu_i, \tag{2}$$

and ρ is the parameter explaining the first order spatial association in the mean model, $\mathbf{W}_i$ is the i-th row of the $n \times n$ weight matrix $\mathbf{W}$, which follows the

assumed first order neighborhood structure (i.e., regions, units or departments share a common border), and $\mathbf{y}$ is the $n \times 1$ vector of the observed values of the response variable under study. Finally, we propose the generalized spatial conditional normal Poisson model, which also allows for modelling spatial neighborhood structures. That is, we assume that $(Y_i|\nu_i)$ follows a Poisson distribution with parameter $\mu_i,\ i = 1, \ldots, n$, where $\nu_i \sim N(0, \sigma_i^2)$. In addition, it is assumed that conditional mean μ_i and the variance terms in the random effect distribution, σ_i^2's, are modelled as functions of some explanatory variables, so that its regression structures are given by

$$\log(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \rho\mathbf{W}_i\mathbf{y} + \nu_i \ \text{ and } \ \log(\sigma_i^2) = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \eta\mathbf{W}_i\mathbf{y}, \tag{3}$$

where $\mathbf{z}_i$ is the $q_\phi \times 1$ vector of explanatory variables for the i-th observation, $\boldsymbol{\gamma}$ is a $q_\phi \times 1$ vector of unknown regression parameters, and η is the parameter explaining the first order spatial association in the dispersion model. In these models, the random factor is associated to the overdispersion generated, for example, from the heteroscedasticity or the possible spatial correlation of higher order than that considered in the model. In the generalized spatial conditional normal Poisson model, if we have a fixed $\sigma_i^2 = \sigma^2$ and keep σ_i^2 unstructured (i.e., do not propose any model for it), we have the spatial conditional normal Poisson regression model with $\tau = \sigma^2$. If, in addition, $\rho = 0$, we have the Poisson normal regression model.

3 Application

Data considered here correspond to the 32 departments (regions, geographical units or states) in Colombia. Variables available for each one of the geographical units are: the number of children under five who died in the five-year period 2000-2005 (i.e., variable ND), the percentage of women over 18 who had suffered any type of physical abuse from their current partners (i.e., variable Viol), the percentage of the population that had basic services not being satisfactorily attended to for the year 2004 (i.e., variable NBI), the resources (in thousands) provided for academic achievement or education and integral attention for young children per household for families coming from the general participation system (i.e., the general plan for the allocation of resources from the central government to the departmental or municipal governments) in the year 2005 (i.e., variable Rec), the percentage of young people (i.e., between 18 and 23 years old) who had access to a higher academic achievement level (i.e., to a higher educational level) in the year 2005 (i.e., variable HE), the percentage of children under one year of age who had the third dose of the polio vaccine applied in the year 2004 (i.e., the variable Vac), the number of mothers (in thousands) that had their last child born between 1999 and 2005 and that underwent a postnatal screening period (i.e., the variable NScree), the number of mothers (in thousands) that had their last child after 1999 (i.e., variable NMother99), and the percentage of mothers who had to pay for the total cost of the postnatal screening period (i.e., the variable Pay). For the weight matrix, $\mathbf{W}$, we assume that the elements w_{ij} is equal to one if region j belongs to the neighborhood of region i, and equal to zero, otherwise. As for the prior distributions, we assume independent normal distribution, $N(0, 10^5)$, for all of the regression parameters, including the spatial association parameters, ρ and η. In the specific application considered here, the behavior

of the chains, all of them with small transient periods, seems to show that the convergence was quickly achieved. Therefore, a burn in period of 2000 iterations is assumed from the 10000 iterations developed in the estimation process. The best fitting model for this data was the generalized spatial conditional normal Poisson model with BIC and DIC values of 200.78 and 251.88, respectively, and mean and variance regression models given by:

$$\log(\mu_i) = \beta_0 + \beta_1 \text{Rec}_i + \nu_i, \quad \nu_i \sim N(0, \sigma_i^2) \tag{4}$$

$$\log(\sigma_i^2) = \gamma_0 + \eta \mathbf{W}_i \mathbf{y}, \tag{5}$$

with the corresponding estimates reported in Table 1.

TABLE 1. Parameter estimates, together with their standard deviations for the mean and variance parameters for the infant mortality data and the generalized spatial conditional normal Poisson model.

	β_0	β_1	γ_0	η
Estimate	3.283	-4.564×10^{-04}	0.522	-6.617×10^{-03}
Standard deviation	0.271	2.091×10^{-04}	0.352	2.069×10^{-03}

4 Conclusions

Results from a simulation study and from the analysis of the infant mortality rates data, for the Poisson case, and to data including the proportion of mothers who underwent a postnatal screening period in Colombia, for the binomial case, conclude that the proposed models fit better than both the previously proposed models (i.e., the ones not modelling the overdispersion in the data sets) or the well known intrinsic conditional autoregressive (ICAR) models. More specifically, for the infant mortality rates data, the best fitting model was the generalized spatial conditional normal Poisson model, with spatial structures and random effects included both in the mean and overdispersion regression models. As for the proportion of mothers who underwent a postnatal screening period, they were the spatial conditional binomial normal and beta binomial models, with spatial structures and random effects included only in the mean regression model. Finally, a sensitivity analysis was performed to assess the effect or influence of the assumed prior distributions for the different regression parameters for the mean and overdispersion parameters have on the resulting parameter posterior estimates.

Acknowledgments: This work was supported by Ministerio de Economía y Competitividad, FEDER, the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group), and Universidad del País Vasco UPV/EHU under research grants MTM2013-40941-P, MTM2016-74931-P, IT-642-13, UFI11/03 and US15/11, and also by the Department of Statistics, National University of Colombia.

References

Cepeda-Cuervo, E., Córdoba, M. and Núñez-Antón, V. (2017). Conditional overdispersed models: application to count area data. *Statistical Methods in Medical Research*. In press.

Hinde, J. and Demétrio, C.G.B. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, **27(2)**, 151 – 170.

Quintero-Sarmiento, A, Cepeda-Cuervo, E. and Núñez-Antón, V. (2012). Estimating infant mortality in Colombia: some overdispersion modelling approaches. *Journal of Applied Statistics*, **39**, 1011– 1036.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, **8(2)**, 158 – 183.

Constructing wildebeest density distributions by spatio-temporal smoothing of ordinal categorical data using GAMs

Elaine A. Ferguson[1], Jason Matthiopoulos[1], Dirk Husmeier[2]

[1] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
[2] School of Mathematics and Statistics, College of Science and Engineering, University of Glasgow, Glasgow, UK

E-mail for correspondence: e.ferguson.2@research.gla.ac.uk

Abstract: Spatio-temporal smoothing of large ecological datasets describing species distributions can be made challenging by high computational costs and deficiencies in the available data. We present an application of a GAM-based smoothing method to a large ordinal categorical dataset on the distribution of wildebeest in the Serengeti ecosystem.

Keywords: GAMs; Ordinal categorical; Smoothing; Spatio-temporal; Wildebeest.

1 Introduction

Spatio-temporal smoothing of species distribution data has many potential uses in ecology; for example, to provide a smooth density function that can be used with gradient matching approaches (Xun et al. 2013) to fit partial differential equation (PDE) models of animal movement. A range of smoothing methods (kernel density estimation, splines, Gaussian processes, etc.) have been developed in the statistical literature. However, the practicalities and expense involved in collecting species distribution data over large areas in the field can mean that the data are not in a form that these methods can readily be applied to. Ordinal categorical data, for example, may be collected when it is infeasible to accurately count all individuals in a population, so that the abundance at each point in space and time is instead estimated as belonging to a broader abundance category. A relatively small number of approaches have been developed

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

for smoothing data of this type, where we need to recover the underlying true density of individuals from the categories (Chu and Ghahramani 2005, Wood et al. 2016). Smoothing large datasets in multiple dimensions can also be made challenging by high computational costs. Methods that allow smoothing of these datasets even when computational resources are limited would therefore be very useful. Here we present an application of a method for applying spatio-temporal smoothing to a large ordinal categorical dataset on the distribution of wildebeest in the Serengeti ecosystem of Tanzania and Kenya.

2 Methods

The wildebeest distribution data, which have been described and utilised in a number of previous studies (Norton-Griffiths 1973, Maddock 1979, Boone et al. 2006, Holdo et al. 2009), were obtained from monthly aerial surveys of the Serengeti ecosystem during the period from August 1969 to August 1972. Each cell in a grid of $25km^2$ cells was assigned to one of five wildebeest abundance categories: 0, 1-25, 26-250, 251-2,500 and >2,500 individuals per $25km^2$. There were 2,576 cells making up the spatial grid, all of which were sampled on 33 occasions during the time period, resulting in a large dataset with a total 85,008 data points.

To smooth the data in time, t, and the two spatial dimensions (x, y) we fitted GAMs (generalised additive models) with a tensor product (composed of cubic regression spline smooths, where overfitting was prevented by penalisation of the integral of the squared second derivatives) between these three variables using the mgcv package (Wood 2011) in R (R Core Team 2015). We used the ordinal categorical GAM method described in Wood et al. (2016), where the linear predictor gives the value of a latent variable, here representing the wildebeest density underlying the ordinal categories. The cut-off points that demarcate the five ordinal categories were specified, and the probability that a point in space and time belongs to a given category equals the probability that the latent variable lies between the corresponding category cut-offs at that point.

In Wood et al. (2016), the latent function can range from $-\infty$ to ∞, but we know that wildebeest density has a minimum 0 and a finite maximum $W_{\max}$. We can introduce these constraints by applying a sigmoidal transformation to the latent function L after the GAM has been fitted, giving a preliminary wildebeest density $\hat{W}$ as follows:

$$\hat{W}(x, y, t) = \frac{W_{\max}}{1 + \exp(-L(x, y, t))} \tag{1}$$

Note that this also required that an inverse sigmoid transform be applied to the category cut-offs $\mathbf{c}$ prior to the GAM fitting:

$$\bar{\mathbf{c}} = -\log\left(\frac{W_{\max}}{\mathbf{c}} - 1\right) \tag{2}$$

$W_{\max}$ was estimated by first assuming that the wildebeest densities in the grid cells assigned to the lower four ordinal categories, which had known upper and lower bounds, were equal to the mid-points of those categories. The sum of the

densities in these lower category cells for each month was then subtracted from the total number of wildebeest W_T known to be in the region from a population count in 1971 (Norton-Griffiths 1973). The remaining wildebeest for each month were assumed to be divided evenly between the cells in the highest ordinal category (which was unbounded above) for that month. We took $W_{\max}$ to be the largest wildebeest density estimated for cells in the highest abundance category over all months.

Even after applying sensible upper and lower bounds to the latent function, large fluctuations in the area under $\hat{W}$ (which represents the total number of wildebeest in the region) can occur over time. This is undesirable, since we expect wildebeest numbers to remain relatively stable at W_T over the time period of interest. We therefore consider the normalised wildebeest density $\bar{W}$, where the total number of animals is maintained at W_T by normalising $\hat{W}$ as follows:

$$\bar{W}(x,y,t) = \frac{\hat{W}(x,y,t)\, W_T}{\int \hat{W}(x,y,t) dx dy} \tag{3}$$

Due to computational time and memory constraints, a sufficiently flexible GAM could not be fitted to the entire large dataset simultaneously. We therefore divided the time series into three contiguous intervals and fitted a GAM in (x,y,t) to each interval separately. Each GAM had 20 knots in the marginal smooth in each spatial dimension, and a number of knots in the marginal smooth in time that was equal to the number of time points present in the data subset to which the GAM was fitted (11 or 12). This resulted in the effective degrees of freedom, which are determined by the degree of penalization (selected during fitting) applied to the integral of the squared second derivatives, being considerably lower than the maximum number available, suggesting that the number of knots was sufficient (Wood 2006). The three GAMs were joined together by averaging at the link times l_i $(i \in 1,2)$, with smoothness being maintained by allowing the influence of each GAM on the others to decline smoothly, according to the parameter σ, as distance from the point of joining increased. For a given point $(\bar{x},\bar{y},\bar{t})$, therefore, we obtain a final estimate of wildebeest density W by

$$W(\bar{x},\bar{y},\bar{t}) = \bar{W}_{GAM_j}(\bar{x},\bar{y},\bar{t}) + \sum_{i=1}^{2} a_i \exp\left(\frac{-(\bar{t}-l_i)^2}{2\sigma^2}\right) m_i(\bar{t}) \tag{4}$$

Here $\bar{W}_{GAM_j}$ is the normalised wildebeest density obtained from the GAM fitted to time interval j, where

$$j = \begin{cases} 1 & \text{if} \quad \bar{t} \le l_1 \\ 2 & \text{if} \quad l_1 < \bar{t} \le l_2 \\ 3 & \text{if} \quad \bar{t} > l_2 \end{cases} \tag{5}$$

The a_i are given by

$$a_i(\bar{x},\bar{y},l_i) = \frac{\bar{W}_{GAM_i}(\bar{x},\bar{y},l_i) - \bar{W}_{GAM_{i+1}}(\bar{x},\bar{y},l_i)}{2} \tag{6}$$

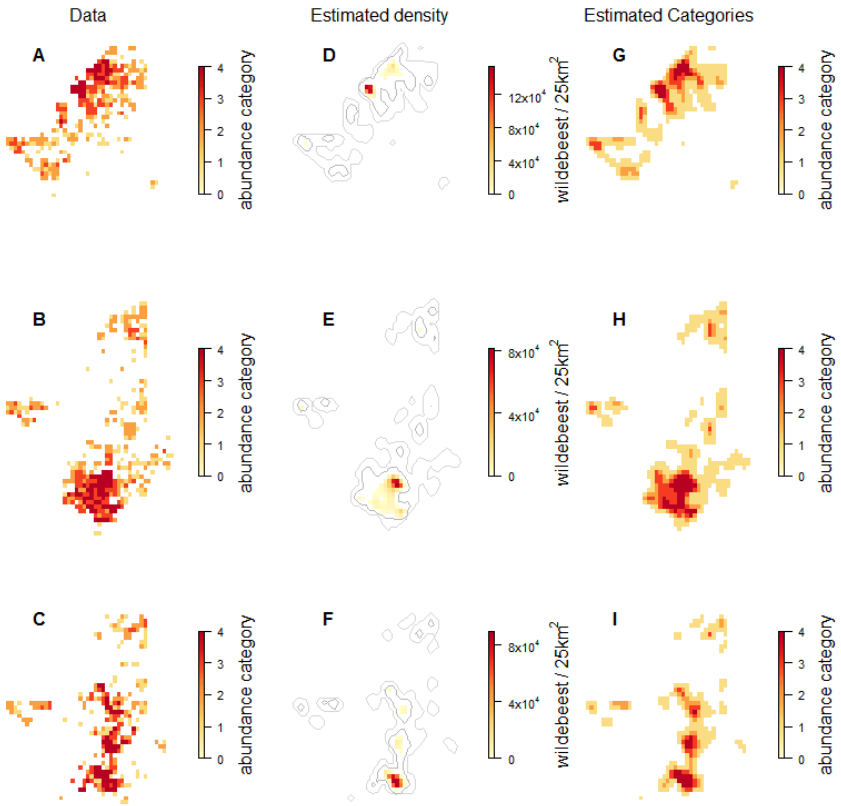


FIGURE 1. Model fit in space at three different time points. **A-C:** The wildebeest spatial distribution data for months 1, 18 and 35. **D-F:** The smooth wildebeest density distribution estimated in space by the model for months 1, 18 and 35. The two contours indicate the boundaries between abundance categories 0, 1 and 2. **G-I:** Estimated wildebeest abundance categories based on **D-F**.

and the m_i, which ensure that the adjustments are made in the correct direction on either side of each link point, are

$$m_i(t) = \begin{cases} -1 & \text{if} \quad \bar{t} \leq l_i \\ 1 & \text{if} \quad \bar{t} > l_i \end{cases} \tag{7}$$

If the influence of the adjoining GAMs declines too slowly with distance from the link points, relative to the rate at which changes occur in $\bar{W}_{GAM_i}$ (i.e. σ is too large), unrealistic negative values of W can occur. We therefore tuned σ by starting with a relatively large value and gradually decreasing it until no negative values of W occurred.

3 Results and Conclusion

The method described was found to successfully produce a smooth function in space that resembles the original data (Figure 1). The resulting function is also observed to be smooth in time, with no evidence that the wildebeest density changes either more slowly or more rapidly around the GAM link times than it does elsewhere in the time period (Figure 2). This suggests that our approach of linking models that have been fitted to subsets of a larger dataset is a promising means of reducing the high computational costs of smoothing large datasets in multiple dimensions. Using this method, we have recovered realistically bounded wildebeest abundance estimates from coarse ordinal categories; an ability that could be useful in the field of ecology where such imperfect data are common. By producing a smooth surface from which spatial and temporal gradients in density can

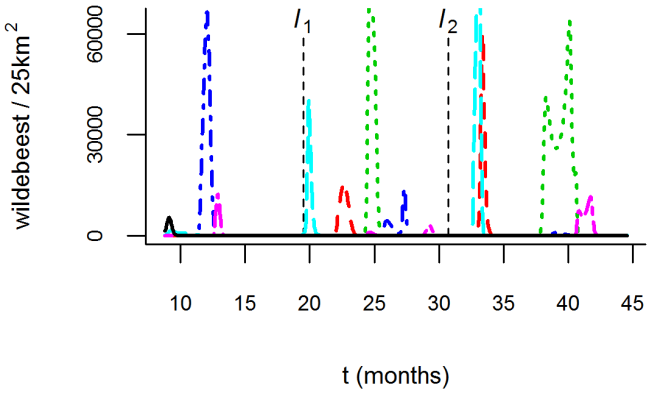


FIGURE 2. Changes in the estimated wildebeest density in six grid cells (indicated by different colours/line types) over the time period of interest. The link times between the three GAMs are indicated by dashed vertical lines.

be calculated, our method also promises to enable statistical inference for PDE models of animal movement using the gradient matching approach of Xun et al. (2013), which we will investigate in future work.

Acknowledgments: Special thanks to Ricardo M. Holdo for providing access to the wildebeest distribution data. E.A.F. is funded by a University of Glasgow Lord Kelvin/Adam Smith PhD scholarship.

References

Boone, R.B., Thirgood, S.J. and Hopcraft J.G.C. (2006). Serengeti wildebeest migratory patterns modeled from rainfall and new vegetation growth. *Ecology*, **87**, 1987 – 1994.

Chu, W. and, Ghahramani, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, **6**, 1019 – 1041.

Holdo, R. M., Holt, R. D. and Fryxell, J. M. (2009). Opposing rainfall and plant nutritional gradients best explain the wildebeest migration in the Serengeti. *The American Naturalist*, **173**, 431 – 445.

Maddock, L. (1979). The migration and grazing succession. In: *Serengeti: dynamics of an ecosystem*, Sinclair, A. R. E., and Norton-Griffiths, M. (editors), Chicago: University of Chicago Press, 104 – 129.

Norton-Griffiths, M. (1973). Counting the Serengeti migratory wildebeest using two-stage sampling. *East African Wildlife Journal*, **11**, 135 – 149.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. CRC Press.

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., Pya, N., and Sfken, B. (2016). Smoothing parameter and model selection for general smooth models. arXiv:1511.03864v2.

Xun, X., Cao, J., Mallick, B., Maity, A., and Carroll, R.J. (2013). Parameter Estimation of Partial Differential Equation Models. *Journal of the American Statistical Association*, **108**, 1009 – 1020.

Spatio-temporal clustering of traffic networks

Ashwini Venkatasubramaniam[123], Ludger Evers[2], Konstantinos Ampountolas[13]

[1] Urban Big Data Centre, University of Glasgow, United Kingdom,
[2] School of Mathematics and Statistics, University of Glasgow, United Kingdom,
[3] School of Engineering, University of Glasgow, United Kingdom

E-mail for correspondence: a.venkatasubramaniam.1@research.gla.ac.uk

Abstract: We present a novel Bayesian clustering method for spatio-temporal data observed on a network and apply this model to cluster an urban traffic network. This method employs a distance dependent Chinese restaurant process (DDCRP) to provide a cluster structure, by incorporating the observed data and geographic constraints of the network. However, in order to fully capture the dependency structure of the data, a conditional auto-regressive model (CAR) is used to model the spatial dependency within each cluster.

Keywords: Network; Spatial; Clustering; Bayesian

1 Introduction

Heterogeneous urban traffic networks with regions of varying congestion levels have unique fundamental properties and *clustering* aids in the division of a city into homogeneous regions. We propose a novel Bayesian clustering technique for spatio-temporal network data which is based on an amalgamation of a distance dependent Chinese restaurant process (DDCRP) and a spatio-temporal conditional autoregressive model (CAR). We assume that we observe a time series of measurements that represent congestion levels aggregated over each junction in the network and the degree of similarity between adjacent junctions can be used to define spatially contiguous clusters. Existing literature relevant to clustering techniques for transportation networks account for spatial constraints but typically do not incorporate changes over time within a cluster. Traditional clustering algorithms such as k-means and probability mixture models also require choices to be made about the number of clusters.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

2 Method

In our model, the road network forms an undirected graph with junctions acting as nodes and road segments between junctions as edges. This graph can be used to define the spatial component of the precision matrix. The spatial precision $\mathbf{\Sigma_S}^{-1}$ within a cluster is modelled as a CAR model and the temporal precision $\mathbf{\Sigma_T}^{-1}$ as an auto-regressive (AR1) model. Other models that incorporate temporal dependency such as the Matern covariance function are also possible. Due to the presence of a grid network topology with a limited number of road segments between junctions (typically not more than four per junction), the precision matrices exhibit sparsity and we utilize a CAR model proposed by Leroux (2000) to define this spatial precision. We define an adjacency matrix $\mathbf{W}$ and the precision matrix $\mathbf{Q} = \mathbf{\Sigma_S}^{-1} = \rho(\text{diag}(W_{k++}) - \mathbf{W}) + (1 - \rho)\mathbf{I}_{n_J}$, where ρ controls how strongly correlated adjacent junctions are, $\text{diag}(W_{k++})$ is a diagonal matrix with elements equivalent to the row sums of $\mathbf{W}$, and $\mathbf{I}_{n_J}$ is an $n_J \times n_J$ identity matrix (n_J = number of junctions). With the presence of a unique observation for every space-time combination (n_J junctions and n_T time points), a covariance matrix $\mathbf{\Sigma}$ can be written as $\mathbf{\Sigma_S} \otimes \mathbf{\Sigma_T}$. Clusters can be obtained by removing edges such that the graph can be partitioned into components not connected to each other. In Figure 1, a network composed of eight nodes can be divided into two clusters such that there are road segments but no links between adjacent junctions of two differing clusters.

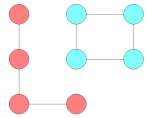


FIGURE 1. Graph showing two clusters formed in the network based on the presence or absence of a road segment between junctions.

Individually removing edges not supported by the data would yield a sparse graph, but would be unlikely to result in a graph with more than one component. Instead, we use a prior to enforce that edges are omitted in a way that leads to a clear partitioning of the graph. We utilize a modified version of the DDCRP first introduced by Blei (2011) that allows our model to incorporate geographic constraints of the network, account for the shape, and determine the number of clusters. The DDCRP makes assumptions of non-exchangeability to account for components of distance such as time, space, etc. In a traditional Chinese restaurant process (CRP) (also a special case of the DDCRP), a restaurant can be assumed to consist of an infinite number of tables. Customers $i = 1 \ldots n$ are individual data points that enter and take a seat at a randomly chosen table k. Tables are deemed to be clusters and after a finite number of customers n_k have been seated, the seating plan represents a partition. In the usual representation of the CRP, customers choose tables. In a DDCRP, clusters are instead determined based on friendships between customers i and j, with a group of friends then sitting at an assigned table z_i, i.e., forming a cluster. Thus, $z(\mathbf{c})$ are table assignments that follow from customer assignments. In a non-sequential DDCRP, clusters arise from some customers choosing to befriend themselves or someone already connected to them, resulting in a redundant assignment. We also mod-

ify the DDCRP to allow customers to befriend more than one customer, which controls for the number of singleton clusters. Let c_i be the index of a customer that is sitting with customer i and we describe the distribution of this customer assignment as:

$$P(c_i = j|\alpha) \propto \begin{cases} 0, j \neq i \text{ and } i \nsim j \\ 1, j \neq i \text{ and } i \sim j \\ \alpha, j = i \end{cases}$$

In our method, customers are junctions and friendships can only occur along road segments between junctions. Since this modified DDCRP suggests a prior over a combinatorial number of junction assignments, the posterior is intractable and inference is carried out using a Metropolis within Gibbs sampler. We assume that the measure of congestion levels $\mathbf{Y}$ follows a Gaussian distribution and the likelihood at partition $z(\mathbf{c})$ gives the product of probabilities calculated for sets of observations at each determined cluster. To account for the spatial dependency within the cluster, we define $\mathbf{\Sigma_S}$ and $\mathbf{\Sigma_T}$ to represent the precision matrix described earlier. Accordingly, the likelihood can be defined as:

$$\ln(P(\mathbf{Y}|\mathbf{\Sigma_S}, \mathbf{\Sigma_T}, \sigma^2, \tau^2)) = -\frac{n_J n_T}{2}\ln(2\pi) - 0.5\ln|\sigma^2\mathbf{I} + \tau^2\mathbf{\Sigma_S} \otimes \mathbf{\Sigma_T}|$$
$$-0.5\text{vec}(\mathbf{Y})^{\mathrm{T}}[\sigma^2\mathbf{I}+\tau^2\mathbf{\Sigma_S} \otimes \mathbf{\Sigma_T}]^{-1}\text{vec}(\mathbf{Y})$$

We can rewrite terms in the likelihood, $\text{vec}(\mathbf{Y})^{\mathrm{T}}[\sigma^2\mathbf{I} + \tau^2\mathbf{\Sigma_S} \otimes \mathbf{\Sigma_T}]^{-1}\text{vec}(\mathbf{Y}) = \text{vec}(\mathbf{Y})^{\mathrm{T}}(\mathbf{\Gamma_S} \otimes \mathbf{\Gamma_T})(\sigma^2\mathbf{I}+\tau^2\mathbf{\Lambda})^{-1}(\mathbf{\Gamma_S^{\mathrm{T}}} \otimes \mathbf{\Gamma_T^{\mathrm{T}}})\text{vec}(\mathbf{Y}) = \text{vec}(\mathbf{\Gamma_T^{\mathrm{T}}Y\Gamma_S})^{\mathrm{T}}(\sigma^2\mathbf{I}+\tau^2\mathbf{\Lambda_S} \otimes \mathbf{\Lambda_T})^{-1}\ \text{vec}(\mathbf{\Gamma_T^{\mathrm{T}}Y\Gamma_S})$, where σ^2 is variance of the noise, τ^2 is a prior variance, $\mathbf{\Lambda_T}$ represents a diagonal matrix of the eigenvalues of $\mathbf{\Sigma_T}$, and $\mathbf{\Gamma_T}$ represents a matrix of the eigenvectors of $\mathbf{\Sigma_T}$. Here, we only need to compute the diagonal of $[\sigma^2\mathbf{I} + \tau^2\mathbf{\Lambda_S} \otimes \mathbf{\Lambda_T}]^{-1}$ on the rotated data $\text{vec}(\mathbf{\Gamma_T^{\mathrm{T}}Y\Gamma_S})^{\mathrm{T}}$. In addition, $\ln|\sigma^2\mathbf{I} + \tau^2\mathbf{\Sigma_S} \otimes \mathbf{\Sigma_T}|$ can be rewritten as $\sum_{t=1}^{n_T}\sum_{s=1}^{n_J} -\ln(\tau^2\mathbf{\Lambda_T}[t,t] \cdot \mathbf{\Lambda_S}[s,s] + \sigma^2\mathbf{I})$. Together, these terms can be evaluated for an efficient solution in $O(n_J^2 n_T + n_T^2 n_J + n_J^3 + n_T^3)$ rather than $O(n_J^3 n_T^3)$. Sampling from this posterior can happen in two phases where we first remove the customer and then consider how the likelihood term can be changed when this customer is replaced. The sampler thus has the potential to change multiple cluster assignments through a single change in customer assignment and using these moves is able to explore the space of possible partitions to determine a partition structure conditional on observed data.

3 Results

Occupancy is defined as the percentage of time that a location on the road is occupied by vehicles. In our example, occupancy data was generated using the AIMSUN microscopic simulator for a network in downtown San Francisco composed of 316 links and 158 junctions. This was recorded over a period of six hours with a sampling frequency of 180 seconds. We cluster the simulated network such that each individual cluster represents a level of occupancy that is distinct from other clusters, as shown in Figure 2.

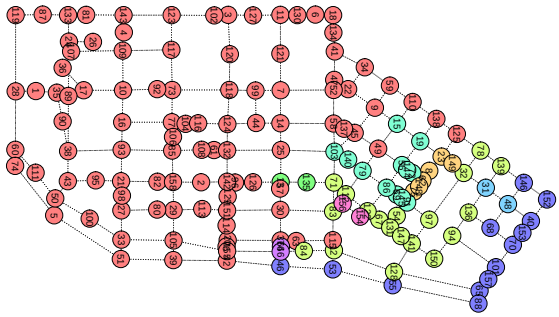


FIGURE 2. Traffic network with clusters that indicate different congestion levels.

This paper proposes a Bayesian clustering algorithm that accounts for spatial constraints and is modelled in a computationally efficient manner on data with varying temporal patterns. Further work seeks to identify clusters that change over time. However, current Kronecker product tricks that enhance efficiency cannot be utilized since spatial precision would change over time.

Acknowledgments: Research funded by the Lord Kelvin Adam Smith scholarship, University of Glasgow, 2014–2018

References

Blei, D. and Frazier, P. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12(Aug):2461 – 2488.

Leroux, B., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179 – 191. Springer.

Bayesian variable selection for identifying the source of food-borne disease outbreaks

Rianne Jacobs[1], Emmanuel Lesaffre[2], Peter Teunis[1], Jan van de Kassteele[1]

[1] National Institute for Public Health and the Environment (RIVM), Bilthoven, NL
[2] L-Biostat, KU Leuven, Leuven, Belgium

E-mail for correspondence: rianne.jacobs@rivm.nl

Abstract: Early identification of contaminated food products is crucial in reducing social and economic burdens of food-borne disease outbreaks. Analytic case-control studies are primarily used in this identification. In this paper, we develop a Bayesian variable selection method to account for misclassified responses and missing covariates. The method, implemented in JAGS/R Software is used to analyse the *Salmonella* Thompson 2012 outbreak data.

Keywords: Bayesian variable selection; Missing value imputation; Scaled logistic regression; Spike-slab variable selection.

1 Introduction

With food chains becoming increasingly complex and food products being transported across the globe with increasing ease, contaminated food products can rapidly cause food-borne disease outbreaks. Such outbreaks have a large social and economic burden on society; see for example the *Salmonella* Thompson 2012 outbreak in the Netherlands (Friesema et al. 2014). Early detection of such outbreaks and the subsequent identification of the contaminated food product(s) is crucial in reducing these burdens.

Identification of contaminated food products is a long, cumbersome process involving several steps which are not clear cut - much like a criminal investigation where information is incomplete, delayed, uncertain and continually updated. Analytic case-control studies are the main epidemiological tool in this process of identification. Once an outbreak has been detected, patients fill out an extensive

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

food consumption questionnaire for some period prior to becoming ill. Simultaneously, controls are sampled matched according to municipality, age and gender and they also fill out the food consumption questionnaire. One can well imagine the practical difficulties that subjects have trying to recall their dietary consumption and the resulting amount of missing values in such data. In addition, although controls are questioned on their symptoms, it is impossible to confirm whether they are indeed true controls (i.e. not infected) or rather asymptomatic infections (i.e. infected but not ill), resulting in misclassification of the response.

The analysis of the questionnaire data typically involves classical logistic regression. Due to the large number of different food products people may have consumed, one often has a variable selection problem, where one attempts to identify relevant exposures. Moreover, in the beginning of an outbreak, the number of covariates (i.e. food products) may very well be close to or even greater than the number of observations. Classical variable selection procedures, i.e. a combination of univariable analysis and stepwise forward or backward selection based on p-values, are most employed (Friesema et al. 2014), thereby ignoring the classical problems of multiple testing and bias they imply. When searching for the cause of an outbreak we, therefore, need a far more sophisticated variable selection procedure.

We argue that the Bayesian approach offers tools to deal with variable selection problems with missing covariates and misclassified responses. Indeed, Bayesian methods allow us to use external information to aid the modelling when data are scarce. This is crucial in the analysis of our case-control data especially in the light of early identification when very few data are yet available. In addition, Bayesian methods provide us with the flexibility to account for the additional problems of missing covariates and misclassified responses which is hard to solve in the frequentist setting. In this paper, we, therefore, develop a Bayesian variable selection method which also accounts for misclassified responses and missing covariates.

2 Data and method

The *Salmonella* Thompson 2012 outbreak data was obtained from a case-control study performed by the National Institute for Public Health and the Environment (RIVM) in which smoked salmon was found to be the source of the outbreak but only after six weeks since the onset of the outbreak. The percentage of missing covariates is high - up to 67% for a given observation. *Salmonella* infections are often asymptomatic, i.e. an infected person does not become ill, implying a sensitivity of less than one, $P(Y = 1|T = 1) < 1$ (Y: observed response, T: true response). Moreover, as a case only entered the dataset if it was twice laboratory-confirmed, we assume that no non-infected person entered the dataset as a case, i.e. $P(Y = 1|T = 0) = 0$, implying that the specificity is one, $P(Y = 0|T = 0) = 1$.

To deal with the problem of misclassified responses, our Bayesian variable selection model extends the logistic regression model. Logistic regression models the probability of an observed success, $P(Y = 1)$, which is equal to a true success, $P(T = 1)$, for a correctly classified response. In our case, however, this equality is contaminated by the sensitivity (Se), $P(Y = 1) = P(T = 1)P(Y = 1|T = 1)$,

resulting in the scaled logistic regression model:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\mu_i) \\ \mu_i &= \pi_i \times \text{Se} \\ \text{logit}(\pi_i) &= \beta_0 + \sum_{j=1}^{p} x_{i,j} \times \beta_j. \end{aligned} \tag{1}$$

The scaled logistic model is, however, unidentifiable without extra information from, e.g., validation data. In a Bayesian setting, one may, in addition also use historical information to feed a prior distribution on the sensitivity.

To incorporate Bayesian variable selection, we apply the stochastic search variable selection (SSVS) method (George and McCulloch, 1993) in which a mixture prior on the parameters, β_j, with one spike and one slab Gaussian component is constructed. The variances of the spike and slab are τ^2 and $c^2\tau^2$, respectively. The spike and slab prior is given by

$$\begin{aligned} \beta_j|\tau^2, c^2 &\sim \gamma_j \text{N}(0, \tau^2 c^2) + (1-\gamma_j)\text{N}(0, \tau^2), \\ \gamma_j|\omega_j &\sim \text{Bernoulli}(\omega_j) \\ \omega_j &\sim \text{Beta}(a_{j,0}, b_{j,0}) \end{aligned} \tag{2}$$

where γ_j is the indicator variable for inclusion of β_j into the model with inclusion probability, ω_j. Parameters $a_{j,0}$ and $b_{j,0}$ are chosen to reflect prior knowledge about the probability that a covariate should be in the model.

The large percentage of missing covariates in the data implies that many of the $x_{i,j}$'s in Eq. 1 are missing. To impute these covariates, we construct a covariate probability model as

$$f(x_1, x_2, ..., x_p) = f(x_1) \sum_{j=2}^{p} f(x_j|x_1, ..., x_{j-1}) \tag{3}$$

with $f(x_j|x_1, ..., x_{j-1})$ a Bernoulli distribution with logit link function. In the case of many covariates, as in our data, it is reasonable to assume that covariate x_j does not depend on all $j-1$ covariates. Similarly to the variable selection of the response model (Eq. 2), we perform variable selection in each of the regression models of the covariate probability model in Eq. 3, applying the 2-level variable selection model of Mitra and Dunson (2010).

3 Data analysis

We applied our Bayesian variable selection model to the *Salmonella* Thompson data. The model was implemented in R Software using JAGS for the MCMC sampling, running 5 chains with a burn-in of 1000 iterations and then a further 4000 iterations per chain. Trace plots showed good mixing. In this analysis, priors for the inclusion probabilities were chosen to favour slightly parsimonious models, namely $\omega_j \sim \text{Beta}(1, 2)$, based on the idea that a food product is not guilty unless proven so by the data. In order to make the scaled logistic model identifiable, we need an informative prior for the sensitivity. In this analysis, we used $\text{Se} \sim \text{Beta}(33, 4)$, assuming a median sensitivity of 0.9 and 5^{th} percentile of 0.8. The

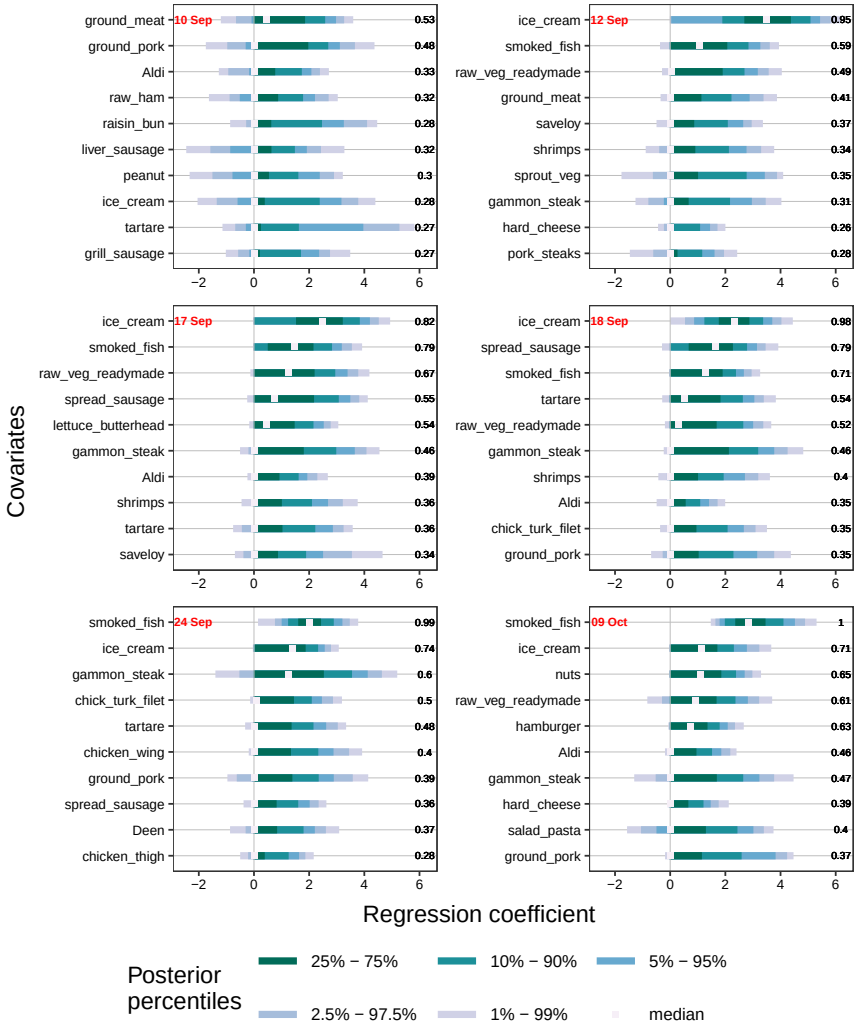


FIGURE 1. Posterior percentiles of regression coefficients and corresponding one-sided posterior inclusion probabilities, $P(\beta_j > 0.05)$, in the analysis of subsets of the *Salmonella* Thompson data simulating the available data at different time points during the outbreak.

variance parameters, τ and c, of the spike and slab components, were chosen such that the two distributions intersect at a practical significance level of 0.05 and that the slab distribution is relatively narrow avoiding excessively large β's, which are very unlikely in practice. The intercept terms of response and covariate models were given a diffuse normal prior distribution. Looking at the posterior inclusion probabilities of each covariate (Fig. 1), the covariate for smoked fish clearly had the highest inclusion probability and, therefore, our model correctly

identified the contaminated food product.

4 Conclusion

In this paper, we developed a method that deals with the problems of variable selection, missing covariates and misclassified responses in the context of source identification in food-borne disease outbreaks. We have shown how a Bayesian analysis allows a relatively easy implementation of these concepts in the re-analysis of the Dutch Salmonella *Thompson* 2012 outbreak data. Moreover, the Bayesian analysis performed better than the standard logistic backward variable selection model (not shown in this paper).

The method presented in this paper constitutes a first attempt at formalizing the methodology necessary for the analysis part of food-borne disease outbreak investigations. Current procedures are very much ad hoc in nature resulting in difficult to interpret and misleading results. The method in this paper is methodologically sound and the analysis results are intuitive and easy to interpret.

References

Friesema, I., de Jong, A., Hofhuis, A., ... and van Pelt, W. (2014). Large outbreak of Salmonella Thompson related to smoked salmon in the Netherlands, August to December 2012. *Euro Surveillance*, **19**, 20918.

George, E.L. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881 – 889.

Mitra, R. and Dunson, D. (2010). Two-Level Stochastic Search Variable Selection in GLMs with Missing Predictors. *The International Journal of Biostatistics*, **6**, 33.

Using extended dgLARS to study Diabetes

Hassan Pazira[1] and Ernst Wit[2]

[1] Johann Bernoulli Institute (JBI), University of Groningen, The Netherlands

E-mail for correspondence: h.pazira@rug.nl

Abstract: Diabetes is one of the most common long-term health conditions. Today, diabetes takes more lives than AIDS and breast cancer combined. It is a leading cause of blindness, kidney failure, amputations, heart failure and stroke. The aim of this study is to use the differential geometric generalization of the LARS algorithm for the double-parameter Gamma GLM with a canonical link function to identify candidate factors that may be associated with diabetes. To compute the solution path faster, we use the improved predictor-corrector algorithm, as proposed in Pazira *et al.* (2017).

Keywords: High-dimensional inference, Gamma GLMs, Least angle regression, Improved predictor-corrector algorithm, Dispersion parameter.

1 Introduction

In the last decade, the cases of people living with *diabetes* jumped almost 50 percent. Worldwide, it afflicts more than 380 million people. And the World Health Organization estimates that by 2030, that number of people living with diabetes will more than double. Today, diabetes takes more lives than AIDS and breast cancer combined. It is a leading cause of blindness, kidney failure, amputations, heart failure and stroke. Living with diabetes places an enormous emotional, physical and financial burden on the entire family.

In this paper we consider the benchmark *diabetes* data used in Efron *et al.* (2004) and Ishwaran *et al.* (2010), among others. The response y is a quantitative measure of disease progression for patients with diabetes one year later. The data includes 10 baseline measurements for each patient, such as *age*, *sex*, *bmi* (body mass index), *map* (mean arterial blood pressure), and six blood serum measurements: *ldl* (high-density lipoprotein), *hdl* (low-density lipoprotein), *ltg* (lamotrigine), *glu* (glucose), *tc* (triglyceride) and *tch* (total cholesterol), in addition to 45 interactions and 9 quadratic terms, for a total of 64 variables for each patient, so

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

that this data has $n = 442$ observations on $p = 64$ variables. The aim of the study is to identify which of the covariates are important factors in disease progression. This diabetes data can be found in the new version of our `dglars` package.

In literature several methods have been proposed to identify variables that can affect the disease. For this kind of problems the number of variables, say p, can be much larger than the sample size n. In this case, it is often assumed that only a small number of covariates (factors) contributes to the response, which leads to assume the sparsity of the model, many elements of the coefficients vector $\boldsymbol{\beta}$ are equal to zero. In recent statistical literature, many variable selection techniques for sparse regression models are based on the penalized likelihood approach to estimate a solution curve embedded in the parameter space and then to find the point that represents the best compromise between sparsity and predictive behaviour of the model. Some important examples are the Least Absolute Shrinkage and Selection Operator (LASSO) estimator (Tibshirani, 1996), the Smoothly Clipped Absolute Deviation (SCAD) method (Fan and Li, 2001), among others.

Differently from the methods cited above, Efron *et al.* (2004) introduced a new method to select important variables in a linear regression model called least angle regression method (LARS). Augugliaro *et al.* (2013) proposed a new approach based on the differential geometrical representation of a GLM. The method, which does not require an explicit penalty function, has been called differential geometric LARS (dgLARS) method because it is defined generalizing the geometrical ideas on which LARS is based. The later authors considered a class of the *exponential family* for a GLM, so that, they assumed that the dispersion parameter is known, $\phi = 1$. They also used the predictor-corrector (PC) algorithm to compute the solution curve.

In this paper we consider the dgLARS method for a larger class of the exponential family, namely the *exponential dispersion family*, when the dispersion parameter ϕ is unknown, and obtain the extended dgLARS estimator for Gamma GLM with arbitrary link function. Aim of this paper is to identify a set of important factors that can affect the diabetes disease using the extended dgLARS.

2 Differential Geometric LARS for general GLM

Let $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)^\top$ be a n-dimensional random vector with independent components. In what follows we shall assume that Y_i is a random variable with p.d.f belonging to an exponential dispersion family (Jorgensen, 1987, 1997), i.e.,

$$p_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}, \qquad y_i \in \mathcal{Y}_i \subseteq \mathbb{R}, \tag{1}$$

where $\theta_i \in \Theta_i \subseteq \mathbb{R}$ is the canonical parameter, $\phi \in \Phi \subseteq \mathbb{R}^+$ is the dispersion parameter, and $a(.)$, $b(.)$ and $c(.,.)$ are given functions. In the following, we assume that each Θ_i is an open set and $a(\phi) = \phi$. We consider ϕ as an unknown parameter. The expected value of $\mathbf{Y}$ is related to the canonical parameter by $\boldsymbol{\mu} = \{\mu(\theta_1), \cdots, \mu(\theta_n)\}^\top$, where $\mu(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}$ is called mean value mapping, and the variance of $\mathbf{Y}$ is related to its expected value by the identity $\mathrm{Var}(\mathbf{Y}) = \phi\mathbf{V}(\boldsymbol{\mu})$, where $\mathbf{V}(\boldsymbol{\mu})$ is an $n \times n$ diagonal matrix with elements, called the variance functions, $V(\mu_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}$. Since μ_i is a reparameterization, model (1) can be also denoted as $p_{Y_i}(y_i; \mu_i, \phi)$.

Following McCullagh and Nelder (1989), a Generalized Linear Model (GLM) is defined by means of a known function $g(\cdot)$, called link function, relating the expected value of each Y_i to the vector of covariates $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ by the identity $g\{E(Y_i)\} = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ where η_i is called the i^{th} linear predictor and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the vector of regression coefficients. In order to simplify our notation we let $\boldsymbol{\mu}(\boldsymbol{\beta}) = \{\mu_1(\boldsymbol{\beta}), \ldots, \mu_n(\boldsymbol{\beta})\}^\top$ where $\mu_i(\boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$. Therefore, the joint probability density function can be written as $p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}), \phi) = \prod_{i=1}^n p_{Y_i}(y_i; \mu_i(\boldsymbol{\beta}), \phi)$. In the following of this paper we shall use $\ell(\boldsymbol{\beta}, \phi; \mathbf{y}) = \log p_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}(\boldsymbol{\beta}), \phi)$ as notation for the log-likelihood function, $\partial_m \ell(\boldsymbol{\beta}, \phi; \mathbf{y}) = \frac{\partial \ell(\boldsymbol{\beta}, \phi; \mathbf{y})}{\partial \beta_m} = \phi^{-1} \partial_m \ell(\boldsymbol{\beta}; \mathbf{y})$ as notation for the m^{th} score function, and $\mathcal{I}_{mn}(\boldsymbol{\beta}, \phi) = E[\partial_m \ell(\boldsymbol{\beta}, \phi; \mathbf{y}) \cdot \partial_n \ell(\boldsymbol{\beta}, \phi; \mathbf{y})] = \phi^{-1} \mathcal{I}_{mn}(\boldsymbol{\beta})$ as notation for a element of the Fisher Information matrix function.

The Rao's score test statistic, given as $r_m(\boldsymbol{\beta}, \phi) = \frac{\partial_m \ell(\boldsymbol{\beta}, \phi; \mathbf{y})}{\sqrt{\mathcal{I}_m(\boldsymbol{\beta}, \phi)}} = \phi^{-1} r_m(\boldsymbol{\beta})$, helps to define $\rho_m(\boldsymbol{\beta})$, the angle between the m^{th} basis function $\partial_m \ell(\boldsymbol{\beta}, \phi; \mathbf{Y})$ and the tangent residual vector $\boldsymbol{r}(\boldsymbol{\beta}, \phi, \mathbf{y}; \mathbf{Y}) = \sum_{i=1}^n (y_i - \mu_i) \frac{\partial \ell(\boldsymbol{\beta}, \phi; \mathbf{y})}{\partial \mu_i}$, defined as follows

$$\rho_m(\boldsymbol{\beta}, \phi) = \arccos \left[\frac{r_m(\boldsymbol{\beta}, \phi)}{\|\boldsymbol{r}(\boldsymbol{\beta}, \phi, \mathbf{y}; \mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}} \right], \tag{2}$$

where $\|\cdot\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}$ is the norm defined on the tangent space $\mathcal{T}_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{M}$, where the set $\mathcal{M}$ is a p-dimensional submanifold of the differential manifold $\mathcal{S}$. From (2), the Rao's score test statistic contains the same information as the angle $\rho_m(\boldsymbol{\beta}, \phi)$. Thereby we can define the extended dgLARS method with respect to the Rao's score test statistic rather than the angle as respects the smallest angle is equivalent to the largest Rao's score test statistic.

The extended dgLARS solution curve, which is denoted by $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma) \subset \mathbb{R}^{k+1}$, with $\gamma \in [0, \gamma^{(1)}]$, whereby $0 \leqslant \gamma^{(p)} \leqslant \cdots \leqslant \gamma^{(2)} \leqslant \gamma^{(1)}$, is defined in the following way: For any $\gamma \in (\gamma^{(k+1)}, \gamma^{(k)}]$, the dgLARS estimator is chosen in such a way that

$$\begin{aligned} \mathcal{A}(\gamma) &= \{a_1, a_2, \cdots, a_k\}, \\ |r_{a_i}(\hat{\boldsymbol{\beta}}(\gamma), \phi)| &= |r_{a_j}(\hat{\boldsymbol{\beta}}(\gamma), \phi)| = \gamma, \qquad \forall a_i, a_j \in \mathcal{A}(\gamma), \\ |r_{a_h^c}(\hat{\boldsymbol{\beta}}(\gamma), \phi)| &< |r_{a_i}(\hat{\boldsymbol{\beta}}(\gamma), \phi)| = \gamma, \qquad \forall a_h^c \in \mathcal{A}^c(\gamma) \text{ and } \forall a_i \in \mathcal{A}(\gamma), \end{aligned} \tag{3}$$

where $\mathcal{A}(\gamma) = \{m \,:\, \hat{\beta}_m(\gamma) \neq 0\}$ is called *active set* and $\mathcal{A}^c(\gamma) = \{m \,:\, \hat{\beta}_m(\gamma) = 0\}$ is the complement of the active set. The new covariate is included in the active set at $\gamma = \gamma^{(k+1)}$ where the following condition is satisfied:

$$\begin{aligned} &\exists a_h^c \in \mathcal{A}^c(\gamma^{(k+1)}) : \\ &\qquad |r_{a_h^c}(\hat{\boldsymbol{\beta}}(\gamma^{(k+1)}), \phi)| = |r_{a_i}(\hat{\boldsymbol{\beta}}(\gamma^{(k+1)}), \phi)|, \forall a_i \in \mathcal{A}(\gamma^{(k+1)}). \end{aligned} \tag{4}$$

From Pazira *et al.* (2017) we know that the value of the estimated dispersion parameter does not change the order of the variables included in the active set while it affects the value of $\ell(\hat{\boldsymbol{\beta}}, \hat{\phi}; \mathbf{y})$, and as a result it affects the value of various information criteria such as AIC or BIC.

For the high-dimensional setting ($p \geq n$), we use the dispersion estimator $\hat{\phi}_P(\gamma)$ at $\gamma \in [0, \gamma_{max}]$ by the Pearson-like dispersion estimator, as proposed by Wood

(2006);

$$\hat{\phi}_P(\gamma) = \frac{1}{n - |\mathcal{A}(\gamma)|} \sum_{i=1}^{n} \frac{(y_i - g^{-1}(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)))^2}{V(g^{-1}(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)))}, \tag{5}$$

where $|\mathcal{A}(\gamma)| = \#\{j : \hat{\beta}_j(\gamma) \neq 0\}$ such that $\hat{\beta}_j(\gamma)$ is the element of the extended dgLARS estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$.

To compute the solution curve, we use the improved predictor-corrector (IPC) algorithm, described in Pazira *et al.* (2017), because this algorithm leads to an decrease in the run times needed for computing the solution curve.

3 Application to Diabetes Dataset

In the recent literature, variable selection techniques, such as LARS and Spike and Slab, were used in a linear regression model applied to the explained diabetes data. While we spotted that, surprisingly, the response is markedly right-skewed which can arise from a non-normal distribution, e.g. Gamma. Therefore, we fit a Gamma regression model for this diabetes data and use the extended dgLARS method by means of the IPC algorithm.

We first apply a number of variable selection methods such as LARS (Efron *et al.*, 2004), LASSO (Tibshirani, 1996), Elastic Net (Zou and Hastie, 2005), and Spike and Slab (Ishwaran *et al.*, 2010) by using the `lars`, `glmnet` and `spikeslab` packages, and then compare the results to the results obtained from the proposed dgLARS method implemented by our `dglars` package. Note that, for the dgLARS method we use the *Gamma* family in our package, while this family is not available in other packages, so that we fit the *Gaussian* family to the data to be able to use these packages.

When we compare the results of the dgLARS Gamma method to the results obtained from other algorithms, we find out the remarkable results. From Table 1 we can see that, the variables selected by the LARS, LASSO and Elastic Net methods are the same, and almost in all models the first 4 variables ($3, 9, 4$ and 7) are the same. Moreover, importantly, all models (except the dgLARS) have the same selected variables just in the different order. While all algorithms (except the dgLARS) select the covariates $37, 12, 22, 27, 33$ and 52 in the first 20 variables, our proposed algorithm does not select them among the top 20 variables. Instead, the dgLARS algorithm by the Gamma model selects several new other variables (indicated in bold in Table 1) which none of the other algorithms do. For instance, the variables $60, 18, 42, 35$ and 40 are selected with the inverse link function.

TABLE 1. The sequences of the top 20 predictors selected by the LARS, LASSO, Elastic Net, Spike and Slab and dgLARS algorithms obtained for diabetes data.

Algorithm	Selected Variables																			
LARS	3	9	4	7	37	20	19	12	22	28	2	10	27	11	30	46	33	52	24	29
LASSO	3	9	4	7	37	20	19	12	22	28	2	10	27	11	30	46	33	52	24	29
Elastic Net	3	9	4	7	37	12	20	19	10	22	28	2	27	30	11	52	46	33	24	29
Spike and Slab	3	9	4	7	2	20	37	19	12	27	52	11	10	22	63	30	24	58	43	5
dgLARS (*inverse*)	3	9	4	7	20	**60**	2	46	**18**	10	**42**	28	11	19	30	**35**	29	**40**	24	63

TABLE 2. A list of the top 20 selected variables and their parameter estimates obtained using dgLARS Gamma method (with inverse canonical link, $\eta_i = -\frac{1}{\mu_i}$) for diabetes data. $|\mathcal{A}_{CV}| = 20$, $|\mathcal{A}_{AIC}| = 24$ and $|\mathcal{A}_{BIC}| = 10$.

	Variable		Coefficient Estimate		
Step	Name	Number	CV	AIC	BIC
1	*bmi*	3	0.0182	0.0187	0.0171
2	*ltg*	9	0.0262	0.0278	0.0205
3	*map*	4	0.0129	0.0136	0.0101
4	*hdl*	7	-0.0145	-0.0159	-0.0105
5	*age* : *sex*	20	0.0067	0.0069	0.0042
6	*hdl* : *ltg*	**60**	0.0046	0.0053	0.0032
7	*sex*	2	-0.0090	-0.0113	-0.0035
8	*map* : *hdl*	46	0.0040	0.0052	0.0011
9	*ltg*^2	**18**	-0.0053	-0.0067	-0.0015
10	*glu*	10	0.0001	-0.0001	0
11	*bmi* : *ltg*	**42**	-0.0026	-0.0032	0
12	*age* : *glu*	28	0.0008	0.0010	0
13	*age*^2	11	0.0021	0.0027	0
14	*glu*^2	19	0.0016	0.0012	0
15	*sex* : *map*	30	0.0012	0.0020	0
16	*sex* : *ltg*	**35**	0.0007	0.0015	0
17	*sex* : *bmi*	29	0.0006	0.0015	0
18	*bmi* : *hdl*	**40**	0.0004	0.0015	0
19	*age* : *ldl*	24	-0.0002	-0.0014	0
20	*tch* : *glu*	63	0	0.0013	0

In Table 2, we also report the sequence of the top 20 variables and their parameter estimates obtained using the dgLARS Gamma method with canonical link, $\eta_i = -\frac{1}{\mu_i}$, based on variable selection methods: AIC, BIC and CV.

As a result, the extended dgLARS method based on a *Gamma* model, with the inverse canonical link function, finds out that the variables "*hdl* : *ltg*", "*ltg*^2", "*bmi* : *ltg*", "*sex* : *ltg*" and "*bmi* : *hdl*" (namely: $60, 18, 42, 35$ and 40) are more important factors in disease progression than the variables "*bmi* : *map*", "*bmi*^2", "*age* : *map*", "*age* : *ltg*", "*sex* : *hdl*" and "*tc* : *tch*" (namely: $37, 12, 22, 27, 33$ and 52).

References

Augugliaro, L., Mineo, A.M., and Wit, E.C. (2013). Differential Geometric Least Angle Regression: A Differential Geometric Approach to Sparse Generalized Linear Models, *J. R. Statist. Soc. B*, **75**, 471 – 498.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *Ann. Statist.*, **32**, 407 – 499.

Fan, J. and Li, R. (2001). Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties. *J. Am. Statist. Ass.*, **96**, 1348 – 1360.

Hastie T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Ishwaran, H., Kogalur, U.B., and Rao, J.S. (2010). spikeslab: Prediction and variable selection using spike and slab regression. *The R Journal.* **2**, 68 – 73.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Liner Models.* London: Chapman and Hall.

Pazira, H., Augugliaro, L., and Wit, E.C. (2017). Extended differential geometric LARS for high-dimensional GLMs with general dispersion parameter. *Submitted*.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *J. R. Statist. Soc. B*, **58**, 267 – 288.

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

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Statist. Soc. B*, **67**, 301 – 320.

Stable Variable Selection with AdaSub

Christian Staerk[1], Maria Kateri[1]

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

E-mail for correspondence: christian.staerk@rwth-aachen.de

Abstract: The recently proposed Adaptive Subspace Method (AdaSub) aims at finding the best model according to a certain variable selection criterion. We illustrate the performance of AdaSub and its stability with respect to changes of its tuning parameters. Furthermore we demonstrate through simulations that in case the model selected by AdaSub does not coincide with that proposed by the criterion, AdaSub often reduces the number of falsely selected variables (false positives) and provides a more stable model.

Keywords: Adaptive Subspace Method; Variable Selection; Linear Regression; Bayesian Information Criterion.

1 Introduction

We consider the problem of variable selection in linear regression models. Classical variable selection criteria include the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), aiming at optimal predictions and identification of the true generating model, respectively. The challenging problem with these ℓ_0-type selection criteria is that they lead to combinatorial and in general NP-hard optimization problems: If there are p possible explanatory variables, then there are 2^p possible models for which the criterion has to be evaluated in a full enumeration. To overcome this prohibitive computational approach, Staerk et al. (2016) propose the Adaptive Subspace Method (AdaSub) which is based on the idea of adaptively solving lower-dimensional sub-problems in order to provide a solution to the original problem. It can be shown that, under certain conditions, AdaSub identifies the best model according to the criterion used (Staerk et al., 2017).

This work focuses on two important issues related to the performance of the AdaSub algorithm. Since AdaSub is a stochastic algorithm in which certain tuning parameters have to be specified, it is crucial to investigate the role of these parameters and their effect on the stability of the model selected by AdaSub. After

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

a brief presentation of the AdaSub algorithm in Section 2, this issue is addressed in Section 3.

The second issue refers to the interesting scenario in which AdaSub does not identify the best model according to the criterion. It is well-known that the BIC is variable selection consistent in the classical asymptotic setting, i.e. the best BIC model coincides with the true generating model with probability tending to one if the sample size n tends to infinity and the number of explanatory variables p is bounded. However, this does not necessarily imply that the best BIC model is always a good approximation to the "truth" for finite sample sizes n. In fact, it will be demonstrated that in small sample size situations the best BIC model tends to select many noise variables (false positives) and to overfit the data if the true underlying model is relatively sparse. In Section 3 we show through simulations that AdaSub can mitigate the problem of selecting many "unstable" variables in such situations.

2 Adaptive Subspace Method (AdaSub)

Let $\{X_j;\ j \in \mathcal{P}\}$ be the set of explanatory variables with index set $\mathcal{P} = \{1, \ldots, p\}$ and let $\mathcal{M} = \mathfrak{P}(\mathcal{P}) = \{S \subseteq \{1, \ldots, p\}\}$ be the corresponding space of linear models. Suppose that we observe data $\mathcal{D} = (X, Y)$ with design matrix X and response Y. Let $C : \mathcal{M} \to \mathbb{R}$ be any model selection criterion for the given data $\mathcal{D}$ (e.g. the BIC). We assume that $C(S) \neq C(S')$ for all $S \neq S'$ and that we want to find the model $S^* \in \mathcal{M}$ that minimizes C, i.e. $S^* := \arg\min_{S \in \mathcal{M}} C(S)$.

Define the map $f_C : \mathcal{M} \to \mathcal{M}$ by $f_C(V) := \arg\min_{S \subseteq V} C(S)$ for $V \in \mathcal{M}$. So $f_C(V)$ denotes the best model according to criterion C among all models included in V. The steps of AdaSub are given by:

(1) Initialize expected search size $q \in (0, p)$, learning rate $K > 0$ and number of iterations $T \in \mathbb{N}$.
(2) For $j \in \mathcal{P}$ initialize $r_j^{(0)} = \frac{q}{p}$.
(3) For $t = 1, \ldots, T$:
   (a) Draw $b_j^{(t)} \sim \text{Bernoulli}(r_j^{(t-1)})$ independently for $j \in \mathcal{P}$.
   (b) Set $V^{(t)} = \{j \in \mathcal{P};\ b_j^{(t)} = 1\}$.
   (c) Compute $S^{(t)} = f_C(V^{(t)})$.
   (d) For $j \in \mathcal{P}$ update $r_j^{(t)} = \frac{q + K \sum_{i=1}^{t} 1_{S^{(i)}}(j)}{p + K \sum_{i=1}^{t} 1_{V^{(i)}}(j)}$, where 1_A denotes the indicator function of a set A.

As the final subset selected by AdaSub one can either (i) choose the "best" sampled model $\hat{S}_\text{b}$ for which $C(\hat{S}_\text{b}) = \min\{C(S^{(1)}), \ldots, C(S^{(T)})\}$, or (ii) consider $\hat{S}_\rho = \{j \in \mathcal{P};\ r_j^{(T)} > \rho\}$ with some threshold $\rho \in (0, 1)$.

3 Choice of Tuning Parameters in AdaSub

In order to illustrate the performance of AdaSub in a high-dimensional set-up and how it is effected by the choice of the tuning parameters K and q, we consider an example with $p = 1000$ and $n = 100$. For this we generate data $\mathcal{D} = (X, Y)$ by simulating $X = (X_{ij}) \in \mathbb{R}^{n \times p}$ with independent rows

$X_{i,*} \sim \mathcal{N}_p(0, \Sigma)$, where $\Sigma_{kl} = 0.3$ for $k \neq l$ and $\Sigma_{kk} = 1$. Furthermore let $\beta^0 = (1, -1, 1, 2, -2, 2, 0, \ldots, 0)^T \in \mathbb{R}^p$ be the true vector of coefficients with active set $S_0 = \{1, \ldots, 6\}$. The response $Y = (Y_1, \ldots, Y_n)^T$ is simulated via $Y_i \overset{\text{ind.}}{\sim} N(X_{i,*}\beta^0, 1)$, $i = 1, \ldots, n$. The criterion C we adopt is the Extended BIC (EBIC) with parameter $\gamma = 0.5$, which is especially suited for high-dimensional situations (Chen and Chen, 2008). The R-package "leaps" (Lumley and Miller, 2009) is used to compute $f_C(V)$ for $V \in \mathcal{M}$.

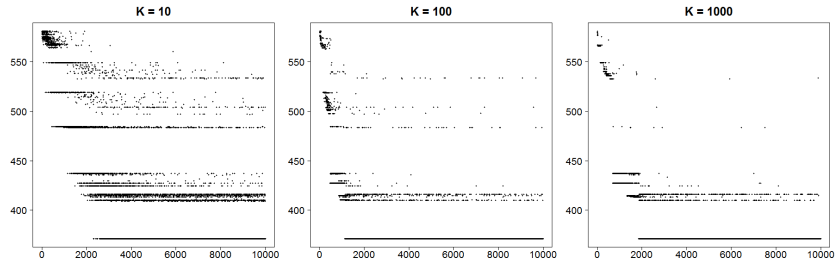


FIGURE 1. AdaSub for high-dimensional example: Plot of the evolution of EBIC($S^{(t)}$) along the iterations (t) for different values of K ($q = 5$ fixed).

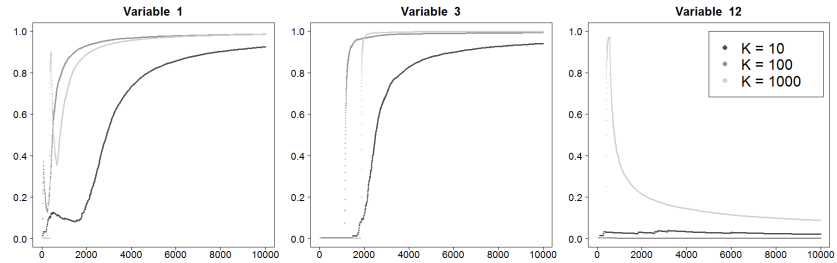


FIGURE 2. AdaSub for high-dimensional example: Plot of the evolution of $r_j^{(t)}$ for $j \in \{1, 3, 12\}$ along the iterations (t) for different values of K ($q = 5$ fixed).

We apply AdaSub with $T = 10,000$ iterations on a dataset simulated as above and fix $q = 5$ as the initial expected search size. Figure 1 and Figure 2 show the evolution of EBIC($S^{(t)}$) and $r_j^{(t)}$ along the iterations t for different values of the learning rate $K \in \{10, 100, 1000\}$. In all three cases the algorithm identifies the correct model $S_0 = \{1, \ldots, 6\}$ with EBIC(S_0) ≈ 371 based on $\hat{S}_{\text{b}}$ and $\hat{S}_\rho$ for $\rho \in [0.2, 0.9]$, i.e. $S_0 = \hat{S}_{\text{b}} = \hat{S}_\rho$.

However, there is a trade-off in choosing $K > 0$: If K is small ($K = 10$), then AdaSub adapts slowly to the information learned about the variables and hence a very diverse range of models is considered. If instead K is large ($K = 1000$), then the algorithm might actually "converge" too fast. Suppose for example that a variable X_j is not chosen when it is first considered in the model search (i.e. $j \in V^{(t)}$ but $j \notin S^{(t)}$), then $r_j^{(t)} = \frac{q}{p+K} \approx 0$ for K very large, so variable X_j will probably not be considered in the model search for a long time. In our case, the choice $K = 100 = n$ seems favourable, for which AdaSub only needed 1123

iterations to find S_0. Nevertheless, an important observation is that the selected model by AdaSub is stable with respect to changes of K, as long as the number of iterations is large enough.

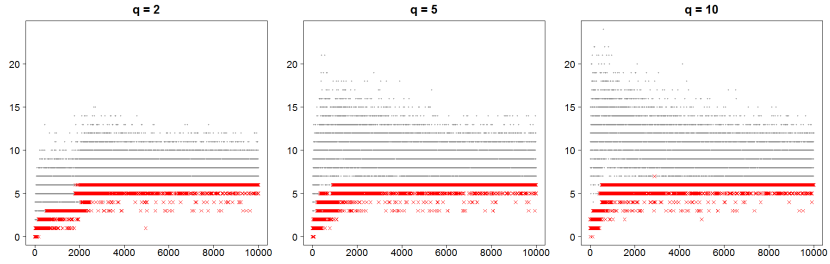


FIGURE 3. AdaSub for high-dimensional example: Plot of the sizes of the sampled sets $V^{(t)}$ (grey dots) and the sizes of the "best" subsets $S^{(t)}$ (red crosses) along the iterations (t) for different values of q ($K = 100$ fixed).

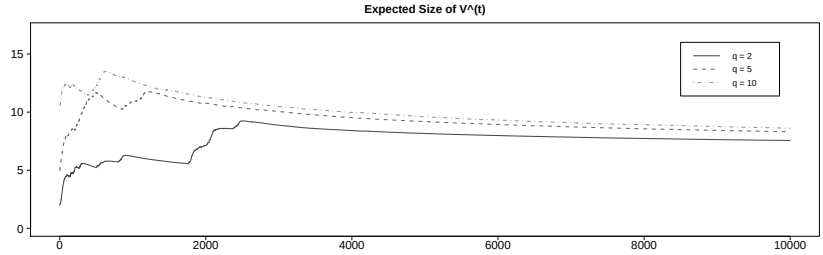


FIGURE 4. AdaSub for high-dimensional example: Plot of the evolution of the expected search size along the iterations (t) for different values of q ($K = 100$ fixed).

We now apply AdaSub with $T = 10,000$ iterations on the same dataset as above for different values of $q \in \{2, 5, 10\}$, while $K = 100$ is fixed. Figure 3 shows the sizes of the sampled sets $V^{(t)}$ and the sizes of the "best" subsets $S^{(t)} = f_C(V^{(t)})$ along the iterations t, while Figure 4 depicts the evolution of the expected sizes of the sets $V^{(t)}$ which are given by $E[|V^{(t)}|] = \sum_{j \in \mathcal{P}} r_j^{(t-1)}$ for $t = 1, \ldots, T$. We can see that the AdaSub algorithm automatically and quickly adjusts the expected search sizes $E[|V^{(t)}|]$ and that the algorithm "converges" against the true underlying model S_0 with six variables, no matter which initial expected search size q is used. Ideally, the tuning parameter q should be chosen in a way such that it reflects the subjectively expected size of the best model $S^* = f_C(\mathcal{P})$ according to criterion C. However, a general observation is that the choice of the tuning parameter q seems not to be as crucial as the proper choice of the learning rate K. A more detailed and refined analysis of theses issues is an interesting topic for future research.

4 Improving Stability of BIC by AdaSub

In order to investigate the stability of BIC, we consider a low-dimensional scenario with $p = 30$, so that the determination of the best BIC model S^* is computationally feasible. The sample size n is increased from 40 to 200 in steps of size 20 and for each value of n we simulate 100 different datasets according to the simulation setup described above, with the following modification: For each dataset, we select $s_0 \in \{0, \dots, 10\}$ and $S_0 \subset \mathcal{P}$ of size $|S_0| = s_0$ randomly; then for $j \in S_0$ we independently simulate $\beta_j^0 \sim \mathcal{U}[-2, 2]$ from the uniform distribution, while we set $\beta_j^0 = 0$ for $j \notin S_0$. Based on these simulated data, we compare the performance of the thresholded model $\hat{S}_\rho$ with threshold $\rho = 0.9$ from AdaSub with the best BIC model S^*. In AdaSub we set $T = 2000$, $q = 5$ and $K = n$.

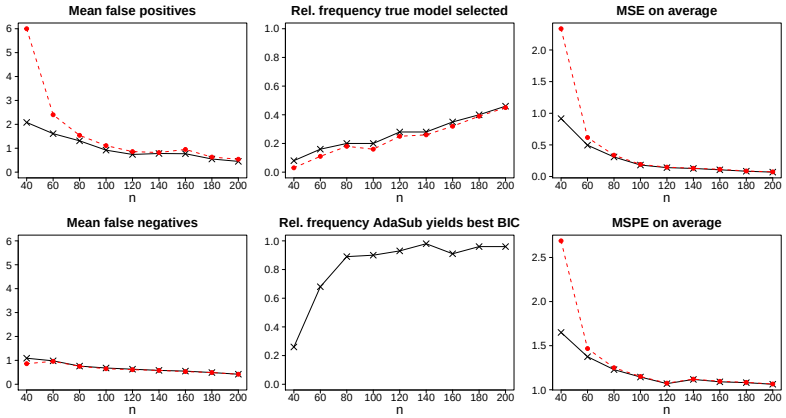


FIGURE 5. Comparison of $\hat{S}_{0.9}$ from AdaSub (solid lines with crosses) with best BIC model S^* (dashed lines with dots) in terms of mean number of false positives/ false negatives, relative frequency of selecting the true model S_0, rel. frequency of agreement between $\hat{S}_{0.9}$ and S^*, Mean Squared Error (MSE) and Mean Squared Prediction Error (MSPE) on independent test set with sample size 100.

Figure 5 shows that the best BIC model S^* tends to select many false positives if the sample size is small. In contrast, the thresholded model $\hat{S}_{0.9}$ selected by AdaSub is often sparser and yields less false positives in situations where the BIC is too liberal (at the prize of a slightly increased mean of false negatives). When the sample size increases, the best BIC model S^* becomes more "stable" and the relative frequency that $\hat{S}_{0.9}$ and S^* agree tends to 1. However, selecting the thresholded model from AdaSub is beneficial for small sample sizes yielding higher relative frequencies of selecting the true model S_0, smaller Mean Squared Errors (MSE) and smaller Mean Squared Prediction Errors (MSPE).

A reason for the undesirable behaviour of the best BIC model is that the discrete nature of the ℓ_0-penalty can lead to "overfitting" of the criterion, since the optimization is carried out among all possible $2^{30} \approx 10^9$ models. This problem has been addressed both by the Statistics community (e.g. Breiman, 1996) and the Machine Learning community (e.g. Loughrey and Cunningham, 2005). However,

the simulation results show that AdaSub can mitigate the "overfitting problem" of BIC in the given situation of a sparse underlying true model.

The tendency that AdaSub selects a sparser model in unstable situations is also observed in additional simulations with different correlation structures of X and different selection criteria C. In ongoing research we want to investigate this phenomenon further and aim to provide theoretical explanations for the promising performance of AdaSub.

References

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24(6)**, 2350 – 2383.

Chen, J. and Chen, Z. (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Spaces, *Biometrika*, **95(3)**, 759 – 771.

Loughrey, J. and Cunningham, P. (2005). Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets, *In Research and Development in Intelligent Systems XXI*, Springer London , 33 – 43.

Lumley, T. and Miller, A. (2009). leaps: Regression Subset Selection. R package version 2.9, *http://CRAN.R-project.org/package=leaps*.

Staerk, C., Kateri, M. and Ntzoufras, I. (2016). An Adaptive Subspace Me-thod for High-Dimensional Variable Selection. In: *Proc. of the 31st International Workshop on Statistical Modelling*, Rennes, 295 – 300.

Staerk, C., Kateri, M. and Ntzoufras, I. (2017). High-Dimensional Variable Selection via Low-Dimensional Adaptive Learning. *(submitted)*

Boosting distributional regression models for multivariate responses

Andreas Mayr[1,2], Janek Thomas[2], Matthias Schmid[3], Florian Faschingbauer[1], Nadja Klein[4]

[1] Friedrich-Alexander-University (FAU) Erlangen-Nürnberg, Germany
[2] Ludwig-Maximilians-University Munich, Germany
[3] Rheinische Friedrich-Wilhelms-University Bonn, Germany,
[4] University of Melbourne, Melbourne Business School, Australia

E-mail for correspondence: andreas.mayr@fau.de

Abstract: We introduce a boosting algorithm for multivariate distributional regression that is able to estimate these complex models while simultaneously selecting the most informative variables in potentially high-dimensional settings. Our proposed method is evaluated empirically and applied in a recent birth cohort study. The aim is to model the growth of children via the combined distribution of height and weight during early childhood to identify possible predictors.

Keywords: correlated responses; gradient boosting; multivariate GAMLSS.

1 Introduction

One of the most popular semi-parametric statistical modelling approaches beyond the classical regression of the mean are generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005). The main idea of GAMLSS is that each parameter of the conditional distribution – not only the expected value – is modelled by its own additive predictor. This flexible framework can be further extended towards multivariate outcomes, in order to model the joined distribution of two or more responses (Klein et al., 2014). We combine this approach with an extended statistical boosting algorithm (Mayr et al., 2012) which allows to estimate statistical models while simultaneously selecting the most influential variables. Besides evaluating its performance in simulations, we use this approach for the joint distribution of height and weight of a German birth cohort study in order to select predictors for the growth of children. Particularly, we are looking for predictors in the clinical information from mother and

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

child at birth and in a questionnaire on the socio-demographic characteristics of the parents (e.g., education).

2 Multivariate GAMLSS

In case of a univariate Gaussian response $Y \sim N(\mu, \sigma^2)$ in GAMLSS we model both the location $\mu = \eta_\mu(x)$ and the scale $\sigma = \exp(\eta_\mu(x))$ of the conditional distribution via additive predictors. In case of a bivariate Gaussian distribution, e.g., for $Y = (\text{weight}, \text{height})^\top \in \mathbb{R}^2$,

$$\begin{pmatrix}\text{weight}\\ \text{height}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

we follow the same principle,leading to five different additive predictors, including one for the correlation parameter ρ.

$$\eta_{\mu_1}(x) = \mu_1, \quad \eta_{\sigma_1}(x) = \log(\sigma_1),$$

$$\eta_{\mu_2}(x) = \mu_3, \quad \eta_{\sigma_2}(x) = \log(\sigma_2), \quad \eta_\rho(x) = \rho/\sqrt{(1-\rho^2)}$$

3 Boosting multivariate GAMLSS

While variable selection is already a complicated issue for GAMLSS, the complexity further increases for multivariate distributions. In order to deal with this issue, we use a further extended version of the gradient boosting algorithm introduced in Mayr et al. (2012). The new algorithm fits in every step $m = 1, ..., m_{\text{stop}}$ all partial derivatives of the joint likelihood one-by-one to the base-learners $h_1(x_1)$,..., $h_p(x_p)$, selecting the best-performing one j^* for each dimension. Afterwards, the best overall update is selected based on the potential increase in the likelihood (Thomas et al., 2017):

$$\begin{aligned}
&\frac{\partial}{\partial\eta_{\mu_1}} l(y_1, y_2, \hat{\mu_1}^{[m]}, \hat{\sigma_1}^{[m]}, \hat{\mu_2}^{[m]}, \hat{\sigma_2}^{[m]}, \hat{\rho}^{[m]}) \overset{\text{select}}{\longrightarrow} j^*_{\mu_1}\\
&\frac{\partial}{\partial\eta_{\sigma_1}} l(y_1, y_2, \hat{\mu_1}^{[m]}, \hat{\sigma_1}^{[m]}, \hat{\mu_2}^{[m]}, \hat{\sigma_2}^{[m]}, \hat{\rho}^{[m]}) \overset{\text{select}}{\longrightarrow} j^*_{\sigma_1}\\
&\frac{\partial}{\partial\eta_{\mu_2}} l(y_1, y_2, \hat{\mu_1}^{[m]}, \hat{\sigma_1}^{[m]}, \hat{\mu_2}^{[m]}, \hat{\sigma_2}^{[m]}, \hat{\rho}^{[m]}) \overset{\text{select}}{\longrightarrow} j^*_{\mu_2} \quad \textbf{best} \overset{\text{update}}{\longrightarrow} \hat{\eta}^{[m+1]}_{\mu_2}\\
&\frac{\partial}{\partial\eta_{\sigma_2}} l(y_1, y_2, \hat{\mu_1}^{[m]}, \hat{\sigma_1}^{[m]}, \hat{\mu_2}^{[m]}, \hat{\sigma_2}^{[m]}, \hat{\rho}^{[m]}) \overset{\text{select}}{\longrightarrow} j^*_{\sigma_2}\\
&\frac{\partial}{\partial\eta_{\rho}} l(y_1, y_2, \hat{\mu_1}^{[m]}, \hat{\sigma_1}^{[m]}, \hat{\mu_2}^{[m]}, \hat{\sigma_2}^{[m]}, \hat{\rho}^{[m]}) \overset{\text{select}}{\longrightarrow} j^*_{\rho}
\end{aligned}$$

Only this best overall-update is finally carried out, leading to data driven variable selection and mode-choice. Note, that the boosting algorithm is fitting the

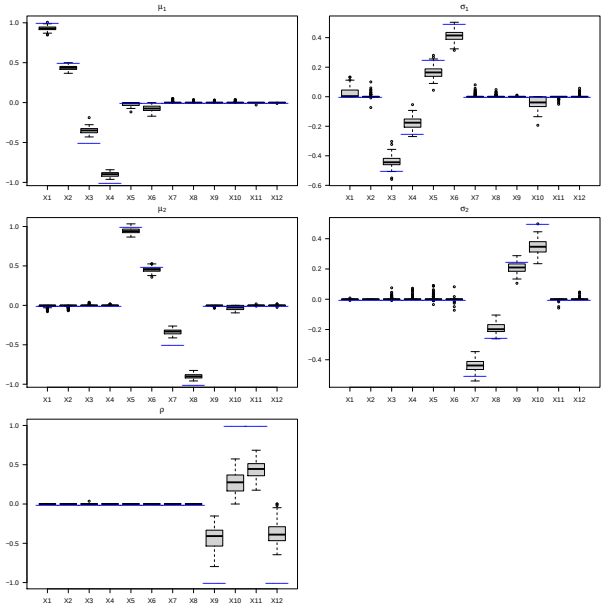


FIGURE 1. Resulting coefficients for 12 partly overlapping informative variables from 100 simulations with $n = 500$ and $p = 1000$. Each plot refers to one additive predictor $(\eta_{\mu_1}, \eta_{\sigma_1}, \eta_{\mu_2}, \eta_{\sigma_2}, \eta_\rho)$. The horizontal lines are the corresponding true values: Generally, the algorithm selects the correct variables – the amount of shrinkage differs among the predictors.

negative gradient of the joint likelihood (with respect to the different additive predictors) and not the actual observations.

An implementation of this extended algorithm is provided via a new option in the R add-on package `gamboostLSS` (Hofner et al., 2016).

4 Simulation

We first carried out a simulation study with $n = 500$ bivariate Gaussian distributed outcomes and $p = 1000$ explanatory variables. Only 12 variables have an actual effect on any of the distribution parameters (see Figure 1). Our results suggest that the algorithm converges to the correct solution and is able to identify a small subset of informative variables in potentially high-dimensional data situations . There is small tendency to falsely include informative variables in both scale parameters; e.g. X_{10}, which actually has only an effect on σ_2 and ρ is often also included for σ_1.

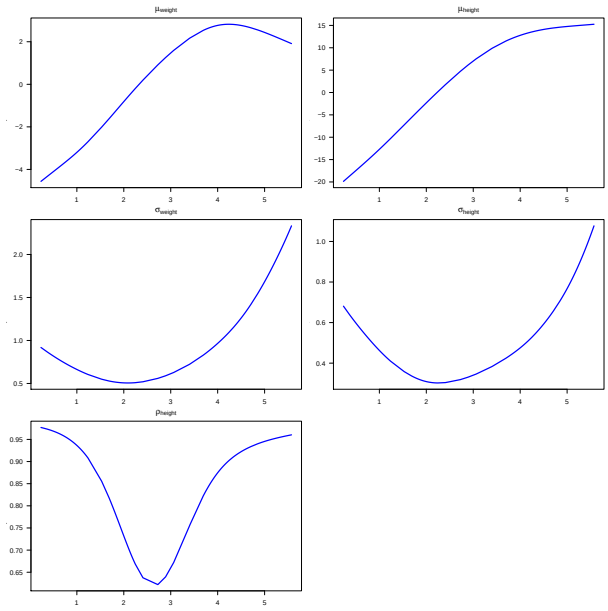


FIGURE 2. Partial effect of the age on the parameters of the joint distribution of weight and height in early childhood.

5 Birth cohort study

We apply our new approach analysing the joint distribution of weight and height in early childhood. The underlying data set is a birth cohort study including 453 children born at the University Hospital Erlangen, with measurements at five different time points. As possible explanatory variables we have the clinical information of the mother and the baby at birth, the weight and the height of the parents and results of a questionnaire on their socio-demographic background. For continuous variables we use P-splines as base-learners while all other variables enter via simple-linear models. To account for the longitudinal structure of the data we include subject-specific random intercepts.

Picking the most influential variables, the results of our models suggest, that the weight in early childhood is influenced to a much greater extend by the socio-demographic background of the parents (e.g., education level) than the height. The latter (μ_{height}) only depends on the parents' height at birth. Generally, the algorithm selected more variables for σ_{weight} and σ_{height} than for the mean parameters (see Table 1) . The correlation between height and weight follows a u-type shape, while the variance increases for both outcomes with the children getting older (Figure 2).

TABLE 1. Selected variables for the different additive predictors and the type of their effect (linear positive, linear negative, smooth) on the growth of children. The results are based on the analysis of a recent birth cohort study with the bivariate outcome $Y = (\text{weight}, \text{height})^\top$.

Variable	μ_{weight}	μ_{height}	σ_{weight}	σ_{height}	ρ
age child	smooth	smooth	smooth	smooth	smooth
birth-weight	smooth		smooth	smooth	
weight father	pos.		pos.		
weight mother	pos.		pos.		
height father	neg.	pos.	neg.	pos.	
height mother	neg.	pos.	neg.	pos.	
female child	neg.				
education mother	neg.		neg		
age mother				neg.	
breast-fed			pos.	neg.	
cesarean-section				neg.	
stress-level (birth)	neg.		neg.	neg.	
gestational age				neg.	neg.
alcohol (pregnancy)				neg.	neg.

6 Discussion

We propose a boosting algorithm to estimate and select GAMLSS distributional regression models (Rigby and Stasinopoulos, 2005) for multivariate outcomes. GAMLSS had been already extended towards multivariate distributions by Klein et al. (2014) in a Bayesian setting.

The proposed boosting algorithm, due to being based on a machine-learning approach, is applicable to high-dimensional data with more candidate variables than observations ($p > n$) – as illustrated in the simulation. It takes advantage of a recent extension (Thomas et al., 2017) of the original algorithm for boosting GAMLSS (Mayr et al., 2012). Following this new approach, tuning of the algorithm (via the number of boosting iterations) boils down to a one-dimensional problem, making the computationally burdensome grid search (Hofner et al., 2016) for the optimal combination of stopping iterations unnecessary.

In the application to the birth cohort study we illustrated how the algorithm simultaneously selects only the most influential variables for the different predictors: although being a relatively low-dimensional data set, this task would have been infeasible for most other variable selection approaches for distributional regression models.

Acknowledgments: The work on this article was supported by the Deutsche Forschungsgemeinschaft (http://www.dfg.de, grant number SCHM 2966/1-2) and the Interdisciplinary Center for Clinical Research of the Friedrich-Alexander Uni-

versity Erlangen-Nürnberg (project J49).

References

Klein, N., Kneib, T., Klasen, S. and Lang, S. (2014). Bayesian Structured Additive Distributional Regression for Multivariate Responses. *Applied Statistics*, **64**, 569-591.

Hofner, B., Mayr, A., Schmid, M. (2016). gamboostLSS: An R Package for Model Building and Variable Selection in the GAMLSS Framework. *Journal of Statistical Software* , **74**(1), 131.

Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012). Generalized additive models for location, scale and shape for high-dimensional data – a flexible approach based on boosting. *Applied Statistics*, **61**(3), 403-427.

Mayr, A., Binder, H, Gefeller, O. and Schmid, M. (2014). The Evolution of Boosting Algorithms - From Machine Learning to Statistical Modelling. *Meth Inf Med*, **53**(6), 419-427.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., Hofner, B. (2017). Stability selection for component-wise gradient boosting in multiple dimensions. *arXiv*: 1611.10171.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**, 507 – 554.

LASSO-type penalization in the framework of Generalized Additive Models for Location, Scale and Shape

Andreas Groll[1], Thomas Kneib[1], Nikolaus Umlauf[2]

[1] Department of Statistics, University of Göttingen, Germany
[2] Department of Statistics, University of Innsbruck, Austria

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

Abstract: We propose a regularization approach for high dimensional data set-ups in the generalized additive model for location, scale and shape (GAMLSS) framework. It is designed for linear covariate effects and is based on L_1-type penalties. The following three penalization options are provided: the conventional least absolute shrinkage and selection operator (LASSO) for metric covariates, and both group and fused LASSO for categorical predictors.

Keywords: GAMLSS; Lasso; Penalization; Variable selection; Variable fusion.

1 Introduction

A model class that has gained increasing attention in recent years is the class of the GAMLSS, introduced by Rigby and Stasinopoulos (2005). In contrast to conventional regression approaches, where usually the expected mean is regressed, in the GAMLSS all distribution parameters (as, for example, the location, scale and shape) can be simultaneously modeled in terms of covariates. However, in high dimensional data set-ups classical fitting procedures for the GAMLSS often become very unstable and methods for variable selection are desirable.

The first ones who addressed the issue of variable selection, i.e. the selection of a reasonably small subset of informative covariates to be included in a particular GAMLSS, were Mayr et al. (2012). They extended boosting techniques, which originated in the machine learning field, to the framework of the GAMLSS. The approach is called **gamboostLSS** and is based on classical gradient boosting, which they successfully adapted to the GAMLSS characteristics. Both variable selection and model choice are naturally available within their regularized regres-

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

sion framework. For an implementation into the statistical software R, see Hofner et al. (2016).

The present work describes a different regularization approach for the high dimensional GAMLSS framework, which is designed for linear covariate effects only and is based on L_1-type penalties. Using adequate penalties, the cases of conventional LASSO for metric covariates, and both the group and fused LASSO for categorical predictors are covered. The implementation of the methods is incorporated into the unified modeling architecture for distributional generalized additive models (GAMs) established in Umlauf et al. (2017a), which exploits the general structure of GAMs and encompasses many different response distributions, estimation techniques, model terms etc. The corresponding R-package **bamlss** (Umlauf et al., 2017b) embeds many different approaches suggested in literature and software and serves as a unified conceptional "Lego toolbox" for complex regression models. Furthermore, within its framework both the implementation of algorithms for complex regression problems and the integration of already existing software are substantially facilitated.

For illustration purposes, the proposed methods are applied to Munich rent standard data, which are used as a reference for the average rent of a flat depending on its characteristics and spatial features.

2 Model specification

Along the lines of Rigby and Stasinopoulos (2005), who regard the GAMLSS as a semiparametric regression-type model with both linear and smooth covariate effects, in the following we focus on the fully parametric model with solely linear effects. Let $\mathbf{y} = (y_1, \ldots, y_n)^\mathrm{T}$ be the response vector with single observations $y_i, i = 1, \ldots, n$, being conditionally independent given a set of covariates. The corresponding conditional density $f(y_i|\boldsymbol{\theta}_i)$ usually depends on several distribution parameters $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{id})^\mathrm{T}$ that commonly represent distribution characteristics like location, scale, shape and/or kurtosis, but generally may be any of the distribution's parameters. After all, the key feature of a GAMLSS is that each distribution parameter θ_k can be modeled by its own predictor η_{θ_k} for $k = 1, \ldots, d$, which, in our case, depends linearly on a set of p_k covariates together with an intercept β_{0k}. Following Mayr et al. (2012), we denote by $g_k(\cdot)$ known monotonic link functions, corresponding to the linear predictor of the submodel of parameter θ_k. Then, a generalized linear model for location, scale and shape is given by the following set of equations

$$g_k(\theta_k) = \beta_{0k} + \sum_{j=1}^{p_k} \mathbf{x}_{jk}^\mathrm{T} \boldsymbol{\beta}_{jk} \; = \; \eta_{\theta_k} \; .$$

As the covariates can be metric and/or categorical, we use the general notation $\mathbf{x}_{jk}^\mathrm{T} \boldsymbol{\beta}_{jk}$ for a single predictor term, which either collects all covariate dummies and regression coefficients corresponding to the jk-th group of variables, if the covariate is categorical, and which reduces to a product of scalar values, if the covariate is metric, i.e. $x_{jk}\beta_{jk}$. Estimation of regression parameters can be based on maximizing the model's log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \left(f(y_i|\boldsymbol{\theta}_i) \right) \; , \tag{1}$$

with vector $\boldsymbol{\beta}$ collecting the effects of all linear predictors $\eta_{\theta_k}, k = 1, \ldots, d$. Note that the log-likelihood (1) depends on the parameters $\boldsymbol{\beta}_{jk}$ through the relations $\theta_{ik} = g^{-1}(\eta_{\theta_{ik}})$. Suitable fitting schemes are implemented in the R-package **gamlss** (Stasinopoulos and Rigby, 2007) and base on the following principle: in each iteration, backfitting steps are successively applied to all distribution parameters, using the submodel fits of previous iterations as offset values for those parameters that are not involved in the current step. However, in high dimensional situations these fitting procedures often become very unstable and methods for variable selection are needed.

3 L_1-type penalization

In the following, three L_1-type penalties are introduced, which are designed for linear covariate effects: the conventional LASSO for metric covariates, and the group and fused LASSO if categorical covariates are present. Altogether, a term $\lambda J(\boldsymbol{\beta})$ is subtracted from the log-likelihood (1), where $J(\boldsymbol{\beta})$ is a combination of (parts of) the three penalty terms from below and λ a tuning parameter, controlling the strength of the penalty.

Classical LASSO: For a metric covariate x_{jk}, following Tibshirani (1996), the absolute value of its (scalar) regression coefficient is penalized, i.e.

$$J_m(\beta_{jk}) = |\beta_{jk}| \,.$$

Group LASSO: For a (dummy-encoded) categorical covariate $\mathbf{x}_{jk}$, the L_2-norm of vector $\boldsymbol{\beta}_{jk}$, which collects all corresponding coefficients, is penalized (compare, e.g., Meier et al., 2008), i.e.

$$J_g(\boldsymbol{\beta}_{jk}) = ||\boldsymbol{\beta}_{jk}||_2 \,.$$

Fused LASSO: Alternatively, for categorical covariates often clustering of categories with implicit factor selection is desirable. Depending on the nominal (first expression) or ordinal scale level (second expression) of the covariate, one of the following two penalties can be used (compare Gertheiss and Tutz, 2010):

$$J_f(\boldsymbol{\beta}_{jk}) = \sum_{l>m} w_{lm}^{(jk)} |\beta_{jkl} - \beta_{jkm}|, \quad \text{or} \quad J_f(\boldsymbol{\beta}_{jk}) = \sum_{l=1}^{c_{jk}} w_l^{(jk)} |\beta_{jkl} - \beta_{jk,l-1}| \,,$$

where c_{jk} is the number of levels of categorical predictor $\mathbf{x}_{jk}$ and $w_{lm}^{(jk)}, w_l^{(jk)}$ denote suitable weights. Choosing $l = 0$ as the reference, $\beta_{jk0} = 0$ is fixed.

All proposed penalties have the attractive property to be able to set the coefficients of single (groups of) covariates to zero and, hence, to perform variable selection. Within the estimation procedure, i.e. the corresponding back-fitting algorithm implemented in **bamlss**, local quadratic approximations of all presented penalty terms are used (see Oelker and Tutz, 2015). Further, note that **bamlss** generally also allows to assign to each linear predictor of the d submodels (corresponding to parameters $\theta_k, k = 1, \ldots, d$) its own penalty term, i.e. a term $\lambda_k J(\boldsymbol{\beta})$. This way, the estimated models become even more flexible, with the drawback that then the grid search for the optimal tuning parameters λ_k has to be carried out on d dimensions and, hence, becomes computationally more demanding.

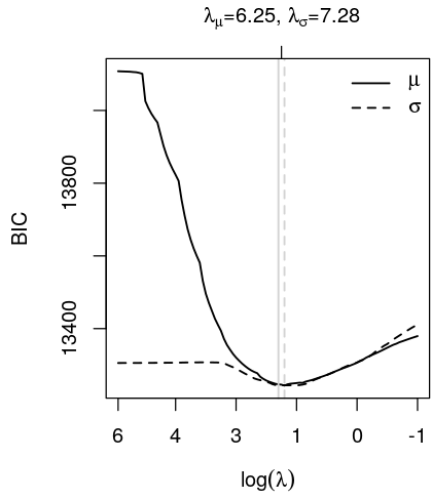


FIGURE 1. Marginal BIC curves for parameters μ and σ, holding the other tuning parameter fix at the respective minimum of the BIC.

4 Application on Munich rental guide data

We now apply the proposed penalization approach to the Munich rent data, which come from 3015 households interviewed for the Munich rent standard 2007. The response is the monthly rent per square meter in Euro, and from a large set of covariates we incorporate a selection of nine factors, both ordered as well as nominal and binary, similar to Gertheiss and Tutz (2010). We fit a Gaussian GAMLSS and use for both distribution parameters, i.e. μ and σ, a combination of the two different fused LASSO penalties introduced above. In order to obtain a flexible fit, the penalty terms of both corresponding linear predictors are assigned with separate tuning parameters λ_μ and λ_σ, respectively.

The optimal tuning parameters are selected by BIC on a 2-dimensional grid. Figure 2 shows the corresponding marginal BIC curves for both μ and σ, holding the other tuning parameter fix at the respective minimum of the BIC. Figure 2 and 3 show the paths of the dummy coefficients of both the ordinal covariate *year of construction* and the nominal *district*, which are penalized by the two different fused LASSO penalties from above. It is seen that with increasing tuning parameters λ_μ and λ_σ, respectively, categories are successively fused, i.e. the coefficients are set equal. In addition, it can be seen that for the ordinal covariate *year of construction* in Figure 2 only neighboring coefficients are fused, while for the nominal factor *district* in Figure 3 any groups of coefficients can be aggregated.

References

Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, **4**(4), 2150 – 2180.

Hofner, B., Mayr, T., and Schmid, M. (2016). **gamboostLSS**: An R package for model building and variable selection in the GAMLSS framework. *Journal*

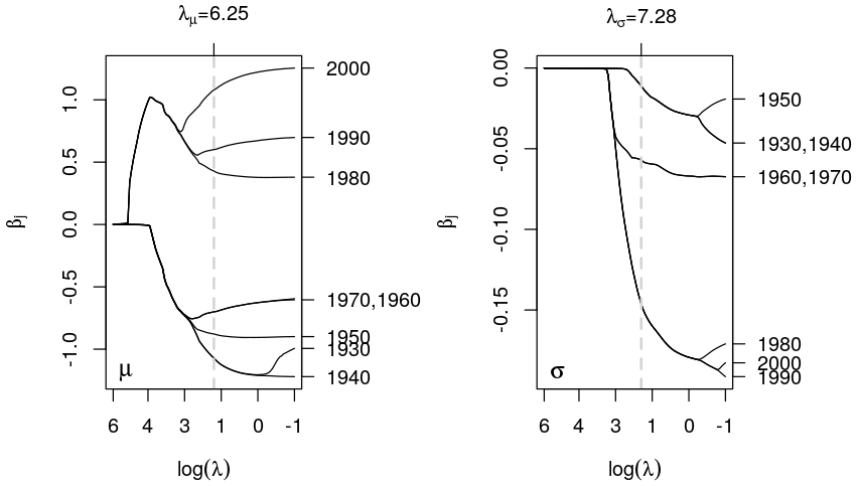


FIGURE 2. Ordinal fused coefficient paths for the year of construction for parameters μ (left) and σ (right); vertical dashed lines: optimal tuning parameters.

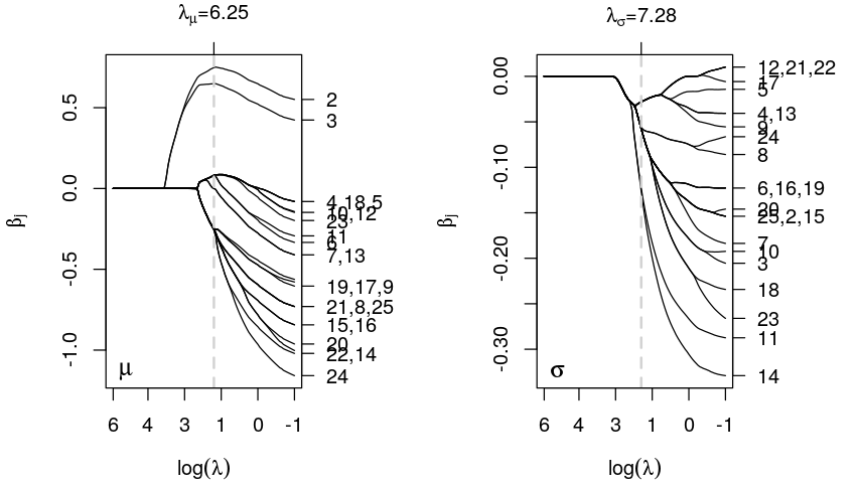


FIGURE 3. Nominal fused coefficient paths for the district effect for parameters μ (left) and σ (right); vertical dashed lines: optimal tuning parameters.

of Statistical Software, **74**(1), 1 – 31.

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**(3), 403 – 427.

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

Oelker, M.-R. and Tutz, G. (2015). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and*

Classification, 1 – 24.

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.

Stasinopoulos, D.M. and Rigby, R.A. (2007) Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**(7).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267 – 288.

Umlauf, N., Klein, N., and Zeileis, A. (2017a). BAMLSS: Bayesian additive models for location, scale and shape (and beyond). Working Paper, Faculty of Economics and Statistics, University of Innsbruck.

Umlauf, N., Klein, N., Zeileis, A. and Köhler, M. (2017b). **bamlss**: Bayesian additive models for location, scale and shape (and beyond). R package version 0.1-2, URL: http://cran.r-project.org/package=bamlss.

Boosting Generalized Additive Models for Location, Scale and Shape for Functional Data

Almond Stöcker[1], Sarah Brockhaus[1], Sophia Schaffer[2], Benedikt von Bronk[2], Madeleine Opitz[2], Sonja Greven[1]

[1] Department of Statistics, LMU Munich, Germany
[2] Soft Condensed Matter Group, LMU Munich, Germany

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

Abstract: We extend Generalized Additive Models for Location, Scale, and Shape (GAMLSS) to regression with functional response. GAMLSS are a flexible model class allowing for modeling multiple distributional parameters at once. The model is fitted via gradient boosting, which provides inherent model selection. We apply the functional GAMLSS to analyze bacterial interaction in *E. coli* and show how the consideration of variance structure fruitfully extends usual growth models.

Keywords: Functional data; GAMLSS; Gradient Boosting.

1 Introduction

We propose a flexible approach to regression with functional response allowing for simultaneously estimating multiple distributional characteristics of response curves. It therefore generalizes usual functional mean regression models.

In functional data analysis (Ramsay and Silverman; 2005), functional response regression aims at estimating covariate effects on response curves. Depending on the application, covariates might effect the response curves in quite different ways. This makes it particularly important to have a general framework for functional regression to draw on.

Rigby and Stasinopolous (2005) introduced GAMLSS extending usual Generalized Additive Models (GAM) to multiple distributional parameters: covariate effects on all parameters can be analyzed simultaneously, while doubtful assump-

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

tions of homoscedasticity can be overcome. At the same time the range of applicable distributions of GAM is extended to non-exponential family distributions.

We develop and implement functional GAMLSS combining a GAMLSS framework developed by Mayr et al. (2012) and a general functional regression framework (Brockhaus et al.; 2016a). For scalar-on-function regression the combination with GAMLSS was discussed by Brockhaus et al. (2016b) providing a selection of different functional covariate effect types. Thus, the framework now provides flexible functional GAMLSS for scalar-on-function, function-on-scalar and function-on-function regression.

We apply the model to analyzing bacterial interaction of two competing *Escherichia coli* bacteria strains: a toxin producing 'C-strain' and a toxin sensitive 'S-strain'. Both are exposed to different external stress levels. Our aim is to model the S-strain growing behavior in dependence of C-strain and stress level. As the growth process can be influenced in various ways, the covariates might effect both location and scale of growth curves.

2 Model formulation

Consider a data scenario with N observations of a functional response Y and respective covariates $\mathbf{X}$. Y is a stochastic process, such that its realized trajectories $y_i : T \to \mathbb{R}$, $t \mapsto y_i(t)$ for $i = 1, ..., N$ present the response curves over an index set T. Let $\mathbf{x}_i = (x_{i,1}, ..., x_{i,p})^{\mathrm{T}}$ denote the collection of the i-th observed covariates. Each single covariate might be categorical, scalar or functional.

We assume that for all $t \in T$ the conditional distribution of the response $F_{Y(t)|\mathbf{X}}$ is known up to the unknown parameters $\vartheta(t) = (\vartheta^{(1)}(t), ..., \vartheta^{(Q)}(t))$. For each parameter an additive regression model is assumed specifying

$$g^{(q)}(\vartheta^{(q)}) = h^{(q)}(\mathbf{x}) = \sum_{j=0}^{J^{(q)}} h_j^{(q)}(\mathbf{x}), \qquad q = 1, ..., Q,$$

where $g^{(q)}$ is a link function for the q-th parameter of interest. The model structure corresponds to the GAMLSS introduced by Rigby and Stasinopoulos (2005). However, covariates and response may now be functions. Correspondingly, also $\vartheta^{(1)}, ..., \vartheta^{(Q)}$ are modeled as functions.

The effect functions $h_j^{(q)}(\mathbf{x}, t)$ are constructed modularly using tensor product bases, which allows for flexible specification of a variety of effect types of scalar and functional covariates (see Brockhaus et al.; 2016a).

The model coefficients are estimated using gradient boosting. It generalizes former model based boosting approaches (Brockhaus et al.; 2016b, Thomas et al.; 2016) to GAMLSS for functional response regression. Gradient boosting is a gradient descend method for model fitting, which aims at minimizing a loss function. This is performed in a component-wise and stepwise procedure and thereby yields automatic model selection, prevents over-fitting and allows for fitting models with more parameters than observations.

3 Analysis of bacterial interaction in *Escherichia coli*

Equilibria in biodiversity stand and fall with bacterial interaction. A comprehensive analysis of underlying processes involves investigating both expectation and variability of growth curves.

Von Bronk et al. (2016) establish an experimental setup with two cohabiting *Escherichia coli* bacteria strains: a colicin producing 'C-strain' and a colicin sensitive 'S-strain'. Single bacteria of the C-strain can liberate the colicin, which kills numerous S-strain bacteria on contact. However, the S-strain might still outgrow the C-strain. The arising population dynamics are influenced by external stress induced with the antibiotic agent Mitomycin C (MitC). At $N = 334$ observation sites, bacteria under consideration are exposed to four different MitC concentrations. Bacterial growth curves $S_i(t)$ of the S-strain and $C_i(s)$ of the C-strain, $i = 1, ..., 334$, are observed over 48 hours.

We model the i-th S-strain curve $S_i(t)$ in dependence on the C-strain growth. To this end, we assume $S_i(t)$ to be gamma distributed with mean and standard deviation $\vartheta_i(t) = (\mu_i(t), \sigma_i(t))^{\mathrm{T}}$. Both mean and standard deviation are modeled analogously as

$$\log \vartheta_i^{(q)}(t) = f_0^{(q)}(t) + f_{MitC_i}^{(q)}(t) + h_0^{(q)}(C_i\,, t) + h_1^{(q)}(C_i', t), \quad q = 1, 2,$$

with historical effects $h_j^{(q)}(C_i, t) = \int_0^t C_i(s)\beta_j^{(q)}(s, t)\, ds$, $j = 0, 1$.

For both mean and standard deviation, the model includes functional intercepts $f_0^{(q)}(t)$ and group specific deviations $f_{MitC_i}^{(q)}(t)$ per MitC level. $C_i(s)$ and its derivative $C_i'(s)$ reflect C-strain spread and growth. In the historical effect, the complete history of the C-strain up to the current time point t may influence the S-strain growth (see (Brockhaus et al.; 2016a)). Smooth functional components are modeled using tensor-product B-Spline basis.

We observe MitC effects on both mean and standard deviation curves of the S-strain. Especially, the standard deviation substantially increases with time and with the MitC concentration.

Regarding the effects of the C-strain, we observe the effect of the derivative C' to be dominant. Moreover, the coefficient function shows temporal regimes. Such regimes were already noticed by von Bronk et al. (2016) and may now be quantitatively analyzed (Figure 1).

4 Discussion

Functional GAMLSS present an extremely flexible model class, which combines two conceptual extensions of GAM. They are of particular interest, if not only the mean curve but rather the full dynamic of the response process is under consideration. In this sense, it provides a natural extension to the usual analysis of bacterial growth.

References

Brockhaus, S., Fuest, A., Mayr, A. and Greven, S (2016b). Signal regression models for location, scale and shape with an application to stock returns. *arXiv:1605.04281*.

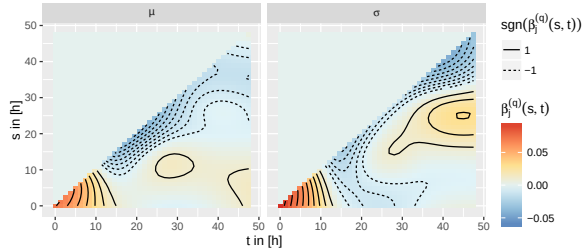


FIGURE 1. Coefficient functions $\beta^{(q)}(s,t)$ for the historical effects of $C_i'(s)$ on mean and standard deviation of S-strain growth curves. The t-axis presents the time line for the response curve, the s-axis the one for the C-strain. For a fixed $s = s_0$, $\beta^{(q)}(s_0,t)$ describes the influence of the covariate curve at that time over the whole S-strain growth curve.

Brockhaus, S., Melcher, M., Leisch, F. and Greven, S (2016a). Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing*, accepted.

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.

Ramsay, J. and Silverman, B.W. (2005). *Functional Data Analysis*. New York: Springer-Verlag.

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

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A. and Hofner, B. (2016). Stability selection for component-wise gradient boosting in multiple dimensions. *Manuscript submitted for publication.*

von Bronk, B., Schaffer, S., Götz, A. and Opitz, M. (2016). Heterogeneous toxin production promotes success in two-strain bacterial competition. *Manuscript submitted for publication.*

Markov-Switching GAMLSS with an Application to Operational Losses

Julien Hambuckers[1], Thomas Kneib[1], Roland Langrock[2], Alexander Sohn[1]

[1] Chair of Statistics, Georg-August-Universität Göttingen, Germany
[2] Chair of Statistics, University of Bielefeld, Germany

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

Abstract: This paper investigates modeling the dynamics of random sums representing the total operational losses where randomness is both in the loss frequencies and the loss sizes such that a compound Poisson process model is employed. To this end, we formulate a Markov-switching generalized additive model for location, scale and shape that allows all parameters of the compound loss distribution to depend on economic covariates in a flexible way while simultaneously allowing this dependence to vary over time according to a hidden state process. Relying on this approach, we analyze a novel dataset of 817 losses resulting from frauds in the Italian bank UniCredit.

Keywords: distributional regression; extreme events; hidden Markov model; generalized Pareto distribution.

1 Markov-Switching Compound Poisson Process Models for Total Losses

We are interested in modeling the temporal dynamics of the distribution of random sums

$$L_t = \sum_{i=1}^{N_t} Y_{i,t},$$

for $t = 1, 2, \ldots$, where N_t is the number of events occurring in period t, and $Y_{i,t}$ is the severity of the i^{th} event occurring during period t. In the actuarial literature, this model is used for modeling the total claim of an insurance company over time whereas in the banking literature it is used to model the total operational loss suffered by a bank. Operational losses are defined by the Basel Committee

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

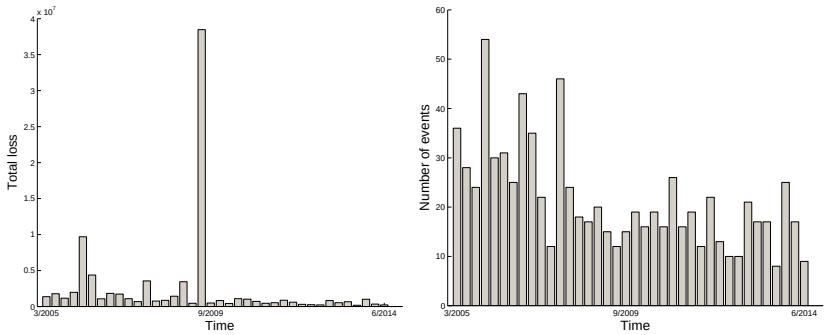


FIGURE 1. Total losses (left) and number of losses per quarter (right).

for Banking Supervision (BCBS) as *direct or indirect losses resulting from inadequate or failed internal processes, people and systems or from external events.* In the financial literature, the distribution of N_t is referred to as the *frequency distribution* whereas the distribution of $Y_{i,t}$ is termed the *severity distribution.*

The distribution of L_t is determined indirectly by making assumptions about the frequency distribution and the severity distribution as well as their dependence. For our application, we will assume

$$\begin{aligned} N_t &\overset{iid}{\sim} \text{Poisson}(\lambda), \\ Y_{i,t} &\overset{iid}{\sim} \text{GPD}(\gamma, \sigma), \end{aligned}$$

where realisations of N_t and $Y_{i,t}$, $\forall\{t, i\}$ are assumed to be independent, $\lambda > 0$ is the frequency parameter of the Poisson distribution for the counts and the severities follow ageneralized Pareto distribution $\text{GPD}(\gamma, \sigma)$ with shape parameter $\gamma \geq 0$, scale parameter $\sigma > 0$ and density

$$f(y; \gamma, \sigma) = \begin{cases} 1 - \left(1 + \gamma \frac{y}{\sigma}\right)^{-1/\gamma}, & \gamma \neq 0 \\ 1 - \exp\left(-\frac{y}{\sigma}\right), & \gamma = 0 \end{cases}$$

In combination, this leads to a compound Poisson process model for the total losses L_t.

We are then interested in

- relating the distributional parameters λ, γ and σ to economic covariates by embedding the compound Poisson model within the framework of generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005), and
- allowing for temporal variation in the regression effects via a Markov switching structure implied by a latent state variable S_t.

We therefore propose Markov-switching GAMLSS that rely on combining a Markov chain structure for the latent state S_t with a GAMLSS specification for the total loss distribution extending the Markov-switching generalized additive models framework by Langrock et al. (2016). For each distributional parameter $\theta \in$

$\{\lambda, \gamma\sigma\}$, we assume an additive predictor

$$g(\theta) = \eta_\theta = x'\beta_\theta^{(S_t)} + \sum_{j=1}^{J} h_{\theta,j}^{(S_t)}(z_j),$$

where $g(\cdot)$ is a monotonic link function, x is a vector of covariates with (state-dependent) linear effects $\beta_\theta^{(S_t)}$, and $h_{\theta,j}(z_j)^{(S_t)}$ are the (state-dependent) non-linear effects of the continuous covariates z_j modelled via penalised splines. All model parameters (relating to the Markov chain for the latent state and to the conditional compound Poisson process given state membership) can then be determined based on penalised likelihood inference with smoothing parameters determined by cross-validation.

2 Application to Operational Losses

We apply the proposed Markov-switching GAMLSS to a novel data set of 817 operational losses at UniCredit, one of the largest European banks. We focus on the particular class of operational losses resulting from external frauds, see Figure 1 for some descriptives on the data set. The collection period of the losses ranges between January 2005 and June 2014. Losses have been scaled by an unknown factor for anonymity reasons. The minimal amount considered is Euro 25,000, so that we have a sample of extreme losses, where the GPD is a reasonable hypothesis for the severity distribution. We have access to the exact date of each loss, meaning that we can assign each loss to a specific year and a specific quarter, and compute the total loss for each quarter. From a regulatory point of view, banks are expected to define the capital reserve for a horizon of one year but we work on a quarterly basis, as we wouldn't have enough time periods to estimate correctly the transition probabilities of the model otherwise.

In our analyses, we rely on the following covariates (see Figure 2):

- To model the frequency parameter λ, we use the percentage of the total revenue coming from fees (PRF) which can be interpreted as a measure of the economic well-being of the bank (with higher ratios implying smaller dependence on market interest rates). The PRF also measures the level of activity of the bank on behalf of clients.
- To model the scale parameter σ, we use the Italian unemployment rate, serving as a proxy for the overall economic performance of Italy, where UniCredit has its main activities.
- To model the shape parameter γ, we consider the values of the VIX index which is a measure of the market volatility, based on put and call options of the S&P500. It is also considered as a barometer of market sentiments.

The estimated covariate effects for the different latent states are shown in Figure 3 while estimated state probabilities are shown in Figure 4 (left).

For operational losses, the capital reserve is derived from the estimated 99.9% quantile of L_t such that this is a quantity of major interest for banking regulators. Figure 4 (right) shows estimates for this quantile derived from our model. One interesting feature is that the Markov-switching GAMLSS, unlike simpler alternatives, never entails quantile breeches on the observed data.

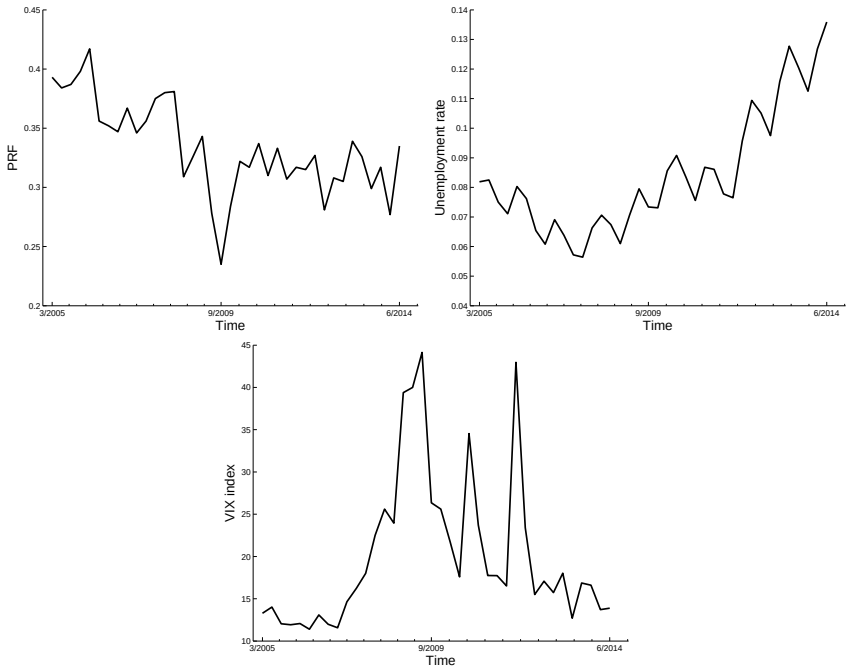


FIGURE 2. Values of the explanatory variables over the considered period. From left to right: lagged values (one quarter) of the PRF, the Italian unemployment rate and the VIX index.

Acknowledgments: Financial support by the German Research Foundation (DFG), Research Training Group 1644 "Scaling Problems in Statistics" and project KN 922/4-2 is gratefully acknowledged. The authors acknowledge Dr. F. Piacenza for providing us with the data.

References

Hambuckers, J., Kneib, T., Langrock, R. and Sohn, A. (2016). A Markov-switching Generalized Additive Model for Compound Poisson Processes, with Applications to Operational Losses Models. Technical Report.

Langrock, R., Kneib, T., Glennie, R. and Michelot, T. (2016). Markov-switching generalized additive models. *Statistics and Computing*, to appear.

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

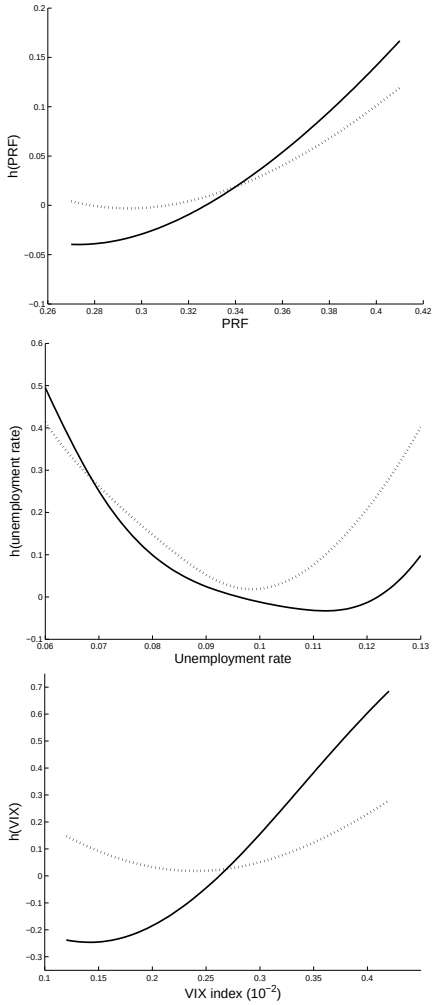


FIGURE 3. Estimated nonparametric effects of the covariates for state 1 (solid line) and state 2 (dotted line).

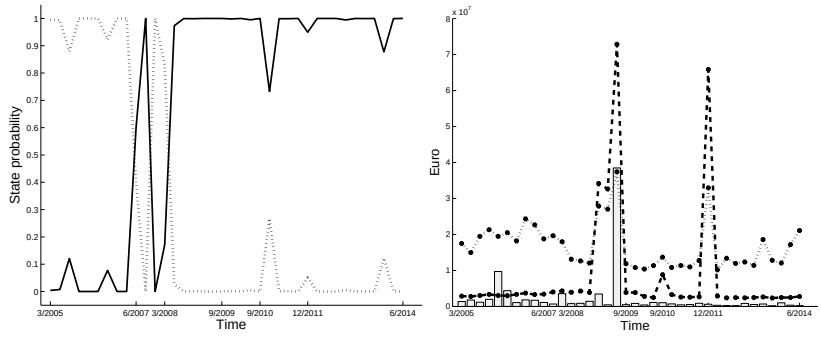


FIGURE 4. State probabilities for the different regimes (left; state 1 as solid line, state 2 as dotted line) and $\hat{Q}_{0.999}(L_t; X_t, S_t)$ over time (right; state 1 as dashed line, state 2 as dotted line; total losses superimposed as grey bars).

Aggregate Claims Modelling via Dynamic Score Driven Models

Mariana Melo[1,2], Cristiano Fernandes[1], Eduardo Melo[2,3]

[1] Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil
[2] SUSEP - Brazilian Insurance Supervisor, Brazil
[3] CPES - Insurance Economics Research Center, ESNS, Brazil

E-mail for correspondence: mariarozo@gmail.com

Abstract: In the insurance industry, the measurement of obligations and the evaluation of actuarial risks depend fundamentally on the aggregate claims distribution. From a mathematical point of view, the aggregated claims variable is a random sum of random variables. However, obtaining the analytical expression for this probability distribution is a hard task. In this paper, a new approach is proposed for the modelling of the aggregated claims predictive distribution. We combine the newly proposed Generalized Autoregressive Score (GAS) models, to specify a non-Gaussian distribution for both the number of claims and for the claims severity. By use of the Fast Fourier Transform (FFT), we are then able to numerically obtain the aggregated claims distribution. The proposed method is applied to real data, provided by a leading Brazilian motor insurer.

Keywords: GAS models; Aggregate Claims; Collective risk model; Random sum.

1 Introduction

In the insurance industry, the measurement of obligations and the evaluation of actuarial risks depend fundamentally on the aggregate claims distribution. New solvency supervision principles have required more sophisticated approaches to assess the claims expenses of an insurer. In addition to obtaining the expected value of these expenses, it is essential to insurers to assess the inherent risk of these cash flows, whose distribution of probability becomes the main objective. From a mathematical point of view, the aggregated claims variable is a random sum of random variables. Therefore, randomness is present on both the values of the individual claims and on the number of incurred claims. However, the aggregate claims distribution, except in rare circumstances, does not present a

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

closed form.

In this paper, a new class of observation driven models proposed by Creal et al. (2013) and Harvey and Chakravarty (2008), named the Generalized Autoregressive Score (GAS) models, is proposed for the modelling of the aggregated claims predictive distribution. By construction GAS models allow parameters of the probability distribution to evolve on time according to an updating equation which use the score as a driving mechanism. We combine the GAS models, applied for the number of claims and also for the claims severity distribution, with the use of Fast Fourier Transform (FFT) to compute the aggregated claims distribution.

Differently from the traditional approach for the number and severity of claims distribution, parameter time varying models allows a more precise forecast once the model follows the parameter dynamic on time. Aggregate claims distribution, except in rare circumstances, does not present a closed form. FFT was chosen to obtain this distribution since it is an efficient numerical method with low computational effort. The proposed method is applied to real data, provided by a leading Brazilian motor insurer. The number of claims was both modeled trough a Poisson distribution and a negative binomial distribution. Claim severity was assumed both gamma or lognormal distributed. In all models appropriate parameters were made time varying according to the score driven mechanism.

2 Models

The aggregate claims represents the total amount paid for all claims incurred in a certain period of time in an insurance portfolio. In the collective risk model, the aggregate losses is represented by a sum of a random number (claim frequency) of individual payments (claim severity). In this work, the claim frequency and claim severity data are both considered as times series, so the aggregate claim at time t is represented by $S_t^* = \sum_{i=1}^{N_t} X_{it}$, $t = 1, ..., T$, where N_t is the random variable that represents the claim frequency at time t and X_{it}, i=1,...,N_t, is the random variable that represents the severity of the claim i at time t. The assumptions are that $X_{1t}, X_{2t}, \ldots$ are independent and identically distributed (iid) at each time t and independent of N_t.

2.1 Score driven models

GAS models provide a general framework for modelling distributions with time varying parameters. The time variation of the parameters is introduced by letting parameters be functions of lagged dependent variables, using the scaled score. It is also possible to introduce exogenous variables on the parameter updating equation. In this framework, the parameters are perfectly predictable given the past information.

For a GAS(p,q) model, the predictive model density and the associated updating equation are given by $y_t \sim p(y_t|f_t, \theta; \mathcal{F}_{t-1})$ and $f_{t+1} = w + \sum_{i=1}^{p} A_i s_{t-i+1} + \sum_{j=1}^{q} B_j f_{t-j+1}$, where $\mathcal{F}_t = \{y_1, \ldots, y_t\}$, f_t is the time-varying parameter vector, θ is the static parameter vector, $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$, w is a vector of constants, A_i and B_j are coefficient matrices and s_t is the scaled score ($s_t = S_t \nabla_t$).

The score is defined as $\nabla_t = \frac{\partial ln\, p(y_t|f_t,\theta)}{\partial f_t}$ and the scaling matrix S_t is, usually, chosen between $I_{t|t-1}^{-1}$, $I_{t|t-1}^{-1/2}$ ou I, where $I_{t|t-1} = \mathrm{E}_{t-1}\left[\nabla_t \nabla_t'\right]$ is the information matrix and I is the identity matrix. Fixed parameters estimation is done by maximum likelihood (ML).

2.2 Claim frequency and severity models

As in most cases in the insurance field, we have used the Poisson and negative binomial distributions for claim frequency (N_t), considering the mean λ_t and parameter p_t time-varying, respectively.

For the claim severity models, following the results of Creal et al. (2014), it is assumed that all severity claims ($X_{i,t}$) at time t are identical distributed and cross-sectionally independent conditional on f_t and on information set $\mathcal{F}_{t-1}$. In this case, the log-likelihood, the score vector and the information matrix at time t take a simple additive form: $l_t = \sum_{i=1}^{n_t} \ln p(x_{i,t}|f_t, \theta^{(2)}; \mathcal{F}_{t-1})$, $\nabla_t = \sum_{i=1}^{n_t} \nabla_{t,i}$ and $I_{t|t-1} = \sum_{i=1}^{n_t} I_{t|t-1,i}$, where n_t is the claim frequency at time t. Gamma and lognormal are popular selections for distributions of motor insurance loss given their appropriate characteristics, as non-negative support, positive skewness and variance proportional to the mean-squared (constant coefficient of variation). In this article, these two distributions are used to model the claims severity. Models descriptions are given in Table 1. For the Gamma model, α is the shape parameter and $\frac{\mu_t}{\alpha}$ is the scale parameter.

TABLE 1. Details for the claim frequency and claim severity models.

Models	f_t	∇_t	$I_{t\|t-1}$
Claim frequency (N_t):			
Poisson(λ_t)	$ln(\lambda_t)$	$n_t - \lambda_t$	λ_t
NegBin(r, p_t)	$ln\left(\frac{p_t}{1-p_t}\right)$	$r(1-p_t) - n_t p_t$	$r(1-p_t)$
Claim severity ($X_{t,i}$):			
Gamma $\left(\alpha, \frac{\mu_t}{\alpha}\right)$	$ln(\mu_t)$	$\alpha\left(\frac{\sum_{i=1}^{n_t} x_{t,i}}{\mu_t} - n_t\right)$	$n_t\alpha$
Lognormal(μ_t, σ^2)	μ_t	$\sum_{i=1}^{n_t} \ln x_{i,t} - n_t\mu_t\sigma^2$	$\frac{n_t}{\sigma^2}$

Our goal is to obtain the aggregate claims distribution. In this sense, since this distribution can not be obtained analytically for the vast majority of cases, the models for series of claims numbers (N_t) and for severity (X_{it}) described previously were used with the FFT technique to obtain the target distribution. The FFT is an efficient method to calculate compound distributions by inverting characteristic functions. Klugman et al.(2012) and Wang (1998) give an explanation of the method in application to aggregate claims distribution.

3 Empirical application

The data base used in this paper consist of motor insurance individual claims of Brazilian leading insurer, occurred between 01/2006 and 02/2014, totalizing 30.959 claims in 98 months. The observations of the first 86 months were used to fit the models, leaving the remaining 12 months for out-of-sample evaluation. Figure 1 presents the data time series (claim frequency, claim severity mean and total aggregate loss). The claim severity values are deflated by a consumer inflation index. All models were run using R Software. We adopted models with $p = q = 1$ on the updating equation of f_t and $S_t = I_{t|t-1}^{-1}$. The models were also fitted with $S_t = I_{t|t-1}^{-1/2}$, however the maximum likelihood optimization proved to be more stable with the first option.

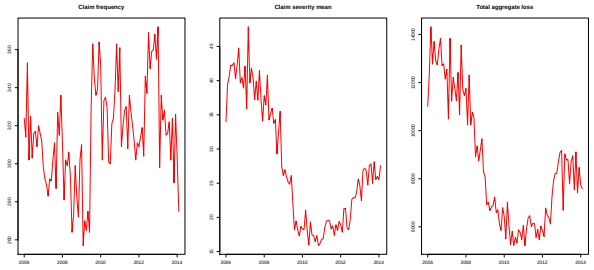


FIGURE 1. Data time series (claim frequency, claim severity mean and total aggregate loss) from 01/2006 to 02/2014.

Quantile residuals were used to evaluate the models. The results for serial autocorrelation tests (Ljung-Box test up to the lag 30), conditional heteroscedasticity tests (Ljung-Box test up to lag 30 on the squared residuals) and normality tests (Jarque-Bera) tests were satisfactory, except for the lognormal model. In order to compare models, the Akaike Information Criteria (AIC), Bayesian Schwarz Criteria (BIC), the log-likelihood value, as well as the MAPE (Mean Absolute Percentage Error) were used. The results are shown on Table 2. Our results, including residuals analyses, indicate that the gamma distribution is more suitable to the severity data. The Poisson and negative binomial models for claim frequency presented similar results but the Poisson model produces a slightly better out-of-sample performance. Also, in comparison with a static benchmark model, where all parameters are fixed in time, our model obtained better results. Figure 2 presents information about the aggregate claims distributions for 12 months ahead, obtained using FFT and the results from Poisson and gamma models for claim frequency and claim severity respectively.

4 Concluding Remarks

Following the methodology presented, we obtained the complete density for the aggregate claims random variable conditioned on time and not only means and upper moments. We used a new framework of time-series models, driven by score

Model	*log-Lik*	AIC	BIC	MAPE
Claim frequency				
Poisson GAS(1,1)	-381.967	771.935	781.752	4.872
Neg. Binomial GAS(1,1)	-380.412	770.825	783.097	4.870
Claim severity				
Gamma GAS(1,1)	-108790.3	217590.6	217602.9	7.042
Lognormal GAS(1,1)	-109255.4	218520.8	218533.1	5.007

TABLE 2. In sample results, using data from 01/2006 to 02/2013, for claim frequency models (Poisson GAS(1,1) and Negative Binomial GAS(1,1)) and for claim severity models (Gamma GAS(1,1) and Lognormal GAS(1,1)).

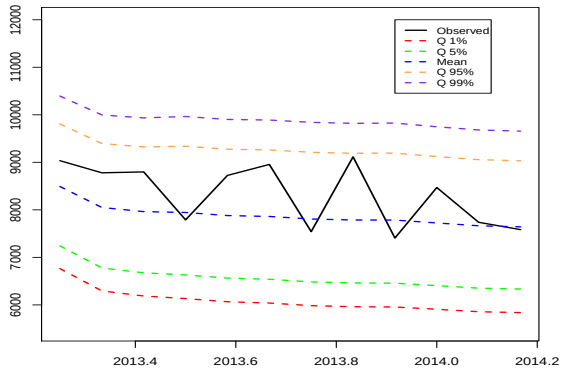


FIGURE 2. Quantiles and means of the aggregate claims distributions obtained from the Poisson GAS(1,1) and Gamma GAS(1,1) models - 03/2013 to 02/2014.

with time-varying parameters, to model the predictive distribution of the claim frequency and the claim severity. In comparison to static fitted distributions, the predictive distributions are more appropriate as they take into account the time dynamics and the parameters uncertainty. To the best of our knowledge, there are few studies on obtaining the predictive distribution of the aggregate claims and most of them use the Bayesian approach. In addition to the application proposed in this article, the presented method can be applied to other lines of business or other problems involving sum of random variables and calculation of compound distributions, as well as developed with different probability distributions that best fit the data.

From the aggregate claims distribution, it is possible to calculate the probability that the insurer has a portfolio loss above a given value, perform premium calculations, company ruin probability, and obtain subsidies for technical provisions and risk-based capital calculations. The applications of our method meet the demands of the current solvency principles and the measurement of insurance contracts by making it possible to obtain predictive distributions, conditioned in time, from the loss of insurance portfolios. Assessment uncertainty in relation

to this stochastic process allows for the practice of risk management practices, either through the acquisition of measures such as VaR, TVaR or others, as well as to calculate risk margins.

Acknowledgments: The first author thanks FUNENSEG for the support.

References

Creal, D., Koopman, S.J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, **28(5)**, 777 – 795.

Creal, D., Schwaab, B., Koopman, S.J., and Lucas, A. (2014). Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics*, **96(5)**, 898 – 915.

Harvey, A. and Chakravarty, T. (2008). *Beta-t-(e) garch*. University of Cambridge, Faculty of Economics.

Klugman, S.A., Panjer, H.H., and Willmot, G.E. (2012). *Loss models: from data to decisions*. John Wiley & Sons.

Wang, S. (1998). Aggregation of correlated risk portfolios: models and algorithms. *In Proceedings of the Casualty Actuarial Society*, **85**, 848 – 939.

Enhancing predictive performance of vulnerability to poverty estimates

Maike Hohberg[1], Katja Landau[1], Thomas Kneib[1], Stephan Klasen[1], Walter Zucchini[1]

[1] University of Göttingen, Germany

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

Abstract: We analyze two modifications to a simple measure of vulnerability to poverty, which defines vulnerability as expected income poverty. First, to model income, we use distributional regression and relate each parameter of the conditional income distribution to a set of covariates. Second, instead of defining a household as vulnerable if its probability of being poor is larger than 0.5, we construct a vulnerability cutoff employing the receiver operating characteristic (ROC). Using panel data from Germany, we find that our new cutoff method considerably increases predictive performance while placing the income regression model into the distributional regression framework improves predictions for some years.

Keywords: vulnerability to poverty, gamlss, ROC

1 Introduction

Although not living in poverty at the moment, many households are extremely vulnerable to events such as job loss, unexpected expenditures, economic downturns, and weather phenomena, and can easily fall below the poverty line. Policymakers are therefore often interested in knowing which people are at risk of poverty in order to adequately design anti-poverty programs. Even though a few empirical applications of vulnerability to poverty measures evaluated predictive performance of their estimates, very little attention has been paid to their improvement.

One popular approach considers vulnerability as expected poverty (Chaudhuri et al., 2002) and defines the vulnerability of a household h as the estimated probability of falling below a poverty threshold z of some normally distributed

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

measure of welfare y_h (e.g. log of income) given a covariate vector $\mathbf{x'}_h$. That is

$$\widehat{\Pr}(\ln y_h < \ln z | \mathbf{x}'_h) = \Phi \left(\frac{\ln z - \mathbf{x}'_h \hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{x}'_h \hat{\boldsymbol{\theta}}}} \right) \tag{1}$$

where $\hat{\boldsymbol{\beta}}$ is a vector of coefficients from a regression specification with log income as the dependent variable. Implying a relationship between higher volatility in income and poverty risk, the variance is allowed to vary with the covariates across households yielding coefficients $\hat{\boldsymbol{\theta}}$. Both coefficient vectors are estimated via a three-step feasible generalized least squares (FGLS) procedure as in Amemiya (1977). The household is then classified as vulnerable if the probability in (1) is equal to or greater than 0.5.

The need for an enhanced method results from three major drawbacks of the standard approach: First, the welfare measure is always assumed to be (log)normally distributed but in many applications income or expenditures do not behave (log)normally. Second, once we depart from the normality assumption, parameters other than mean and variance could be modeled to capture the full effect on the whole conditional income distribution. Third, setting the vulnerability cutoff at 0.5, neglects the variability a household faces. If the expected income equals the poverty line, the normal density function in equation (1) yields 0.5 independent of the standard deviation (McCarthy et al., 2016). Additionally, this classification does not always perform well in terms of prediction (Bergolo et al., 2012; Celidoni, 2013).

We tackle all of these drawbacks by two modifications: Distributional regression is aiming at the first and second point. Within this flexible framework, different distributions can be selected to model income, and all parameters of this distribution are related to a structured additive predictor which can incorporate non-linear effects. This is equivalent to generalized additive models for location scale and shape (GAMLSS, Rigby & Stasinopoulos, 2005). However, we prefer the term "(structured additive) distributional regression" as some distributions neither have location nor scale parameters but potentially only shape parameters (Klein et al., 2015). Our approach to use the ROC to determine the vulnerability cutoff focuses on the third drawback of the traditional method. We thus directly address recent criticism of the traditional vulnerability threshold raised by e.g. Bergolo et al.(2012) and McCarthy et al. (2016) and propose a endogenous cutoff which improves predictive precision.

2 Estimation strategy

2.1 Data

To demonstrate the benefits of our modifications we use data of 15 years, namely from 1993 to 2008 of the German Socio-Economic Panel (SOEP), which offers comprehensive coverage of household characteristics and income enabling us to retrospectively observe whether a household did become poor or not.

2.2 Model

Placing the income regression within the framework of structured additive distributional regression, the conditional income distribution is given by a density conditioned on parameters ϑ_{hk}, $k = 1, \ldots, K$. Each of the K parameters is itself dependent on the covariates including variables such as past income, number of children and elderly in the household h, education and gender of the household's head. We thus write

$$g_k(\vartheta_{hk}) = \eta_h^{\vartheta_k} = \mathbf{x}_h' \boldsymbol{\beta}_h^{\vartheta_k} + \sum_{j=1}^{J_k} f_j^{\vartheta_k}(\boldsymbol{\nu}_h) \tag{2}$$

where g_k is a link function, $\eta_h^{\vartheta_k}$ the predictor for the k-th parameter, the vector $\mathbf{x}_h'$ contains covariates which are assumed to have a linear effect, $\boldsymbol{\beta}_h^{\vartheta_k}$ are the corresponding coefficients for these covariates, and $f_j^{\vartheta_k}(\boldsymbol{\nu}_h)$ are smooth functions of J_k continuous covariates $\boldsymbol{\nu}_h$ which have non-linear effects. More precisely, for the covariate past income we relax the restrictive assumption of a linear effect and use P(enalised)-splines (Eilers & Marx, 1996) to flexibly model its relationship to the dependent variable. As conditional distributions, we use in addition to the lognormal, the Burr distribution. The parameters are simultaneously estimated via a back-fitting algorithm that maximizes the penalized likelihood avoiding the stepwise FGLS procedure of the traditional approach. The models are implemented in the software R (R Core Team, 2016) using the package `gamlss` (Stasinopoulos & Rigby, 2007).

2.3 Constructing a vulnerability cutoff using the ROC

After fitting the models, a cutoff (the vulnerability line) is specified such that households whose predicted income falls below the cutoff are classified as vulnerable. To put it more precisely, we first calculate predicted incomes from the models above. We then determine which households are actually poor in the next year by using the observed incomes and a predefined poverty line. After that, we use each of the predicted incomes as a hypothetical vulnerability line. Households with an income below the hypothetical vulnerability line will be declared vulnerable. For each possible cutoff, that is each predicted income, we construct the ROC by comparing the actual poor and non-poor with vulnerable and non-vulnerable households. A higher cutoff naturally leads to a higher true positive rate (TPR) as more households are declared vulnerable but also increases the false positive rate (FPR). Depending on the costs of "false alarms" or "missing a household", a prerequisite can now be arbitrarily established. We decided to define the predicted income which leads to a TPR of 80 percent as the vulnerability line.

With the method described so far, we determine a cutoff *ex post*. To estimate vulnerability as a forward looking perspective on poverty, we rely on the past vulnerability lines to make a prediction for the following year. Thus, the predictions will not exactly meet a TPR of 80 percent but will be close to it.

3 Main results and concluding remarks

We compare how well the traditional and our proposed models predict poverty status using the TPR, FPR, and scoring rules. The models differ with respect to the distribution of the response (and hence parameters that are related to covariates), nonlinearities in the covariate effects, and the vulnerability cutoff.

We find that for some years using distributional regression does improve predictive performance mainly due to incorporating non-linear effects of past income. We suppose that effects of modeling shape parameters are relatively small due to the use of data from an industrialized country where social safety nets allow households to cope with idiosyncratic shocks. Since modeling scale and shape parameters should account for idiosyncratic shocks, it is possible that for Germany those idiosyncratic shocks are either little prevalent in the data or households are able to cope reasonably well with them. Nonetheless, distributional regression is an attractive alternative to estimate vulnerability to poverty as it estimates mean and variance simultaneously and is able to incorporate non-linear effects. These effects seem important in our dataset especially when looking at incomes at the lower end of the distribution.

Regarding our new cutoff method, we find that it largely outperforms the traditional approach. Figure 1 shows the improvement in TPR and FPR when both of our proposed modifications are applied but with the new cutoff being the main driver. The plane's division is based on our arbitrary prerequisites or, to put it bluntly, which TPR and FPR are acceptable for us for being a good prediction. The result for the predictions of each year is represented by a point. Best predictions lie in the 4th quadrant, worst predictions in the 2nd quadrant.

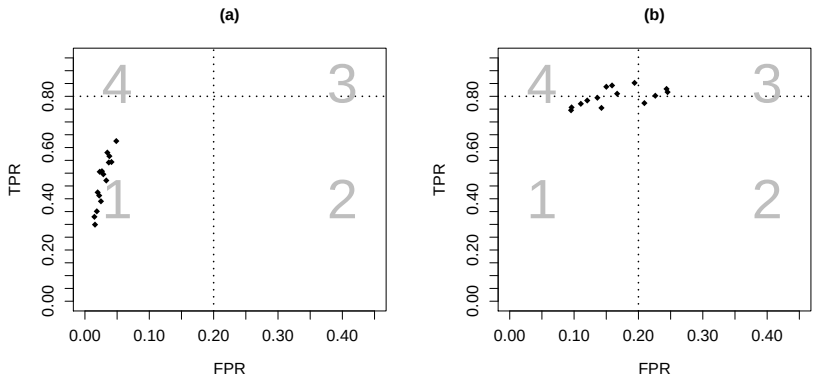


FIGURE 1. Plots of accuracy for different models. Plot (a) shows results for the traditional approach, i.e. 0.5 poverty probability cutoff, log normal distribution, no splines. Plot (b) represents the modified model, i.e. new cutoff, 3-parameter Burr distribution, spline for past income.

The low TPRs for the traditional approach are somewhat surprising and we put further effort in investigating why we obtain these striking differences in predictive performance. Looking at the effects of past income on current income,

we find that for the most part the relationship is roughly linear whereas non-linearities occur especially at the bottom end of the income distribution where the effects are much lower and can even be negative. These effects are not accounted for in models assuming a linear relationship. On the other hand, models that do incorporate non-linear effects also have difficulties in predicting very low incomes as our dataset only comprises about 10-14 percent poor households. Hence, the poor observations only have a small impact on the overall models' prediction ability, and thus the models predict overly optimistic incomes resulting in too few households being classified as vulnerable.

Our new cutoff method is able to mitigate these weaknesses of the models since the cutoff is determined endogenously and constructed in a way to fulfill a prescribed TPR. In contrast, the traditional method relies heavily on the model specification and its prediction abilities can be weak, especially at the lower extremes of the income distribution. As the 0.5 probability was advocated for developing countries, where the share of the poor is generally higher than in Germany, it is possible that the traditional cutoff performs better in other settings. One might argue that the 0.5 cutoff simply needs to be set lower in our case. We tested several other probability cutoffs but the results where still not satisfactory yielding very high FPRs.

Since our new cutoff method is constructed to improve predictive performance, can mitigate weaknesses in the income generating model specification, and is easy to implement, it is a useful tool if researchers or policymakers are particularly interested in correctly identifying vulnerable households rather than in measuring overall vulnerability.

References

Amemiya, T. (1977). The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model *Econometrica*, **45**, 955 – 968.

Bergolo, M., Cruces, G., and Ham, A. (2012). Assessing the Predictive Power of Vulnerability Measures: Evidence from Panel Data for Argentina and Chile *Journal of Income Distribution*, **21**, 28 – 64.

Celidoni, M. (2013). Vulnerability to Poverty: An Empirical Comparison of Alternative Measures. *Applied Economics*, **45**, 1493 – 1506.

Chaudhuri, S., Jalan, J., and Suryahadi, A. (2002). Assessing Household Vulnerability to Poverty from Cross-sectional Data: A Methodology and Estimates from Indonesia *Discussion Paper Series 0102-52*, Department of Economics, Columbia University.

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

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). Bayesian Structured Additive Distributional Regression with an Application to Regional Income Inequality in Germany *Annals of Applied Statistics*, **9**, 1024 – 1052.

McCarthy, N., Brubaker, J., and de La Fuente, A. (2016). Vulnerability to Poverty in rural Malawi *Policy Research Working Paper WPS7769*, The World Bank.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized Additive Models for Location, Scale and Shape *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507 – 554.

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R *Journal of Statistical Software*, **23**.

A heteroscedastic probabilistic temperature forecasting model incorporating spread-error correlation and high-resolution forecasts

Annette Möller [1], Jürgen Groß [2]

[1] Institute of Applied Stochastics and Operations Research, Clausthal University of Technology, Germany,
[2] Institute for Mathematics and Applied Informatics, University of Hildesheim, Germany

E-mail for correspondence: annette.moeller@tu-clausthal.de

Abstract: Ensembles of deterministic numerical weather predictions require correction with respect to forecast bias and dispersion properties. A prominent approach is the use of statistical postprocessing models, yielding calibrated and sharp predictive probability distributions. A recently developed postprocessing method based on an AR-model for the forecast errors is now extended to incorporate a heteroscedastic variance parameter, accounting for the well-known spread-error correlation of forecast ensembles. Making use of a high-resolution forecast provides the model with additional information leading to improved properties of the predictive distribution.

Keywords: Statistical postprocessing model; predictive probability distribution; autoregressive process; high-resolution forecast; spread-error correlation.

1 Introduction

The method of choice to quantify and assess sources of uncertainty in deterministic numerical weather prediction (NWP) models is the employment of forecast ensembles (Leutbecher and Palmer, 2008). However, ensembles of NWP forecasts typically exhibit biases and dispersion errors, thus require statistical postprocessing to improve reliability and forecast skill (Gneiting and Katzfuss, 2014).

Recently, Möller and Groß (2016) developed an ensemble postprocessing model for temperature, based on autoregressive information present in the forecast errors of the raw ensemble members.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

We present an extended version of the model in Möller and Groß (2016), where the variance parameter is defined to depend on the ensemble spread, thus leading to a heteroscedastic model accounting for the well-known spread-error correlation (Gneiting et al., 2005), stating that there exists a positive association between ensemble spread and ensemble mean forecast error. Further, an additional high-resolution forecast is utilized for the postprocessing model. This single forecast was obtained by running the respective NWP model with higher resolution than used for the other ensemble members. Thus, the high-resolution forecast provides more detailed information on small-scale processes not captured by the lower-resolution forecast ensemble.

2 A heteroscedastic postprocessing model

Let $\{x_1(t), \ldots, x_m(t)\}$ denote an ensemble of NWP forecasts for a univariate (Gaussian distributed) weather variable $Y(t)$ at a fixed location. Further, let $x_{\text{hres}}(t) = x_{m+1}(t)$ denote an additional high-resolution forecast of $Y(t)$. The respective forecast error series are given by $e_i(t) = Y(t) - x_i(t)$, $i = 1, \ldots, m+1$. Möller and Groß (2016) assume that the individual series $\{e_i(t)\}$, $i = 1, \ldots, m$ follow a weakly stationary AR(p) process (Shumway and Stoffer, 2006). Fitting autoregressive models to each individual error series yields an AR-adjusted ensemble $\{\widetilde{x}_1(t), \ldots, \widetilde{x}_m(t)\}$. We now follow the idea of Möller and Groß (2016) and Gneiting et al. (2005) and assume the predictive distribution to be Gaussian. However, in the extended model, the high-resolution forecast is considered as additional covariate, that is

$$Y(t)|x_1(t), \ldots, x_m(t), x_{\text{hres}}(t) \sim \mathcal{N}(\mu(t), \sigma^2(t)) \,. \tag{1}$$

To account for the fact that the members $x_1(t), \ldots, x_m(t)$ form an exchangeable group, while the high-resolution forecast has somewhat different properties, the parameters of the Gaussian distribution are defined as (weighted) sums of the respective group-wise parameters. That is, $\mu(t) = \frac{1}{2}\left(\mu_{\text{ens}}(t) + \mu_{\text{hres}}(t)\right)$, and $\sigma(t) = \frac{1}{2}\left(\sigma_{\text{ens}}(t) + \sigma_{\text{hres}}(t)\right)$. Here, $\mu_{\text{ens}}(t)$ is estimated by the mean $\overline{\widetilde{x}}(t) = \frac{1}{m}\sum_{i=1}^{m} \widetilde{x}_i(t)$ of the AR-adjusted forecast ensemble $\{\widetilde{x}_1(t), \ldots, \widetilde{x}_m(t)\}$. To explicitly incorporate the spread-error correlation of ensemble forecasts, our model for $\sigma^2(t)$ combines the estimated dispersion of the error process retrieved solely from the past of each member $x_i(t)$ individually with the spread of the AR-adjusted members. Specifically, the standard deviation $\sigma_{\text{ens}}(t)$ is parameterized as

$$\sigma_{\text{ens}}(t) = w \sqrt{\frac{1}{m}\sum_{i=1}^{m} \gamma_i^2(t)} + (1 - w)\sqrt{S^2(t)}. \tag{2}$$

Here, $S^2(t)$ is the empirical variance of the AR-adjusted forecast ensemble $\{\widetilde{x}_1(t), \ldots, \widetilde{x}_m(t)\}$ and $\gamma_i^2(t) = \text{Var}(e_i(t))$ is the variance of the $AR(p)$ process assumed for the forecast error $e_i(t) = Y(t) - x_i(t)$ of ensemble member $x_i(t)$. The parameters $\gamma_i(t)$ can directly obtained from the fitted $AR(p)$ model, while the weights w are estimated by the so-called minimum-CRPS approach (Gneiting et al, 2005).

The high-resolution parameters $\mu_{\text{hres}}(t)$ and $\sigma_{\text{hres}}(t)$ are estimated in the same way. However, the above formulas are reduced to the $(m+1)$-st summand, thus are

TABLE 1. Verification statistics averaged over 76 stations and 4341 test dates.

	CRPS	DSS	RMV	Var(PIT)
EMOS(ENS)	0.906	2.152	1.491	0.095
EMOS(ENS,HRES)	0.876	2.153	1.435	0.093
H-AR-EMOS(ENS)	0.903	2.013	1.532	0.088
H-AR-EMOS(ENS,HRES)	0.873	1.936	1.502	0.085

solely based on the AR-adjusted high-resolution forecast $\widetilde{x}_{m+1}$. Although the part of the predictive variance corresponding to the ensemble spread is zero for a single ensemble member, the variance $\gamma_i^2(t)$ of the $AR(p)$ error series corresponding to an individual member $x_i(t)$ can still be computed - on the basis of past values. This fact highlights an additional advantage of the autoregressive postprocessing model introduced by Möller and Groß (2016), as standard postprocessing models are not able to estimate the variance parameter based only on a single ensemble member.

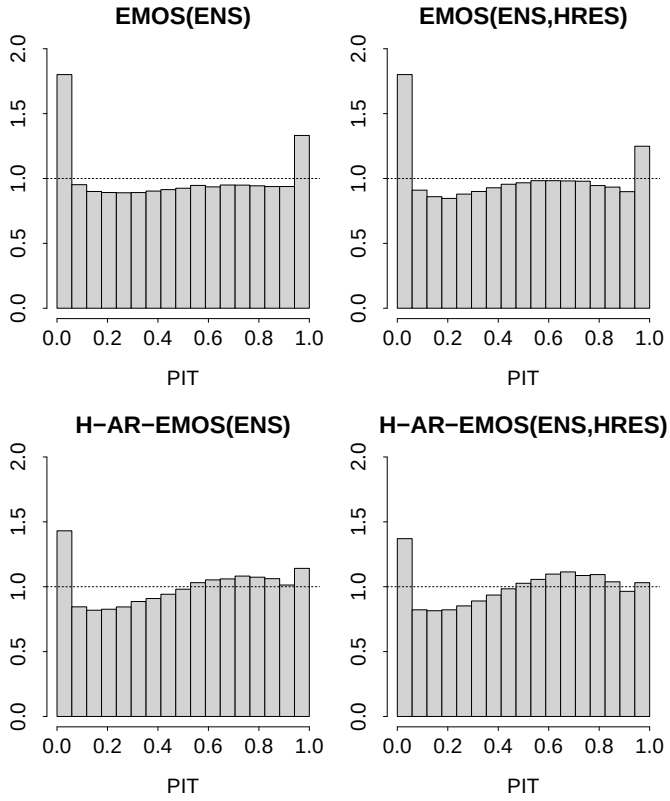


FIGURE 1. PIT histograms aggregated over 76 stations and 4341 test dates.

As described in Möller and Groß (2016), the proposed model can be combined with other predictive distributions (e.g. based on EMOS) in a spread-adjusted linear pool (Gneiting and Ranjan, 2013), to further improve predictive performance. The basic approach of combining the group-wise mean and variance parameters of the regular members and the high-resolution forecast by a (weighted) sum constitutes a simple 2-group model. The assumption of equal weights in both groups can be relaxed leading to weights estimated from training data, which provides a more data driven version of the 2-group model. The SLP combination proposed in Möller and Groß (2016) allows a further generalization of this approach. For each of k ensemble member groups a separate predictive distribution (e.g. obtained by the heteroscedastic AR-EMOS model, however, any postprocessing model could be used) is fitted and the distributions of the k groups are then SLP-combined.

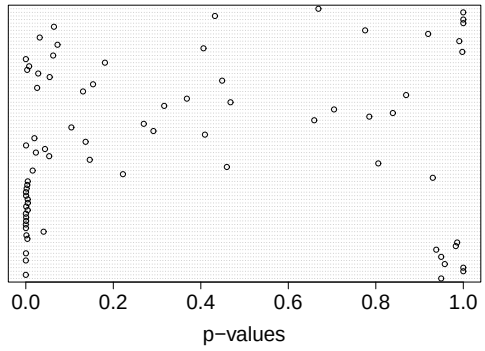


FIGURE 2. p-values of the 76 station-wise Diebold-Mariano tests for CRPS values of H-AR-EMOS vs. EMOS.

3 Case study with ECMWF temperature forecasts

We employ 24-h ahead forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF, see e.g. Buizza et al., 2007) for 2-m surface temperature over Germany. A data set over a period of 12 years is considered, ranging from 2002-01-01 to 2014-03-20 and containing forecasts and observations at 76 stations in Germany resulting in a total of 329916 forecast cases.

The ECMWF ensemble consists of 50 exchangeable members, a control forecast and a high-resolution forecast. In the original analysis of Möller and Groß (2016), only the 50 regular members where considered, while the two additional forecasts have not been utilized. For the analysis presented here, we make explicit use of the high-resolution forecast.

Table 1 presents results for the continuous ranked probability score and the Dawid-Sebastiani score (CRPS and DSS; see, e.g. Wilks, 2011; Gneiting and Katzfuss, 2014) for the EMOS model, and the heteroscedastic AR-EMOS model (named H-AR-EMOS here). It is clearly visible that incorporating the high-resolution forecast into any of the models improves the predictive performance substantially. Further, the H-AR-EMOS model substantially improves on the the standard EMOS approach in terms of the DSS, while only a small improvement is visible in the CRPS.

The last two columns of Table 1 present the root mean variance (RMV) and the variance of the probability integral transform (PIT) values. The RMV is used as sharpness measure for the predictive distribution, while the variance of the PIT values assesses the dispersion properties of the distribution, and thus its calibration. Neutral dispersion, corresponding to a uniform PIT histogram, is indicated by a PIT variance equal to $\frac{1}{12} = 0.0833$, the variance of the uniform distribution on $[0, 1]$ (see Gneiting and Ranjan, 2013). In case the predictive distribution matches the distribution the observations are drawn from, the respective PIT values would (approximately) follow a uniform distribution. In line with the principle "maximizing sharpness of the predictive distribution subject to calibration" (Gneinting and Katzfuss, 2014), both aspects of the distribution should be investigated in conjunction. That is, although EMOS has a slightly sharper predictive distribution this is at the expense of calibration, thus the better calibrated H-AR-EMOS model is preferable. Table 1 shows that the heteroscedastic AR-EMOS exhibits the best dispersion properties, with the model incorporating the high-resolution forecast closest to the variance indicating neutral dispersion. On the contrary, the EMOS models are slightly sharper than AR-EMOS, however, at the expense of calibration.

Figure 1 presents the corresponding PIT histograms of the considered models. Their appearance is in line with the values obtained for the PIT variance. While the EMOS models exhibit strong underdispersion (indicated by the strongly pronounced U-shape), the H-AR-EMOS models are able to correct the dispersion properties further. In addition to this, it can be noted that incorporating the high-resolution forecast improves the dispersion properties of both models compared to the version not using the high-resolution forecast.

To asses the significance of the difference in CRPS values a (one-sided) Diebold-Mariano test (Diebold and Mariano, 1995) is performed for the comparision H-AR-EMOS(ENS,HRES) vs. EMOS(ENS,HRES). Figure 2 shows the resulting 76 p-values. When looking at the Figure, it can be seen that for a large portion of the stations the test is highly significant, indicating that H-AR-EMOS indeed yields an improvement over EMOS.

Acknowledgments: We are grateful to the European Centre for Medium-Range Weather Forecasts (ECMWF) and the German Weather Service (DWD) for providing forecast and observation data, respectively.

References

Buizza, R., Bidlot, J.R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G. and Vitart, F. (2007). The new ECMWF VAREPS (variable resolution ensemble prediction system). *Quarterly Journal of the Royal Meteorological Society*, **133**, 681 – 695.

Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, **13**, 253 – 263.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, **1**, 125 – 151.

Gneiting, T., Raftery, A.E., Westveld III, A.H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098 – 1118.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, **7**, 1747 – 1782.

Leutbecher, M. and Palmer, T.N. (2008). Ensemble forecasting. *Journal of Computational Physics*, **227**, 3515 – 3539.

Möller, A. and Groß, J. (2016). Probabilistic temperature forecasting based on an ensemble AR modification. *Quarterly Journal of the Royal Meteorological Society*, **142**, 1385 – 1394.

Shumway R.H., Stoffer D.S. (2006). *Time Series Analysis and Its Applications. With R Examples.* Second Edition. Springer.

Wilks, D.S. (2011). *Statistical Methods in the Atmospheric Sciences.* Academic Press.

Probabilistic temperature post-processing using a skewed response distribution

Manuel Gebetsberger[1], Georg J. Mayr[1], Reto Stauffer[2], Achim Zeileis[2]

[1] Institute of Atmospheric and Cryospheric Sciences, Univ. of Innsbruck, Austria
[2] Department of Statistics, Univ. of Innsbruck, Austria

E-mail for correspondence: Manuel.Gebetsberger@uibk.ac.at

Abstract: Weather forecasts are typically based on numerical weather prediction models, where possible forecast errors can be corrected by statistical post-processing methods. A common quantity to develop, test, and demonstrate new post-processing methods is the surface air temperature which is frequently assumed to follow a Gaussian distribution. Nevertheless, the classical Gaussian distributional regression models which use only few covariates are not able to account for all local features leading to strongly skewed residuals. The authors demonstrate two approaches to overcome this problem: assuming a skewed response distribution to directly account for skewness, and extending the classical Gaussian distributional regression model by including all available information from the weather prediction model in combination with a boosting variable selection method.

The preliminary findings show satisfying results especially for an Alpine station by either using a generalized logistic type I distribution with few covariates, or using all available information plus an appropriate variable selection procedure. Both approaches are able to improve the predictions with respect to overall performance and calibration. Furthermore, similar results can be achieved for non-Alpine sites although with smaller improvements.

Keywords: Temperature; Ensemble; Forecast; Skewed; Distributional regression

1 Introduction

Weather forecasts are typically generated by numerical weather prediction (NWP) models. Nowadays, ensemble prediction systems (EPSs) are widely used where

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

multiple NWP runs with slightly perturbed initial conditions and parameterizations try to capture the forecast uncertainty. Nevertheless, it was found that such models often show systematic errors due to simplified physical equations, insufficient resolution, and unresolved processes (Bauer et al. 2015).

One possibility to correct for these errors are statistical post-processing methods. These methods have been tested extensively for various forecast quantities, especially surface air temperature. Distributional regression models (Klein et al. 2015) are one common post-processing strategy. Corrected probabilistic forecasts are obtained by modeling the parameters of a response distribution using linear predictors including covariates provided by an EPS.

Statistical models using the Gaussian assumption are generally able to improve the EPS output and to produce well-calibrated probabilistic temperature forecasts (Gneiting et al. 2015). However, this is not true for all geographical locations. Atmospheric processes associated with unresolved topographical features such as wintry cold pools are per definition not present in the EPS model and often lead to a skewed distribution. If these effects (e.g. wintry cold pools, strong valley heating during summer) cannot be depicted by the regression model, the residuals might be (strongly) skewed leading to inappropriately calibrated probabilistic forecasts.

To overcome possible skewness-related errors, this study assesses two different strategies for an Alpine and two non-Alpine stations. The benefit of a skewed response distribution will be evaluated using different distributional regression models. In addition, it is tested whether additional covariates from the EPS remove the need of a skewed response distribution. Focus will be on the Alpine station where pronounced skewness is visible in observed surface air temperature records.

2 Regression Framework

Distributional regression models can be expressed in a general form:

$$y \sim \mathcal{D}\big(h_1(\theta_1) = \eta_1,\ \ldots,\ h_K(\theta_K) = \eta_K\big) \tag{1}$$

where $\mathcal{D}$ represents the parametric distribution for the response y with θ_k, $k = 1, \ldots, K$ parameters. The parameters are linked to the additive predictors η_k using a monotone link function h_k (e.g., identity, logit, log). Each parameter $h_k(\theta_k)$ can be expressed by an additive predictor of the form:

$$\eta_k = \eta_k(\mathbf{x}, \beta_k) = f_{1k}(\mathbf{x}, \beta_{1k}) + \cdots + f_{Pk}(\mathbf{x}, \beta_{Pk}) \tag{2}$$

including various (possibly non-linear) functions f_{pk}, $p = 1, \ldots, P$.

3 Data and Model Setup for Comparison

Results for three different sites and $+12h$–$+96h$ forecasts in 6-hourly intervals are shown: Innsbruck located in the European Alps, Munich in the Alpine foreland, and Hamburg in a plain environment. These sites are selected to ensure spatial

independence of observation-forecast data pairs and to investigate the influence of different topographical environments. The covariates are based on forecasts from the European Centre for Medium-Range Weather Forecasts. The EPS provides (discrete) probabilistic forecasts consisting of 50 exchangeable ensemble members for a wide range of atmospheric variables. Two types of regression models are estimated: "simple" models which only use $2m$ temperature forecasts, and "complex" models using all available information (110 covariates) plus a smooth cyclic spline effect representing the seasonality as a function of the day of the year (*season*). Table 1 shows a summary of the models and their specification. We make use of the logistic and generalized logistic distribution in order to test distributions with slightly heavier tails compared to the classical Gaussian distribution. Additionally, the generalized logistic distribution type I (Johnson et al. 1995) is able to account for possible skewness and has the distribution function

$$F(x) = 1/(1 + \exp(-(x - \mu)/\sigma))^{\zeta}, \tag{3}$$

where ζ defines the additional shape parameter that has to be estimated. Due to the large number of covariates in the "complex" models, a likelihood gradient boosting approach (*R* package **bamlss**, Umlauf et al. 2017, R package version 0.1-1, https://r-forge.r-project.org/projects/bayesr/) is used for variable selection and coefficient estimation. A 10-fold block-wise cross-validation is further performed to ensure full out-of-sample results.

The overall performance of the models is evaluated on the ignorance score (Ign), and PIT histograms to visually assess calibration (Gneiting et al. 2015). To quantify the information provided by individual PIT histograms the reliability index $RI = \sum_{i=1}^{I} |\kappa_i - \frac{1}{I}|$ is chosen (Feldmann et al. 2015). It accounts for absolute deviations of each PIT-bin from perfect uniform calibration, where I defines the number of bins and κ_i the relative fraction in each bin. Ideally, RI values are close to zero. Additionally, the sharpness of the predictive distributions is verified in terms of average width of predictions intervals which should be as small as possible.

4 Results and Discussion

The response distribution for surface temperature is often assumed to be Gaussian in statistical ensemble post-processing studies. This is obviously not appropriate for the investigated Alpine station used in this study. The simple Gaussian

TABLE 1. Covariates used for distribution parameters μ, σ, ζ for all response distributions. "mean"/"sd": ensemble means and standard deviations, and "*season*": smooth seasonal effect. Covariates in brackets only for "complex" models.

Response Distribution	location μ	scale σ	shape ζ
Gaussian	mean (+*season*)	sd	—
Logistic	mean (+*season*)	sd	—
Generalized Logistic (I)	mean (+*season*)	sd	*season* (+mean)

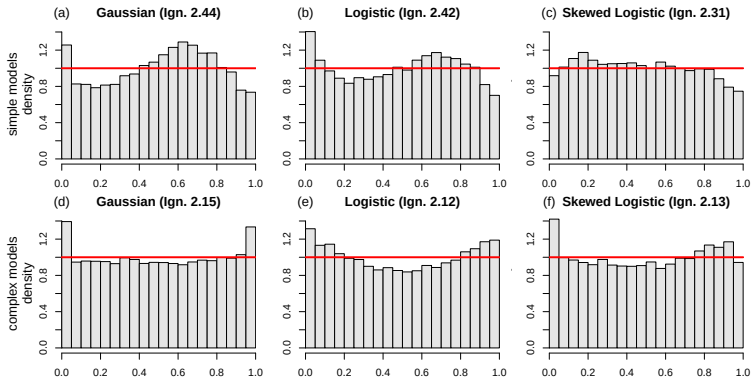


FIGURE 1. PIT histograms for the Alpine site. Left to right: Gaussian, logistic, and skewed logistic models. Top column shows the "simple" models, bottom column the "complex" models. Ignorance shown in brackets. Horizontal line indicates perfect calibration.

regression model shows uncalibrated and strongly skewed predictions (Fig. 1a). The skewed logistic distribution could clearly improve the predictive performance, indicated by a smaller ignorance and an almost uniformly distributed PIT histogram (Fig. 1c). This improvement is also visible for all individual lead times in terms of ignorance (Fig. 2a) and calibration (Fig. 2d).

A better overall performance in ignorance can be achieved using more complex models which can select from a large number of EPS covariates. Although additional covariates remove the need of a skewed response distribution, the complex Gaussian model still shows large errors in the tails (Fig. 1d), which might result from considerably smaller prediction intervals compared to simple Gaussian model (Fig. 2g).

Overall best ignorance for the Alpine site was achieved using all available covariates and a logistic response distribution (Fig. 1e), however, there is a distinct trade-off between perfect calibration, lowest possible ignorance, and also computational time. While simple models suggest the use of a skewed distribution, complex models perform very similar on all three distributions and imply that the Gaussian assumption might be sufficient. Nevertheless, these complex models using a cross-validated boosting procedure are roughly ten times more expensive in terms of computational time.

Similar results could also be found for the Alpine foreland site (Fig. 2b,e,h) and the topographically plain site (Fig. 2c,f,i), if only with smaller magnitude of improvements due to the two strategies and more visible at shorter lead times. At these two sites, the raw EPS has generally a better forecast performance due to less unresolved distinct local features in the EPS forecasts. Since better EPS forecasts lead to more informative covariates, the overall predictive performance of the statistical models is best at the plain site (Fig. 2c) and worst in the Alpine environment (Fig. 2a).

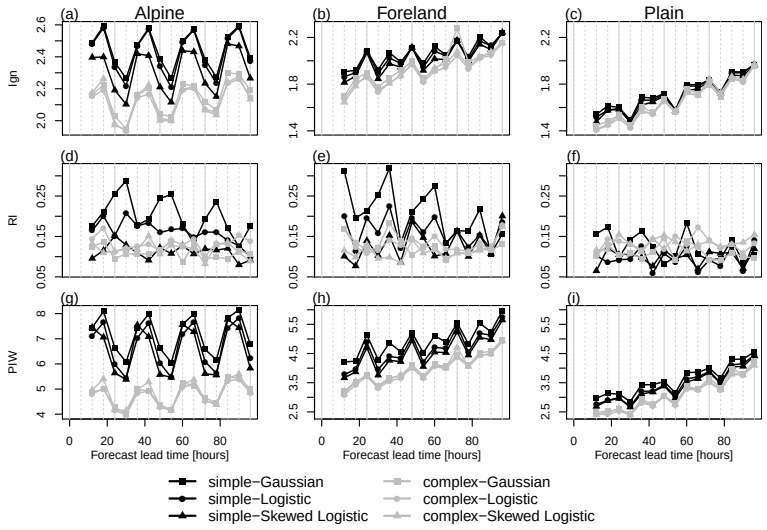


FIGURE 2. Verification measures for an Alpine, a foreland, and a plain site (left to right) using "simple" models (black lines) and "complex" models (grey lines). From top to bottom the ignorance score (Ign), reliability index (RI), and mean width of the 80% prediction interval (PIW) are shown, evaluated at forecasting lead times $+12h$–$+96h$. The 80% PIW defines the width between the predicted 0.1 and 0.9 quantile. Note that the Alpine site has a different range on the y-axis.

5 Conclusion

The results shown highlight the importance of an appropriate distributional assumption for post-processed surface air temperature forecasts, especially when only the corresponding temperature covariate is used in the regression model. A skewed logistic assumption improves calibration and forecast performance for all tested sites, but more pronounced in the Alpine environment.

Compared to this skewed regression model, complex models cannot further improve overall calibration but clearly obtain smaller prediction intervals which also lead to better forecast performance. Generally, the difference in forecast performance between the chosen distributional assumptions for the complex models is rather small. This indicates that the Gaussian assumption is already suitable given the large number of covariates. Nevertheless, these complex models are computationally roughly 10 times more expensive, and hence can be difficult to implement in an operational system.

The authors are currently extending this study by including additional sites to analyze the most influential covariates. A small set of such covariates might already lead to an improvement of similar magnitude, while considerably cutting down the computational costs. While the Gaussian assumption looks suitable for the complex models considered, it has to be investigated whether this is still true if the models only contain the most important covariates, especially for stations located in a complex environment.

Acknowledgments: Austrian Science Fund (FWF; TRP 290-N26), and Autonome Provinz Bozen, Abteilung für Bildungsförderung, Universität & Forschung (ORBZ110725). Results partly achieved using the HPC infrastructure LEO of the University of Innsbruck.

References

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, **525**, 47 – 55.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review*, **143**, 955 – 971.

Gneiting, T., Balabdaoui, F., and Raftery, A.E. (2015). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society*, **69(2)**, 243 – 268.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions., Volume 2. 2nd edition.* New York: Wiley & Sons.

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *The Annals of Applied Statistics*, **9(2)**, 1024 – 1052.

Umlauf, N., Klein, N., and Zeileis, A. (2017). *BAMLSS: Bayesian additive models for location, scale and shape (and beyond).* Working Papers, Faculty of Economics and Statistics, University of Innsbruck. URL: http://EconPapers.repec.org/RePEc:inn
05.

Boosting multivariate Gaussian models for probabilistic temperature forecasts

Thorsten Simon[1], Nikolaus Umlauf[2], Georg J. Mayr[1], Achim Zeileis[2]

[1] Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Austria
[2] Department of Statistics, University of Innsbruck, Austria

E-mail for correspondence: thorsten.simon@uibk.ac.at

Abstract: Many weather prediction tasks are multivariate problems, e.g., predicting several quantities (such as temperature and precipitation) for a particular time or predicting a single quantity over time. In the latter case, a state-of-the-art method is to fit several marginal prediction models and then combine them using ensemble copula coupling (ECC). As an alternative approach, we propose to fit a single multivariate Gaussian model where all parameters (means, variances, and correlations) can be expressed by additive models. For estimation of the resulting large number of parameters a gradient boosting algorithm is employed. Results for a case study show equal performance with respect to marginal predictive distributions and better performance with respect to the full multivariate distribution in comparison to nonhomogeneous Gaussian regressions (NGRs) combined with ECC.

Keywords: boosting; additive models; multivariate Gaussian; weather prediction.

1 Introduction

To obtain calibrated weather forecasts for several lead times, e.g., predicting temperature 12 h, 36 h, 60 h, ... in advance, the output of numerical weather prediction (NWP) ensemble systems is often postprocessed using statistical models. One popular choice for temperature forecasts is nonhomogeneous Gaussian regression combined with ensemble copula coupling (NGR-ECC, see Schefzik *et al.*, 2013). In a first step (NGR) linear models are employed for the location and scale parameter of a Gaussian distribution for several lead times separately. In a second

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

step (ECC) the predicted quantiles are reordered according to the raw ensemble output, in order to preserve the covariance structure of the NWP ensemble.

This study aims at extending NGR in the following way: Estimating predictive distributions for several lead times and their correlations simultaneously. The location, scale and correlation parameters of a multivariate Gaussian (MVN) will be expressed by GAM-type additive predictors η. Estimating multivariate distributions with these specifications is a complex task for dimensions higher than 2 (Klein *et al.*, 2015). Gradient boosting can offer an attractive solution to fit a MVN with additive predictors for all parameters as only first derivatives of the log-likelihood with respect the predictors are required. Another benefit of boosting is that selection and shrinkage of coefficients can be obtained.

2 Methods

The log-likelihood of the multivariate Gaussian for a k-dimensional observation $\mathbf{y} = (y_1, y_2, \ldots, y_k)^{\mathrm{T}}$ can be parameterized as follows,

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{y}) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}),$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)^{\mathrm{T}}$ denotes the vector of the mean parameters and $\boldsymbol{\Sigma}$ denotes the covariance matrix. The latter can be decomposed into $\boldsymbol{\Sigma} = \mathbf{D}\boldsymbol{\Omega}\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with the standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_k$ on the diagonal, and $\boldsymbol{\Omega}$ is the correlation matrix with the elements ρ_{ij}.

The parameters μ_i, σ_i and ρ_{ij} are linked to their predictors $\eta_{\mu i}$, $\eta_{\sigma i}$ and $\eta_{\rho ij}$ by the identity, the log and the rhogit function, respectively. The partial derivatives with respect to the predictors of μ_i and σ_i are,

$$\frac{\partial l}{\partial \eta_{\mu i}} = \sum_{j=1}^{k} \varsigma_{ij}(y_j - \mu_j) \quad \text{and} \quad \frac{\partial l}{\partial \eta_{\sigma i}} = -1 + \tilde{y}_i \sum_{j=1}^{k} \omega_{ij}\tilde{y}_j,$$

where ς_{ij} and ω_{ij} denote the elements of the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ and inverse correlation matrix $\boldsymbol{\Omega}^{-1}$, respectively. Additionally, $\tilde{y_i} = (y_i - \mu_i)/\sigma_i$. The partial derivative with respect to the predictor of ρ_{ij} is

$$\frac{\partial l}{\partial \eta_{\rho ij}} = \left[-\frac{1}{2}\omega_{ij} + \frac{1}{2}\left(\sum_{m=1}^{k}\omega_{im}\tilde{y}_m\right)\left(\sum_{m=1}^{k}\omega_{jm}\tilde{y}_m\right)\right] \times \left(1+\eta_{\rho ij}^2\right)^{-\frac{3}{2}}.$$

In order to fit the model a gradient boosting algorithm is applied as implemented by Umlauf *et al.* (2017). The algorithm is an iterative procedure. The number of iterations $m_{\max}$ has to be defined in advance. In each step the coefficients of the term which would contribute most to maximizing the log-likelihood are updated by the proportion ν of the local estimate of the coefficients. Thus, the boosting algorithm results in $m_{\max}$ distinct sets of coefficients. The optimal set of coefficients is selected by out-of-sample validation. A generic description of gradient boosting is given by Mayr *et al.* (2012).

3 Application in weather prediction

A case study is presented for predicting temperature in Innsbruck, Austria (47.260°N, 11.357°E). Six lead times are considered, 12 h, 36 h, ..., 132 h in advance. Data is on hand from January 2011 to December 2016 leading to a sample size of roughly 2150. The ensemble predictions of the European Centre for Medium-Range Weather Forecasts (ECMWF) serve as NWP input. The additive predictors are declared as follows,

$$\eta_{\mu i} = \alpha_{i,0} + \alpha_{i,1} * \text{mean}(\texttt{ens}_i) + f_{i,cc}(\texttt{yearday}),$$

$$\eta_{\sigma i} = \beta_{i,0} + \beta_{i,1} * \log(\text{sd}(\texttt{ens}_i)) + g_{i,cc}(\texttt{yearday}),$$

$$\eta_{\rho ij} = \gamma_{ij,0} + \gamma_{ij,1} * \text{cor}(\texttt{ens}_i, \texttt{ens}_j),$$

where indices $i, j \in \{1, 2, \ldots, 6\}$ refer to the lead times 12 h, 36 h, ..., 132 h, respectively. $\texttt{ens}_i$ denotes the raw ECMWF ensemble temperature forecast, and $f_{i,cc}(\texttt{yearday})$ and $g_{i,cc}(\texttt{yearday})$ are cyclic smooth functions modeled by splines to account for annual cycles. The model is trained on the period 2011–2015. Validation on the data of year 2016 leads to an optimal set of coefficients after $m_{\text{opt}} = 5000$ iterations where $\nu = 0.05$.

Figure 1 displays the fitted nonlinear functions $f_{i,cc}(\cdot)$ and $g_{i,cc}(\cdot)$, which contribute to the additive predictors $\eta_{\mu i}$ and $\eta_{\sigma i}$, respectively.

The coefficients describing $f_{i,cc}(\cdot)$ were not selected within the first 5000 iterations. Thus, $f_{i,cc}(\cdot)$ remains flat. This suggests that the bias between the model temperature and observations is constant throughout the year.

$g_{i,cc}(\cdot)$ (Fig. 1, right) contributes to the predictor of σ_i on the log-scale. $g_{i,cc}(\cdot)$ reveals an annual cycle with two peaks. One occurs in January and one in June/July. The fitted effects for all lead times vary only slightly among eachother.

However, the main focus of this study is to model the correlation structure of temperature between the lead times, 12 h, 36 h, ..., 132 h. Figure 2 summarizes the distribution of the fitted correlations. All parameters on the first diagonal next to the main diagonal of Ω vary around 0.74. The parameters on the second diagonal vary around 0.46. Thus, the fitted correlation matrixes exhibit a structure similar to a symmetric Toeplitz matrix or or even close to the correlation matrix of a AR-process.

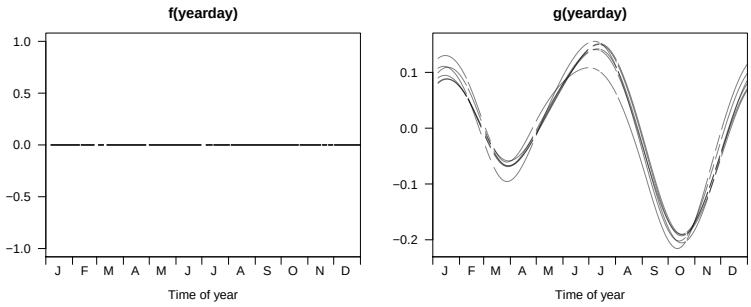


FIGURE 1. Nonlinear effects for all lead times. **Left:** $f_{i,cc}(\texttt{yearday})$ contributing to $\eta_{\mu i}$. **Right:** $g_{i,cc}(\texttt{yearday})$ contributing to $\eta_{\sigma i}$.

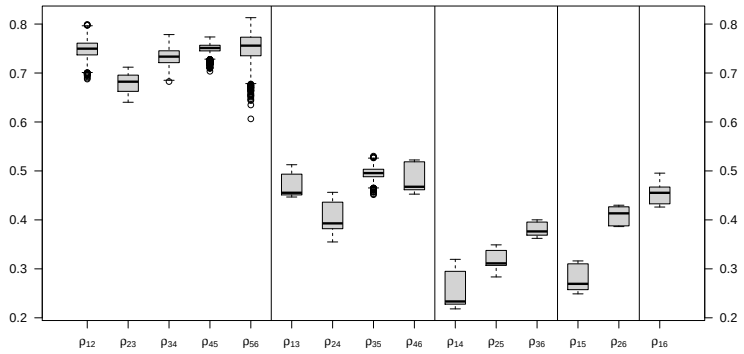


FIGURE 2. Fitted correlation coefficients ρ_{ij} for all days in 2016. Each box-and-whisker plot indicates the distribution of one correlation parameter over all sample cases.

The range indicated by the box-and-whisker plots (Fig. 2) suggests that the values of the intercepts $\gamma_{ij,0}$ are more important for determining the structure of the correlation matrix than the coefficients of the linear terms, $\gamma_{ij,1}$. Thus, only a small part of the correlation structure modeled by the numerical ensemble can be retained by the statistical model.

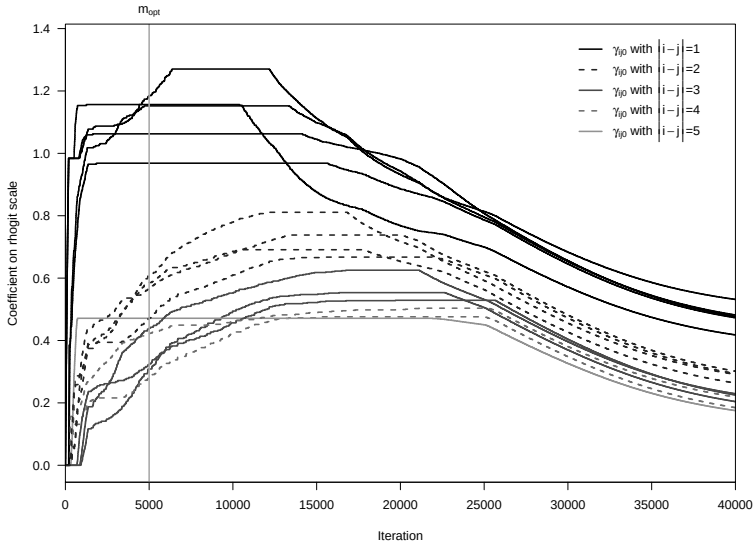


FIGURE 3. Boosting paths for the coefficients of the predictors of the correlation parameters on the rhogit scale.

Figure 3 illustrates how the values of the coefficients $\gamma_{ij,0}$ develop during the iterative boosting procedure. The intercepts $\gamma_{ij,0}$ are selected before the $\gamma_{ij,1}$ are

selected. However, after 15000–25000 iterations the values of the intercepts start dropping, which might be caused by an overfitting of the location parameter μ_i.

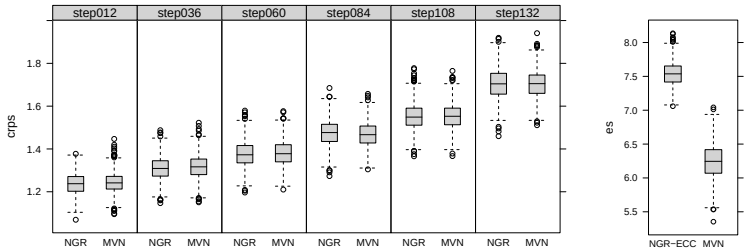


FIGURE 4. Out-of-sample scoring. **Left:** continuous rank probability score (CRPS) for univariate NGR models and marginal predictive distributions of the boosted MVN model. **Right:** energy score (ES) of the NGR-ECC and the boosted MVN.

Figure 4 compares the performance of the proposed method to the performance obtained by the state-of-the-art method NGR-ECC. The NGR models are also fitted via boosting. The marginal predictive distributions of the boosted MVN model are compared to NGR models fitted for every single lead time separately via the continuous ranked probability score (CRPS, Gneiting and Raftery, 2007). The left panel of Figure 4 reveals that the two models perform equally well with respect to their marginal distributions. The multivariate performance of the models is assessed in terms of the energy score (ES, Gneiting and Raftery, 2007). The boosted MVN outperforms the NGR-ECC in our case (Figure 4, right).

4 Conclusions

This study suggests to fit multivariate Gaussian distributions via gradient boosting, where additive predictors can be defined for all location, scale and correlation parameters. A case study in the field of weather forecasting shows promising results.

Further investigations are needed to fully understand the potential of boosting MVN with additive predictors. There are alternative ways to parameterize the correlation matrix, i.e., modeling the parameters of its inverse or its Cholesky decomposition (Pourahmadi, 2011).

In the present case one could assume an AR-process among the response variable as they are temporally ordered. This kind of parameterization is implicated by the findings in this study (cf. Fig. 2). However, more research is needed to examine whether a parsimonious or flexible parameterization is superior in this kind of application.

Depending on the problem changing the parameterization can have an effect on the required iterations until convergence of the boosting algorithm, and could yield different results when a shrunken version of the model is selected.

Acknowledgments: We acknowledge the funding by the Austrian Research Promotion Agency (FFG) project LightningPredict (Grant No. 846620).

References

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, **102**, no. 477, 359 – 378.

Klein, N., Kneib, T., Klasen, S. and Lang, S. (2015). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society, Series C*, **64**, Part 4, 569 – 591.

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**, Part 3, 403 – 427.

Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, **26**, no. 3, 369 – 387.

Schefzik, R., Thorarinsdottir, T.L., and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, **28**, no. 4, 616 – 640.

Umlauf, N., Klein, N., and Zeileis, A. (2017). BAMLSS: Bayesian additive models for location, scale and shape (and beyond). *Working Papers, Faculty of Economics and Statistics*, University of Innsbruck.

Generalization of the Whittle likelihood for nonparametric spectral density estimation

Claudia Kirch[1], Matthew Edwards[2], Alexander Meier[1], Renate Meyer[2]

[1] Otto-von-Guericke University, Magdeburg, Germany
[2] University of Auckland, New Zealand

E-mail for correspondence: meyer@stat.auckland.ac.nz

Abstract: Most nonparametric Bayesian approaches use Whittle's likelihood to estimate the spectral density as the main nonparametric characteristic of stationary time series, as e.g. Choudhuri et al. (2004) and Rosen et al (2012). But as shown in Contreras-Cristan et al. (2006), the loss of efficiency of the nonparametric approach using Whittle's likelihood can be substantial. We show that the Whittle likelihood can be regarded as a special case of a nonparametrically corrected parametric likelihood which gives rise to a robust and more efficient Bayesian nonparametric spectral density estimate based on a generalized Whittle likelihood. Its frequentist properties are investigated in a simulation study. Applications to LIGO gravitational wave data and the El Niño Southern Oscillation phenomenon will be described.

Keywords: Bayesian nonparametrics; stationary time series; spectral density estimation; Bernstein polynomial prior; gravitational waves.

1 Introduction

Most Bayesian nonparametric approaches to time series analysis are based on Whittle's likelihood approximation (Whittle, 1957), as e.g. Choudhuri et al. (2004) and Rosen et al. (2012). We will show that the Whittle likelihood can be regarded as the likelihood of a parametric working model, namely iid Gaussian, which has been nonparametrically corrected in the frequency domain. Borrowing an idea from a periodogram bootstrap for time series in Kreiss and Paparoditis (2003), we propose a generalization of the Whittle likelihood that uses a more realistic parametric working model, e.g. an AR(p) model, again nonparametrically corrected in the frequency domain and suggest a Bayesian semi-parametric

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

approach to spectral density estimation.

2 Generalized Whittle Likelihood

Let $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$ be a real stationary time series with mean zero and absolutely summable autocovariance function $\gamma(h)$. Then the spectral density exists, is given by the Fourier transform of the autocovariance function $f(\lambda) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k)\mathrm{e}^{-\mathrm{i}k\lambda}$ for $0 \leq \lambda \leq 2\pi$, is bounded and periodic, i.e. $f(2\pi - \lambda) = f(\lambda)$. Nonparametric estimation of the spectral density is based on the periodogram $I_n(\lambda) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} Z_t \mathrm{e}^{-\mathrm{i}t\lambda}\right|^2$. It is well known that the periodograms evaluated at the Fourier frequencies $\lambda_j = \frac{2\pi j}{n}$ are asymptotically independent and Exponential$(f(\lambda_j))$ distributed for $j = 0, \ldots, N = \lfloor (n-1)/2 \rfloor$. This gives rise to Whittle's likelihood approximation (Whittle, 1957)

$$p^W(\mathbf{Z}_n|f) \propto \exp\left\{-\sum_{j=0}^{N}\left(\log f(\lambda_j) + \frac{I_n(\lambda_j)}{f(\lambda_j)}\right)\right\}.$$

This approximation is exact only in the case of Gaussian white noise in which case $f(\lambda_j) = \sigma^2/2\pi$ but yields a good approximation also in the case of non-Gaussian time series, see e.g. Shao and Wu (2007). Now consider that we start with a $N(0, \sigma^2)$ working model. We then Fourier transform the time series, correct in the frequency domain by multiplication with a correction matrix $C_n^{1/2} = \mathrm{diag}(\ldots, \frac{f(\lambda_i)}{\sigma^2/2\pi}, \ldots)^{1/2}$, and inverse Fourier transform into the time domain. Then the density of the transformed time series $F_n^T C_n^{1/2} F_n \mathbf{Z_n}$ induced by the Gaussian iid working model is the Whittle likelihood. This gives rise to a generalization: instead of a Gaussian iid working model, we take a more realistic parametric working model, e.g. an AR(p) model, and define a correction matrix $C_n = C_n(f, f_{param}) = \mathrm{diag}(\ldots, \frac{f(\lambda_i)}{f_{param}(\lambda_i)}, \ldots)$ where f_{param} denotes the spectral density of the parametric working model. Then we call the density of the transformed time series $F_n^T C_n^{1/2} F_n \mathbf{Z}_n$ the *generalized Whittle likelihood*:

$$p^C_{param}(\mathbf{Z}_n|f) \propto \det(C_n)^{-1/2} p_{param}(F_n^T C_n^{-1/2} F_n \mathbf{Z}_n).$$

It can be shown that the classical Whittle likelihood is a special case, the nonparametrically corrected likelihood of a Gaussian AR(0) model. Furthermore, if the model is correctly specified, then $p^C_{param} = p_{param}$ but even in the misspecified case, the periodogram associated with the generalized Whittle likelihood is asymptotically unbiased for the true spectral density. For a proof, we refer to the preprint by Kirch et al. (2017).

3 Bayesian Spectral Density Estimation Using the Generalized Whittle Likelihood

In the following, we propose a semi-parametric Bayesian approach to spectral density estimation using the generalized Whittle likelihood and specify the prior distribution. We assume a Gaussian AR(p) working model and use a nonparametric Bernstein-Dirichlet prior for the spectral density as in Choudhuri et al.

(2004), see also Petrone (1999). However, we put the prior not on the spectral density f itself but on the ratio $c_\eta(\lambda) = f(\lambda)/f_{param}(\lambda; \mathbf{a})^\eta$ where $\mathbf{a}$ denotes the parameters of the AR(p) working model and $\eta \in [0, 1]$ captures the confidence in the parametric model. Assuming a stationary mean zero Gaussian time series with absolutely summable autocovariance function and a spectral density that is bounded away from zero, one can show that for fixed Gaussian AR(p) working model and a fixed confidence parameter η the generalized Whittle likelihood, the classical Whittle likelihood, and the true likelihood are all mutually contiguous. Furthermore, under some technical assumptions on the prior, the posterior distribution is consistent, i.e. the posterior distribution computed using the generalized Whittle likelihood will concentrate in a neighbourhood of the true spectral density for increasing length of the time series. For a detailed proof, see Kirch et al. (2017).

In practical applications, we will also estimate the parameters $\mathbf{a}$ of the AR(p) working model and the confidence parameter η jointly with the nonparametric spectral density. Therefore, prior specification is completed by putting a Uniform(0,1) prior on η and Uniform(-1,1) priors on the partial autocorrelations corresponding to the autoregressive parameters $\mathbf{a}$ of the AR(p) working model to ensure stationarity and causality. The degree p of the AR(p) model is fixed and determined in a preceding parametric AR(p) model selection procedure by the value of p minimizing the deviance information criterion (DIC), see Spiegelhalter et al. (2002) and Meyer (2016).

We sample from the joint posterior distribution using an adaptive version of a MH-within-Gibbs sampler, see also Roberts and Rosenthal (2009). Details regarding posterior computation can be found in Kirch et al. (2017) and software is available in the R package `beyondWhittle` on CRAN, see Meier et al. (2017).

4 Simulation Study

We briefly summarize the major findings of a comprehensive simulation study in Kirch et al. (2017). We generated data from various Gaussian ARMA(p, q) time series and compared the performance w.r.t. the average integrated absolute error (IAE) and the coverage probability of a uniform 95% credible interval (UCI) of the Bayesian spectral density estimate based on a parametric AR(p) likelihood, the classical Whittle likelihood, and the generalized Whittle likelihood, i.e. a nonparametrically corrected likelihood with an AR(p) working model where the order p was chosen by minimization of DIC. In the correctly specified AR(p) scenario, of course the parametric estimator performed best. But the estimator based on the generalized Whittle likelihood benefits from the correctly specified AR(p) model and performed only marginally inferior to the parametric model whereas the estimator based on the Whittle likelihood had much higher IAE and lower UCI coverage. In the misspecified MA(q) scenario, the parametric estimator had highest IAE and lowest coverage as expected, whereas the nonparametric estimator based on the Whittle likelihood now had lowest IAE. However, the generalized Whittle estimator was only marginally inferior w.r.t. IAE but had even higher UCI coverage.

5 Applications

The Southern Oscillation Index (SOI), the monthly standardized anomaly of the mean sea-level pressure difference between Tahiti and Darwin (available form the Australian Bureau of Meteorology) is one of the key atmospheric indices for gauging the strength of El Niño events and their potential impacts on the Australian region. We use the mean-centered SOI time series, shown in Figure 1, to illustrate the effect of the choice of the AR(p) working model on the spectral density estimate.

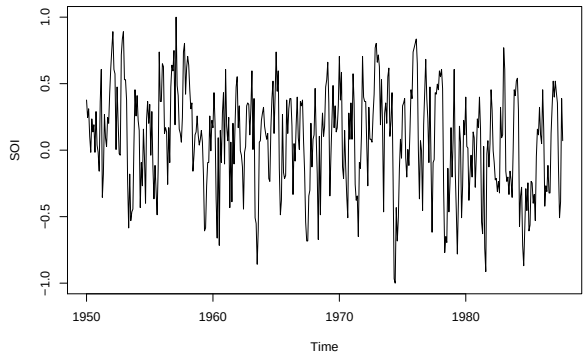


FIGURE 1. Plot of the SOI index from 1950 to 1987.

Figure 2 shows the negative maximum log likelihood (NLL) value for different AR(p) models.

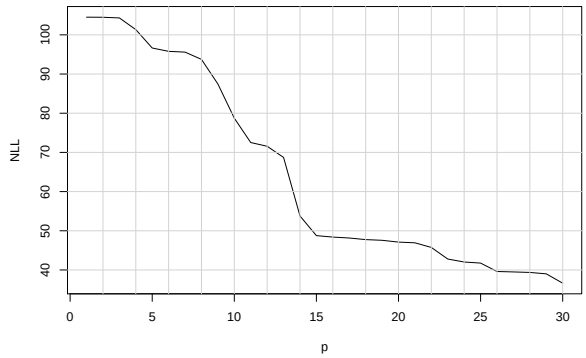


FIGURE 2. Negative maximum log likelihood for various AR(p) models of SOI data.

Figure 3 shows the log-periodogram overlaid by the posterior median spectral density estimates of the parametric AR(p) model and generalized Whittle likelihood for the autoregressive orders $p = 0, 5, 11$ and 15 that correspond to "elbows"

in the scree-like plot of the negative maximum log likelihood values in Figure 2. For $p = 0$, the Bernstein polynomials of the nonparametric correction cannot yet

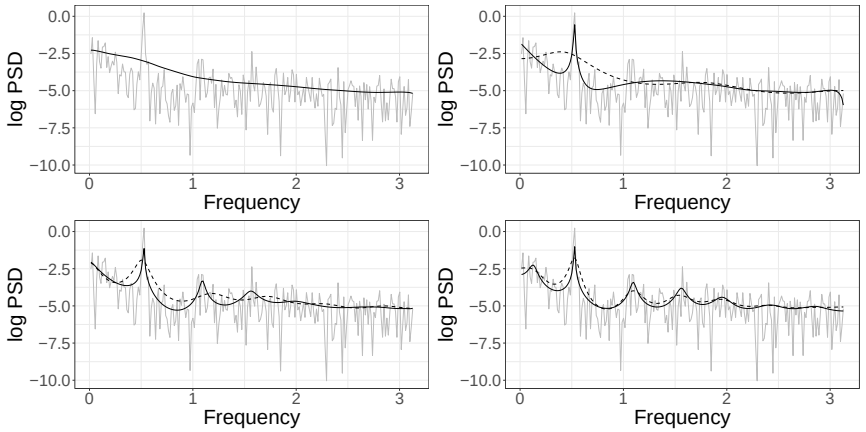


FIGURE 3. Posterior median spectral density estimates of the generalized Whittle likelihood (solid black) and AR (dashed black) for the SOI data on a logarithmic scale, for different autoregressive orders $p = 0, 5, 11, 15$.

capture the main peak whereas for $p = 5$ they can. With $p = 11$, minor peaks are also well estimated with the generalized Whittle but not with the classical likelihood. The estimators of a parametric AR(15) and the nonparametrically corrected model are very similar but all peaks of the AR(15) estimator are sharpened even further via the nonparametric correction.

Because of space constraints, we refer to Kirch et al. (2017) for an application to LIGO gravitational wave data that we aim to discuss in detail at the conference.

Acknowledgments: This work was supported by DFG grant KI 1443/3-1. We also thank the New Zealand eScience Infrastructure (NeSI) and the Universitätsrechenzentrum (URZ) Magdeburg for their high performance computing facilities, and the Centre for eResearch at the University of Auckland and Jörg Schulenburg for their technical support.

References

Choudhuri, N., Ghosal, S. and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *JASA*, **99**, 1050 – 1059.

Contreras-Cristán, A., Gutiérrez-Peña, E. and Walker, S.G. (2006). A note on Whittle's likelihood. *Communications in Statistics – Simulation and Computation*, **35**, 857 – 875.

Kirch, C., Edwards, M., Meier, A., and Meyer, R. (2017). Beyond Whittle: Nonparametric correction of a parametric likelihood with a focus on Bayesian time series analysis. `https://arxiv.org/abs/1701.04846`.

Kreiss, J.-P., and Paparoditis, E. (2003). Autoregressive-aided periodogram bootstrap for time series. *Annals of Statistics*, **31**, 1923 – 1955.

Meier, A., Kirch, C., Edwards, M., and Meyer, R. (2017). beyondWhittle: Bayesian Spectral Inference for Stationary Time Series. R package.

Meyer, R. (2016). Deviance Information Criterion (DIC). N. Balakrishnan, P. Brandimarte, B. Everitt, G. Molenberghs, W. Piegorsch, and F. Ruggeri (Eds.). *Wiley StatsRef: Statistics Reference Online*, Wiley.

Petrone, S. (1999). Random Bernstein Polynomials. *Scandinavian Journal of Statistics* **26**, 373 – 393.

Roberts, G.O. and Rosenthal, J.S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18**, 349 – 367.

Rosen, O., Wood, S., and Stoffer, D.S. (2012). Adaptive Spectral Estimation for Nonstationary Time Series. *JASA*, **107**, 1575 – 1589.

Shao, X. and Wu, B.W. (2007). Asymptotic spectral theory for nonlinaer time series. *Annals of Statistics*, **35**, 1773 – 1801.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *JRSSB*, **64**, 583 – 639.

Whittle, P. (1957). Curve and periodogram smoothing. *JRSSB*, **19**, 38 – 63.

Model-based clustering for populations of networks

Mirko Signorelli[1,2], Ernst C. Wit[1]

[1] Johann Bernoulli Institute, University of Groningen, Netherlands
[2] Department of Statistical Sciences, University of Padova, Italy

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

Abstract: We propose a clustering method for sequences of graphs. The method exploits mixtures of generalized linear models and allows to detect subpopulations of graphs in a model-based fashion. We provide an Expectation-Maximization algorithm with Simulating Annealing (EMSAGC) for model estimation.

Keywords: EMSAGC; network; graph; mixtures of generalized linear models; model-based clustering.

1 Introduction

The last decades have witnessed a growing interest in network science. Although statistical network models traditionally aimed at modelling relations encoded in a single network, the increasing availability of relational data has encouraged the collection of several instances of the same network. Examples include longitudinal sequences of networks, which represent the dynamic evolution of a complex system, and populations of networks, where each network represents the state of a system for a given statistical unit.

Since it is reasonable to expect networks within such longitudinal or cross-sectional sequences of networks to be similar to a certain degree, it is clear that modelling each network separately would be a dispersive and ineffective strategy. Instead, by jointly modelling them we can borrow information between similar graphs and achieve a much more parsimonious answer.

For these reasons, we propose a novel method capable to characterize the joint distribution of sequences of networks and to cluster networks, which relies on mixtures of generalized linear models (Grün and Leisch, 2008). This approach combines the flexibility of mixture models with the fact that many popular network models can be estimated by specifying a generalized linear model. We also

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

propose EMSAGC, an estimation algorithm based on the integration of the EM with simulated annealing.

2 Model-based clustering of networks

We consider a sequence of K graphs $\mathcal{S} = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_K\}$, where each graph $\mathcal{G}_k = (V, E_k)$, $k \in \{1, ..., K\}$, defines a specific set of edges E_k between the same set of v vertices V, and it is represented by the adjacency matrix Y_k. We represent $\mathcal{S}$ with an array $\mathbf{Y}$ of size $v \times v \times K$. An entry y_{ij}^k in $\mathbf{Y}$ refers to the presence (and intensity) or absence of edge $e_{ij} \in E_k$.

In principle, we could imagine that each graph $\mathcal{G}_k$ is drawn from a different distribution $f(Y|\theta_k)$, $k \in \{1, ..., K\}$ with parameter vector θ_k - to wit, $Y_k \sim f(Y|\theta_k)$. In the presence of many networks, however, this would result in a cumbersome modelling exercise, leading to K different models.

Instead, we postulate the existence of $M \leq K$ subpopulations of graphs $\mathcal{S}_1, ..., \mathcal{S}_M$ within $\mathcal{S}$, each with density $f(Y|\theta_m)$, $m \in \{1, ..., M\}$. We denote by $Z_k \in \{1, ..., M\}$ the latent identifying label of graph $\mathcal{G}_k$, such that $Z_k = m$ if $\mathcal{G}_k \in \mathcal{S}_m$. Therefore, we view each graph $\mathcal{G}_k \in \mathcal{S}$ as a random draw from a mixture model whose components are the densities $f(Y|\theta_m)$

$$Y_k \sim \sum_{m=1}^{M} \pi_m f(Y|\theta_m), \tag{1}$$

with mixing proportions $\pi_m = Pr(Z_k = m)$, $m \in \{1, ..., M\}$ denoting the prior probabilities that a graph belongs to the mth subpopulation $\mathcal{S}_m$. Clearly, $\pi_m \geq 0 \ \forall m$ and $\sum_m \pi_m = 1$. The likelihood of $\mathcal{S}$ is then

$$L(\mathbf{Y}, Z|\Theta) = \prod_{k=1}^{K} Pr(Y_k|Z_k, \Theta) Pr(Z_k|\Theta) = \prod_{k=1}^{K} \pi_{Z_k} f(Y_k|\theta_{Z_k}), \tag{2}$$

where $\Theta = (\theta_1, ..., \theta_M)$ and $Z = (Z_1, ..., Z_K)$.

Many popular network models, such as the p_1 and p_2 models (Holland and Leinhardt, 1981; van Duijn et al., 2004) as well as stochastic blockmodels and their extensions (Wang and Wong, 1987), assume edges (or, for directed graphs, dyads) to be statistically independent and, thus, they can be estimated with a generalized linear model (or, in the case of the p_2 model, a generalized linear mixed model). If one of such models is employed, each of the densities $f(Y|\theta_m)$ in Equations (1) and (2) can thus be estimated by specifying a generalized linear model. This can be done by considering densities f from exponential dispersion families, and modelling the conditional expectation of each edge y_{ij} as

$$\eta_{ij} = g[E(y_{ij}|x_{ij}, \theta_m)] = x_{ij}\beta, \tag{3}$$

where x_{ij} indicates a set of nodal or edge-specific attributes and g is the link function. Clearly, the density $f(Y|\theta_m)$ can then be obtained as

$$f(Y|\theta_m) = \prod_{i<j} f(y_{ij}|\theta_m). \tag{4}$$

With these specifications, model (1) defines a mixture of generalized linear models (Grün and Leisch, 2008).

Within this framework, it is possible to cluster graphs according to features such as node degree, reciprocation, membership of groups and nodal or edge-specific covariates. For example, if the interest is simply on nodal degree, a mixture of p_1 models can be considered; if a partition of nodes into groups is known, a mixture of stochastic blockmodels can be specified as well (we provide an example of this in Section 4). More generally, if one would simply like to cluster graphs without assuming a specific network model, they can specify a model with one parameter for each pair of nodes, i.e., $\eta_{ij}^k = \gamma_{ij}^m$. Note, however, that this approach does not allow to cluster graphs according to their tendency towards transitivity, since Exponential Random Graph Models cannot be specified as generalized linear models.

3 Model estimation

3.1 EMSAGC: Expectation-Maximization algorithm with Simulated Annealing for Graph Clustering

In principle, maximization of (2) could be performed with a simple implementation of the EM algorithm. However, in our simulations this often resulted in low accuracy of the clustering method, in particular for the cases of binary graphs and of sparse edge-valued graphs (Signorelli, 2017). Therefore, we propose EMSAGC, an estimation algorithm that integrates the EM with Simulating Annealing and is capable to improve the clustering accuracy. Simulated Annealing (Eglese, 1990) is a strategy that avoids the risk to get trapped in a local optimum of an objective function f by proposing a move from the current local maximum $\hat{x}$ to a proposal $\tilde{x}$, and by allowing a positive probability to accept the move even when $f(\tilde{x}) < f(\hat{x})$. The algorithm consists of the following iterative steps:

1. for $k \in \{1, ..., K\}$ and $m \in \{1, ..., M\}$, define the initial probabilities $p_{km}^1 = Pr(z_{km} = 1)$. Denote by P^1 the $K \times M$ matrix which collects these probabilities;
2. for $t = 1, 2, ...$:
   - □ **M step with Simulated Annealing**.
     - M1. Given P^t, estimate M network models (specified as GLMs) with weights given by $\left(p_{1m}^t, ..., p_{Km}^t\right)$ for the m-th component and derive $\hat{\Theta}^t$.
     - M2. If $t \geq 2$ and $L(\mathbf{Y}, Z|\hat{\Theta}^t) \leq L(\mathbf{Y}, Z|\hat{\Theta}^{t-1})$, consider the alternative state $\tilde{P}^t$ and estimate $\tilde{\Theta}^t$:
       - ⋆ if $L(\mathbf{Y}, Z|\tilde{\Theta}^t) \geq L(\mathbf{Y}, Z|\hat{\Theta}^t)$, set $\hat{\Theta}^t = \tilde{\Theta}^t$ and $P^t = \tilde{P}^t$.
       - ⋆ if $L(\mathbf{Y}, Z|\tilde{\Theta}^t) < L(\mathbf{Y}, Z|\hat{\Theta}^t)$, set $\hat{\Theta}^t = \tilde{\Theta}^t$ and $P^t = \tilde{P}^t$ with probability equal to

$$\left(\frac{\log L(\mathbf{Y}, Z|\tilde{\Theta}^t)}{\log L(\mathbf{Y}, Z|\hat{\Theta}^t)}\right)^{1/T(t)}, \tag{5}$$

       where $T(t) = \frac{1}{\log t}$.

□ **E step**. Given $\hat{\Theta}^t$, derive P^{t+1} as

$$p_{km}^{t+1} = \frac{Pr(\mathcal{G}_k|\hat{\theta}_m^t)}{\sum_{j=1}^{M} Pr(\mathcal{G}_k|\hat{\theta}_j^t)}. \quad (6)$$

3. Choose the best solution within the sequence $\{\hat{\Theta}^1, \hat{\Theta}^2, ..., \}$, i.e.

$$\hat{\Theta}^{EMSAGC} = \text{argmax}_{t=1,2,...} L(\mathbf{Y}, Z|\hat{\Theta}^t). \quad (7)$$

3.2 A comparison of the performance of the EM and EMSAGC algorithms

We have assessed the performance of the EM and EMSAGC algorithms on simulated sequences of undirected networks with two mixture components. We have considered both binary graphs and edge-valued graphs, employing mixtures of logistic and of Poisson generalized linear models to estimate the clusters. When a sequence of edge-valued graphs whose edge values follow a Poisson distribution is considered, the EM algorithm already results in a very good clustering accuracy.

Instead, as Table 1 shows, when we consider sequences of binary graphs and of graphs whose edge values are overdispersed (i.e., they follow a negative binomial distribution) the accuracy of the EM algorithm is often low, especially when either K or v are very small. In these cases, application of EMSAGC results in a higher accuracy of the retrieved clusters, yielding a substantial improvement over the performance of the EM algorithm. Full details on the simulation settings can be found in Signorelli (2017).

4 Example application

We consider 10 daily interaction networks between employees of the French Institute for Public Health Surveillance, collected by Génois et al. (2015). In partic-

TABLE 1. Comparison of the accuracy obtained with the EM and EMSAGC algorithms, for binary graphs and edge-valued graphs with overdispersed degree distribution. Each cell displays the percentage of correctly clustered graphs.

		Binary graphs		Sparse edge-valued graphs	
K	v	EM	EMSAGC	EM	EMSAGC
10	10	72%	74%	67%	86%
10	50	68%	99%	65%	100%
10	100	65%	98%	66%	100%
20	10	92%	100%	81%	92%
20	50	85%	100%	91%	100%
20	100	81%	100%	84%	100%
50	10	100%	100%	96.4%	98%
50	50	100%	100%	100%	100%
50	100	100%	100%	100%	100%

ular, we focus on those employees who have registered interactions for at least 6 days; each of them belongs to one of four departments (DISQ, DMCT, DSE and SRH). It is reasonable to assume that interactions are affected both by department affiliation and by individual features. Therefore, we consider a mixture of degree-corrected stochastic blockmodels with two components, where the number of daily interactions between node i from department r and node j from department s follows a Poisson distribution whose mean depends, on the log-scale, both on nodal effects α_i, α_j and on block-interaction parameters ϕ_{rs}.

The application of the EMSAGC algorithm allows to detect a variation in the pattern of interactions among departments from the first 7 days (cluster 1) to the remaining 3 days (cluster 2) considered in the study. The estimate of the block-interaction parameters in the two clusters are shown in Table 2. Overall, we find two changes in the pattern of interaction across departments. On the one hand, members of DISQ and DSE are more active within their department in the first 7 days considered, but then they interact more with each other in the remaining 3 days. On the other hand, employees in DMCT and SRH seem to follow the opposite pattern: in the last 3 days, they reduce interactions between their departments and are more active within their own department.

References

Eglese, R. (1990). Simulated annealing: a tool for operational research. *European journal of operational research*, 46(3), 271–281.

Genois, M. et al. (2015). Data on face-to-face contacts in an office building suggests a low-cost vaccination strategy based on community linkers. *Network Science*, 3(03), 326–347.

TABLE 2. Comparison of block-interactions between clusters. Employees in departments DSE and DISQ interact more within their department in the first 7 days, and more between each other in the last 3 days (corresponding parameters are emphasised in bold). Conversely, employees in DMCT and SRH interact more with each other in the first 7 days, and within their own departments in the last 3 days (corresponding parameters are underlined).

Block-interaction parameter	Estimates	
	Cluster 1	Cluster 2
DMCT	0.71	0.86
DSE	**0.92**	**0.58**
DISQ	**1.43**	**1.10**
SRH	1.83	2.01
DMCT-DSE	-0.15	-0.12
DMCT-DISQ	-0.22	-0.18
DMCT-SRH	-0.34	-0.56
DSE-DISQ	**-0.24**	**0.03**
DSE-SRH	-0.53	-0.49
DISQ-SRH	-0.96	-0.96

Grün, B. and Leisch, F. (2008). Finite mixtures of generalized linear regression models. In: *Recent advances in linear models and related areas*, 205-230. Springer.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373), 33–50.

Signorelli, M. (2017). *Inferring community-driven structure in complex networks.* Doctoral dissertation, University of Groningen. Retrieved from *http://hdl.handle.net/11370/5e39 168a-49f9-95e2-98860de508a9.*

van Duijn, M. A. J., Snijders, T. A. B., Zijlstra, B. H. J. (2004). p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2), 234–254.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397), 8–19.

Generalized Method of Moments for Estimating the Parameters of Stochastic Actor-oriented Models

Tom A. B. Snijders[1,2], Viviana Amati[3], Felix Schönenberger[3]

[1] Department of Sociology, University of Groningen, The Netherlands
[2] Nuffield College and Departement of Statistics, University of Oxford, UK
[3] Department of Computer and Information Science, University of Konstanz, Germany

E-mail for correspondence: tom.snijders@nuffield.ox.ac.uk

Abstract: Stochastic actor-oriented models are models for dynamic network data, collected by observing a network and a behavior over time in a panel design. We present an estimator based on the generalized method of moments, using more statistics than parameters. For some examples we study the resulting gain in statistical efficiency. We discuss algorithmic issues that have to be solved to obtain a stable procedure.

Keywords: Network panel data; Stochastic actor-oriented models; Generalized method of moments; Selection and Influence processes.

1 Stochastic actor-oriented models

Stochastic actor-oriented models (SAOMs) (Snijders, 2001; Steglich et al., 2010) are applied to analyze network panel data $(x, z)_{t_1}, \ldots, (x, z)_{t_M}$, defined as observations of a network x and a behavior z observed at $M > 1$ time points. The network x is a digraph; the behavior z is an ordinal discrete variable with integer values, defined at the level of actors.

In SAOMs the network and the behavior are jointly the two dependent variables. The dependence of network dynamics and that of behavioral dynamics on the network-behavioral configuration are referred to as "selection" process and "influence" process, respectively. SAOMs allow to distinguish the role of selection and influence mechanisms in explaining the network and behavioral changes over time.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

Following the approach originally proposed by Holland and Leinhardt (1977), SAOMs assume that the time series $(x,z)_{t_1},\ldots,(x,z)_{t_M}$ is the outcome of a continuous-time Markov chain. The observed panel data are assumed to be the result of a sequence of unobserved changes happening between two consecutive observation time points t_{m-1} and t_m. For simplicity, we treat only the case $M=2$.

The series of changes is modeled in the following way. At a random moment, one actor is selected and has the opportunity to make either a network or a behavioral change. Due to the Markov assumption, the time between two consecutive opportunities for a change must be exponentially distributed. The corresponding rate parameters are called $\lambda_m^{[X]}$ and $\lambda_m^{[Z]}$ for network and behavioral changes, respectively.

Probabilities of changes are determined by so-called *evaluation functions*, defined separately for the two variables. For variable $V=X$ or Z, the evaluation function is given by

$$f_i^{[V]}(\beta^{[V]},x,z)=\sum_{k=1}^{K^{[V]}}\beta_k^{[V]}\,s_{ik}^{[V]}(x,z) \tag{1}$$

where $s_{ik}^{[V]}(x,z)$ for $V=X,Z$ are suitable statistics of which a description can be found, e.g., in Steglich et al. (2010).

When an actor gets the opportunity for a network change, he can create one new outgoing tie, terminate one existing outgoing tie, or do nothing. When an actor gets the opportunity for a behavioral change, he can increase or decrease his behavior by one unit, or do nothing. Changes are modeled according to a multinomial logit model. Given the current state (x,z), the probability $p_{(x',z)}^{[X]}$ that, for a network change, the next state is (x',z), and the probability $p_{(x,z')}^{[Z]}$ that, for a behavioral change, the next state is (x,z'), is given by

$$p_{(x',z)}^{[X]}=\frac{\exp\big(f_i^{[X]}(\beta^{[X]},x',z)\big)}{\sum_{x''}\exp(f_i^{[X]}(\beta^{[X]},x'',z))}\quad\text{and}\quad p_{(x,z')}^{[Z]}=\frac{\exp\big(f_i^{[Z]}(\beta^{[Z]},x,z')\big)}{\sum_{z''}\exp\big(f_i^{[Z]}(\beta^{[Z]},x,z'')\big)}.$$

The vector of parameters $\theta=(\lambda^{[X]},\lambda^{[Z]},\beta^{[X]},\beta^{[Z]})$ of SAOMs is usually estimated using the method of moments (MoM), solving the equation

$$E_\theta\big[s\big(X(t_2),Z(t_2)\big)-s\big(x(t_2),z(t_2)\big)\;\big|\;\big(X(t_1),Z(t_1)\big)=\big(x(t_1),z(t_1)\big)\big]=0\;. \tag{2}$$

Statistics $s(X,Z)$ modeling influence (dependent variable: behavior) are distinguished from those modeling selection (dependent variable: network) by using cross-lagged variables. Statistics suggested by (1), providing information about $\beta^{[X]}$ and $\beta^{[Z]}$, respectively, are

$$\sum_i s_i^{[X]}\big(X(t_2),Z(t_1)\big)\quad\text{and}\quad\sum_i s_i^{[Z]}\big(X(t_1),Z(t_2)\big)\;.$$

The solution of (2) is the MoM estimate for θ. It can be approximated using a stochastic algorithm (Snijders, 2001) based on Monte Carlo approximation of the expected values and the Robbins-Monro step (Robbins and Monro, 1951). The algorithm is implemented in the library `RSiena` (`http://www.stats.ox.ac.uk/ snijders/siena/`) of the R software.

2 Generalized method of moments for SAOMs

It is natural that we would like to include as much information as possible in the estimation process. When analyzing the co-evolution of networks and behaviors, additional information is provided by the non-cross-lagged statistics $s_i^{[V]}(X(t_2), Z(t_2))$ for $V = X, Z$, where both dependent variables are taken at t_2. Taking these into account may improve the estimation, especially when changes in both the behavior and the set of outgoing ties are observed for a significant proportion of actors. We refer to these changes as *simultaneous changes*. Here, we propose to include the non-cross-lagged statistics in the estimation process to get a better estimate for θ, following the generalized method of moments (GMoM; Hansen, 1982) estimation.For notational brevity, we denote $(X, Z) = \big(X(t_2), Z(t_2)\big)$ and (x, z) likewise.

Including more statistics than parameters results in an over-identified system of moment equations. The GMoM estimate for θ is the value $\widehat{\theta}$ minimizing the distance between the expected values of the statistics and their sample counterpart:

$$F(X, Z; \theta) = E_\theta[s^*(X, Z) - s^*(x, z)]^{\mathrm{T}} W E_\theta[s^*(X, Z) - s^*(x, z)] \tag{3}$$

where $s^*(X, Z)$ is the vector containing the cross-lagged as well as non-cross-lagged statistics, the expectation is conditional given $\big(X(t_1), Z(t_1)\big)$ as in (2), and W is a positive semi-definite matrix. An optimal choice for W is the inverse of the variance-covariance matrix of $s^*(X, Z)$.

From (3) and the convexity of $F(X, Z; \theta)$, it follows that the estimate $\widehat{\theta}$ is the value solving

$$\frac{\partial}{\partial \theta} F(X, Z; \theta) = B_\theta E_\theta[s^*(X, Z) - s^*(x, z)] \,=\, 0 \tag{4}$$

where $B_\theta = \partial\big(E_\theta[s^*(X, Z) - s^*(x, z)]^{\mathrm{T}} W\big)/\partial\theta$.

Equation (4) is equal to equation (2) except for the matrix B_θ. Therefore, $\hat{\theta}$ can be computed using a modification of the stochastic approximation algorithm used also for the MoM. Main changes in the algorithm are related to the estimation and the inclusion of the matrix B_θ in the Robbins-Monro step. Based on the econometric literature, we implemented both the efficient two-step and the iterated two-step GMoM estimators and tested their performance. However, here we report only the results based on the iterated two-step estimator.

3 Simulations

Several simulations were conducted to investigate the relative efficiency of the GMoM estimator with respect to the MoM estimator. A nested simulation design was used: For several specification of the SAOMs, 500 network panel data were generated and estimated twice by both methods. Such a design has the advantage of accounting for the variation in both the simulations and the stochastic algorithm.

For illustrative purposes, we consider a simple model specification based on empirical results deriving from the estimation of SAOMs in friendship networks and on the use of only one non-cross-lagged statistic as described in the following.

The simulated network panel data comprises a binary directed network and a behavior with 5 ordinal categories, both simulated at 2 points in time, as well as a constant binary actor covariate. Those were generated using a SAOM specified by basic effects describing the network structure (e.g. density and reciprocity), the impact of an actor covariate (e.g. popularity and activity) and the distribution of the behavior (linear and quadratic effects). We also included in the model the behavior similarity effect and the average similarity effect. These two model homophily whose outcome is the similarity of two connected actors.

The similarity effect models selection, i.e. the tendency of actor to form ties with actors that are similar to themselves, whereas the average similarity effect models influence, i.e. the tendency of actors which are tied to become similar in their behavior. The corresponding cross-lagged statistics $s_{sim}^{[X]}$ and $s_{av\_sim}^{[Z]}$ are defined, respectively, as

$$s_{sim}^{[X]} = \sum_{ij} x_{ij}(t_2) \left(\frac{\Delta - |z_i(t_1) - z_j(t_1)|}{\Delta} - \overline{sim}_z \right)$$

$$[Z]_{av\_sim} = \sum_{ij} \frac{1}{\sum_j x_{ij}(t_1)} x_{ij}(t_1) \left(\frac{\Delta - |z_i(t_2) - z_j(t_2)|}{\Delta} - \overline{sim}_z \right) ,$$

where Δ is the range of the behavioral variable and $\overline{sim}_z$ is the average of the observed similarity scores $\frac{\Delta - |z_i - z_j|}{\Delta}$.

The non-cross-lagged statistic related to the statistics above is defined as:

$$s_{sim}^{*} = \sum_{ij} x_{ij}(t_2) \left(\frac{\Delta - |z_i(t_2) - z_j(t_2)|}{\Delta} - \overline{sim}_z \right) \quad .$$

and used here to compare the relative efficiency of the GMoM to the MoM.

4 Results

The root mean squared errors of the estimators were compared using a one-sided Wilcoxon-signed ranked test (alternative hypothesis: $RMSE_{GMoM} < RMSE_{MoM}$) to evaluate the relative efficiency of the GMoM to the MoM.

Table 1 shows the results for two models differing in the value of the rate parameters, and consequently in the mean percentages of simultaneous changes (25% and 45% for lower and higher rates, respectively). The comparison suggests that the GMoM outperforms the MoM and the gain in efficiency mainly affects the parameters related to the effects for those a non-cross-lagged statistics was considered. The gain is due to both the smaller deviations and the higher efficiency of the GMoM estimates, as illustrated in Figure 1, which contrasts the absolute deviations and the estimated standard errors of the GMoM (y-axis) and the MoM (x-axis) estimates for the parameters of the behavior similarity and the average similarity effects.

More generally, the set of simulations we performed indicates that the gain in efficiency is negligible when i) the proportion of simultaneous changes is rather low, and ii) the non-cross-lagged statistics are non-informative, i.e.,

		MoM		GMoM		
	θ	Est.	RMSE	Est.	RMSE	Sig.
Selection						
Rate	3.50	3.40	0.84	3.41	0.83	
Density	-2.20	-2.34	0.80	-2.31	0.74	*
Reciprocity	1.80	1.85	0.35	1.85	0.35	
Transitive triplets	0.25	0.21	0.10	0.22	0.10	
3-cycles	-0.27	-0.24	0.21	-0.24	0.20	
Transitive Ties	0.67	0.71	0.34	0.72	0.34	
Outdegree popularity (sqrt)	-0.60	-0.66	0.44	-0.65	0.42	
Covariate-related popularity	0.19	0.23	0.51	0.22	0.48	
Covariate-related activity	0.48	0.59	0.39	0.58	0.37	
Same covariate	0.61	0.73	0.36	0.72	0.34	
Behavior similarity	3.00	3.40	1.83	3.27	1.37	**
Influence						
Rate	1.50	1.55	0.92	1.54	0.91	
Linear	0.05	0.06	0.28	0.06	0.28	
Quadratic	0.10	0.08	0.16	0.08	0.16	
Average similarity	6.00	5.69	3.28	5.65	3.01	**
Selection						
Rate	7.00	6.66	1.94	6.70	1.81	**
Density	-2.20	-2.27	0.82	-2.24	0.67	**
Reciprocity	1.80	1.84	0.35	1.84	0.32	
Transitive triplets	0.25	0.22	0.08	0.22	0.08	
3-cycles	-0.27	-0.23	0.18	-0.22	0.17	
Transitive Ties	0.67	0.71	0.31	0.71	0.30	
Outdegree popularity (sqrt)	-0.60	-0.71	0.46	-0.70	0.40	
Covariate-related popularity	0.19	0.31	0.53	0.29	0.44	**
Covariate-related activity	0.48	0.60	0.42	0.58	0.33	**
Same covariate	0.61	0.71	0.35	0.70	0.28	
Behavior similarity	3.00	3.52	2.49	3.36	1.39	**
Influence						
Rate	2.00	2.00	1.28	2.00	1.24	*
Linear	0.05	0.07	0.25	0.06	0.25	
Quadratic	0.10	0.08	0.13	0.08	0.13	
Average similarity	6.00	5.34	3.36	5.41	3.00	**

TABLE 1. Simulation results, 2 waves for 50 actors: true value of the parameter (θ), mean parameter estimates (Est.), and root mean squared errors (RMSE). The last column (Sig.) reports the significance of the Wilcoxon signed rank test comparing the RMSE of the two estimators (* p-value < 0.05, **p-value < 0.01).

highly correlated with the cross-lagged statistics. In these cases, the MoM is preferred to the GMoM since the MoM algorithm is faster. Furthermore, in accordance with applications of the GMoM in economics, we observed that the algorithm based on the iterated two-step estimator is more prone to instability when the non-cross-lagged statistics are non-informative.

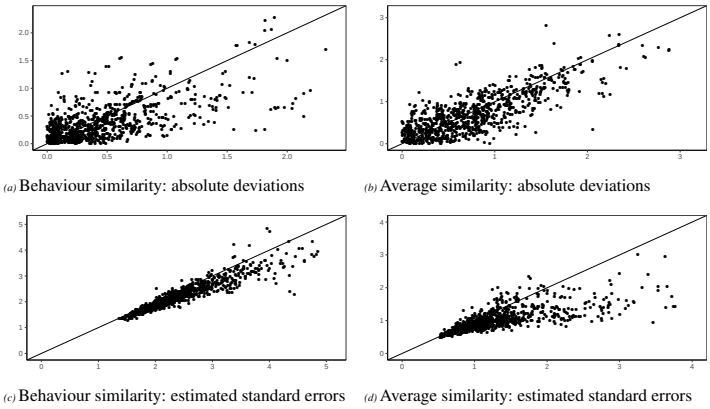
Acknowledgments: Part of this research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n. 319209.

References

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029 – 1054.

Holland, P. W. and Leinhardt, S. (1977). A Dynamic Model for Social Networks. *Journal of Mathematical Sociology*, **5(1)**, 5 - 20.

Robbins, H. and Monro, S. (1951). A stochastic approximation method.



(a) Behaviour similarity: absolute deviations (b) Average similarity: absolute deviations

(c) Behaviour similarity: estimated standard errors (d) Average similarity: estimated standard errors

FIGURE 1. Comparison between the absolute deviations and the estimated standard errors of the MoM (x-axis) and the GMoM (y-axis) for the behavior similarity and the average similarity effects for the network panel data simulated with the lower rates. The line in the plot represents the diagonal.

The Annals of Mathematical Statistics, **22**, 400 - 407.

Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, **31(1)**, 361 – 395.

Steglich, C. E. G., Snijders, T. A. B., and Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, **40(1)**, 329 – 393.

Stochastic block models for social network data: inferential developments

Francesco Bartolucci[1], Maria Francesca Marino[2], Silvia Pandolfi[1]

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

E-mail for correspondence: silvia.pandolfi@unipg.it

Abstract: Stochastic block models have known a flowering interest in the social network literature. They provide a tool for discovering communities and identifying clusters of individuals characterized by similar social behaviors. In this framework, full maximum likelihood estimates are not achievable due to the intractability of the likelihood function. For this reason, several approximate solutions are available in the literature. Here, we propose a new and more efficient approximate method for estimating model parameters, which has a hybrid nature in the sense that it exploits different features of existing methods. The proposal is illustrated by a Monte Carlo simulation study.

Keywords: Classification likelihood; Composite likelihood; EM algorithm; Random graphs; Variational inference.

1 Introduction

Stochastic Block Models (SBMs; Snijders and Nowicki, 1997, Nowicki and Snijders, 2001) represent an important tool of analysis in the social network literature when the focus is on discovering communities and clustering individuals with respect to their social behavior. According to the SBM specification, each individual in the network belongs to one of k distinct blocks, corresponding to the categories of a discrete latent variable, and the probability of observing a connection between two units only depends on their block memberships. Despite the simplicity of the model, Maximum Likelihood (ML) inference remains problematic due to the intractability of the likelihood function.

Some approximate solutions are available in the literature. These are mainly based on classification likelihood (Choi *et al.*, 2012), variational approximation

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

(Daudin *et al.*, 2008), or on a composite likelihood approach based on all triples in the network (Ambroise and Matias, 2012). In this work, we propose a new and more efficient approximate method for estimating model parameters, which has a hybrid nature as it is based on a classification likelihood but has features in common also with full likelihood and with variational and composite likelihood inference. We illustrate the potential of the proposed approach by an intensive simulation study.

The reminder of this paper is structured as follows. The SBMs are introduced in the following section, together with ML estimation of the model parameters and the alternative approximate inferential solutions. In Section 3 we illustrate the proposed hybrid estimation method, whereas in Section 4 we describe the main results of the simulation study.

2 Stochastic block models

In this section we briefly outline the main assumptions of the SBMs and we introduce the alternative approaches for parameter estimation.

2.1 Model assumptions

Let $\mathbf{Y}$ denote a socio-matrix involving n individuals/nodes and whose generic element, Y_{ij}, is equal to 1 if there is a tie between i and j and is equal to 0 otherwise, with $i, j = 1, \ldots, n$. We focus on *undirected* networks where $Y_{ij} = Y_{ji}$, $i \neq j$, with probability 1, although the extension to the directed case is straightforward. Moreover, self cycles are excluded by design.

In order to analyze social interactions existing between the nodes in the network, SBMs assume the existence of k unobserved blocks, which are described by the node-specific latent variables U_i, $i = 1, \ldots, n$, having discrete support points $1, \ldots, k$ and corresponding mass probabilities $p(U_i = u) = \pi_u$, $u = 1, \ldots, k$. Also, SBMs postulate a *local independence assumption* between nodes: conditional on the latent variables U_i and U_j, responses Y_{ij} are assumed to be independent Bernoulli random variables with success probability $\psi_{u_i u_j} = p(Y_{ij} = 1 | U_i = u_i, U_j = u_j)$. Therefore the response variables only depend on the block memberships of individuals i and j.

Despite the simplicity of the above assumptions, ML inference for this class of models remains problematic. This is because the manifest distribution of the entire network $\mathbf{Y}$ is obtained by marginalizing out all latent variables and this marginalization is computationally tricky. In particular, we have

$$p(\mathbf{Y}) = \sum_{\mathbf{u}} p(\mathbf{Y} \mid \mathbf{u}) p(\mathbf{u}),$$

where $\mathbf{u}$ denotes a realization of the random vector $\mathbf{U} = (U_1, \ldots, U_n)'$, and

$$\begin{aligned} p(\mathbf{Y} \mid \mathbf{u}) &= \prod_{i<n} \prod_{j>i} p(y_{ij} \mid U_i = u_i, U_j = u_j), \\ p(\mathbf{u}) &= \prod_{i<n} \pi_{u_i}. \end{aligned}$$

Therefore, to estimate the model parameters collected in $\boldsymbol{\theta}$, approximate methods that are alternative to full ML are needed, even dealing with small networks.

2.2 Maximum likelihood approach

With very small networks, the log-likelihood $\ell(\boldsymbol{\theta}) = \log p(\boldsymbol{Y})$ may be maximized via the Expectation-Maximization (EM) algorithm, which is based on the following complete data log-likelihood function related to the classification log-likelihood:

$$\ell^*_{cl}(\mathbf{u}, \boldsymbol{\theta}) = \sum_{i<n} \sum_{j>i} \log \left[p(y_{ij} \mid U_i = u_i, U_j = u_j)\, \pi_{u_i} \pi_{u_j} \right].$$

The EM algorithm maximizes $\ell(\boldsymbol{\theta})$ by alternating two steps until convergence. The **E-step** (Expectation) consists in computing the conditional expected value of $\ell^*_{cl}(\mathbf{u}, \boldsymbol{\theta})$ given the observed data and the current value of the parameters. At the **M-step** (Maximization) we update the parameter vector $\boldsymbol{\theta}$ by maximizing the expected value previously computed.

2.3 Available approximate estimation methods

Among the available approximate methods, it is important to mention the classification likelihood approach (Choi *et al.*, 2012). In this framework, the realizations of the random vector $\mathbf{U}$ are considered as fixed parameters to be estimated. Two steps are alternated until convergence: **C-step** (Classification), which consists in maximizing the classification likelihood

$$\ell_{cl}(\mathbf{u}, \boldsymbol{\theta}) = \sum_{i<n} \sum_{j>i} \log p(y_{ij} \mid U_i = u_i, U_j = u_j)$$

with respect to $\mathbf{u}$ by iteratively moving nodes from one block to another as in a k-means algorithm; **M-step** (Maximization), in which we update $\boldsymbol{\theta}$ by maximizing $\ell_{cl}(\mathbf{u}, \boldsymbol{\theta})$ with $\mathbf{u}$ being fixed at the previous value.

An alternative estimation method is based on a variational approximation of the model likelihood (Daudin *et al.*, 2008), in which parameter estimates are derived by maximizing a lower bound of the log-likelihood function. This method is computationally fast but may lead to non-optimal estimators and may suffer from local maxima solutions that strongly depend on the starting values of the estimation algorithm. Moreover, it may be quite slow to converge.

Finally, it is worth mentioning the composite likelihood approach introduced by Ambroise and Matias (2012), see also Bartolucci et al. (2015), in which parameter estimates are derived by maximizing a composite likelihood function based on the probability of all triples in the network. This approach leads to an estimator with properties similar to those of the standard ML estimator but it may suffer from identifiability problems. Even in this case, the algorithm may be slow to converge.

3 Proposed hybrid ML inference

The proposed hybrid ML method lies in between full ML and classification likelihood inference discussed in the previous section. More in detail, this method is based on the following hybrid log-likelihood function:

$$\ell_{hyb}(\tilde{\mathbf{u}}, \boldsymbol{\theta}) = \sum_{i<n} \log \sum_{u_i} p\left(\mathbf{y}_i \mid U_i = u_i, \mathbf{U}_{(-i)} = \tilde{\mathbf{u}}_{(-i)}\right) \pi_{u_i},$$

where $\mathbf{y}_i$ is the vector of all observed responses y_{ij}, $j \neq i$, for unit i, $\tilde{\mathbf{u}}$ is a realization of $\mathbf{U}$ considered as a vector of fixed discrete parameters in the set $\{1, \ldots, k\}$, and $\tilde{\mathbf{u}}_{(-i)}$ is the corresponding subvector without the i-th element. Moreover, $p(\mathbf{y}_i \mid U_i = u_i, \mathbf{U}_{(-i)} = \tilde{\mathbf{u}}_{(-i)})$ denotes the conditional distribution of $\mathbf{y}_i$ given the underlying latent variables.

The proposed method is based on alternating three steps until convergence. The **C-step** (Classification) consists in maximizing $\ell_{hyb}(\tilde{\mathbf{u}}, \boldsymbol{\theta})$ with respect to $\tilde{\mathbf{u}}$ by iteratively moving nodes from one block to another as in a k-means algorithm. At the **E-step** (Expectation) we compute the conditional expected value of the complete data log-likelihood function corresponding to $\ell_{hyb}(\tilde{\mathbf{u}}, \boldsymbol{\theta})$, and denoted by $\ell^*_{hyb}(\tilde{\mathbf{u}}, \boldsymbol{\theta})$, given the current value of $\boldsymbol{\theta}$ and $\tilde{\mathbf{u}}$. At the **M-step** (Maximization), $\boldsymbol{\theta}$ is updated by maximizing the expected value of $\ell^*_{hyb}(\tilde{\mathbf{u}}, \boldsymbol{\theta})$.

The inferential method we propose is related to the classification likelihood method because we explicitly consider the partition defined by the latent variables as depending on parameters to be estimated. Moreover, it has features in common with a composite likelihood method based on the response configuration of each node i. Finally, it relies on an optimization algorithm with structure and numerical complexity similar to that of the variational approach, while being typically faster to converge.

4 Simulation study

To assess the performance of the proposed approach, we carried out a simulation study in which this approach is compared with the available alternatives, that is, the k-means algorithm, the variational, the composite, and the classification likelihood approaches. We considered $B = 100$ samples generated from an SBM with varying number of groups ($k = 2, 3, 4, 5$) and equal proportions ($\pi_u = 1/k$, $u = 1, \ldots, k$), varying sample sizes ($n = 20, 50, 100$), and varying intra-group (ψ_{in}) and inter-group (ψ_{out}) connectivity parameters. In particular, we assumed three different models: *high intra-group connectivity* (M1, with $\psi_{\text{in}} = 0.30$ and $\psi_{\text{out}} = 0.03$); *high inter-group connectivity* (M2, with $\psi_{\text{in}} = 0.03$ and $\psi_{\text{out}} = 0.30$); *no structure* (M3, with $\psi_{\text{in}} = 0.55$ and $\psi_{\text{out}} = 0.45$). Then, we set $\psi_{u_i u_j} = \psi_{\text{in}} \times \alpha$ if $u_i = u_j$ and $\psi_{u_i u_j} = \psi_{\text{out}}$ if $u_i \neq u_j$ with $\alpha \sim \text{Unif}(0.5, 1.5)$ for all models.

The agreement between the estimated and the true latent structure is evaluated in terms of number of correctly classified nodes (*#ccn*) with respect to the true partition.

With reference to the most sensible scenarios, Figure 1 reports the boxplots of the values of *#ccn* obtained in the $B = 100$ samples by the alternative inferential approaches under comparison.

From the figure we observe that, for all methods under comparison, the quality of clustering worsens as k increases, due to the higher uncertainty on the latent structure, while it improves when the community structure in the data is more evident, that is, when moving from model M3 to M1. Moreover, the proposed approach seems to recover better the true clustering of the nodes in all simulation settings we considered, especially when dealing with more complex model structures. Lastly, the hybrid ML algorithm we propose generally requires a lower computational time to reach the converge and seems to be less sensitive to the adopted starting rule. In this respect, it represents an appealing alternative to the available approximate ML methods for parameter estimation in the SBM framework.

Further developments will be devoted to investigate the theoretical properties of the proposed method, with the aim of extending the approach also to the longitudinal social network framework.

References

Ambroise, C. and Matias, C. (2012). New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B,* **74**, 3 – 5.

Bartolucci, F., Marino, M.F., and Pandolfi, S. (2015). Composite likelihood inference for hidden Markov models for dynamic networks. *MPRA Paper No. 67242*.

Choi, D.S., Patrick, J.W., and Airoldi, E.M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*, **99**, 273 – 284.

Daudin, J.J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, **18**, 173 – 183.

Nowicki, K. and Snijders, T.A.B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, **96**, 1077 – 1087.

Snijders, T.A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, **14**, 75 – 100.

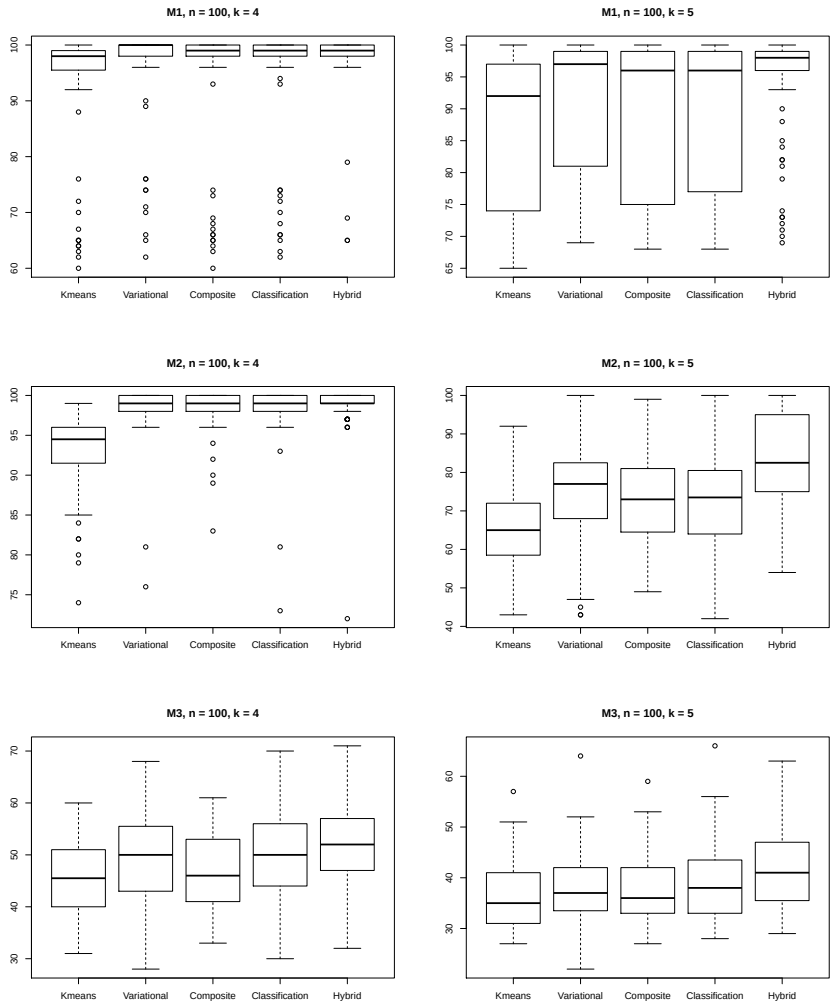


FIGURE 1. Number of correctly classified nodes (#*ccn*) under models M1, M2, and M3.

A sequentially coupled non-homogeneous dynamic Bayesian network model with segment-specific coupling strengths

Mahdi Shafiee Kamalabad[1], Marco Grzegorczyk[1]

[1] Johann Bernoulli Institute (JBI), Rijksuniversiteit Groningen, The Netherlands

E-mail for correspondence: m.shafiee.kamalabad@rug.nl

Abstract: We propose a generalised version of a non-homogeneous dynamic Bayesian network (NH-DBN) model with sequentially coupled interaction parameters. Unlike the earlier model, our new model does not assume that the segment-specific interaction parameters are coupled with the same coupling strength. Our new model introduces segment-specific coupling strengths, and we show that this generalization can lead to improved network reconstruction accuracies.

Keywords: Non-homogenous dynamic Bayesian networks; sequential-information coupling; segment-specific coupling strengths, RJMCMC

1 Introduction

In the field of systems biology non-homogeneous Dynamic Bayesian Networks (DBNs) have become a popular model class for learning regulatory networks from gene expression data. In many applications the intensities of the regulatory processes (i.e. the network parameters) undergo temporal changes. The conventional homogeneous DBN models, which assume that the network parameters stay constant over time, therefore often lead to biased results and erroneous conclusions. Non-Homogeneous DBNs (NH-DBNs) combine homogeneous DBNs with multiple change point processes. Loosely speaking, a set of changepoints divides the temporal data into disjunct segments and the data points within each segment are modelled by separate DBNs. To allow for some information sharing among segments it is often assumed that the same network structure applies to all segments and that only the network parameters are subject to temporal changes. Often the number and locations of those changepoints are unknown so that they have to be inferred together with the network structure and the segment-specific network parameters.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

The standard NH-DBN models assume a priori that the segment-specific network parameters are independent, and thus, they do not allow for any information sharing: the network parameters have to be learned independently in each segment and uninformative prior distributions are imposed.

A NH-DBN model with sequentially coupled network parameters was proposed in Grzegorczyk and Husmeier (2012). The key idea is to take into account that the network parameters often evolve gradually in time and that they are often similar for neighbouring segments. Grzegorczyk and Husmeier (2012) therefore proposed to use the posterior expectation of the network parameters of a segment h as prior expectation for the next segment $h+1$. For each segment $h > 1$ this yields an informative network parameter prior, which is built from the Bayesian inference result obtained for the preceding segment $h-1$. In the model of Grzegorczyk and Husmeier (2012) there is one single coupling parameter for each network node and all node-specific network parameters are coupled with the same coupling strength among all segments. A low (high) coupling parameter leads to peaked (vague) network parameter priors for all segments, where each segment-specific prior is centred around the posterior mean of its preceding segment. This is a potential bottleneck of the approach, since the similarity of the segment-specific network parameters can vary over time. There might be changepoints where the network parameters change only slightly, so that an informative (peaked) prior should be used, as well as changepoints with drastic network parameter changes, where an uninformative (vague) prior would be more appropriate. To address this shortcoming we propose a novel generalised sequentially coupled NH-DBN model with segment-specific coupling parameters. In the new model the coupling strengths δ_h $(h > 1)$ between two neighbouring segments $h-1$ and h are segment-specific.

Our new 'generalised' sequentially coupled NH-DBN model can be thought of as a continuous version of the 'partially sequentially coupled' NH-DBN model, which we proposed one year ago at IWSM 2016; see Shafiee and Grzegorczyk (2016). In our earlier work we introduced discrete binary indicator variables δ_h which indicate for each segment h whether it is coupled to the preceding one ($\delta_h = 1$; peaked prior around posterior of preceding segment) or uncoupled from the preceding one ($\delta_h = 0$; vague uninformative priors around 0). The new model replaces the segment-specific binary indicator variables $\delta_h \in \{0, 1\}$ (uncoupled vs. coupled) by segment-specific continuous variables $\delta_h \in \mathbb{R}^+$, and thus models the coupling strengths between all neighbouring pairs of segments $(h-1, h)$ continuously.

2 Methodological details

Let y be the response variable in a segment-wise linear Bayesian regression model. Given a set of k covariates, $\pi = \{X_1, \ldots, X_k\}$, we assume that the data can be divided into H disjunct segments with segment-specific regression coefficient vectors $\mathbf{w}_h$. Furthermore, let $\mathbf{y}_h$ and $\mathbf{X}_{\pi,h}$ denote the response vector and the design matrix for segment h, where $\mathbf{X}_{\pi,h}$ includes a first column of 1's for the intercept. We then have for $h = 1, \ldots, H$

$$\mathbf{y}_h \sim \mathcal{N}(\mathbf{X}_{\pi,h}\mathbf{w}_h, \sigma^2\mathbf{I}) \tag{1}$$

where σ^2 is the noise variance with $\sigma^{-2} \sim GAM(\nu, \nu)$ and $\nu = 0.01$. While the standard uncoupled NH-DBN uses independent Gaussian priors for all H regression coefficient vectors, $P(\mathbf{w}_h) \sim \mathcal{N}(\mathbf{0}, \delta\sigma^2\mathbf{I})$, the sequentially coupled model from Grzegorczyk and Husmeier (2012) uses the prior

$$P(\mathbf{w}_h) = \begin{cases} \mathcal{N}(\mathbf{0}, \delta\sigma^2\mathbf{I}), & h = 1 \\ \mathcal{N}(\mathbf{m}_{h-1}, \lambda\sigma^2\mathbf{I}) & h > 1 \end{cases} \tag{2}$$

where

$$\mathbf{m}_h = \left(\lambda^{-1}\mathbf{I} + \mathbf{X}_{\pi,h}^T\mathbf{X}_{\pi,h}\right)^{-1}\left(\lambda^{-1}\mathbf{m}_{h-1} + \mathbf{X}_{\pi,h}^T\mathbf{y}_h\right) \quad (h \geq 1)$$

The prior of the first segment has a special parameter δ which is different from the coupling parameter λ, as the coefficients of the first segment $h = 1$ cannot be coupled to a preceding segment and thus must get the uninformative mean vector $\mathbf{m}_0 = \mathbf{0}$. Only for $h > 1$ the prior mean is the posterior mean of the preceding segment $E[\mathbf{w}_{h-1}|\mathbf{y}_{h-1}, \mathbf{X}_{\pi,h-1}, \lambda_{h-1}] = \mathbf{m}_{h-1}$, where the noise variance σ^2 has been marginalized out. In our new generalised sequentially coupled model we replace equation (2) by:

$$P(\mathbf{w}_h) = \mathcal{N}(\mathbf{m}_{h-1}, \lambda_h\sigma^2\mathbf{I}) \tag{3}$$

where $\mathbf{m}_0 = \mathbf{0}$ and λ_1 refers to the δ parameter in equation (2). The only difference between equations (2) and (3) is then that the single coupling parameter λ for $h > 1$ has been replaced by segment-specific coupling parameter $\lambda_2, \ldots, \lambda_H$. Our generalised model is identical to the original sequentially coupled model for $H \leq 2$ and more flexible than the original model for $H > 2$. The $\mathbf{w}_h$ posterior means of our model are

$$\mathbf{m}_h = \left(\lambda_h^{-1}\mathbf{I} + \mathbf{X}_{\pi,h}^T\mathbf{X}_{\pi,h}\right)^{-1}\left(\lambda_h^{-1}\mathbf{m}_{h-1} + \mathbf{X}_{\pi,h}^T\mathbf{y}_h\right) \quad (h \geq 1)$$

We follow Grzegorczyk and Husmeier (2012) and use the hyperpriors

$$\delta^{-1} \sim GAM(\alpha_1, \beta_1),\ \lambda_1^{-1} \sim GAM(\alpha_1, \beta_1),\ \lambda_h^{-1} \sim GAM(\alpha_h, \beta_h)\ \ (h > 2)$$

with $\alpha_1 = 2$, $\beta_1 = 0.2$, $\alpha_h = 3$, and $\beta_h = 3$. The segment-specific regression coefficient vectors, $\mathbf{w}_h$, the noise variance, σ^2, and the segment-specific coupling parameters, λ_h, can then be sampled from their full conditional distributions (Gibbs sampling steps)

$$\mathbf{w}_h \sim N(\mathbf{m}_h, \sigma^2(\lambda_g\mathbf{I} + \mathbf{X}_{\pi,h}^T\mathbf{X}_{\pi,h})^{-1})$$

$$\lambda_h^{-1} \sim Gam\left(\alpha + \frac{k+1}{2}, \beta + \frac{1}{2}\frac{1}{\sigma^2}[\mathbf{w}_h - \mathbf{m}_{h-1}]^T[\mathbf{w}_h - \mathbf{m}_{h-1}]\right)$$

$$\sigma^{-2} \sim Gam\left(0.01 + \frac{T}{2}, 0.01 + \frac{1}{2}\sum_{h=1}^{H}(\mathbf{y}_h - \tilde{\mathbf{m}}_h)^T\tilde{\mathbf{\Sigma}}_h^{-1}(\mathbf{y}_h - \tilde{\mathbf{m}}_h)\right)$$

where $\tilde{\mathbf{m}}_h = \mathbf{X}_{\pi,h}\mathbf{m}_{(h-1)}$, $\tilde{\mathbf{\Sigma}}_h = \mathbf{I} + \lambda_h\mathbf{X}_{\pi,h}\mathbf{X}_{\pi,h}^T$, and T is the total number of data points. i.e. the length of the response vector $\mathbf{y} := (\mathbf{y}_1^T, \ldots, \mathbf{y}_H^T)^T$.

The marginal likelihood with all parameters, except for the λ_h hyperparameters, integrated out can then be given in closed form

$$p(\mathbf{y}_h|\pi,\lambda_h) = \frac{\Gamma(\frac{T_h+\nu}{2})}{\Gamma(\frac{\nu}{2})} \cdot \frac{\nu^{\nu/2}}{\pi^{T_h/2}\det(\tilde{\mathbf{\Sigma}}_h)^{1/2}} \cdot (\nu+\Delta^2)^{-\frac{1}{2}(T+\nu)}$$

where $\Delta^2 = (\mathbf{y}_h - \tilde{\mathbf{m}}_h)^T \tilde{\mathbf{\Sigma}}_h^{-1}(\mathbf{y}_h - \tilde{\mathbf{m}}_h)$ is the squared Mahalanobis distance, and T_h is the length of the vector $\mathbf{y}_h$.

Inferring the covariate sets and the data segmentations:
The objective is to sample the covariate sets, π, and the data segmentations from the posterior distribution. For that we use the Metropolis Hastings sampling moves, described in Grzegorczyk and Husmeier (2012). The covariate sets π are modified by adding, deleting or substituting single variables from π, where a priori every set π with up to three covariates is assumed to be equally likely while the prior probability of sets with more than three covariates is set to zero ('fan-in equal to 3'). The adaptation of those moves to our model is straightforward. For the segmentations we again follow Grzegorczyk and Husmeier (2012) and employ a changepoint process where the distance between neighbouring changepoints is assumed to follow a negative Binomial distribution with parameters $p \in [0,1]$ and $r = 1$. In our study we implement 9 different values: $p = 0.02, 0.025, 0.03, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3$. The changepoints are inferred via changepoint birth, death and re-allocation moves, where unlike in Grzegorczyk and Husmeier (2012) and Shafiee and Grzegorczyk (2016), those moves here become Reversible Jump Markov Chain Monte Carlo (RJMCMC) moves, as the number of *continuous* $\delta_h \in \mathbb{R}^+$ parameters now varies with the number of segments H. We implement the RJMCMC variants of the three changepoint moves in the standard way by re-sampling the involved δ_h parameters from their Inverse-Gamma prior distributions. The determinant of the Jacobian transition matrix is then equal to 1.

Networks: For a domain with n nodes we apply the regression model to each node y^i $(i = 1,\ldots,n)$ separately. The potential regulator sets of y^i are all subsets $\pi_i \subset \{y^1,\ldots,y^{i-1},y^{i+1},\ldots,y^n\}$ of the other nodes. As the interactions are subject to a time delay, the segment-specific design matrices $\mathbf{X}_h^i$ are built from the values of the covariates at the preceding time points. Merging the n parent sets gives an interaction network $\mathcal{G} := \{\pi_1,\ldots,\pi_n\}$. There is an edge from node y^j to node y^i in $\mathcal{G}$ if (and only if) $y^j \in \pi_i$.

Network reconstruction accuracy: With the MCMC sampling scheme, we can generate samples of networks $\mathcal{G}^r = \{\pi_1^r,\ldots,\pi_n^r\}_{r=1,\ldots,R}$, and we average across those networks to obtain for each individual edge $j \to i$ $(j,i \in \{1,\ldots n\} : j \neq i)$ a marginal posterior probability $\hat{e}_{j,i} = \frac{1}{R}\sum_{r=1}^{R} I_{j\to i}(\mathcal{G}^r)$, where $I_{j\to i}(\mathcal{G}^r) = 1$ if $j \in \pi_i^r$, and $I_{j\to i}(\mathcal{G}^r) = 0$ otherwise. When the true edges are known, $e_{i,j} \in \{0,1\}$, the network reconstruction accuracies can be quantified in terms of areas under the precision-recall curves (AUPRC).

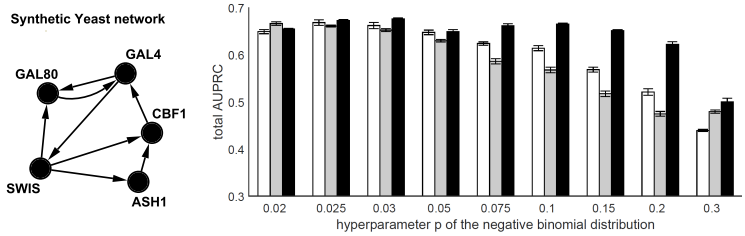


FIGURE 1. **Network reconstruction accuracy for S. cerevisiae (yeast).** AUPRC histograms for the three NH-DBN models. The white bars refer to the NH-DBN model with sequentially coupled parameters from Grzegorczyk and Husmeier (2012), the grey bars refer to the standard uncoupled NH-DBN model, and the black bars refer to the generalised model, proposed here. There is a histogram for each of 9 different hyperparameters of the negative Binomial prior on the distance between neighbouring changepoints. The heights of the bars are the average AUCPR values, averaged across 10 independent MCMC simulation results and the error bars indicate standard deviations.

3 Data

Cantone et al. (2009) synthetically designed a network of five genes in Saccharomyces cerevisiae (yeast) and measured expression levels of these genes in vivo with quantitative real-time PCR at 37 time points over 8 hours. In about the middle of this time period, they changed the environment by switching the carbon source from galactose to glucose. We pre-process those data as described in Grzegorczyk and Husmeier (2012), including a gene-specific standardization to mean 0. The goal of our empirical study in Section 4 is to infer the true network, shown in in Figure 1, from the data.

4 Yeast network reconstruction

We show that the novel generalised NH-DBN model with segment-specific coupling parameters reaches a higher network reconstruction accuracy than the standard uncoupled NH-DBN model and the original sequentially coupled NH-DBN model, proposed by Grzegorczyk and Husmeier (2012). To this end, we cross-compared the network reconstruction accuracies of those models on the yeast gene expression data set, described in Section 3. For all three models we ran MCMC simulations with 9 different parameters p of the negative Binomial prior (on the distances between changepoints), so as to compare the models for varying numbers of changepoints: Higher p values lead to more changepoints and thus make the information-coupling scheme more important. We ran each MCMC simulation for $R = 100,000$ iterations, and our convergence diagnostics suggested that this simulation length is sufficient for all three models. The results of our empirical study are summarized in Figure 1. It can be seen that the new model, overall, shows the best performance. It performs significantly better than the two competing models for high p parameters (i.e. when many changepoints are inferred). For low p parameters the performance between the two sequentially coupled model does not differ, since the two models do not differ if there is only up to one single changepoint per gene ($H \leq 2$ segments).

5 Conclusion

Our results suggest that the new NH-DBN model with segment-specific coupling strength parameters can reach a significantly better network reconstruction accuracy than the two competing NH-DBN models. The new model can be thought of as a continuous version of the partially sequentially coupled NH-DBN model, proposed in Shafiee and Grzegorczyk (2016). The focus of our future research will be on cross-comparing those two models systematically and on determining their relative merits and shortcomings.

References

Cantone I. et al. (2009) *A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches* Cell, 137, 172-181.

Grzegorczyk, M., and Husmeier, D. (2012). *A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology.* Statistical Applications in Genetics and Molecular Biology, 11(7), online article.

Shafiee Kamalabad, M. and Grzegorczyk, M. (2016). *A non-homogeneous dynamic Bayesian network model with partially sequentially coupled network parameters.* Proceedings of the International Workshop on Statistical Modelling (IWSM2016), vol. 1, 139-144, Rennes, France.

Multiple imputation for high-dimensional data using sequential penalized regression

Faisal Maqbool Zahid[1], Christian Heumann[1]

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

E-mail for correspondence: faisal-maqbool.zahid@stat.uni-muenchen.de

Abstract: Missing data occurs in almost every field of research. Multiple imputation (MI) is a statistical technique which has become increasingly popular to generate multiple data sets without missing values which can then be analyzed using the standard statistical methods for complete data.The sequential regression imputation, building one model for each variable with missing data, is a practical approach under MAR to get multiply imputed data sets. The existing algorithms and software tools either perform poorly or fail to respond for high-dimensional data structures. The question of the best strategy for multiple imputation for high-dimensional data is still not clearly answered in the literature. In this paper we are proposing an MI technique based on sequential penalized regression to address this issue. The proposed technique uses L1 and L2 penalties and thus gives a penalized version of the sequential regression approach with high-dimensional data. The performance of the proposed technique is examined in a simulation study with covariates from different distributions. Mean Squared Imputation Error (MSIE) is used to study the imputation performance of the proposed algorithm as compared to the other existing approaches.

Keywords: Missing data; Multiple imputation; Sequential regression imputation; Penalization; Conditional distribution; High-dimensional data.

1 Introduction

Missing data occurs in almost every field of research. Multiple imputation (MI) has become increasingly popular for imputing the missing data in recent years due to its flexibility. Multiple imputation replaces the missing value with more than one plausible values drawn from their predictive distributions conditional on the observed data. As a result, one gets multiple imputed data sets which can then be analyzed independently using standard methods for complete data and re-

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

sults from all imputed data sets are combined using Rubin's rule (Rubin (1987)). Harel and Zhou (2007) and Horton and Kleinman (2007) provide a nice review of theory and implementation of different MI approaches with a comparison of different software packages for MI. In case of high-dimensional data the number of variables p can be close to the or even greater than the sample size n. In such case the usual likelihood estimates in the imputation model may not converge or even do not exist for $p > n$ case. The use of regularization techniques becomes indispensable in such cases with high-dimensional data. There is still a gap in the literature that needs to be filled relating to the development of efficient MI algorithms for high-dimensional data. Although Zhao and Long (2016) proposed Bayesian lasso regression for multiple imputation with normally distributed data, they considered only one variable with missing data in their simulation setting. This approach is also computationally very expensive and even becomes practically infeasible with an increasing number of variables with missing data. In this paper we are proposing regularized regression for model selection and parameter estimation for the imputation model. Each variable with missing values is assumed to have a different distributional form and is imputed with its own imputation model using L1 penalty (for model selection) and if needed also L2 penalty (for fitting the model with selected covariates).

2 Regularization

Regularization methods that are derived from MLE are based on the penalized log-likelihood $l_p(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) - \frac{\lambda}{2} J(\boldsymbol{\beta})$.

Ridge Regression: Ridge regression uses the quadratic penalty $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} {\beta_j}^2$. It shrinks the parameter estimates towards zero and each other. The shrinkage parameter $\lambda \geq 0$ controls the amount of shrinkage. Ridge regression produces biased estimates as $\hat{\boldsymbol{\beta}}^{\boldsymbol{\star}} = (\mathbf{X}^T\mathbf{W}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{W}\ \mathbf{D}^{-1}[\mathbf{y} - \boldsymbol{\mu}]$, with $\text{cov}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\star}}) = \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\star}}) = (\mathbf{X}^T\mathbf{W}\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{X})\ (\mathbf{X}^T\mathbf{W}\mathbf{X} + \lambda\mathbf{I})^{-1}$. Here mean μ_i is related to the linear predictor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ as $\mu_i = h(\eta_i)$. $\mathbf{D}$ and $\mathbf{W}$ are $n \times n$ diagonal matrices. The ith diagonal element of $\mathbf{D}$ is $D_i = \partial h(\eta_i)/\partial \boldsymbol{\eta}$. The ith element of $\mathbf{W}$ is given as $w_i = \sigma_i^2 = \text{var}[h(\eta_i)]$ for $i = 1, \ldots, n$.

Lasso Regression: Lasso (least absolute shrinkage and selection operator) uses L1 penalty $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$. Lasso penalty set some coefficients exactly equal to 0 with large enough value of λ and hence provides a parsimonious model with $p >> n$. The lasso penalty will perform imputation model selection for high-dimensional data in our algorithm.

3 MI using regularization

Let $\mathbf{X}_{n\times p}$ be a data matrix consisting of p variables (complete or with missing values) corresponding to a set of n subjects. The variables in the data matrix can assume different distributional forms. Let all the variables in the data matrix be divided into two mutually exclusive groups as $\mathbf{X} = (\mathbf{X}^{\text{miss}}_{n\times q}, \mathbf{X}^{\text{comp}}_{n\times(p-q)})$ where $\mathbf{X}^{\text{miss}}$ contains q, out of p, variables with some missing values and $\mathbf{X}^{\text{comp}}_{n\times(p-q)}$ covers rest of the $p - q$ complete variables without any missing information. Here q may be

equal to p in which case $\mathbf{X} = \mathbf{X}^{\text{miss}}$. The proposed algorithm can be described as follows:

- **Step 1: Initialization**

Repeat Step 2 till convergence

- **Step 2: Convergence,** for $j = 1, 2, \ldots, q$
  a) Fit $\mathbf{x}_j^{\text{miss}} \sim \mathbf{X} = (\mathbf{X}^{\text{comp}}, \mathbf{X}_{-j}^{\text{miss}})$ using L1 penalty.
  b) compute $\hat{\boldsymbol{\beta}}$ from $\mathbf{x}_j^{\text{miss}} \sim \mathbf{X}^{\text{select}}$.
  c) update missing values of $\mathbf{x}_j^{\text{miss}}$ with $\hat{\boldsymbol{\mu}} = \mathbf{X}_{m_j \times p}^{\text{select}} \cdot \hat{\boldsymbol{\beta}}$.

Repeat step 3 for m times with $\mathbf{X}$ achieved at convergence.

- **Step 3: Imputation**, Repeat for $j = 1, 2, \ldots, q$,
  a) perform steps 2(a,b) to compute $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}})$.
  b) compute $\hat{\boldsymbol{\beta}}_{\text{new}}$ = a random sample from MVN$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}}))$ to impute the missing values with

$$\begin{cases} \mathbf{X}_{m_j \times p}^{\text{select}} \cdot \hat{\boldsymbol{\beta}}_{\text{new}} + \text{noise} & \text{if } \mathbf{x}_j^{\text{miss}} \text{ is normal,} \\ \text{rbinom}(h(\mathbf{X}_{m_j \times p}^{\text{select}} \cdot \hat{\boldsymbol{\beta}}_{\text{new}})) & \text{if } \mathbf{x}_j^{\text{miss}} \text{ is binary, and} \\ \text{rpois}(h(\mathbf{X}_{m_j \times p}^{\text{select}} \cdot \hat{\boldsymbol{\beta}}_{\text{new}})) & \text{if } \mathbf{x}_j^{\text{miss}} \text{ is a count variable.} \end{cases}$$

  c) update $\mathbf{X}_{n \times q}^{\text{miss}}$.

4 Simulation study

A simulation study was performed with $S = 100$ data sets using sample size $n = 100$ and $p = 200$. The covariates are drawn from a p-dimensional multivariate normal distribution with mean $\mathbf{0}$ and $\rho^{|j-k|}$ (with $\rho = 0.8$) covariance among $\mathbf{x}_j$ and $\mathbf{x}_k$. The MAR (Missing at

TABLE 1. Overall MSIE for imputed values with proposed algorithm, mice and VIM. The MIlasso1 represents the results when step 3 of algorithm is also used for the convergence. The numbers within brackets represent number of samples (out of 100) for which a particular algorithm didn't respond. The bold face values represent the best results for a particular approach.

miss %	MIlasso1	MIlasso2	mice	VIM
10	810.66	**776.01**	1370.29	1305.94
20	1642.06	**1568.48**	2415.59	2656.38
30	2488.91	**2374.67**	3227.52	4060.23
40	3347.97	**3203.18**	3966.40	5553.46
60	(14)5025.75	(14)**4848.28**	(50)5469.01	(20)8778.00

Random) mechanism was considered with $10\%, 20\%, 30\%, 40\%$, and 60% missing values in each of $q = 100$ covariates selected at random. One third variables were

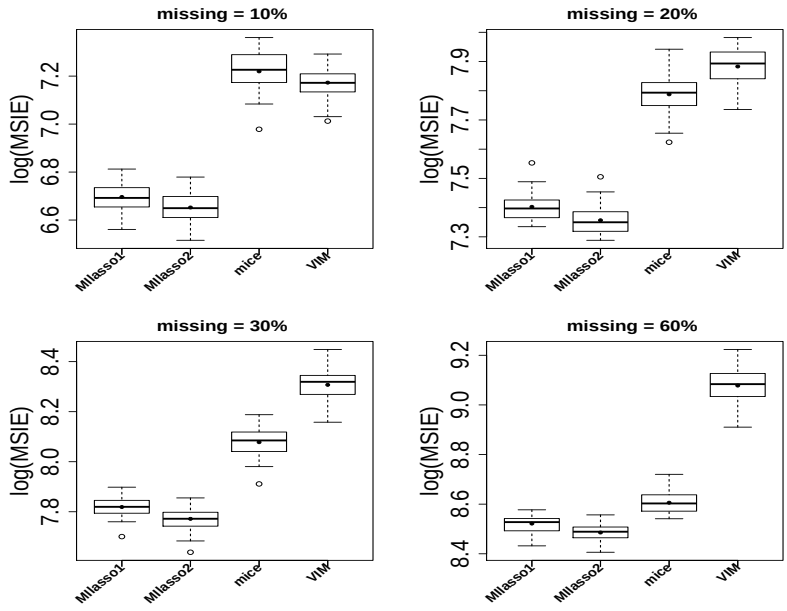


FIGURE 1. Simulation Study: Box plots for log(MSIE)

binary and one third were taken as count variables randomly. The performance of proposed algorithm is compared with existing software packages `mice`, `VIM` and `Amelia`. The Mean Squared Imputation Error (MSIE) is used for comparison and is given as: $\text{MSIE} = \frac{1}{S}\sum_S \left[\frac{1}{m}\sum_j \left\{\sum_{i=1}^{m_j}(x_{ij}^* - x_{ij})^2\right\}\right]$, where x_{ij}^* represents the ith imputed value for observed value x_{ij} of variable $\mathbf{x}_j$. The results are shown in Table 1. The values within brackets here show the number of samples for which a particular algorithm didn't work. The results of `Amelia` are not shown in the table because it did not respond in most cases because of its sensitivity to the correlation structure and high dimension. The results of log(MSIE) are also presented in the form of box plots in Figure 1. The MIlasso1 represents the results when step 3 of algorithm is also used for the convergence. The box plots show that our proposed algorithm performs better than its competitors. The MSIEs were also splitted according to the distribution of the missing variables to examine how well the different algorithms impute under different distributional forms. These results of such split, not shown here, also showed consistent performance in favor of the proposed algorithm.

References

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys* New York, NY: Wiley.

Horton, N. J., Kleinman, K. P. (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, **61**, 79 – 90.

Harel, O. and Zhou, X. (2007) Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, **26**, 3057 – 3077.

Zhao, Y. and Long, Q. (2016) Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research* , **25 (5)**, 2021 – 2035.

REML in nonlinear mixed-effects models with heavy-tailed distributions

José Clelto Barros Gomes[1], Cibele Maria Russo[2]

[1] Departamento de Estatística, Universidade Federal do Amazonas, Brazil
[2] Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil

E-mail for correspondence: clelto@ufam.edu.br

Abstract: This work proposes the estimation of nonlinear mixed-effects models for continuous data. We present a restricted maximum likelihood estimator (REML), based on the integration of the fixed effects and a stochastic simulation procedure. A comparison of REML estimator with maximum likelihood (ML) estimator is presented. Random effects and errors are assumed to follow normal, slash or Student-t distributions. A pharmacokinetic dataset are used to illustrate the methods.

Keywords: Nonlinear mixed models, restricted maximum likelihood, MCEM algorithm

1 Introduction

Nonlinear mixed-effects models (NLME) aim to model the nonlinear relationship between the response variable and the covariates through the parameters. The nonlinear relationship can occur either in fixed or random effects or only in one of them. Russo (2009), for example, worked with a nonlinear marginal model, where random effects are added linearly to the model. Mixed-effects models are usually proposed for problems with correlated data, such as longitudinal or repeated measures data, with applications in several areas such as epidemiological, pharmacokinetic industry, economics and agriculture. Several examples can be found in Pinheiro and Bates (2000). These models usually bring a great challenge in the parameters estimation of the fixed effects and variance components. Most authors use the ML estimate or the REML estimate. Unlike the ML estimates, the REML is preferred since it produces less biased estimates for variance components. It takes into account the degrees of freedom lost when estimating

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

the fixed effects parameters, which does not occur with the ML estimator. This is especially more evident in small samples, since for large samples the estimates do not differ much. The development of REML will have a formulation as proposed by Meza et al. (2007) and Monte Carlo Expectation - Maximization (MCEM) algorithm will be used to estimate the parameters.

2 The model formulation

The scale mixture normal (SMN) distributions have a stochastic representation

$$\boldsymbol{Y} = \boldsymbol{\mu} + \kappa(\boldsymbol{U})^{1/2}\boldsymbol{Z},$$

where $\boldsymbol{Z} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ is independent of the mixture variable $U \sim H(u;\nu)$ with H a cumulative distribution function that specify the SMN model, and ν is a scalar or vector parameter indexing the distribution of U, $\kappa(U)$ is the weight function. The NLME can be written in two stage. The first stage consists of n nonlinear regression models

$$\boldsymbol{y}_i = g(\boldsymbol{\varphi}_i, \boldsymbol{X}_i) + \boldsymbol{\epsilon}_i,\ i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{y}_i$ $(n_i \times 1)$ denotes the response vector for subject i, $\boldsymbol{X}_i$ is a matrix of explanatory variables, n is the number of subjects, n_i the number of observations of subject i and $\boldsymbol{\epsilon}_i$ the within-subject errors. In the second stage the fixed and random effects are included to the model through the vector $\boldsymbol{\varphi}_i$ $(n_i \times 1)$ that can be written linearly by

$$\boldsymbol{\varphi}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{b}_i, \tag{2}$$

where $\boldsymbol{\beta}$ $(p \times 1)$ denotes the vector of fixed effects, $\boldsymbol{b}_i$ $(q \times 1)$ denotes the vector of random effects and $\boldsymbol{A}_i$ $(q \times p)$ is a known design matrix . From (1) and (2) we assume

$$\begin{pmatrix} \boldsymbol{\epsilon}_i \\ \boldsymbol{b}_i \end{pmatrix} \sim \text{SMN}_{n_i+q}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{pmatrix}; H \right),$$

where $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_i$ are positive-definite dispersion matrices. Here we assume $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{\tau}) = \text{diag}(\boldsymbol{\tau})$ is a diagonal matrix and let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q)^\top$ be the elements and $\boldsymbol{\Sigma}_i = \sigma^2\boldsymbol{I}_{n_i}$, with $\sigma^2 > 0$ a scalar. For the sake of simplicity we use $g(\boldsymbol{\beta}, \boldsymbol{b}_i)$ to represent $g(\boldsymbol{\varphi}, \boldsymbol{X}_i) = g(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{b}_i, \boldsymbol{X}_i)$. We assume to errors and random effects the normal, Slash$_\nu$ and Student-t$_\nu$ distributions.

3 Estimation

Assuming a vague prior for the fixed effects $\boldsymbol{\beta}$, REML estimation to variance components, $\boldsymbol{\tau}$ and σ^2, can be obtained integrating out the fixed effects, besides integrate in $\boldsymbol{b}$, given by

$$\begin{aligned} f(\boldsymbol{y}; \boldsymbol{\tau}, \sigma^2) \quad &= \quad \prod_{i=1}^n \int_0^\infty \int_{\mathbb{R}^q} \int_{\mathbb{R}^p} \phi_{n_i}(\boldsymbol{y}_i | g(\boldsymbol{\beta}, \boldsymbol{b}_i), \kappa(u_i)\boldsymbol{\Sigma}_i) \\ &\quad \times \phi_q(\boldsymbol{b}_i | \boldsymbol{0}, \kappa(u_i)\boldsymbol{D})\, d\boldsymbol{b}_i\, d\boldsymbol{\beta}\, dH(u;\nu), \end{aligned}$$

where $\phi_n(\cdot|\boldsymbol{\mu}, \Sigma)$ denotes n-dimensional normal probability density function with parameters $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$. The complete data of the model are given by $\boldsymbol{y}_c = (\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{u})$,

and its likelihood function can be written as

$$\begin{aligned} L_c(\boldsymbol{\theta}^*|\boldsymbol{y}_c) &= \prod_{i=1}^n p(\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{u}_i|\boldsymbol{\theta}^*) \\ &= \prod_{i=1}^n f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{b}_i, u_i, \sigma^2) f(\boldsymbol{b}_i|u_i, \boldsymbol{\tau}) h(u_i|\nu), \end{aligned}$$

where $\boldsymbol{\theta}^* = (\boldsymbol{\tau}, \sigma^2)$. The complete likelihood function is given by

$$\begin{aligned} \ell_c(\boldsymbol{\theta}^*) &= \sum_{i=1}^n \ell(\boldsymbol{\theta}^*; \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, u_i) \\ &= \sum_{i=1}^n \{\log f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{b}_i, u_i, \sigma^2) + \log f(\boldsymbol{b}_i|u_i, \boldsymbol{\tau}) + \log h(u_i|\nu) + C\}, \end{aligned}$$

C assumes a constant value.

To estimate the parameters we use MCEM algorithm, which facilitates the inference and consists in two steps. In the E Step we obtain the expectation in Q-function

$$Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{*(t)}) = E[\ell_c(\boldsymbol{\theta}^*)|\boldsymbol{y}; \boldsymbol{\theta}^{*(t)}]$$

and in the M Step we want to find $\boldsymbol{\theta}^{*(t+1)}$ such that maximize $Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{*(t)})$,

$$Q(\boldsymbol{\theta}^{*(t+1)}|\boldsymbol{\theta}^{*(t)}) \geq Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{*(t)}),$$

the two steps are repeated till convergence. For more details of this algorithm see Wang (2007).

The Step E of the ith individual in the $(t+1)$th iteration can be written as

$$\begin{aligned} Q_i(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{*(t)}) &= E[\ell(\boldsymbol{\theta}^*; \boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, u_i)|\boldsymbol{y}_i, \boldsymbol{\theta}^{*(t)}] \\ &= \int_0^\infty \int_{\mathbb{R}^q} \int_{\mathbb{R}^p} \{\phi_{n_i}(\boldsymbol{y}_i|g(\boldsymbol{\beta}, \boldsymbol{b}_i), \kappa(u_i)\boldsymbol{\Sigma}_i) \\ &\quad \times \phi_q(\boldsymbol{b}_i|\boldsymbol{0}, \kappa(u_i)\boldsymbol{D}) f(\boldsymbol{\beta}, \boldsymbol{b}_i, u_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{*(t)})\} \, \mathrm{d}\boldsymbol{\beta} \, \mathrm{d}\boldsymbol{b}_i \, \mathrm{d}u_i. \end{aligned}$$

One can use the Gibbs sampler algorithm with Metropolis-Hastings steps to generate samples of $[\boldsymbol{\beta}, \boldsymbol{b}_i, u_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{*(t)}]$ by sampling of the full conditional distributions from $[u_i|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{y}_i, \boldsymbol{\theta}^{*(t)}]$, $[\boldsymbol{b}_i|\boldsymbol{\beta}, u_i, \boldsymbol{y}_i, \boldsymbol{\theta}^{*(t)}]$ and $[\boldsymbol{\beta}|\boldsymbol{b}_i, u_i, \boldsymbol{y}_i, \boldsymbol{\theta}^{*(t)}]$.

Assuming that $\{(\boldsymbol{\beta}^{(1)}, \boldsymbol{b}_i^{(1)}, u_i^{(1)}), \ldots, (\boldsymbol{\beta}^{(m_i)}, \boldsymbol{b}_i^{(m_i)}, u_i^{(m_i)})\}$ is a random sample of size m_i from $[\boldsymbol{\beta}, \boldsymbol{b}_i, u_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{*(t)}]$, the E Step at the $(t+1)$th iteration can be written as

$$\begin{aligned} Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{*(t)}) &= \sum_{i=1}^n Q_i(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{*(t)}) \\ &= \sum_{i=1}^n \left[\tfrac{1}{m_i} \sum_{j=1}^{m_i} \ell(\boldsymbol{\theta}^*; \boldsymbol{y}_i, \boldsymbol{\beta}^{(j)}, \boldsymbol{b}_i^{(j)}, u_i^{(j)}) \right] \\ &\propto \sum_{i=1}^n \sum_{j=1}^{m_i} \tfrac{1}{m_i} \left[-\tfrac{n_i}{2} \log(\sigma^2) - \tfrac{\kappa^{-1}(u_i^{(j)})}{2\sigma^2} \|\boldsymbol{y}_i - g(\boldsymbol{\beta}_i^{(j)}, \boldsymbol{b}_i^{(j)})\|^2 \right] \\ &\quad + \sum_{i=1}^n \sum_{j=1}^{m_i} \tfrac{1}{m_i} \left[-\tfrac{1}{2} \log(|\boldsymbol{D}|) - \tfrac{\kappa^{-1}(u_i^{(j)})}{2} \boldsymbol{b}_i^{(j)\top} \boldsymbol{D}^{-1} \boldsymbol{b}_i^{(j)} \right]. \end{aligned} \tag{3}$$

In M Step, the REML for the variance components, which are solutions from (3) are given by

$$\widehat{\sigma}^{2(t+1)} = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{1}{m_i} \left[\kappa^{-1}(u_i^{(j)}) \|\boldsymbol{y}_i - g(\boldsymbol{\beta}_i^{(j)}, \boldsymbol{b}_i^{(j)})\|^2 \right],$$

and

$$\widehat{\boldsymbol{D}}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{m_i}\left[\kappa^{-1}(u_i^{(j)})\text{diag}\left(\boldsymbol{b}_i^{(j)}\boldsymbol{b}_i^{(j)\top}\right)\right].$$

The fixed effects in the context of REML can be obtained by $\widehat{\boldsymbol{\beta}} = E(\boldsymbol{\beta}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^*)$.

The standard error estimates follow the expressions developed in Louis (1982), with the observed information matrix written as

$$-E\left\{\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{u})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right\}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^*} - Var\left\{\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{u})}{\partial\boldsymbol{\theta}}\right\}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^*}.$$

The expectation and variance are computed with respect to $f(\boldsymbol{\beta}, \boldsymbol{b}_i, u_i, |\boldsymbol{y}_i, \boldsymbol{\theta}^{*(t)})$.

4 Application

As an illustration, consider the study of the pharmacokinetics of the drug theophylline used in the treatment of asthma (Pinheiro e Bates, 2000). The drug was given orally to twelve patients, whose blood samples were collected in 11 times points after administration, within a maximum of 25 hours, and the serum concentration of the substance (in mg/L) was measured (Figure 1). The first-order compartment model is usually reparameterizated in terms of the logarithm of the clearance and constant rates. Thus, the model can be written as: $C_t = \frac{Dose\exp(lK_e+lK_a-lC_l)}{\exp(lK_a)-\exp(lK_e)}\{\exp[-\exp(lK_e)t] - \exp[-\exp(lK_a)t]\}$, where C_t denotes the concentration observed at the time t (mg/L), t is the time (in hours), $Dose$ is the dose, $lK_e = \log(K_e)$, $lK_a = \log(K_a)$ and $lC_l = \log(C_l)$. K_e is the constant elimination rate (1/hour), K_a is the constant absorption rate (1/hour) which describes how the drug is absorbed from the intestine into the bloodstream, C_l is the clearance rate (L/kg) representing the volume of blood from which the drug is eliminated per unit of time.

REML and ML estimates for variance components parameters are presented in Table 1. Note that the estimates are close, and we highlight the REMLs of the variance components, which are generally larger than the ML estimates. A based-REML fixed effects estimates are also shown as well as the ML estimator.

5 Discussion and remarks

REML and ML estimates to variance components in NLME were obtained supposing to the within-subject errors and random effects are normal, Slash and Student-t distributed. Here, we set four degrees of freedom for Student-t and Slash. We run the MCEM algorithm for an initial 11,000 iteration, in the end of the process we discarded half of them and thinning it in 100 to obtain the convergence. The process to REML was slower than ML to converge. In future research we will investigate REML estimator by Monte Carlo simulation.

Acknowledgments: The first author thanks to Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM, Brazil. Second author thanks to Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, Brazil.

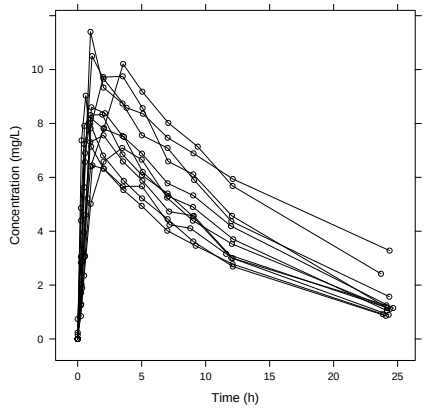


FIGURE 1. Serum concentration (mg/L) in the times of the drug administration to twelve patients.

TABLE 1. REML and ML estimates (standard errors) for theophylline serum concentration data

Par	Normal		Student_4		Slash_4	
	ML	REML	ML	REML	ML	REML
lK_e	-2.460	-2.458	-2.434	-2.460	-2.443	-2.457
	(0.050)	(0.057)	(0.040)	(0.047)	(0.047)	(0.053)
lK_a	0.401	0.493	0.342	0.547	0.414	0.473
	(0.056)	(0.058)	(0.046)	(0.049)	(0.052)	(0.055)
lC_l	-3.206	-3.229	-3.137	-3.230	-3.195	-3.222
	(0.035)	(0.039)	(0.028)	(0.032)	(0.032)	(0.036)
σ^2	0.501	0.504	0.296	0.313	0.322	0.327
	(0.063)	(0.064)	(0.040)	(0.043)	(0.043)	(0.043)
τ_1	0.001	0.001	0.001	0.001	0.000	0.001
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
τ_2	0.441	0.490	0.509	0.557	0.368	0.416
	(0.007)	(0.007)	(0.005)	(0.006)	(0.005)	(0.005)
τ_3	0.029	0.031	0.023	0.027	0.020	0.023
	(0.007)	(0.007)	(0.006)	(0.006)	(0.005)	(0.005)

References

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*. Series B (Methodological), 226 – 233.

Meza, C., Jaffrézic, F., and Foulley, J.L. (2007). Reml estimation of variance parameters in nonlinear mixed effects models using the saem algorithm. *Biometrical Journal*, **49**, 876 – 888.

Pinheiro, J.C. and Bates, D.M. (2000). Mixed-Effects Models in S and S-Plus. New York: Springer.

Russo, C.M., Paula, G.A., and Aoki, R. (2009). Influence diagnostics in nonlinear mixed-effects elliptical models. *Computational Statistics and Data Analysis*, **53**, 4143 - 4156.

Wang, J. (2007). EM algorithms for nonlinear mixed effects models. *Computational Statistics & Data Analysis*, **51**, 3244 – 3256

Parameter Inference in the Pulmonary Circulation of Mice

L. Mihaela Paun[1], M. Umar Qureshi[2], Mitchel Colebank[2], Mansoor A. Haider[2], Mette S. Olufsen[2], Nicholas A. Hill[1], Dirk Husmeier[1]

[1] School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8SQ, UK
[2] Department of Mathematics, NC State University, Raleigh, NC 27695, USA

E-mail for correspondence: l.paun.1@research.gla.ac.uk

Abstract: This study focuses on parameter inference in a pulmonary blood circulation model for mice. It utilises a fluid dynamics network model that takes selected parameter values and aims to mimic features of the pulmonary haemodynamics under normal physiological and pathological conditions. This is of medical relevance as it allows monitoring of the progression of pulmonary hypertension. Constraint nonlinear optimization is successfully used to learn the parameter values.

Keywords: Pulmonary hypertension; Parameter Inference; Constraint Nonlinear Optimization; Partial Differential Equations; Windkessel model.

1 Introduction

Pulmonary hypertension (PH) is a leading cause of right heart failure. It involves vascular remodelling including stiffening of the large and small arteries. Clinically, PH is diagnosed by analysing blood pressure (BP) measured invasively in the large pulmonary arteries. However, key parameters, including arterial stiffness, cannot be measured in vivo. This creates the need for methods to estimate parameters indirectly from the measured haemodynamic blood flow and pressure data. This study uses a 1D fluid dynamical network model that predicts blood flow and pressure in the large pulmonary arteries (for details see Qureshi et al., 2017). The model is used to predict blood flow and pressure in healthy and hypoxic mice, for which data were acquired invasively (Tabima et al., 2012).

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

The method discussed here is not specific to mice, but can easily be extended to analysis of similar data from humans for whom repeated invasive procedures are required for diagnosis and treatment. The ultimate goal, and hence the motivation behind inferring the parameters, is to minimise the number of invasive procedures for PH patients, as well as to assist clinicians in devising better treatment strategies. Thus, this study focuses on inference of key parameters pertinent to disease detection and treatment. We show that using this statistical method, BP prediction is improved in healthy and hypoxic mice compared to the reference prediction obtained by using the best parameter guesses in Qureshi et al. (2017). This leads to enhanced reliability of key parameter estimates obtained using the model.

2 Mathematical Model

The 1D fluid-structure model is derived from the incompressible axisymmetric Navier–Stokes equations for a Newtonian fluid, coupled with a constitutive wall model predicting stiffness of the blood vessels. In addition, assuming that the vessels are cylindrical and the wavelength is significantly longer than their radii, conservation of mass and momentum give

$$\frac{\partial A}{\partial t} + \frac{\partial q}{\partial x} = 0, \qquad \frac{\partial q}{\partial t} + \frac{\partial}{\partial x}\frac{q^2}{A} + \frac{A}{\rho}\frac{\partial p}{\partial x} = -\frac{2\pi\mu r}{\delta}\frac{q}{A}, \tag{1}$$

where x (cm) and t (s) are the axial and temporal coordinates, p (mmHg) is blood pressure, q (ml/s) is blood flow rate, A (cm^2) is the cross-sectional area, $\delta = 40$ μm is the thickness of Stokes-layer in velocity profile, $\rho = 1.055$ g/ml is the blood density and $\mu = 0.0528$ cm^2/s is the viscosity. Assuming the arterial walls are homogeneous, isotropic and thin, the pressure-area relation is given by

$$p - p_0 = \frac{4}{3}\frac{Eh}{r_0}\left(1 - \sqrt{\frac{A_0}{A}}\right) \implies c^2(p) = \frac{A}{\rho}\frac{\partial p}{\partial A} = \frac{2}{3\rho}\frac{Eh}{r_0}\sqrt{\frac{A_0}{A}}, \tag{2}$$

where c (cm/s) is the wave speed, A_0 is the vessel cross-sectional area and r_0 (cm) the vessel radius at the pressure p_0.

The arterial network geometry, including length and radii for the 13 largest vessels in the pulmonary vasculature is obtained from a micro CT image of a healthy mouse lung (see Figure 1(a)). To solve the equations, boundary conditions are specified at the inlet and outlet vessels in the network. The system is driven by imposing an invasively measured flow profile at the inlet of the main pulmonary artery (MPA), while conservation of blood flow and continuity of pressure are enforced across the bifurcations. For the outflow boundary conditions, 3-element Windkessel models (two resistors R_1, R_2 and a capacitor C) are attached to the outlet of each of the seven terminal arteries in the network. The outflow boundary conditions account for the lumped effects of pulmonary haemodynamics beyond the truncated network of large arteries. The Windkessel model relates flow and pressure in the time domain over a cardiac cycle of length T via the input impedance $Z(\omega)$ by:

$$Z(\omega) = R_1 + \frac{R_2}{1 + i\omega C R_2} \implies q(L,t) = \frac{1}{T}\int_0^T p(L, t-\tau)Z(\tau)d\tau, \tag{3}$$

where R_1 and R_2 are the proximal and distal vascular resistances beyond each truncated artery, $R_1 + R_2$ is the total resistance and C is the total compliance of the vascular bed. The model takes a number of parameters as input and predicts the flow and pressure at different locations along the large pulmonary arteries. One of these parameters is the arterial stiffness, which significantly increases during the PH and which can be expressed as follows: $\frac{Eh}{r_0} = f_1 \exp(f_2 r_0) + f_3$, where E and h are the Young's modulus and thickness of the arterial wall, and f_1 (g cm/s^2), f_2 (cm^{-1}), f_3 (g cm/s^2) are the material parameters.

3 Methodology

Let the statistical model be defined by $y_i = f(x_i; \boldsymbol{\theta}) + \epsilon_i$, where $y_i \in \mathbf{y}$ are the noisy measured flow and pressure, $f(.)$ describes the system behaviour that comes from numerically solving the fluids model, $\boldsymbol{\theta}$ are the parameters that we wish to infer from the observed flow and pressure and they are bounded, $x_i \in \mathbf{x}$ denote other input variables and $\boldsymbol{\epsilon}$ are the errors, which we assume are i.i.d and follow a Gaussian distribution. The objective function to be minimised using Constraint Nonlinear Optimization is the Residual Sum of Squares,

$$RSS = \sum_i (y_i - f(x_i; \boldsymbol{\theta}))^2. \tag{4}$$

Under the assumption of i.i.d. Gaussian errors, the log likelihood of the data takes the form

$$\log(L) = -n \log \sqrt{2\pi\sigma^2} - \frac{RSS}{2\sigma^2}. \tag{5}$$

A Sequential Quadratic Programming (SQP) gradient-based method is used to minimise the RSS (Boggs et al., 2000).

4 Simulations

Simulations are set up to mimic experimental waveforms, which are recorded in the MPA in healthy and hypoxic mice (Tabima et al., 2012). The parameter set to be inferred initially includes $\boldsymbol{\theta} = (\frac{Eh}{r_0}, r_1, r_2, c_1)$, where r_1, r_2 are resistances ($r_1 = (1 - 0.5r_1)R_{01}, r_2 = (1 - 0.5r_2)R_{02}$, R_{01} and R_{02} are nominal resistances) and c_1 capacitance ($C = (1 - 0.5c_1)C_0, C_0$ is nominal capacitance) used to predict parameters assigned at the outlet and $\frac{Eh}{r_0}$ is the elastance used to predict stiffness in all vessels. The parameter set is subsequently extended to include a tapering factor, ζ, for the large vessels in the network, as there was evidence of vessel radii decreasing along their length, but this was not quantified during the segmentation process. Since the parameters are on different scales ($\frac{Eh}{r_0} \in [2.424 \times 10^5, 6.85 \times 10^6]$, $r_1, r_2, c_1 \in [-2, 2]$, and $\zeta \in [0, 1.2]$), to avoid having an ill-conditioned problem induced by a high condition number in the Hessian matrix, we rescale the parameters to have the same order of magnitude (Yang et al., 2010). Certain parameter configurations violate the model assumptions; these are marked by setting RSS to a high value (10^{10}). The RSS is calculated for pressure and we aim to find the set of parameters that minimise the RSS. The initial parameter values are uniformly drawn from a Sobol sequence to ensure a good coverage of

the multidimensional parameter space (Bratley et al., 1988). The algorithm is iterated until it satisfies the convergence criterion, i.e. $|\boldsymbol{\theta_i} - \boldsymbol{\theta_{i+1}}| < 10^{-11}$. One forward simulation of the mathematical model takes 13 seconds to complete. The optimization problem required 3 hours on average to reach convergence of parameter estimates.

5 Results and Discussion

Regardless of the initial value, the algorithm converged for both the healthy and the hypoxic mouse studied. Figure 1 shows our optimised pressure waveform, plotted alongside the measured and the reference pressure for the 4D optimization problem. Panel (d) shows the pressure fit for the hypoxic mouse. The optimized fit predicts data better than nominal parameter values, supported by a significantly smaller RSS than the one between the reference and the measured pressure (panel (b)). For the healthy mouse (panel (c)), the simulated pressure closely follows the measured pressure except near the peak, where an offset is registered. Nevertheless, in this case too, a clear improvement is achieved over the reference pressure. We hypothesise that this peak shift is a consequence of (i) the model specifying the elastic behaviour of the blood vessels and/or the boundary conditions, (ii) uncertainty of the geometry measurements which are not specific to a given mouse, (iii) a combination of (i) and (ii). The overall model prediction appears better for the hypoxic than the healthy mouse. When the tapering parameter

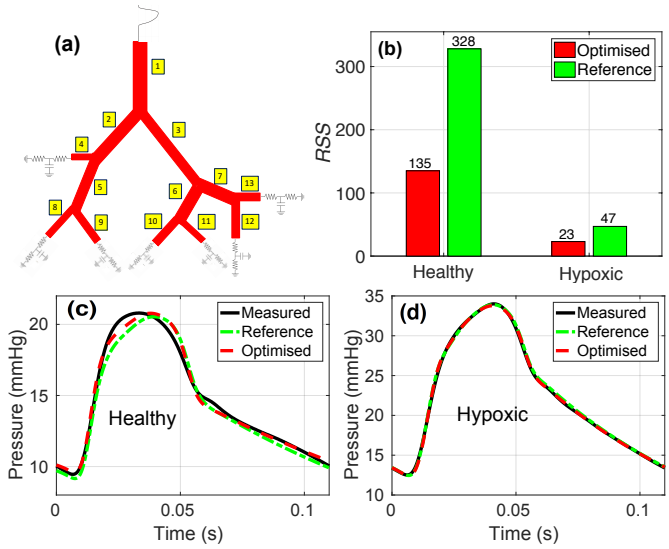


FIGURE 1. (a) The arterial network for the fluid dynamical model, (b) Comparison of RSS between reference and optimised pressure simulations, (c) & (d) Comparison of simulated pressure using reference and optimised parameters values for the healthy and hypoxic mice.

is included in the analysis, a reduction of 31% is registered in the RSS for the control mouse. These results are summarised in Figure 2. Panel (b) illustrates the

optimised pressure waveform for the 5D problem plotted alongside that for the 4D case and the measured pressure is superimposed. While the deviation from the measured data has decreased, the offset in the peak value is still present. In order to select between the two competing models, the 5D model, which includes a tapering parameter, and the 4D one, model selection using the Akaike Information Criterion - AIC (Akaike, 1971) and Bayesian Information Criterion - BIC (Schwarz, 1978) can be performed. However, the estimate of the error variance, σ^2 needed to calculate the log likelihood of the data (5) is not available[1]. Hence, we take the inverse approach and calculate what error variance would make us favour the 5D model over the 4D model. Calculations indicate that if $\sigma^2 < 21$, i.e. signal-to-noise ratio, $SNR > 2.40$, then the 5D model is preferred according to the AIC; if $\sigma^2 < 6.06$, i.e. $SNR > 8.30$, then the 5D model is favoured according to the BIC.

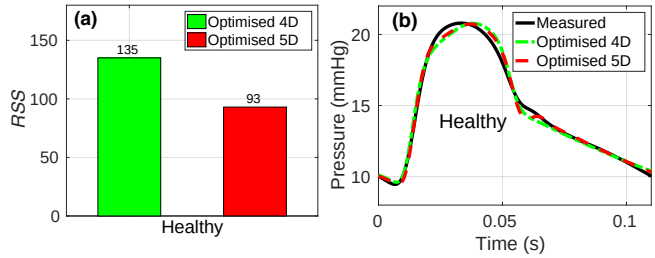


FIGURE 2. (a) Comparison of RSS between the optimised pressure simulations for the healthy mouse, (b) Comparison of simulated pressure using 4D and 5D optimised parameters values for the healthy mouse.

In conclusion, parameters have successfully been inferred for this fluid-structure model. Future work will include improvements in the model to capture a more realistic elastic behaviour of the vessel wall. This may address the alignment issue observed near the peak pressure for the healthy mouse by better controlling the steepness of the pressure. Additionally, to account for the uncertainty associated with the geometry measurements, we will estimate them as part of the inference procedure. Finally, we also aim to apply the statistical methods presented here to a population of mice, as well as to data from human patients.

Acknowledgments: This work is part of the research programme of the Centre for multiscale soft tissue mechanics with application to heart & cancer (SofT-Mech), funded by the Engineering and Physical Sciences Research Council (EPSRC) of the UK, grant reference number EP/N014642/1. Olufsen, Haider and Qureshi were supported by National Science Foundation (NSF-DMS # 1615820). Data were made available by N. Chesler, Department of Biomedical Engineering, University of Wisconsin, Madison.

[1]In principle, we could infer the variance, but due to the slight model mismatch apparent from Figures 1 and 2, the results would be misleading.

References

Akaike, H. (1971), Information theory and an extension of the maximum likelihood principle. In: *Proc. 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR*. 267 – 281.

Boggs, P.T. and Tolle, J.W. (2000), Sequential quadratic programming for large-scale nonlinear optimization. *Journal of Computational and Applied Mathematics*, **124(12)**, 123 – 137.

Bratley, B. (1988), Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator. *ACM Trans Math Softw*, **14(1)**, 88 – 100.

Qureshi, M.U. et al. (2017), Simulating effects of hypoxia on pulmonary haemodynamics in mice. In: *Proc CMBE*. 271 – 274.

Schwarz, G. (1978), Estimating the Dimension of a Model. *Ann. Statist.*, **2(6)**, 461 – 464.

Tabima, D.M. et al. (2012), Persistent vascular collagen accumulation alters hemodynamic recovery from chronic hypoxia. *J. Biomech*, **45(5)**, 799 – 804.

Yang, K.W. and Lee, T.Y. (2010), Heuristic scaling method for efficient parameter estimation. *Chemical Engineering Research and Design*, **88(56)**, 520 – 528.

The Pairwise Expectation Maximization Algorithm for Fitting Parameter-Driven Models

Xanthi Pedeli[1], Cristiano Varin[1]

[1] Ca' Foscari University of Venice, Italy

E-mail for correspondence: xanthi.pedeli@unive.it

Abstract: The likelihood function is often intractable in discrete-valued time series models. Several simulation-based methods to approximate the likelihood have been proposed in the literature. Although simulation methods may make the estimation problem feasible, they are often computationally intensive and the quality of the numerical approximations may be difficult to assess. In this paper we employ a pairwise likelihood approach, that is a composite likelihood based on pairs of observations, for the estimation of parameter-driven models for time series of counts. Maximization of the pairwise likelihood is carried out with a pairwise version of the expectation maximization algorithm.

Keywords: Expectation maximization algorithm; Pairwise likelihood; Parameter-driven models; Surveillance; Time series of counts.

1 Pairwise fitting of parameter-driven models

Parameter-driven models (Cox, 1981) are often considered to extend generalized linear models for time series analysis of counts. The observations Y_t are assumed to be independent random variables conditionally on a latent process U_t, $t = 1, \ldots, n$. The conditional mean of the observations is $\mathrm{E}(Y_t|U_t) = g(x_t^{\mathrm{T}}\beta + U_t)$, where $g(\cdot)$ is a link function, x_t is a vector of covariates with associated regression parameters β and U_t is an autoregressive and moving average process. For these models, maximum likelihood estimation is usually intractable because the likelihood is the n-dimensional integral

$$L(\theta; \mathbf{y}) = \int_{\mathbb{R}^n} \prod_{t=1}^{n} f(y_t|u_t; \theta) f(u_1, \ldots, u_n; \theta) du_1 \ldots du_n \qquad (1)$$

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and the model parameter θ includes the regression parameters β and the parameters of the latent process. Several simulation-based strategies for approximate likelihood inference have been suggested in the literature, see, for example, Davis and Dunsmuir (2016) and the references therein.

In this paper, we focus on an alternative non-simulation approach based on the idea of pairwise likelihood (Varin et al., 2011) that allows for the replacement of the n-fold integration in (1) with a set of two-dimensional integrals. We consider estimators obtained by maximizing the pairwise likelihood of order d that involves only pairs of observations that are far apart no more than d lags,

$$L_p^{(d)}(\theta; \mathbf{y}) = \prod_{t=l+1}^{n} \prod_{l=1}^{d} \int_{\mathbb{R}^2} f(y_{t-l}|u_{t-l}; \theta) f(y_t|u_t; \theta) f(u_{t-l}, u_t; \theta) du_{t-l} du_t.$$

It can be shown that the distance d that maximize the efficiency of the estimators derived from the pairwise likelihood depend on the order of the latent autoregressive and moving average process.

Instead of a direct maximization of the pairwise likelihood, we develop a pairwise version of the expectation maximization algorithm. The expectation step of the algorithm requires to compute the conditional expected value of a set of bivariate complete log-likelihoods,

$$Q(\theta|\hat{\theta}^{(i)}) = \sum_{t=l+1}^{n} \sum_{l=1}^{d} \mathrm{E}\left\{ \log f(y_{t-l}, y_t, U_{t-l}, U_t; \theta) \Big| y_{t-l}, y_t; \hat{\theta}^{(i)} \right\}.$$

The above function can be efficiently approximated with double Gauss-Hermite quadrature. The maximization step is replaced with a conditional maximization of each model parameter with the other parameters fixed at their previous values. The virtue of the conditional maximization is that simple closed-form expressions for the maxima are available, thus making the implementation of the algorithm particularly convenient.

Simulation studies not shown in this short paper indicate that the pairwise fitting method compare well in terms of statistical and numerical efficiency with popular approaches for inference in parameter-driven models implemented in `R` (R Core Team, 2016).

2 Application

We illustrate our fitting method using the `meningo.age` data set included in the `R` package `surveillance` (Höhle *et al.*, 2016). The data concern monthly counts of meningococcal infections in France during the period 1985–1997 subdivided into four age groups, see Figure 1. The total number of monthly observations is $n = 156$.

The standard analysis of this type of surveillance data assumes that infection counts Y_t for each age group are marginally distributed as independent Poisson random variables with mean $\exp(\eta_t)$ specified in way to account for annual seasonality and linear trend,

$$\eta_t = \beta_0 + \beta_1 \cos\left(2\pi \frac{t}{12}\right) + \beta_2 \sin\left(2\pi \frac{t}{12}\right) + \beta_3 \frac{t}{156}. \tag{2}$$

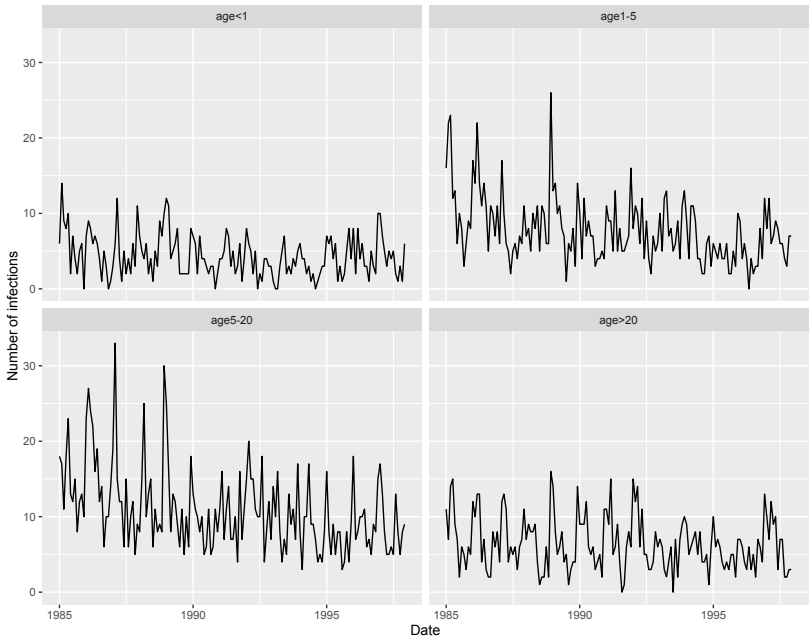


FIGURE 1. Monthly counts of meningococcal infections in France for the period 1985–1997. Source: `R` package **`surveillance`** (Höhle *et al.*, 2016).

In order to handle for the presence of serial correlation, we use the pairwise expectation maximization algorithm for fitting the parameter-driven model that assumes that Y_t follows a Poisson distribution with mean $\exp(\eta_t + U_t)$, where the linear preditor η_t is specified as in (2) and U_t is an autoregressive model of order one.

The parameter estimates and the corresponding standard errors obtained with the standard analysis and the parameter-driven model are displayed in Table 1. Results of the standard analysis indicate significant seasonality and decreasing linear trend for all age groups. The fitted parameter-driven models confirm the conclusions of the standard analysis for all age groups apart from infants (age < 1), where there is significant autocorrelation and no evidence of a decreasing trend of infections. The application illustrates how ignoring serial correlation in surveillance data can lead to misleading inference about sensible quantities such as trends.

Acknowledgments: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 699980.

References

Cox, D.R. (1981) Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics* **8**, 93–115.

TABLE 1. Parameter estimates (standard errors) for models fitted to the montlhy counts of meningococcal infections in France for the period 1985–1997.

Standard Poisson regression				
	age < 1	age $1-5$	age $5-20$	age > 20
β_0	1.64 (0.07)	2.37 (0.05)	2.71 (0.05)	1.95 (0.06)
β_1	0.17 (0.06)	0.16 (0.04)	0.12 (0.04)	0.21 (0.05)
β_2	0.37 (0.06)	0.31 (0.04)	0.26 (0.04)	0.43 (0.05)
β_3	−0.43 (0.13)	−0.75 (0.10)	−0.71 (0.09)	−0.31 (0.11)
Parameter-driven				
	age < 1	age $1-5$	age $5-20$	age > 20
β_0	1.60 (0.13)	2.33 (0.07)	2.67 (0.07)	1.93 (0.07)
β_1	0.16 (0.08)	0.15 (0.05)	0.11 (0.05)	0.21 (0.05)
β_2	0.37 (0.07)	0.31 (0.06)	0.25 (0.05)	0.43 (0.05)
β_3	−0.42 (0.25)	−0.72 (0.12)	−0.69 (0.11)	−0.28 (0.13)
ϕ	0.73 (0.35)	0.32 (0.37)	0.21 (0.24)	0.48 (0.43)
σ	0.19 (0.07)	0.21 (0.06)	0.23 (0.04)	0.13 (0.06)

Davis, R.A. and Dunsmuir, W.T.M. (2016). State space models for count time series. In: *Handbook of Discrete-Valued Time Series*, Chapman & Hall/ CRC.

Höhle, M., Meyer, S. and Paul, M. (2016). `surveillance`: temporal and spatio-temporal modeling and monitoring of epidemic phenomena. `R` package version 1.13.0.

R Core Team (2016). `R`: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–41.

Estimating abrupt change models with covariate-dependent changepoint

Salvatore Fasola[1], Stefania La Grutta[1], Vito M.R. Muggeo[2]

[1] Institute of Biomedicine and Molecular Immunology - CNR, Palermo, Italy
[2] Department of Economics, Business and Statistical Sciences, Palermo, Italy

E-mail for correspondence: salvatore.fasola@ibim.cnr.it

Abstract: We discuss the problem of estimating a changepoint expressed as a linear function of covariates in piecewise constant regression models. Parameters involved in the change-point function are unbounded, therefore conventional grid search techniques result to be unfeasible for estimation. We propose an iterative algorithm which is able to work in this framework, and illustrate it by an application to a real dataset.

Keywords: Linear changepoint; Model linearization; Iterative algorithm.

1 Introduction

Changepoint detection is a commonly addressed problem, especially in Econometrics, Biology, Genetic and Bioinformatics. A piecewise constant regression model with a single changepoint can be expressed as

$$\mu_i = \beta_0 + \beta_1 I(x_i > \psi), \tag{1}$$

where the mean value μ of the response variable Y changes abruptly at the value ψ of the covariate x. Most of the algorithms and applications developed are concerned with estimation of the number and locations of several changepoints. The most widespread approach is grid search based on dynamic programming (Jackson et al., 2005), which can yield the global solution with a $O(n)$ computational cost for any number of breaks (Maidstone et al., 2016).

In this paper we consider the problem of possible heterogeneity in the changepoint, which means to replace in (1) ψ with ψ_i. Some papers address the problem by introducing random effects in the model (Muggeo et al. 2014, Jackson and Sharples 2004). However, no papers address the problem of modelling possible

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

heterogeneity due to covariates, namely when $\psi_i = \psi(\boldsymbol{v}_i; \theta)$. Grid search algorithms are difficult to use, since the parameter θ has unbounded support. In this paper we consider an abrupt change model with covariate-dependent changepoint, $\psi(\boldsymbol{v}_i; \theta) = \theta_0 + \theta_1 v_i$ and introduce a heuristic, iterative algorithm for estimation.

2 Methods

We consider the following model where the mean level of the response Y exhibits a jump at given values ψ of the explanatory variable x. However, ψ is supposed to depend linearly on a single additional covariate v:

$$\mu_i = \beta_0 + \beta_1 I(x_i > \theta_0 + \theta_1 v_i). \tag{2}$$

If the point (v, x) lies below the straight line $\psi = \theta_0 + \theta_1 v$, the mean level of Y is β_0, while it shifts instantaneously to $\beta_0 + \beta_1$ above. Estimating (2) by searching for $\hat{\theta}_1$ and $\hat{\theta}_2$ on a grid is obviously very difficult, since these parameters do not have a finite support. We therefore propose to use the identity

$$I(x_i > \theta_0 + \theta_1 v_i) = \frac{1}{2} \frac{x_i - \theta_0 - \theta_1 v_i}{|x_i - \theta_0 - \theta_1 v_i|} + \frac{1}{2} \tag{3}$$

for $x_i \neq \theta_0 + \theta_1 v_i$, which substituted in (2) gives

$$\begin{aligned}
\mu_i &= \beta_0 + \beta_1 \left(\frac{1}{2} \frac{x_i - \theta_0 - \theta_1 v_i}{|x_i - \theta_0 - \theta_1 v_i|} + \frac{1}{2} \right) \\
&= \beta_0 + \beta_1 \left(\frac{1}{2} \frac{x_i}{|x_i - \theta_0 - \theta_1 v_i|} + \frac{1}{2} \right) + \\
&\quad + \ (-\beta_1 \theta_0) \left(\frac{1}{2} \frac{1}{|x_i - \theta_0 - \theta_1 v_i|} \right) + (-\beta_1 \theta_1) \left(\frac{1}{2} \frac{v_i}{|x_i - \theta_0 - \theta_1 v_i|} \right) \\
&= \beta_0 + \beta_1 z_i + \gamma_0 w_{i0} + \gamma_1 w_{i1},
\end{aligned} \tag{4}$$

where

$$\gamma_0 = -\beta_1 \theta_0 \quad \text{and} \quad \gamma_1 = -\beta_1 \theta_1. \tag{5}$$

Note the 'working' covariates

$$z_i = \left(\frac{1}{2} \frac{x_i}{|x_i - \tilde{\theta}_0 - \tilde{\theta}_1 v_i|} + \frac{1}{2} \right),$$

$$w_{i0} = \left(\frac{1}{2} \frac{1}{|x_i - \tilde{\theta}_0 - \tilde{\theta}_1 v_i|} \right),$$

and

$$w_{i1} = \left(\frac{1}{2} \frac{v_i}{|x_i - \tilde{\theta}_0 - \tilde{\theta}_1 v_i|} \right) \tag{6}$$

enter model (4) linearly, with $\tilde{\theta}_0$ and $\tilde{\theta}_1$ meaning approximate values.

Formulas above suggest the following simple iterative algorithm:

1. choose *starting values* $\tilde{\theta}_0$ and $\tilde{\theta}_1$;

2. compute the *working covariates* (6);
3. estimate the *working linear model* (4) and extract $\hat{\beta}_1$, $\hat{\gamma}_0$ and $\hat{\gamma}_1$;
4. *update* the parameter values via $\hat{\theta}_0 = -\hat{\gamma}_0/\hat{\beta}_1 \quad \text{and} \quad \hat{\theta}_1 = -\hat{\gamma}_1/\hat{\beta}_1$;
5. set $\tilde{\theta}_0 = \hat{\theta}_0$ and $\tilde{\theta}_1 = \hat{\theta}_1$ and *iterate* 2 to 4 until convergence.

A possible strategy for initializing the algorithm is to set starting values $\tilde{\theta}_1 = 0$, and $\tilde{\theta}_0 = \bar{x}$, where $\bar{x}$ is the sample mean of the x_is. Unfortunately, simple application of the aforementioned algorithm will not work properly. In fact, the likelihood typically exhibits many local optima. Besides, denominators of the working covariates (6) go to zero when $x_i \approx \tilde{\theta}_0 + \tilde{\theta}_1 v_i$, namely when points (v_i, x_i) are close to the approximate straight line. In the next section we illustrate a simple point scaling that improves feasibility of the algorithm.

3 Scaling v and x values

The idea is moving points (v_i, x_i) away from the straight line $x = \tilde{\theta}_0 + \tilde{\theta}_1 v$. Note that, when $\tilde{\theta}_1 = 0$, we can simply scale the covariate x, since the approximate change-point $\tilde{\theta}_0$ is unique. Therefore, we operate a rotation of the points (v_i, x_i) to reduce to this suitable situation. At this aim, we consider the angle defined by the approximate straight line, namely $\tilde{\rho} = \arctan(\tilde{\theta}_1)$, and apply the following transformation:

$$\begin{pmatrix} v_i^* \\ x_i^* \end{pmatrix} = \tilde{\Lambda} \begin{pmatrix} v_i \\ x_i \end{pmatrix}, \tag{7}$$

where $\tilde{\Lambda}$ is the rotation matrix

$$\tilde{\Lambda} = \begin{pmatrix} \tilde{\lambda}_{11} & \tilde{\lambda}_{12} \\ \tilde{\lambda}_{21} & \tilde{\lambda}_{22} \end{pmatrix} = \begin{pmatrix} \cos(\tilde{\rho}) & \sin(\tilde{\rho}) \\ -\sin(\tilde{\rho}) & \cos(\tilde{\rho}) \end{pmatrix}. \tag{8}$$

Rewriting the rotated covariate x^* according to (7) gives

$$\begin{aligned} x^* &= \tilde{\theta}_0^* \\ \tilde{\lambda}_{21} v + \tilde{\lambda}_{22} x &= \tilde{\theta}_0^* \\ \tilde{\lambda}_{22} x &= \tilde{\theta}_0^* - \tilde{\lambda}_{21} v \end{aligned}$$

and

$$x = \frac{\tilde{\theta}_0^*}{\tilde{\lambda}_{22}} - \frac{\tilde{\lambda}_{21}}{\tilde{\lambda}_{22}} v,$$

so that

$$\tilde{\theta}_0 = \frac{\tilde{\theta}_0^*}{\tilde{\lambda}_{22}} \quad \text{and} \quad \tilde{\theta}_1 = -\frac{\tilde{\lambda}_{21}}{\tilde{\lambda}_{22}}, \tag{9}$$

and, conversely,

$$\tilde{\theta}_0^* = \tilde{\lambda}_{22} \tilde{\theta}_0 \quad \text{and} \quad \tilde{\theta}_1^* = 0. \tag{10}$$

Note that the slope of the rotated line $\tilde{\theta}_1^*$ is 0 by construction, and this makes easier to operate the scaling. We use a scaling factor $c \in (0, 1)$ and compute a lower, $\tilde{\theta}_0^{*-}$, and a upper, $\tilde{\theta}_0^{*+}$, 'threshold' value:

$$\tilde{\theta}_0^{*-} = \tilde{\theta}_0^* - c(\tilde{\theta}_0^* - x_{(1)}^*), \qquad \tilde{\theta}_0^{*+} = \tilde{\theta}_0^* + c(x_{(n)}^* - \tilde{\theta}_0^*).$$

We therefore consider the standard linear trasformation

$$x'_i = x^*_{(1)} + (x^*_i - x^*_{(1)})(1 - c) \tag{11}$$

for $x^*_i \in [x^*_{(1)}, \tilde{\theta}^*_0]$, and

$$x'_i = \tilde{\theta}^{*+}_0 + (x^*_i - \tilde{\theta}^*_0)(1 - c) \tag{12}$$

for $x^*_i \in (\tilde{\theta}^*_0, x^*_{(n)}]$. Figures 1 and 2 show the scalement mechanism through an example on a set of $n = 100$ randomly generated covariate pairs, assuming $v \sim \mathcal{N}(2, 3^2)$ and $x \sim \mathcal{N}(5, 6^2)$. Supposing the starting straight line given by $x_i = 8 + 2v_i$ (Figure 1, left panel), the rotated straight line, on the rotated points, is $x^* = 3.58$ (Figure 1, right panel). The left panel in Figure 2 represents the rotated points without scaling x^*, namely when $c = 0$, while the right panel shows the effect of a scaling factor $c = 0.1$; note the scaling induces a point-free interval (dashed lines) in the neighbourhood of $\tilde{\theta}^*_0$ (solid line). Finally, we use x', $v' = v^*$ and the rotated straight line to compute auxiliary covariates (6) and fit the working linear model (4).

We stress that, despite the multiple transformations induced to the covariates, estimates of the mean levels β_0 and β_1 are substantially unaffected, while the straight line parameter estimates, of course, are. In particular, the rotation only induces a reparametrization according to (9) and (10), while the scaling should also favour to skip some spurious optima. Note that c should be reduced (halved for example) throughout iterations any time the likelihood decreases.

4 Real data example

To illustrate, we apply the proposed algorithm to the airquality dataset shipped with the R environment. The dataset consists of 154 daily observations concerning some air quality values in New York from May 1, 1973 to September 30, 1973. Tropospheric ozone is an atmospheric pollutant, and its concentration represents

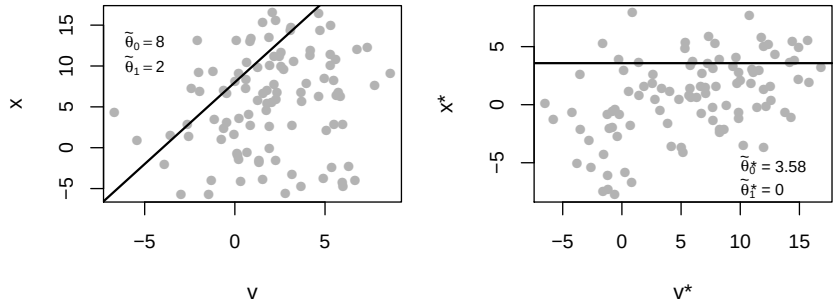


FIGURE 1. Example of starting straight line in a toy dataset (left panel) and rotated points and straight line (right panel).

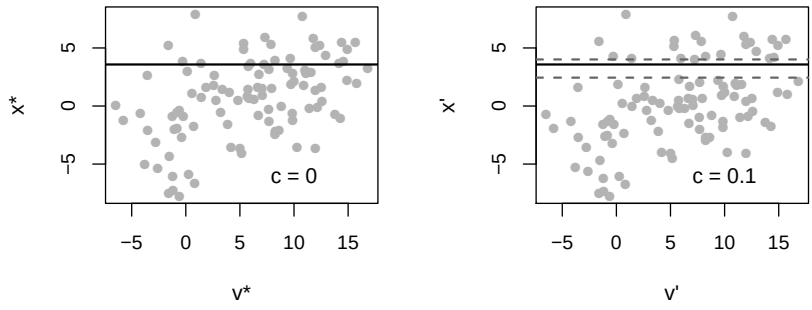


FIGURE 2. Only rotated (left panel) vs rotated and scaled data (right panel).

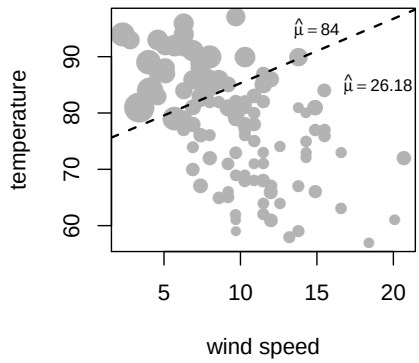


FIGURE 3. Airquality dataset: wind-temperature scatterplot with point size proportional to the ozone level and estimated linear change-point (dashed line).

a common variable of interest in environmental science. The ozone levels may depend on many factors, including atmospheric agents, such as temperature and wind. Therefore we model the mean levels of ozone (Y, parts per billion) as a function of temperature (x, degrees Fahrenheit) and wind (v, average speed in miles per hour). Figure 3 displays the wind-temperature scatterplot with point sizes proportional to the ozone levels for the $n = 116$ complete records. From a visual inspection the concentration of the pollutant seems to increase abruptly in the top-left region of the plot; in particular, a threshold temperature value appears to be approximately 80 degrees. Estimating model (2), assuming the Y_is to be independent and Gaussian, could yield additional information.

To initialize the algorithm we choose $\tilde{\psi} = 80$ ($\tilde{\theta}_1 = 0$) as starting guess. The fitted regression equation is $\hat{\mu}_i = 26.18 + 57.82\ I(x_i > 72.83 + 1.24 v_i)$. Figure 3 displays the estimated straight line on the scatterplot. The mean ozone level appears to increase abruptly from about 26 to 84 p.p.b as temperature goes beyond some

'critical' value. In the absence of wind, the 'critical' value is about 73 degrees. Wind seems to have a 'positive' effect in limiting the ozone levels: in fact, as the wind increases, temperature has to increase further to cause a 'jump' in the ozone levels. The BIC provides good evidence supporting the choice of this regression model.

References

Jackson, B., Scargle, J.D., Barnes, D., Arabhi, S., et al. (2005). Hierarchical generalized linear models. *An algorithm for optimal partitioning of data on an interval*, **12**, 105 – 108.

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2016). On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, **27**, 1 – 15.

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

Jackson, C.H., and Sharples, L.D. (2004). Models for longitudinal data with censored changepoints. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53**, 149 – 162.

Identifying influential observations in complex Bayesian mediation models

Šárka Rusá[1], Arnošt Komárek[1], Emmanuel Lesaffre[2], Luk Bruyneel[3]

[1] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[2] Leuven Biostatistics and Statistical Bioinformatics Centre, University of Leuven, Belgium
[3] Leuven Institute for Healthcare Policy, University of Leuven, Belgium

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

Abstract: Although increasingly complicated (moderated) mediation models are being employed in practice, most of existing mediation literature has not dealt with model diagnostics. We propose a Bayesian approach to the detection of influential observations (or sets of observations). Importance sampling with weights which take advantage of the dependence structure in mediation models is utilized in order to estimate the case-deleted posterior means of the parameters. The method is applied to the ordinal measurements of patients' willingness to recommend hospitals collected on patients in a large European study to answer the research question whether the outcome depends on recorded system-level features in the organization of nursing care, and whether the related effect is mediated by two measurements of nursing care left undone and possibly moderated by nurse education.

Keywords: Bayesian mediation models; Influential observations, Importance sampling.

1 Introduction

Recent advances in the literature on mediation models extended the simple three variable (outcome, mediator, regressor) mediation model to complex models with multiple mediators, non-continuous outcome, multi-level setting and other generalizations. The Bayesian approach, which we will also adopt in this paper, is often utilized in such complicated models because the generalizations are more

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

straightforward. However, no general model diagnostic tool for identification of influential observations is readily available for these models and we deem the development of such instruments as necessity. In this paper, we propose the use of importance sampling with weights which take advantage of the dependence structure in mediation models in order to estimate the case-deleted posterior means of the model parameters and other characteristics hence allowing for identification of influential observations. Suitable importance weights can be used for leave-one-out diagnostics, as well as leaving out all observations from a specific hospital. The methods shown here on one specific mediation model can be used in the same fashion in any mediation model with or without latent variables.

2 The RN4CAST Study and the Research Question

The Registered Nurse Forecasting (RN4CAST) study (Sermeus et al., 2011) is a cross-sectional survey of patients and nurses conducted in 11 European countries, in which the patients and nurses are further clustered in hospitals and nursing units. The data collected in 2009 – 2010 during this FP7-funded project contain information on various hospital characteristics such as nurse staffing, nurse education, number of beds, etc. On top of that, the nurses were interviewed on various aspects of the nursing care such as their well-being, satisfaction with their job, their willingness to recommend the hospital and overtime work. The patients provided information on their satisfaction with hospital care and hospital rating.

In our previous work (Rusá et al., 2017), we presented a multi-level moderated mediation model with ordinal outcome in order to evaluate on how the patients' willingness to recommend the hospital (patient's satisfaction) relates to the system-level features in the organization or nursing care and whether the association is mediated by two measurements of nursing care left undone and possibly moderated by nurse education. In this paper, we aim to explore if there exist hospitals which are influential with respect to such a model. In particular, we want to investigate the change in the indirect effects.

3 Multi-Level Moderation Mediation Model with Ordinal Outcome

We denote the patient outcome (patients' willingness to recommend the hospital) by Y_{ijk} which represents the value of the outcome measured on an ordinal scale $0, \ldots, L$ on the i-th patient from hospital j and country k. It is assumed that there exist unknown thresholds $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)^{\mathrm{T}}$, and a latent variable $Y^{\star}_{ijk}$ such that

$$Y_{ijk} = l, \qquad \text{iff} \quad \alpha_l < Y^{\star}_{ijk} \leq \alpha_{l+1}, \qquad l = 0, \ldots, L,$$

$-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_L < \alpha_{L+1} = \infty$.

The purpose of the model studied in Rusá et al. (2017) was to ascertain whether there was a significant effect of nurse staffing, the quality of nurse work environment and other hospital characteristics ($\boldsymbol{X}_{jk}$) on the patient outcome ($\boldsymbol{Y}_{ijk}$) and if it was mediated by two mediators of care left undone ($M_{1,jk}$ and $M_{2,jk}$).

Moreover, we assessed the moderating effect of nurse education (these terms are included in the covariate vector $\boldsymbol{X}_{jk}$).

In order to deal with the multi-level structure of the RN4CAST data, we included the hospital-level random effects (U_{jk}). Since countries were not sampled but chosen and there are just a handful of countries, we have considered their effects as fixed. Consequently, the considered model equation for the latent outcome variables had the following form:

$$Y^{\star}_{ijk} = \beta_k \; + \; \boldsymbol{X}^{\top}_{jk}\boldsymbol{\beta}_X \; + \; \boldsymbol{M}^{\top}_{jk}\boldsymbol{\beta}_M \; + U_{jk} \; + \; \varepsilon_{ijk}, \tag{1}$$

$$M_{t,jk} = \gamma_{t,k} \; + \; \boldsymbol{X}^{\top}_{jk}\boldsymbol{\gamma}_{t,X} \; + \; \xi_{t,jk}, \quad t = 1,\, 2, \tag{2}$$

$$U_{jk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0,\, \sigma^2_{hospital}\big), \quad \varepsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0,\, \sigma^2_{\varepsilon}\big),$$

$$\boldsymbol{\xi}_{jk} = \big(\xi_{1,jk},\, \xi_{2,jk}\big)^{\top} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_T\big(\mathbf{0},\, \boldsymbol{\Sigma}_{\xi}\big).$$

We assume that $U_{jk}, \varepsilon_{ijk}, \xi_{1,jk}, \xi_{2,jk}$ are mutually independent. For β_k, $\boldsymbol{\beta}_X$, $\boldsymbol{\beta}_M$, $\gamma_{1,k}$, $\gamma_{2,k}$, $\boldsymbol{\gamma}_{1,X}$, $\boldsymbol{\gamma}_{2,X}$, we used vague normal priors. As for $\boldsymbol{\Sigma}_{\xi}$, we will parametrize it using the variance of $\xi_{1,jk}$ denoted by $\sigma^2_{\xi 1}$, the variance of $\xi_{2,jk}$ denoted by $\sigma^2_{\xi 2}$ and their correlation ρ. For the standard deviations $\sigma_{hospital}$, $\sigma_{\varepsilon}\, \sigma_{\xi 1}$, $\sigma_{\xi 2}$, we specify improper uniform priors on $(0, \infty)$. A noninformative prior was considered for ρ: $\rho \sim U(-1, 1)$ (a uniform distribution on $(-1, 1)$). The following noninformative prior was used for the thresholds (Song and Lee, 2012, p.117): $p(\alpha_2, \ldots, \alpha_{L-1}) \propto C$, for $\alpha_1 < \alpha_2 < \ldots < \alpha_{L-1} < \alpha_L$. The samples from the posterior distributions were obtained using the Stan software. The inference is based on MCMC methods, mainly the Hamiltonian Monte Carlo and the Metropolis algorithm.

4 The Detection of Influential Observations

Besides the estimation of the parameters in our model, it may be of interest to assess the influence of some (sets of) observations on the parameter estimates. In case of our application, it makes sense to evaluate the change in the estimates of indirect effects based on the data without a specific hospital. To this end, importance sampling has been utilized in the literature for different sorts of models, see e.g. Bradlow and Zaslavsky (1997). We have generalized this approach to our moderated mediation model with ordinal outcome as sketched below.

Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, be the vectors containing all the model parameters from equations (1) and (2) including the augmented data, respectively. The importance weights needed to evaluate the influence of a hospital j in a country k are defined as

$$w^{\star}_{jk}\left(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{m}\right) = \left[p\left(\boldsymbol{Y}_{jk}|\boldsymbol{Y}^{\star}_{jk}, \boldsymbol{\alpha}\right) p\left(\boldsymbol{Y}^{\star}_{jk}|\boldsymbol{\theta}_1, \boldsymbol{U}_{jk}\right) p\left(\boldsymbol{m}_{jk}|\boldsymbol{\theta}_2\right)\right]^{-1}, \tag{3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\, \boldsymbol{\theta}_2)$ and $\boldsymbol{Y}^{\star}_{jk}$ contains all latent data $Y^{\star}_{ijk}$ in hospital j in a country k. Further, $\boldsymbol{m}$ and $\boldsymbol{Y}^{\star}$, respectively, contain all the mediator measurements $\boldsymbol{m}_{jk}$ and all latent data $\boldsymbol{Y}^{\star}_{jk}$, respectively. It can be shown that we can use the standardized importance weights $w^{(n)}_{jk} \propto w^{\star}_{jk}\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{Y}, \boldsymbol{m}\right)$, $n = 1, \ldots, N$ to estimate the hospital-deleted posterior mean of some function $h(\boldsymbol{\theta})$ using the MCMC

sample $\boldsymbol{\theta}^{(n)}$ from the corresponding posterior distribution as $\sum_{n=1}^{N} w_{jk}^{(n)} h(\boldsymbol{\theta}^{(n)})$ and the value of the cumulative distribution function of $h(\boldsymbol{\theta})$ without hospital j in country k at a point a as $\sum_{n=1}^{N} w_{jk}^{(n)} \mathbb{I}\{h(\boldsymbol{\theta}^{(n)}) \leq a\}$.

However, it turns out that such weights might be too unstable which could lead to erroneous conclusions. This motivates us to define more flexible weights which allow us to choose a suitable function $g\left(\boldsymbol{Y}_{jk}^{\star}, \boldsymbol{Y}, \boldsymbol{m}\right)$ such that some terms in (3) almost cancel out when we multiply the original weights (3) with $g\left(\boldsymbol{Y}_{jk}^{\star}, \boldsymbol{Y}, \boldsymbol{m}\right)$. To be more specific, let

$$w_{(g)jk}^{\star}\left(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{m}\right) = \left[p\left(\boldsymbol{Y}_{jk}|\boldsymbol{Y}_{jk}^{\star}, \boldsymbol{\alpha}\right) p\left(\boldsymbol{Y}_{jk}^{\star}|\boldsymbol{\theta}_1, \boldsymbol{U}_{jk}\right) p\left(\boldsymbol{m}_{jk}|\boldsymbol{\theta}_2\right)\right]^{-1} g\left(\boldsymbol{Y}_{jk}^{\star}, \boldsymbol{Y}, \boldsymbol{m}\right), \tag{4}$$

for some $g\left(\boldsymbol{\theta}_{jk}, \boldsymbol{Y}, \boldsymbol{m}\right) > 0$. Then we can suggest more appropriate and stable importance weights, e.g. we can set

$$g\left(\boldsymbol{Y}_{jk}^{\star}, \boldsymbol{Y}, \boldsymbol{m}\right) = p\left(\boldsymbol{Y}_{jk}^{\star}|\bar{\mu}_{jk}\right), \tag{5}$$

where $\bar{\mu}_{jk}$ is the posterior mean of $\mu_{jk} = \beta_k + \boldsymbol{X}_{jk}^{\top}\boldsymbol{\beta}_X + \boldsymbol{M}_{jk}^{\top}\boldsymbol{\beta}_M + U_{jk}$.

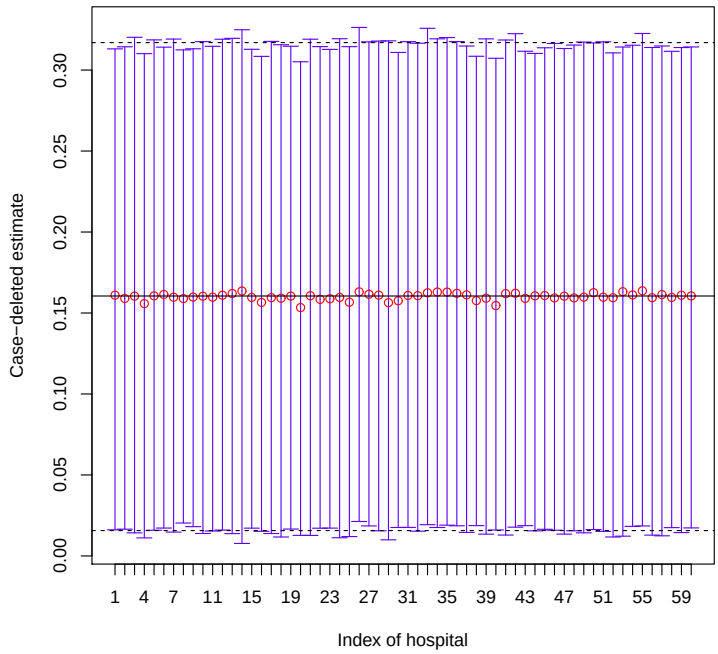


FIGURE 1. The estimate of the indirect effect of nurse working environment on the patient outcome without hospital j in Belgium (red points) and the credible intervals (whiskers) compared to the original point estimate (black solid line) and CI (dashed lines).

Let us note that the estimate computed with the importance weights (4) (that is $\sum_{n=1}^{N} w_{(g)jk}^{(n)} h(\boldsymbol{\theta}^{(n)})$) still converges in probability to the case-deleted posterior mean of $h(\boldsymbol{\theta})$.

To illustrate the method described above, we computed the hospital-deleted estimates of the indirect effect of nurse working environment for all Belgium hospitals. The hospital-deleted estimates based on the importance weights (4) with g defined as in (5) and the original estimate are depicted in Figure 1 together with the corresponding 95%-credible intervals. We can conclude that no estimate of the indirect effect is excessively influenced by any Belgium hospital. Similarly, we can estimate the country-deleted indirect effect with the importance weights given by the product of the importance weights corresponding to the hospitals in the the country. The estimates of the country-deleted indirect effects are shown in Figure 2. We conclude that none of these estimates differ from the posterior mean based on the whole dataset considerably so the model is stable with respect to the influence of particular hospitals and countries.

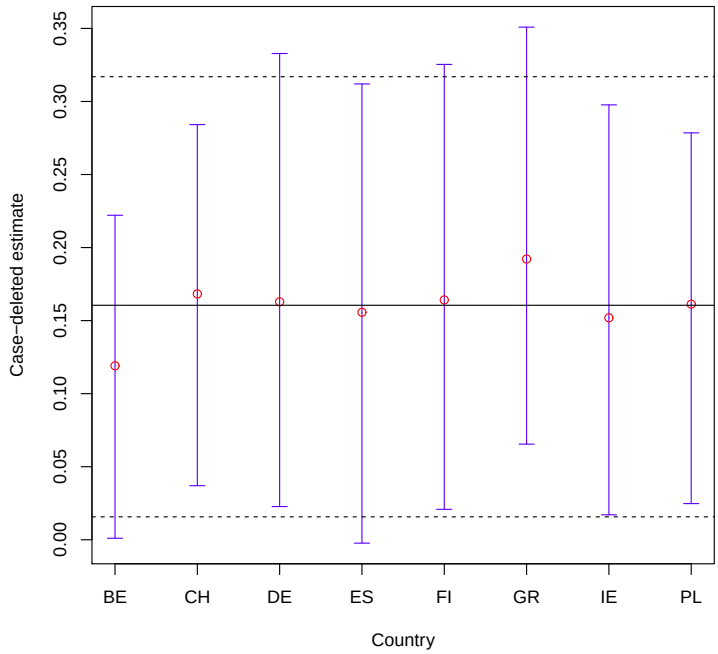


FIGURE 2. The estimate of the indirect effect of nurse working environment on the patient outcome without country k (red points) and the credible intervals (whiskers) compared to the original point estimate (black solid line) and CI (dashed lines).

Acknowledgments: The work was supported by the research grants GAUK 305–10/250689 and GACR 15-04774Y. Support from the Interuniversity Attraction Poles Programme (IAP-network P7/06), Belgian Science Policy Office, is also gratefully acknowledged.

References

Bradlow, E.T. and Zaslavsky, A.M. (1997). *Case Influence Analysis in Bayesian Inference. Journal of Computational and Graphical Statistics*, **6**(3), 314 – 331.

Rusá, Š. , Komárek, A., Lesaffre, E., Bruyneel, L. (2017). Multi-level moderation mediation model with ordinal outcome. *Submitted*.

Sermeus, W., Aiken, L. H., Van den Heede, K. et al. (2011). Nurse forecasting in Europe (RN4CAST): Rationale, design and methodology. *BMC Nursing*, **10**(6), 643 – 664.

Song, X. Y. and Lee, S.Y. (2012). *Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences*. Chichester, West Sussex: John Wiley.

Penalized complexity priors for varying coefficient models

Massimo Ventrucci[1], Maria Franco-Villoria[2], Håvard Rue[3]

[1] University of Bologna, Italy
[2] University of Torino, Italy
[3] KAUST, Saudi Arabia

E-mail for correspondence: massimo.ventrucci@unibo.it

Abstract: We discuss an illustrate the use of penalized complexity priors for varying coefficient models, introducing a natural base model choice that corresponds to a constant coefficient (linear model).

Keywords: Varying Coefficient models, PC prior; INLA; RW1.

1 Introduction

Varying coefficient models (VCMs, Hastie and Tibshirani (1993)) are useful in the presence of an *effect modifier*, a variable that "changes" the effect of a covariate of interest on the response. In practice, VCMs gain flexibility with respect to standard linear models by allowing one or more regression coefficients to vary over a covariate such as time or space. Consider the simple case where there are n observational units indexed by $i = 1, \ldots, n$ and one covariate x_i whose effect on the response y_i depends on another variable z_i; the latter could be a continuous variable (e.g. temperature) or a time/space index (day, region, etc). Assuming y_i belonging to the exponential family, the linear predictor of a generalized VCM is

$$\eta_i = \alpha + f(z_i)x_i \qquad i = 1, \ldots, n. \tag{1}$$

We follow a Bayesian hierarchical framework where the varying coefficient $f(z_i)$, $z_1 \leq \ldots \leq z_n$, in Eq. (1) is described by a vector of random effects $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^{\mathsf{T}}$ distributed at prior as a Gaussian Markov Random Field (GMRF, Rue and Held (2005)). A GMRF is a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and a sparse precision $\boldsymbol{Q}(\boldsymbol{\tau})$ that depends on some hyper-parameters $\boldsymbol{\tau}$

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

and whose non zero pattern specifies conditional dependencies among neighbouring random effects.

2 PC priors for varying coefficient models

Model (1) turns into a generalized linear regression model when the varying coefficient $f(z)$ is constant over z. We denote this model as the *base model*. Therefore, the VCM in equation (1) can be seen as a flexible extension of the base model. If we consider $f(z)$ in terms of the vector of random effects $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\mathsf{T}$ introduced in Section 1, the base model can be obtained setting the hyper-parameters $\boldsymbol{\tau}$ to a particular value (in Section 2.1 an example is discussed).

Elicitation of priors for precision parameters is a long standing topic in the literature on hierarchical Bayesian models. Simpson et al. (2016) recently introduced a new framework for building priors in Bayesian hierarchical additive models, denoted as *Penalized Complexity (PC) priors*. PC priors are computed based on specific principles in which a model component is seen as a flexible parametrization of a base model. The idea is to penalize model complexity, defined in terms of distance from the base model, in such a way that the base model is favoured unless the available data support a more flexible one. Let ξ denote the flexibility parameter and the base model be at $\xi = 0$, the complexity introduced by $\xi > 0$ is measured using the Kullback-Leibler divergence (Kullback and Leibler, 1951),

$$KLD(f||g) = \int f(x;\xi) \log\left(\frac{f(x;\xi)}{g(x;\xi=0)}\right) dx \tag{2}$$

for flexible model f and base model g. The PC prior is defined as an exponential distribution on the (transformed) KLD distance scale $d(\xi) = d(f||g) = \sqrt{2KLD(\xi)}$,

$$\pi(d(\xi)) = \lambda \exp(-\lambda d(\xi)) \tag{3}$$

The PC prior in the original parameter scale ξ follows by a change of variable transformation. For more details on PC priors we refer to the paper by Simpson et al. (2016). In the next section we focus in particular on the PC prior for the precision of a random walk, which is a suitable smoothing prior in the context of VCM.

2.1 The random walk case

A useful parametrization for the varying coefficient in Model (1) is $\theta_i = \beta + \delta_i$, where δ_i indicates deviation from the constant slope β at value z_i. We focus on the simple case of a random walk of order 1 (RW1) prior on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\mathsf{T}$, conditionally on precision τ, the joint density is specified as $\pi(\boldsymbol{\theta}|\tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n-1}(\theta_{i+1} - \theta_i)^2\right)$. The RW1 is a first-order intrinsic Gaussian Markov Random Field (Rue and Held (2005) ch. 3) and describes deviations from an arbitrary overall level. In the context of a VCM it is natural to interpret the latter as the constant slope β. In this sense, the RW1 characterizes a varying coefficient in a very intuitive way: it is a prior that shrinks towards a natural base model given by $f(z_i) = \beta$, $\forall i = 1, ..., n$, with τ controlling the amount of shrinkage. When $\tau^{-1} = 0$ we have $f(z_i) = \beta$,

which implies the linear regression model, $\eta_i = \alpha + \beta x_i$. For $\tau^{-1} > 0$, $f(z_i)$ incorporates higher degree of complexity w.r.t. the constant slope, leading to the flexible VCM. For a generic Gaussian Random effect conditional on τ, the PC prior for τ is the Gumbel type 2 with density (Simpson et al. 2016),

$$\pi(\tau) = 0.5\lambda\tau^{-1.5}\exp(-\lambda\tau^{-0.5}); \tag{4}$$

The parameter λ in Eq. (4) can be selected through a user-defined scaling approach. The user can encode the information available (at prior) on the degree of flexibility of the VCM model with respect to the base model. Simpson et al (2015) suggest eliciting the probability of a tail event regarding the marginal standard deviation, i.e. $Pr(1/\sqrt{\tau} > U) = a$, which yields $\lambda = -\log(a)/U$.

2.2 An illustrative example

We use a well known example from the seminal paper by Hastie and Tibshirani (1993). The data set includes 88 measurements from ethanol-fuelled engines on nitric oxide and nitrogen dioxide concentration (NO_x), equivalence ratio (E) and compression ratio (C). An initial exploratory analysis suggests that the effect of E on NO_x is highly non-linear and that the effect of C on NO_x may depend on E. Hastie and Tibshirani (1993) proposed the following model:

$$NO_x = f_1(E) + f_2(E)C + \epsilon,$$

where both the intercept and slope are continuous functions of E, with ϵ a normal error with noise variance σ^2. Both f_1 and f_2 are modelled with a RW1 prior conditional on precisions τ_1 and τ_2, respectively. To avoid scaling issues inherent in RW models, the precision matrix of the RW1 has been rescaled as described in Sørbye and Rue (2013). We assign the PC prior in Eq. (4) to each τ_i, with the λ_i selected following the practical rule of thumb suggested by Simpson et al. (2016): given the scale of the varying intercept and varying slope are expected to be different, we scale the PC prior for τ_1 such that $Pr(U_1 > 3) = 0.01$ (i.e. $\lambda_1 \approx 1.5$), and the one for τ_2 such that $Pr(U_2 > 0.3) = 0.01$ (i.e. $\lambda_2 \approx 15$). Note that the higher λ_i, the greater the penalty for deviating from the base model. PC priors can be implemented in `INLA` (Rue et al., 2009). We further assume $\pi(\sigma^2) = \text{InverseGamma}(a, b)$, with $a = 1$ and $b = 5 \cdot 10^{-5}$, since the posterior for the noise variance is usually robust to the prior.

Fig. (1) displays the fitted varying intercept and slope, with clear indication of variation over the range of the effect modifier E. The different scale of f_1 and f_2 justifies the use of different λ's for the two components. From Table (1) we see that the posterior $\pi(\tau_2|y)$ is robust for different parametrization of the PC prior. Similar results, not shown here, hold for $\pi(\tau_1|y)$.

3 Concluding remarks

Elicitation of a prior for the hyper-parameters $\boldsymbol{\tau}$ is a crucial aspect for practitioners who wish to specify a Bayesian VCM model, as $\boldsymbol{\tau}$ regulates the degree of flexibility of the VCM w.r.t. the base model. Regardless the chosen model for

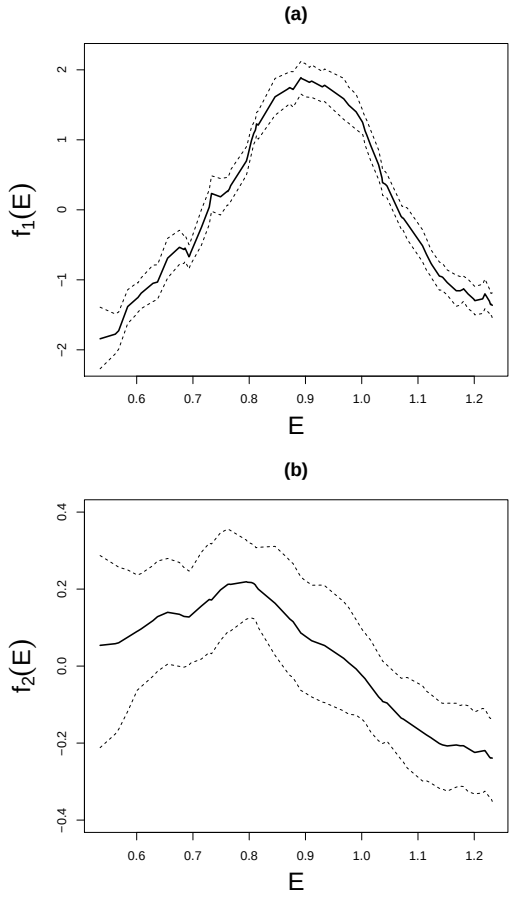


FIGURE 1. Fitted varying intercept (a) and slope (b).

TABLE 1. Posterior summaries for the precision of $f_2(E)$, τ_2, for different PC prior parameters (first column).

PC prior parameters;	$\pi(\tau_2\|y)$:	`0.025quant`	`mean`	`0.975quant`
$U_2 = 0.3$; $a = 0.01$ ($\lambda_2 = 15$)		16.76	69.25	189.22
$U_2 = 0.2$; $a = 0.01$ ($\lambda_2 = 23$)		16.69	69.1	188.8
$U_2 = 0.1$; $a = 0.01$ ($\lambda_2 = 46$)		16.61	68.94	188.52

the varying coefficient, a suitable PC prior shrinking to a sensible base model can always be defined through the application of predefined principles. PC priors avoid over-fitting by construction allowing the VCM model to arise only if data requires it. This is a desirable property as varying coefficients are typically over-parametrized models needing regularization. It is worth investigating the use of

PC priors in more complex models involving spatially varying coefficients.

References

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79 – 86.

Hastie, T. and Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society, Series B*, **55(4)**, 757 – 796.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman & Hall/CRC.

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(2)**, 319 – 392.

Simpson, D., Rue, H., Riebler, A., Martins, T. G. and Sorbye, S. H. (2016). Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). To appear in *Statistical Science*.

Sørbye, S.H. and Rue, H. (2013). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, **8**, 39 – 51.

Confidence intervals for posterior intercepts, with application to the PIAAC literacy survey

Jochen Einbeck[1], Elizabeth Gray[12], Nick Sofroniou[3], Antonio Hermes Marques da Silva Junior[14], Jacob Gledhill[15]

[1] Durham University, UK
[2] University of Bath, UK
[3] Institute for Employment Research, University of Warwick, UK
[4] Universidade Federal do Rio Grande do Norte, Natal, Brazil
[5] Department for Communities and Local Government, UK

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

Abstract: For variance component models, it is often the posterior estimate of the random effect ('posterior intercept') rather than the estimate of the fixed effect parameters, which is of main interest. This is the case, for instance, when ranking region–wise mortality rates (where the crude, regional rates are unreliable due to small observed counts) or for the construction of educational league tables from complex sample surveys. However, in order to be able to decide whether two cluster–level units can *actually be distinguished*, it is clear that one needs a measure of variability of these posterior intercepts. We present an exploration of methods to address this issue which appears to be still undeveloped in the context of the model class considered.

Keywords: Nonparametric maximum likelihood; Empirical Bayes shrinkage; Bootstrap

1 Posterior intercepts

Consider variance component models of type

$$\mu_{ij} = E(y_{ij}|z_i) = h(x_{ij}^{\mathrm{T}}\beta + z_i), \tag{1}$$

where μ_{ij} is the expected response for unit j in cluster i, x_{ij} are the fixed effect covariates which may depend on i, j, or both, and z_i is the random effect

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

operating at the cluster level. If no assumption on the random effect distribution is made, then estimation can be carried out via 'nonparametric maximum likelihood' (Aitkin, 1999). Briefly, the marginal likelihood is approximated by a discrete mixture, the parameters of which are estimated alongside with the fixed effect parameters via the EM algorithm, yielding estimates $\hat{\beta}$, $\hat{z}_1, \ldots, \hat{z}_k$ with masses $\hat{\pi}_1, \ldots, \hat{\pi}_K$. Denote by $\hat{\theta}$ the collection of these estimates, and by $y_i = (y_{i\cdot})$ the set of response values for cluster i. Aitkin (1999) suggested to estimate the mean of the posterior distribution $z_i|y_i$ via 'Empirical Bayes Predictions'

$$\tilde{z}_i = \sum_{k=1}^{K} w_{ik}\hat{z}_k, \tag{2}$$

where $w_{ik} = \hat{P}(k|\hat{\theta}, y_i)$ are the posterior probabilities ('responsibilities') that observation i stems from component k, which can be computed via Bayes' theorem from the parameter estimates $\hat{\theta}$ of the last M step. The quantity of interest are these posterior intercepts, $\tilde{z}_i$.

2 PIAAC data

The PIAAC survey of adult skills was carried out from 01/08/2011 to 31/03/2012 by the OECD in 24 countries (or sub–country entities), and was designed to assess the proficiency of adults in the key competencies of literacy, numeracy, and problem–solving in technology–rich environments. We focus here on the 'literacy' output variable, with six possible outcomes for an assessed individual. We dichotomized this variable as 'people reaching level 3 or above', with 'level 2 and below' being considered as low–skilled, which corresponds to the key European Commission policy marker used to demarcate poor basic skills in the complementary PISA survey carried out at 15–years of age (Eurostat, 2016). As covariates we will use gender, as well as a factor for age (covering the intervals 16–24, 25–34, 35–44, 45–54, and 55+), though we have also explored more complex models using employment status and reading habits which are not reported here. This leads to a (rescaled) logistic regression model $y_{ij} \sim \text{Bin}(n_{ij}, \mu_{ij})/n_{ij}$ where n_{ij} is the (effective) sample size of the jth subpopulation (defined by the covariate combination of interest) for country i, and function $h(\cdot)$ in (1) is the logistic function. Data were extracted using the PIAAC explorer. A model with age*gender interaction and $K = 5$ turned out to capture the upper-level heterogeneity well (Table 1 right).

3 Uncertainty of posterior intercepts

3.1 Analytic approximation

We initially approach the problem analytically, considering the weights w_{ik} in (2) as constants (which, strictly, they are not, since they depend on the parameters estimated in the last M-step). It follows then from (2) that

$$\text{Var}(\tilde{z}_i) = \sum_{k=1}^{K} w_{ik}^2 \text{Var}(\hat{z}_k) + \sum_{j\neq k} w_{ij}w_{ik}\text{Cov}(\hat{z}_j, \hat{z}_k) \tag{3}$$

where the variances and covariances are available from the fitted model according to standard GLM theory. Clearly, the covariance terms cannot be naively omitted since the positions of the $\hat{z}'_k s$ are strongly correlated. However, as $\sum_{k=1}^{K} w_{ik} = 1$ for all i, it is clear that $0 \leq w_{ij} w_{ik} \leq 1/4$ for all pairs $j \neq k$. In addition, it is often (but not always) the case that after EM convergence observations are classified to one of the components with probability equal or close to 1, in which case $w_{ij} w_{ik} \approx 0$. [To exemplify this point, Table 1 gives an excerpt of the matrix $W = (w_{ik})$ for the gender*age model with $K = 5$ components.] In either case, it is clear that the product $w_{ij} w_{ik}$ will be very small for all (or almost all) i, j, k with $j \neq k$, so that the 'naive' approximation

$$\mathrm{Var}(\tilde{z}_i) \approx \sum_{k=1}^{K} w_{ik}^2 \mathrm{Var}(\hat{z}_k) \tag{4}$$

will usually be a good one. Confidence intervals for the posterior intercepts are then obtained from either (3) or (4) via $\tilde{z}_i \pm q\sqrt{\mathrm{Var}(\tilde{z}_i)}$ where q is an appropriate quantile for which we use the 97.5% Gaussian quantile, 1.96.

3.2 NPML–Bootstrap

In order to assess the variability in a potentially more realistic way, we also developed a bootstrap routine which proceeds in two layers. Specifically, for $i = 1, \ldots, n$,

(i) from the set of mass points $\hat{z}_1, \ldots, \hat{z}_k$ draw a masspoint $\check{z}_i$ with probability w_{ik};

(ii) generate new $\check{y}_{ij} \sim \mathrm{Bin}(n_{ij}, \check{\mu}_{ij})/n_{ij}$, where $\check{\mu}_{ij}$ is defined in the natural way via (1), using $\hat{\beta}$ and $\check{z}_i$.

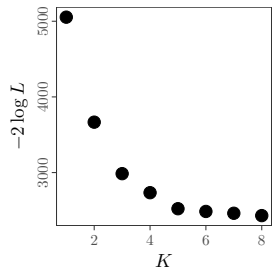
Having $\check{y}_{ij}$, we refit the model, yielding a new set of n posterior intercepts. Repeating these steps M times we have a bootstrap sample of estimates for posterior intercepts. Therefore, by taking the standard deviation of these we have an estimate for their variability.

3.3 Results

Figure 1 (left) gives the $\tilde{z}_i$ along with 'full' confidence intervals (3) and bootstrapped confidence intervals (using $N = 9999$). The analytic and simulation–based intervals are very similar, with minor differences only recognizable for a small subset of countries. The naive intervals (4) cannot be visually distinguished from the full intervals. Therefore, we provide in Figure 1 (right) the ratio of the widths of the naive and full intervals, as well as the bootstrapped and full intervals. All ratios are very close to 1, with slightly larger deviations for the bootstrapped intervals. We also see that five groups of countries can be robustly distinguished (since the corresponding intervals do not overlap), with Japan being the sole best–performing country.

TABLE 1. Left: Excerpt of matrix $W = (w_{ik})$ (4.d.p.) for age*gender model with $K = 5$; right: $-2 \log L$ as a function of K.

k	1	2	3	4	5
Australia	0	0	0	1	0
Austria	0	0.0023	0.9977	0	0
Canada	0	0	1	0	0
...					
Japan	0	0	0	0	1
Netherlands	0	0	0	1	0
...					
$\hat{z}_k$	-0.490	0.011	0.273	0.622	1.307



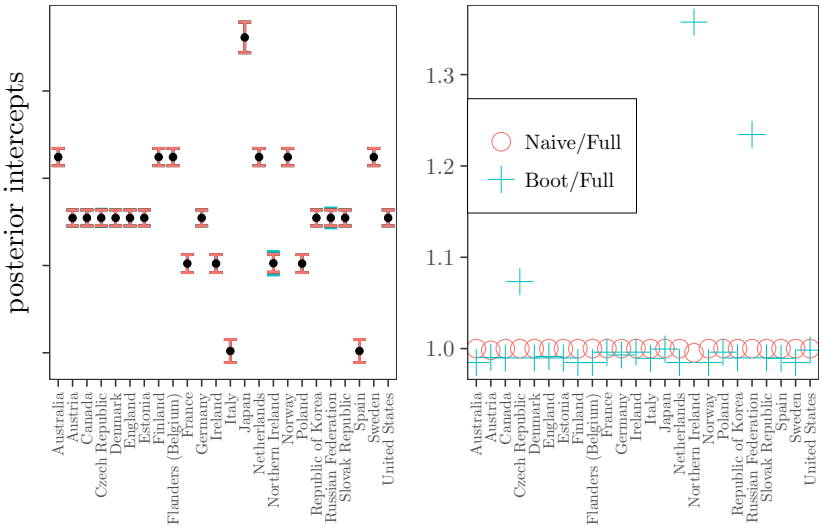


FIGURE 1. Left: Posterior intercepts [black dots] with analytic ('full') [inner interval; red in the online version] and bootstrapped intervals [outer; turquoise]; right: relative width of intervals.

4 Uncertainty of posterior probabilities

4.1 Sampling from posterior likelihood

A potential issue with the methodology discussed so far is that by plugging the ML parameter estimates $\hat{\theta}$ into the expression for w_{ik}, the uncertainty in those estimates is ignored. Hence, the 'certainty' of mass point allocation when taking the w_{ik} at face value can be considered as overstated. To address this problem, Aitkin et al. (2014) suggested the following procedure based on the concept of posterior likelihood (Aitkin, 2010):

a) Assuming flat priors for θ, the posterior distribution $p(\theta|y_1, \ldots, y_n)$ is proportional to the likelihood, $L(\theta)$. Hence, one can take M random draws $\hat{\theta}^{[m]}$, $m = 1, \ldots, M$, from $L(\theta)$.

b) Compute $w_{ik}^{[m]} = P(k|\hat{\theta}^{[m]}, y_i)$, $m = 1, \ldots, M$.

For our purposes, one would then proceed further,

c) Apply step (i) in the algorithm in subsection 3.2 using $w_{ik}^{[m]}$ instead of w_{ik} in the m–th bootstrap repetition.

However, the implementation is non–straightforward and will require a computationally expensive MCMC analysis, involving Gibbs samplers for each k, alternate draws between different types of model parameters, and ad–hoc solutions to the starting value and the label switching problems. Also, having already carried out a full EM procedure, a full–blown MCMC analysis only for the sake of analyzing the posterior intercepts feels rather out of scale. Hence, a simpler alternative idea is considered below.

4.2 Sensitivity assessment via EM process trail

As stated above, an EM algorithm has already been executed. As such, in this process, a series of 'draws' from the full likelihood $L(\theta)$ has been obtained. Assume that, in EM iteration $s = 1, \ldots, S$, we have obtained parameter estimates $\hat{\theta}^{[s]}$ with associated weight matrices $W^{[s]}$ and likelihoods $L^{[s]} \equiv L(\hat{\theta}^{[s]})$. Hence, we possess S draws from $L(\theta)$, including the final iteration, which corresponds to the MLE $\hat{\theta}^{[s]} \equiv \hat{\theta}$. While it is clear that these S draws in no way represent the correct shape of $L(\theta)$, the matrices $W^{[s]}$ can still be used to assess the *sensitivity* of the NPML–Bootstrap to imprecision in the $w_{ik}{}'s$, especially as some of the estimates along the EM process trail correspond to really 'bad' likelihoods (that is, estimates which have a likelihood of effectively 0 to be sampled in part a) of the Aitkin routine).

4.3 Results

For the data and model at hand, the number of required EM iterations turned out to be $S = 6$, and Figure 2 (left) shows $L(\hat{\theta}^{[s]})$ as a function of s. Figure 2 (right) shows the interval length of the NPML–bootstrapped confidence intervals when using $w_{ik}^{[s]}$, $s = 1, \ldots, 5$ relative to that using $w_{ik} \equiv w_{ik}^{[6]}$. It is clear that for all s corresponding to appreciable likelihoods the difference is less than 10%, and even for posterior weights corresponding to really poor likelihoods the increase is generally not more than 50%, indicating robust upper bounds for the uncertainty in this process.

The present paper has demonstrated the utility of bootstrap methods to characterize the sometimes substantial uncertainty in cluster–level estimates which commonly arises in league-table comparisons. While no claim is made that the relative magnitudes of the different intervals will in general behave in the manner of this particular case study, the tools proposed to arrive at this judgement are applicable for arbitrary two–level problems.

References

Aitkin, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. *Biometrics*, **55**, 117 – 128.

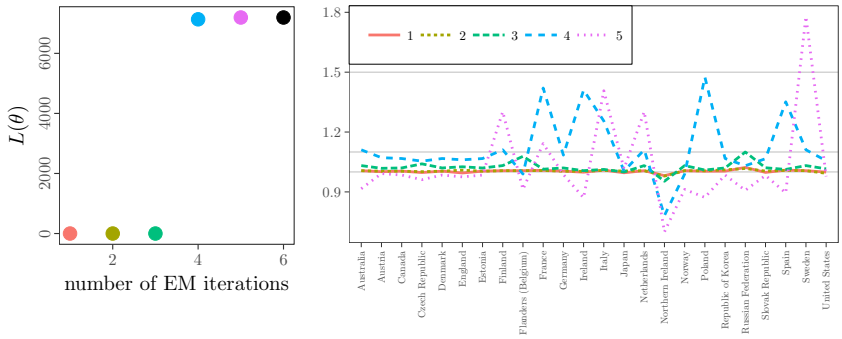


FIGURE 2. Left: Maximum Likelihood $L(\hat{\theta}^{[s]})$ versus EM iteration s; right: interval length using $w_{ik}^{[s]}$, $s = 1, \ldots, 5$ relative to using $w_{ik} \equiv w_{ik}^{[6]}$. The value s is given in the legend.

Aitkin, M. (2010). *Statistical Inference — an Integrated Bayesian/Likelihood Approach.* Chapman & Hall/CRC:Boca Raton.

Aitkin, M., Duy, V. and Francis, B. (2014). Statistical Modelling of the group structure for social networks. *Social Networks*, **38**, 74–87.

Eurostat (2016). Smarter, greener, more inclusive? Indicators to support the Europe 2020 strategy. *Publications Office of the European Union*, European Commission, Luxembourg.

Dealing with response styles in finite mixture models

Rosaria Simone[1], Gerhard Tutz[2]

[1] Department of Political Sciences, University of Naples Federico II, Italy
[2] Ludwig–Maximilians- Universität München, Germany

E-mail for correspondence: rosaria.simone@unina.it

Abstract: The objective of the present work is to propose a mixture model for ordinal data that is able to properly identify response styles with a tendency to choose middle or extreme categories, while accounting for uncertainty. The proposal is grounded on CUB models' paradigm and it is tested on a survey collected at the University of Naples Federico II, in which sports preferences are investigated.

Keywords: Uncertainty; Response Styles; Finite Mixture Models

1 Introduction and preliminaries

The data generating process yielding respondents to produce an ordinal evaluation over m categories out of their latent perception can be effectively modeled as the combination of two unobserved components: the actual *feeling* towards the item, and the *uncertainty* accounting to the inherent indeterminacy of every decision process, see Iannario and Piccolo (2010). Then, the response variable R_i of the i-th subject has probability distribution given by:

$$Pr(R_i = r|\mathbf{x}_i) = \pi Pr(Y_i = r|\mathbf{x}_i) + (1-\pi)Pr(U_i = r), \quad r = 1, \ldots, m. \qquad (1)$$

where $Pr(Y_i = r|\mathbf{x}_i)$ gives the preference part of the model, and $Pr(U_i = r)$ corresponds to the uncertainty - see Tutz *et al.* (2016). In particular, CUB models (D'Elia and Piccolo, 2005) structure the response variable R_i as a mixture between a shifted Binomial with parameter ξ for the preference part-$b_r(\xi)$- and a uniform distribution over the support $\{1, \ldots, m\}$:

$$Pr(R_i = r|\xi_i, \pi) = \pi b_r(\xi_i) + (1-\pi)\frac{1}{m}, \quad r = 1, \ldots, m, m > 3. \qquad (2)$$

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

The starting point of the extension proposed here are the findings in Gottard *et al.*(2016), where the authors introduce varying uncertainty in CUB models, experimenting with alternative choices to the uniform distribution. In this context, a dedicated analysis to response-styles is pursued, by presenting an extension of CUB models where the uncertainty component itself is designed to deal with symmetrical response styles, either with reference to extreme or to middle-scale categories. Similar considerations are explored in Tutz and Schneider (2016) with the Beta-Binomial distribution. Our proposal here is to adjust the uncertainty distribution by making use of the transformation:

$$Pr(U_i = r) = \frac{Pr(S_i = r) + c}{1 + m\,c}, \tag{3}$$

where $Pr(S_i = r)$ is a suitable probability distribution specifying the response styles attitude, and $c \neq 0$ is a chosen constant. A flexible tool to work on this task turns out to be a combination of the uniform distribution-needed to maintain the focus on heterogeneity and indecision due to the item- with the discretized Beta distribution.

Let $X \sim Beta(\alpha, \beta)$ be a Beta distributed random variable, with density:

$$f_X(x) = \frac{1}{B(\alpha, \beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\, dt, \qquad x \in [0, 1],$$

and the Euler Beta Function defined by:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Then we say that a discrete random variable $D = D(\alpha, \beta)$ over the support $\{1, \ldots, m\}$ has the Discretized Beta distribution (of parameters α, β) if:

$$Pr(D = r|\alpha, \beta) = Pr\left(\frac{r-1}{m} \leq X \leq \frac{r}{m}|\alpha, \beta\right), \quad r = 1, \ldots, m,$$

that is, if D is obtained from $X \sim Beta(\alpha, \beta)$ by dividing its range into m equal sub-intervals and then by integrating over them. When $\alpha = \beta$, the distribution is symmetric; in this case, when $0 < \alpha < 1$, the distribution is convex with two modes located at the extreme categories, whereas the distribution is concave and concentrated around the middle categories if $\alpha > 1$. When $\alpha = \beta = 1$, D collapses to the discrete uniform distribution on the given support.

By choosing $c = \frac{1}{m}$ in (3), and D as model for the (symmetric) response style component, we propose to adjust the uncertainty specification in CUB models by using a *C*ombination of a *L*ifted *U*niform and a shifted *B*inomial (CLUB, for short) distribution:

$$Pr(R = r|\xi, \pi, \alpha) = \pi b_r(\xi) + (1-\pi)\left(\frac{Pr(D = r|\alpha) + \frac{1}{m}}{2}\right), \qquad r = 1, \ldots, m. \tag{4}$$

Figure 1 shows, for fixed π and ξ parameters and $m = 7$, the effect on the CUB probability distribution (black solid lines) of including the response style component as determined by CLUB models, both for $\alpha < 1$ (top panels) and for $\alpha > 1$ (bottom panels).

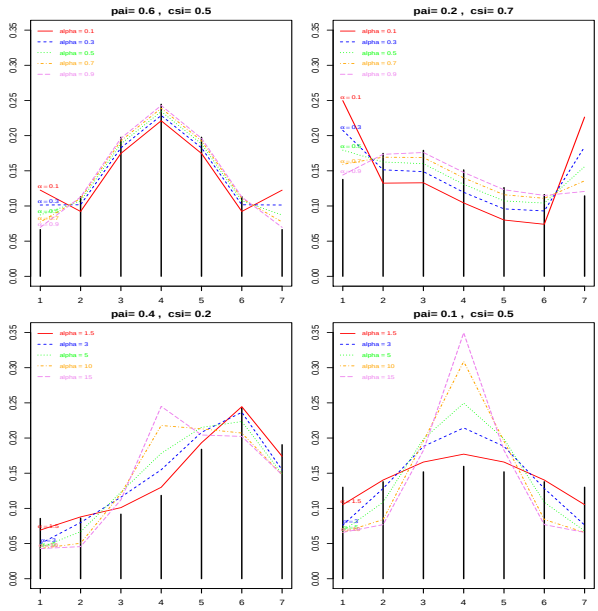


FIGURE 1. Extension from CUB to CLUB

Estimation of CLUB models is performed via Maximum Likelihood methods (ML) and requires a specific implementation of the EM algorithm -McLachlan and Krishnan (1997). Since CUB models are nested into CLUB (for $\alpha = 1$), the Likelihood Ratio Test is performed to check if the improvement of the fit is significant. Also, the symmetrized Kullback-Leibler divergence is computed to assess the discrepancy between observed and fitted distributions. Analytical derivation of the variance-covariance matrix of parameter estimates can be cumbersome due to the unavailability of a closed form for the discretized Beta distribution. Thus, standard errors are obtained either from numerical procedures or by sampling techniques like the bootstrap. The whole analysis is run in the R environment: the code is available upon request from the Authors, and makes use of Package CUB -Iannario *et al.*(2016)- for CUB models estimation.

2 An application to Sport preferences

The performances of CLUB models are discussed on a survey collected at University of Naples Federico II in May 2016. The questionnaire has been filled by $n = 647$ respondents, asked, among others questions, to rank their preferences for 8 sports: Football, Jogging, Volleyball, Tennis, Boxing, Swimming, Cycling and Basketball. Thus, marginally, each of these variables corresponds to rating measurements on a $m = 8$ points scale, of the type: *Rate your preference for the given sport, with 1 = absolutely preferred, and 8 = not at all preferred.* The frequency distributions of the ordinal variables Football and Jogging are U-shaped, thus giving evidence for an extreme category response styles, for which CLUB

models provide an impressive fit. For all items but Boxing and Swimming, there is empirical evidence that supports the inclusion of a middle categories response styles. Figure 2 displays, for each sport variable, the barplots of the observed frequency distributions, over-plotted against both the estimated CLUB distribution and the estimated CUB distribution.

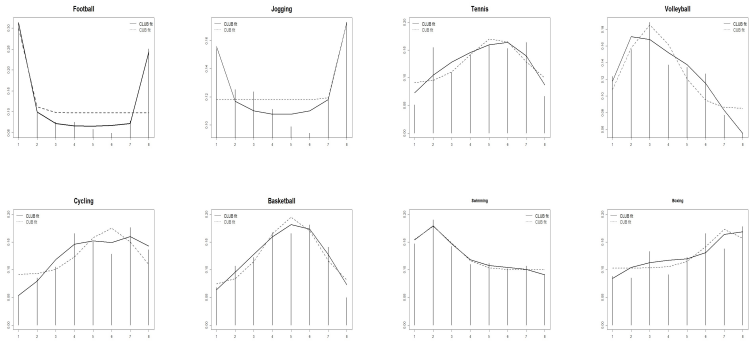


FIGURE 2. Frequency distributions of sport variables, with CLUB and CUB estimated probabilities

Table 1 reports parameter estimates (standard errors in parentheses) for both CLUB and the nested CUB models fitted to the sport variables, the corresponding log-likelihoods and likelihood ratio test statistics for the null $H_0 : \alpha = 1$ against the alternative $H_1 : \alpha \neq 1$, which will be asymptotically χ^2_1 distributed. Non significant values at level 0.05 are highlighted in bold.,

Sport	CLUB			Loglik CLUB	CUB		Loglik CUB	Loglik saturated	LRT
	$\hat{\pi}$	$\hat{\xi}$	$\hat{\alpha}$		$\hat{\pi}$	$\hat{\xi}$			
Football	0.098 *(0.032)*	0.928 *(0.050)*	0.100 *(0.040)*	-1195.342	0.217 *(0.024)*	0.99 *(0.009)*	-1266.47	-1192.766	142.25
Jogging	0.019 *(0.026)*	0.010 *(0.084)*	0.638 *(0.097)*	-1334.546	0.057 *(0.022)*	**0.004**	-1339.187	-1332.782	9.282
Tennis	0.118 *(0.046)*	0.259 *(0.056)*	2.087 *(0.291)*	-1317.944	0.275 *(0.049)*	0.392 *(0.033)*	-1328.925	-1304.701	21.96
Basket	0.181 *(0.079)*	0.351 *(0.061)*	0.543 *(2.352)*	-1303.905	0.414 *(0.048)*	0.430 *(0.021)*	-1309.782	-1299.535	11.75
Cycling	0.208 *(0.033)*	0.114 *(0.020)*	0.713 *(3.490)*	-1314.433	0.267 *(0.049)*	0.322 *(0.040)*	-1329.76	-1311.405	30.65
Volleyball	0.227 *(0.038)*	0.833 *(0.0267)*	2.920 *(0.593)*	-1313.173	0.319 *(0.046)*	0.685 *(0.028)*	-1320.619	-1310.847	14.89
Swimming	0.199 *(0.036)*	0.847 *(0.029)*	1.180 *(0.200)*	-1327.975	0.202 *(0.037)*	0.830 *(0.026)*	-1328.448	-1327.271	**0.946**
Boxing	0.165 *(0.037)*	0.092 *(0.040)*	1.440 *(0.307)*	-1331.424	0.177 *(0.040)*	0.157 *(0.036)*	-1332.456	-1322.767	**2.065**

TABLE 1. Parameter estimates for both CLUB and CUB models

Table 2 reports the (normalized) dissimilarity index between the observed (relative) frequency distribution of each of the sport variables and the estimated probability distributions both for CLUB and CUB models. Similarly, the symmetrized Kullback-Leibler divergence is computed. Both these indexes are strongly reduced when switching from CUB to CLUB. The better performances of the new proposal are supported also by the striking closeness between the maximized log-likelihood and the saturated one (see Table 1).

Sport	Dissimilarity		Symm. Kullback-Leibler	
	CLUB	CUB	CLUB	CUB
Football	0.028	0.166	0.004	0.107
Jogging	0.026	0.050	0.003	0.010
Tennis	0.074	0.095	0.020	0.038
Basket	0.045	0.068	0.007	0.016
Cycling	0.042	0.100	0.005	0.029
Volleyball	0.039	0.065	0.004	0.015
Swimming	0.021	0.029	0.001	0.002
Boxing	0.071	0.082	0.013	0.015

TABLE 2. Dissimilarity and Symmetrized Kullback-Leibler divergence between estimated probabilities and observed (relative) frequencies

2.1 CLUB models with covariates

CLUB model estimation can be enhanced by the inclusion of subjects' characteristics $\mathbf{z}_i$ to explain the occurrence and the effect of the response styles, to be linked to parameter α by means of a log link:

$$log(\alpha_i) = \nu_0 + \nu' \mathbf{z}_i, \qquad i = 1, \ldots, n.$$

For illustrative purposes, let Z be a dichotomous factor with level $Z = 0$ if respondent prefers individual sports and $Z = 1$ if he/she prefers team sports. Then, for Volleyball preferences:

$$log(\alpha_i) = \underset{(0.242)}{1.315} \; \underset{(0.352)}{-0.933} \, Z_i, \qquad i = 1, \ldots, n,$$

implying that the tendency to place Volleyball in the middle of the ranking is strongly associated with preferences for individual sports: indeed, α decreases from $\alpha^{(0)} = 3.723$ when $Z = 0$ to $\alpha^{(1)} = 1.46$ when $Z = 1$ (parameter estimates are significant according to the Wald test, and the log-likelihood attains an estimated maximum of -1310.07; see Table 1 for comparisons). For Jogging preferences, instead, one has:

$$log(\alpha_i) = \underset{(0.196)}{-0.764} + \underset{(0.261)}{0.735} Z_i, \qquad i = 1, \ldots, n,$$

implying that the tendency to place Jogging at the extreme ranks is strongly associated with preferences for individual sports: indeed, since α increases from $\alpha^{(0)} = 0.466$ when $Z = 0$ to $\alpha^{(1)} = 0.971$ when $Z = 1$ (again, parameter estimates are significant according to the Wald test, and the log-likelihood attains an estimated maximum of -1330.07; see Table 1 for comparisons).

3 Conclusions and future developments

The case study here discussed, as well as a simulation study run to validate the proposal, confirms the adequacy of CLUB models to properly identify response-style effects that are otherwise hidden by the uncertainty measure. Mis-specification yields biased estimates of the feeling parameter. The framework here explored to deal with response styles can be combined in a more general setting, currently under scrutiny, to analyze repeated measurements while modeling subjective heterogeneity.

References

D'Elia A. and Piccolo D. (2005). A mixture model for preference data analysis. *Comput. Stat. Data An.*, **49**, 917 – 934.

Gottard, A., Iannario, M. and Piccolo, D. (2016). Varying uncertainty in CUB models. *Advances in Data Analysis and Classification*, **10**(2), 225 – 244.

Iannario, M. and Piccolo, D. (2010). A New Statistical Model for the Analysis of Customer Satisfaction. *Quality Technology & Quantitative Management*, **7**(2), 149 – 168.

Iannario, M., Piccolo, D. and Simone, R. (2016). CUB: A Class of Mixture Models for Ordinal Data. (`R` package version 1.0). http://CRAN.R-project.org/package=CUB.

McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. New York: J.Wiley & Sons.

Tutz, G. and Schneider, M. (2016) Mixture Models for Ordinal Responses with a Flexible Uncertainty Component. *Technical Report*.

Tutz, G., Schneider, M., Iannario, M. and Piccolo, D. (2016). Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification*, DOI:10.1007/s11634-016-0247-9.

Extending the inferential capability of a generalised partial credit model using Bayesian computation: An application to an international disability survey developed by WHO and the World Bank

Sujit Sahu[1], Mark Bass[1], Carla Sabariego[2], Alarcos Cieza[3], Carolina Fellinghauer[4], Somnath Chatterji[5]

[1] Southampton Statistical Sciences Research Institute, University of Southampton, UK
[2] Department of Medical Informatics, Biometry and EpidemiologyIBE, Ludwig-Maximilians-University (LMU), Munich, Germany
[3] Department for Management of Noncommunicable Diseases, Disability, Violence and Injury Prevention, World Health Organization, Geneva 1211, Switzerland
[4] Swiss Paraplegic Research, Nottwil 6207, Switzerland
[5] Department of Information, Evidence and Research, World Health Organization, Geneva 1211, Switzerland

E-mail for correspondence: S.K.Sahu@soton.ac.uk

Abstract: Generalised partial credit models are ubiquitous in many applications in the health and medical sciences that use item response theory. Such polytomous item response models have a great many uses ranging from assessing and predicting an individual's ability to ordering the items to test the effectiveness of the test instrumentation. By implementing these models in a full Bayesian framework, computed through the use of Markov chain Monte Carlo methods, this article extends their inferential capability in three distinct ways. First, the models are extended to include covariate effects thus allowing simultaneous estimation of regression and item parameters. Secondly, full Bayesian methods for ranking the items using the Fisher information criterion (FIC) are developed. This allows us to fully propagate and ascertain uncertainty in the inferences by calculating the item specific FIC which facilitates the item ordering. Lastly, we propose a Monte Carlo method for predicting the ability score of a new individual by approximating the relevant Bayesian predictive distribution. Data from a Model Disability

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

Survey carried out in Sri Lanka by the World Health Organization (WHO) and the World Bank are used to illustrate the methods. The proposed approaches are shown to provide simultaneous model based inference for all aspects of disability which can be explained by environmental and socio-economic factors.

Keywords: Bayesian Methods, Education Testing, Hierarchical Modelling, Item ranking, Item Response Theory

1 Extended abstract

This paper sets out to achieve three extended inferential tasks when the GPCMs are employed. The first extension enables estimation of the item parameters adjusted for the covariate effects. Inference for the covariate effects has also been illustrated with the WHO and the World Bank Model Disability Survey data. A Bayesian approach based on a single model, in contrast to a stage-wise procedural estimation method, allows us to accurately assess the uncertainties not only for the item parameters but also for the regression parameters. This paper illustrates the methods with the main effects model only, but the methods can also be applied to higher order interaction models.

The second inferential extension task ranks the items so that a brief version of the disability survey with fewer items can be prepared. Using an expected FIC we have developed a method for ranking the items. It is up to the practitioner to decide how many items can be afforded in the reduced survey and we acknowledge that there may be other practical considerations which may influence the final item choice. The proposed method will guide item selection based on a desired percentage of information that must be present in the reduced survey.

The third inferential extension is the MCMC based methodology to predict the ability scores of new individuals whose data are observed after model fitting has already been performed. The methodology uses all the relevant covariate information of each new individual so that the best possible Bayesian estimates are obtained. The proposed prediction methodology has been empirically verified by re-estimating the ability scores of a large number (1000) of new individuals by fitting the model to the full data set. Close agreement between the predicted scores and the estimated scores based on all the data shows the effectiveness of the new methodology.

These three methodological extensions allow us to extract a lot more information from the data than what has been possible before. Using a unified model it is concluded that the main effects of the three covariates: gender, age and income are significant in the presence of, hence accounting for, the latent ability, item discriminatory and item difficulty parameters. In addition, the main advantage of the unified model also lies in its ability to make coherent inference on item ordering and ability score prediction for new individuals. By eliminating a stage-wise procedure for the three different inferential tasks, the developed Bayesian methodology proposes a rigorous and coherent inference framework wherever GPCM models are to be used in practice. This framework ensures coherency by having the correct and mutually consistent levels of uncertainty in the three different inferential tasks.

Response Styles in the Partial Credit Model

Gunther Schauberger[1], Gerhard Tutz[1], Moritz Berger[2]

[1] LMU Munich, Germany
[2] University of Bonn, Germany

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

Abstract: Ignoring the response style in the modelling of ordinal responses in psychological measurement can lead to poor and considerably biased estimates of item parameters. This work focuses on the modelling of a tendency to extreme or middle categories in item response data with Likert scales. An extension of the Partial Credit Model is proposed that explicitly accounts for this specific response style. The proposed model contains person-specific response style parameters which are estimated using the framework of generalized mixed linear models. The method is applied to a real data example to illustrate the effect of ignoring response styles.

Keywords: Partial credit model; Likert-type scales; Rating scales; Response styles.

1 The Partial Credit Model

Let the categories $0, \dots, k$ represent graded agree-disagree attitudes with a natural symmetry like *strongly disagree, ..., strongly agree* where $Y_{pi} \in \{0, 1, \dots, k\}$, $p = 1, \dots, P$, $i = 1, \dots, I$ denotes the ordinal response of person p on item i. Then the partial credit model (PCM) as proposed by Masters (1982) is denoted by

$$\log\left(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r-1)}\right) = \theta_p - \delta_{ir}, \quad r = 1, \dots, k$$

where θ_p is the person parameter and $(\delta_{i1}, \dots, \delta_{ik})$ are the item parameters of item i. It can be seen that the model is locally (given response categories $r-1$, r) a binary Rasch model with person parameter θ_p and item difficulty δ_{ir}.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

2 The Partial Credit Model with Response Style

Response styles that account for a tendency to extreme categories or a tendency to middle categories are modeled by modifying the parameter δ_{ir}. An additional person parameter γ_p is introduced that describes the response style of the person. The resulting partial credit model with response style (PCMRS) has the form

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right)=\theta_p+(m-r+0.5)\gamma_p-\delta_{ir},\quad r=1,\ldots,k$$

for an odd number of categories (k is even) and

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right)=\theta_p+(m-r)\gamma_p-\delta_{ir},\quad r=1,\ldots,k$$

for an even number of categories. Here, m represents the middle category where is $m=k/2$ for an odd number of categories and $m=[k/2]+1$ for an even number of categories. The parameter γ_p can be seen as a shifting of thresholds. If γ_p is positive , the thresholds δ_{ir} shift away from each other and, therefore, result in a higher tendency for person p to middle categories. The extreme case $\gamma_p\to\infty$ yields $P(Y_{pi}=m)\to 1$. If γ_p is negative one has the reverse effect; the person has a tendency to the extreme categories. For illustration we show

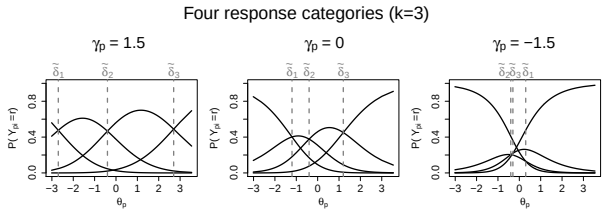


FIGURE 1. Probabilities $P(Y_{pi}=r)$ against θ_p for positive, negative γ_p and $\gamma_p=0$; in the upper panels the number of categories is three, in the lower panels it is four.

in Figure 1 the probabilities of response categories as functions of the person abilities for varying response style parameter γ_p in the case $k=3$ (four response categories). It is immediately seen that for $\gamma_p=-1.5$ the extreme categories have much higher probabilities than for $\gamma_p=0$. The inverse is seen for $\gamma_p=1.5$. It is noteworthy that the probabilities of adjacent categories are now equal, that is, $P(Y_{pi}=r)=P(Y_{pi}=r-1)$, if $\theta_p=\tilde{\delta}_{ir}$ where $\tilde{\delta}_{ir}=\delta_{ir}-(m-r+0.5)\gamma_p$ is the new (person-specific) threshold parameter. The basic concept, explicit modelling of a tendency to middle or extreme categories, has been used before by Tutz and Berger (2016) who considered just one item and modelled the effect of covariates.

The model is estimated using marginal likelihood estimation. For the person parameters, a two-dimensional normal distribution $N(\mathbf{0},\boldsymbol{\Sigma})$ is assumed. The diagonals of the matrix $\boldsymbol{\Sigma}$ contain the variance of the response style parameters σ^2_γ and the variance of the person effects, σ^2_θ, the off diagonals are the covariances between response style and location effects, $cov_{\gamma\theta}$. Therefore, one also allows for a correlation between both person parameters. In the marginal likelihood approach,

the common density of the person parameters is integrated out numerically by Gauss-Hermite integration.

3 Illustrative example

In our applications we use data from the standardiziation sample of the Freiburg Complaint Checklist (FCC) (Fahrenberg, 2010). The FCC is a questionnaire that is used to assess physical complaints of adults, we will focus on the scale *tenseness*. The data set contains 2070 participants (2032 complete cases). Each of the 9 items from the scale *tenseness* is measured on a 5-point response scale that refers to the frequency of the complaint: "never", "about 2 times a year", "about 2 times a month", "approximately 3 times a week" or "almost every day".

Figure 2 shows the estimates for the item parameters. The solid lines represent estimates for the PCMRS and the dashed lines represent estimates for the PCM. The estimates for the PCMRS are more stable across the categories. Figure 3

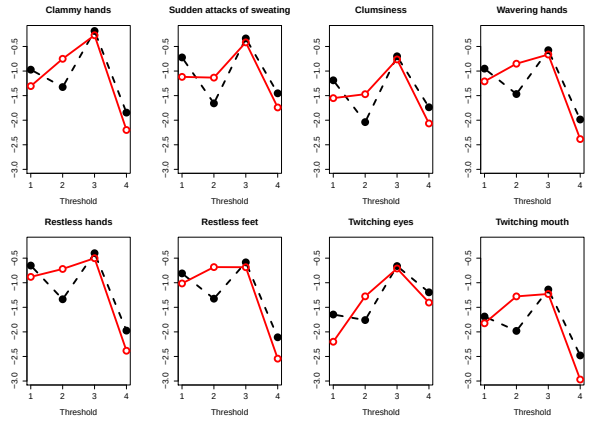


FIGURE 2. Estimates of item parameters for items on tenseness (FCC), separately for each item. Solid line represents estimates for PCMRS, dashed line represents estimates for PCM.

shows the response curves, exemplary for the item "Clammy hands", for three different response style parameters (columns) along the person parameter θ_p. Overall, the last category has very high probabilities, whereas the first category has very low probabilities. In particular, for a person with a rather high tendency to the extremes ($\gamma_p = -\hat{\sigma}_\gamma$), the probability for the last category is above 0.6 throughout the whole range of person parameters.

In a regular PCM, the estimated variance of the person parameters $\hat{\sigma}^2 = 0.546$ while the estimated covariance matrix between person and response style parameters for the PCMRS is

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\sigma}_\theta^2 & \hat{cov}_{\gamma\theta} \\ \hat{cov}_{\gamma\theta} & \hat{\sigma}_\gamma^2 \end{pmatrix} = \begin{pmatrix} 0.449 & 0.263 \\ 0.263 & 1.172 \end{pmatrix}.$$

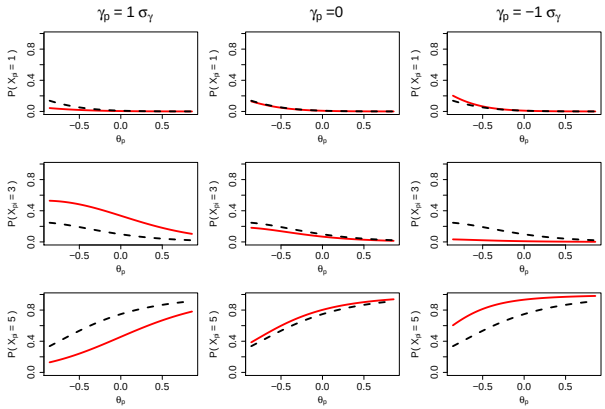


FIGURE 3. Response curves for item "Clammy hands" (FCC) along person parameter θ for categories 1, 3 and 5 and different response styles. Solid lines represent estimates for PCMRS, dashed lines represent estimates for PCM.

The variance (i.e. magnitude) of the response style effects is quite high. Therefore, ignoring the response style in such a case would lead to severely biased item estimates. The covariance of 0.263 between both person parameters refers to a noteworthy correlation of 0.36.

4 Concluding Remarks

The proposed PCMRS model has several advantages, in particular in comparison to mixture models as for example proposed by Eid and Rauber (2000). Mixture models assume that respondents come from different latent classes. Different item response models are fitted within different classes, some may represent the substantive trait, some may represent response style behaviour. One of the problems with mixture models is always the number of classes, which is unknown. Typically one gets quite different models if one fits, for example, two or three classes, since all the parameters change when considering one more class. If one has chosen a number of classes it is still difficult to interpret the difference between classes and explain what feature is represented by a class, it might be a response style or some other dimension that is involved when responding to items. Since the classes are not pre-specified, for example, by explicitly modelling response style behaviour, there is much uncertainty involved and the interpretation of the model within classes often tends to be vague. In contrast, in the PCMRS model the explicit modelling of the response style allows to decide if it is present, and if, how strong it is.

References

Eid, M. and M. Rauber (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, **16(1)**, 20–30.

Fahrenberg, J. (2010). Freiburg Complaint Checklist [Freiburger Beschwerdenliste (FBL)]. *Göttingen, Hogrefe*.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47(2)**, 149 – 174.

Tutz, G. and M. Berger (2016). Response styles in rating scales - simultaneous modelling of content-related effects and the tendency to middle or extreme categories. *Journal of Educational and Behavioral Statistics*, **41(3)**, 239 – 268.

Semi-parametric ordinal regression models for continuous scales

Maurizio Manuguerra[1], Gillian Heller[1], Jun Ma[1]

[1] Statistics Department, Macquarie University, Sydney, Australia

E-mail for correspondence: maurizio.manuguerra@mq.edu.au

Abstract: We extend the ordinal regression model for continuous scales by introducing non-parametric terms in the linear predictor. The model parameters are estimated using constrained optimisation of the penalised likelihood and the penalty parameters are automatically selected via maximisation of their marginal likelihood. An application of the model in a study on alternative chemotherapy treatments for advanced breast cancer is shown. The outcome of interest, the quality of life of patients, is modelled with fixed effects, smoothing terms and individual random effects and its association with the treatment arm over the chemotherapy cycle is estimated. The methodology has been implemented in an R package, ordinalCont, available on CRAN.

Keywords: Ordinal regression; Continuous scale; VAS.

1 Introduction

Continuous self-rating scales are commonly used to evaluate the intensity of outcomes which are intangible and difficult to measure, such as pain and quality of life. Two examples of such scales are the Visual Analogue Scale (VAS) and the Linear Analog Self-Assessment (LASA) scale, used in the pain literature and in quality of life (QoL) studies respectively. Subjects are typically given a linear scale of 100 mm and asked to put a mark where they perceive themselves. The relevance of continuous self-rating scales in the pain literature has been described in Heller et al. (2016), where the frequent use of suboptimal methods in the analysis of VAS and the superior power of ordinal regression analysis (Manuguerra and Heller, 2010) are discussed. In this paper we extend the formulation of the ordinal regression model for continuous scales, including fixed effects, random effects and smoothing terms, explain how to compute the constrained maximum penalised likelihood (MPL) estimates of the regression coefficients and show how

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

to automatically select the optimal smoothing parameters. We also show an example in which the new formulation of the model allows for a more insightful analysis of the data.

2 Methods

2.1 Ordinal regression for continuous scores

In the ordinal regression model for continuous scales, the covariates are modelled on a latent scale W. This is linked to the observed ordinal scale V by an increasing function g (the "g function") such that $W = g(V)$. We assume a stochastic model for W of the form of $W(\eta) = h^\star(\eta) + \epsilon$, where $h^\star(\eta)$ is the deterministic part of the model that depends on the covariates η, and ϵ is a random error term. In the following, we will assume the standard logistic distribution for ϵ, but other distributions, such as the normal, can be used. The cumulative distribution function for the score V can then be written as:

$$\gamma(v|\eta) = P(W \leq g(v)|\eta) = P(\epsilon \leq g(v) - h^\star(\eta)) = \frac{e^{h(v,\eta)}}{1 + e^{h(v,\eta)}} \tag{1}$$

where $h(v, \eta) = g(v) - h^\star(\eta)$ and $g(v) = \sum_u \theta_u \Psi_u(v)$ is modelled with monotonic I-splines (Ramsey, 1988). This guarantees an increasing g function when $\theta_u \geq 0, \forall u$. Assuming p fixed effects, r random effects and s smoothing terms, the predictor can be written as:

$$h^\star(\eta) = X\beta + \sum_{j=1}^{r} Z_j b_j + \sum_{l=1}^{s} \Phi_l(y_l)\vartheta_l = \sum_{m=1}^{\mathrm{T}} A_m \eta_m \tag{2}$$

where $X^{\mathrm{T}} = \left\{X_1^{\mathrm{T}}, \ldots, X_p^{\mathrm{T}}\right\}$ is a $p \times n$ design matrix for the fixed effects, Z_j is the design matrix for the j^{th} random effect and Φ_l is the matrix containing the basis functions (B-splines) evaluated at the values of the covariate y_l. Here η is the set of parameters $\left\{\eta_1^{\mathrm{T}}, \ldots, \eta_t^{\mathrm{T}}\right\}^{\mathrm{T}} = \left\{\beta^{\mathrm{T}}, b^{\mathrm{T}}, \vartheta^{\mathrm{T}}\right\}^{\mathrm{T}}$, with $\beta^{\mathrm{T}} = \left\{\beta_1, \ldots, \beta_p\right\}$, $b^{\mathrm{T}} = \left\{b_1^{\mathrm{T}}, \ldots, b_r^{\mathrm{T}}\right\}$, $\vartheta^{\mathrm{T}} = \left\{\vartheta_1^{\mathrm{T}}, \ldots, \vartheta_s^{\mathrm{T}}\right\}$, $t = p + r + s$ and A_m is the m^{th} element in $\left\{X_1, \ldots, X_p, Z_1, \ldots Z_r, \Phi_1, \ldots \Phi_s\right\}$.

2.2 Penalised likelihood estimation

The log-likelihood for subject i is obtained by differentiating equation (1):

$$\ell_{0i} = \log\Big(\sum_{u=1}^{m} \theta_u \Psi'_u(v)\Big) + h(v, \eta) - 2\log\left(1 + e^{h(v,\eta)}\right) \tag{3}$$

The model parameters we want to estimate are $\eta = \left\{\eta_1^{\mathrm{T}}, \ldots, \eta_t^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\theta = \left\{\theta_1, \ldots, \theta_m\right\}^{\mathrm{T}}$. Their constrained MPL estimates are defined by

$$(\hat{\eta}, \hat{\theta}) = \arg\max_{\eta,\theta}\Big\{\ell_p = \sum_{i=1}^{n} \ell_{0i} - \lambda_g J_g(\theta) - \sum_{m=1}^{t} \lambda_m J_m(\eta_m)\Big\} \tag{4}$$

where the J's are penalty functions, such as roughness penalties, the λ's are smoothing parameters and the first p terms in the last sum account for penalty terms relative to the fixed effects, and are then zero. The roughness penalties can be written in general as $J(\theta) = \theta^{\mathrm{T}} R\theta$ where the R terms are square matrices with j, k elements given by $\int \Psi_j''(y)\Psi_k''(y)dy$. For the random effect terms, R is a unit matrix and the penalty term is $J(b) = b^T R_b b = b^T b$, which implies normally distributed random effects b with variance $\sigma_b^2 = \frac{1}{2\lambda_b}$.

2.3 Estimation of model parameters

The estimation procedure iterates through two steps repeated until convergence. First, given the current values of the λ's, η and θ are estimated, then given the current value of θ and η the λ's are estimated. We comment that this is a constrained optimisation as $\theta \geq 0$.

The Karush-Kuhn-Tucker (KKT) necessary conditions for the constrained MPL estimation of θ and η are:

$$\begin{aligned} &\frac{\partial \ell_p}{\partial \eta_m} = 0 \text{ for any } m \\ &\frac{\partial \ell_p}{\partial \theta_u} = 0 \text{ if } \theta_u > 0 \text{ and } \frac{\partial \ell_p}{\partial \theta_u} < 0 \text{ if } \theta_u = 0 \end{aligned} \tag{5}$$

Equations (5) are solved iteratively, with the unconstrained parameters η estimated using a Newton method and the positively constrained parameters θ estimated using the multiplicative iterative (MI) algorithm (Ma, 2010).

$\boldsymbol{\eta}$: Newton step Given the estimated values of η and θ at iteration k, the values of each $\eta_m \in \eta$ at iteration $k+1$ are obtained with the Newton algorithm:

$$\eta_m^{(k+1)} = \eta_m^{(k)} + \omega^{(k)} \Big(\nabla^2_{\eta_m} \ell_p(\eta^{(k)}, \theta^{(k)}) \Big)^{-1} \Big(\nabla_{\eta_m} \ell_p(\eta^{(k)}, \theta^{(k)}) \Big) \tag{6}$$

where $\omega^{(k)} \in (0, 1]$ is a line search step size that guarantees that $\ell_p(\eta^{(k+1)}, \theta^{(k)}) \geq \ell_p(\eta^{(k)}, \theta^{(k)})$.

The first and the second derivative of the penalised log-likelihood (4) with respect to η are:

$$\begin{aligned} \nabla_{\eta_m} \ell_p &= A_m^T \Big(1_n - 2\gamma(v) \Big) - 2\lambda_m R_m \eta_m \\ \nabla^2_{\eta_m \eta_n} \ell_p &= A_m^T D A_m - 2\delta_{m,n} \Big(\lambda_m R_m \Big) \end{aligned} \tag{7}$$

where D is the diagonal matrix with $\mathrm{diag}(D) = -2\frac{e^h}{(1+e^h)^2}$ and $\delta_{m,n} = 1$ if $m = n$, and 0 otherwise.

θ: MI algorithm Given the estimated values of η at iteration $k+1$ and θ at iteration k, the values of θ at iteration $k+1$ are obtained with the MI algorithm. When $\theta > 0$, by the second KKT condition (5):

$$\nabla_\theta \ell_p = \Psi^T \Big(1_n - 2\gamma(v)\Big) + \Psi'^T E - 2\lambda_\theta R_\theta \theta = 0 \tag{8}$$

where E is a diagonal matrix with $\text{diag}(E) = \frac{1}{g'(v)}$ and 0 otherwise.,
Following Ma (2010), it is easy to derive the updating algorithm:

$$\theta^{(k+1)} = \theta^{(k)} + \omega_\theta^{(k)} s^{(k)} \nabla_\theta \ell_p(\eta^{(k+1)}, \theta^{(k)}) \tag{9}$$

where $s^{(k)} = \frac{\theta^{(k)}}{2\Psi^{\mathrm{T}}\gamma(v) + 2\lambda_\theta [R_\theta \theta]^+}$ and a line search of size $\omega^{(k)} \in (0, 1]$ is introduced to guarantee that $\ell_p(\eta^{(k+1)}, \theta^{(k+1)}) \geq \ell_p(\eta^{(k+1)}, \theta^{(k)})$.

2.4 Automatic smoothing parameter estimation

Fixing the values of η to the most current estimate, the optimal smoothing parameters are the roots of the partial derivatives with respect to each λ_m of the marginal posterior computed integrating out θ and η from (4). The log-posterior can be written as:

$$\Phi = \sum_{i=1}^{n} \ell_{0i} - \Big(\frac{p_\theta}{2}\log(\sigma_\theta^2) + \frac{1}{2\sigma_\theta^2}\theta^T R\theta\Big) - \sum_{m=1}^{t}\Big(\frac{p_m}{2}\log(\sigma_m^2) + \lambda_m \eta_m^T A_m \eta_m\Big) \tag{10}$$

and the marginal log-posterior is obtained integrating out θ and η:

$$\Phi_{\text{marg}} = \log \int \exp\Big(\Phi(\theta, \eta; \sigma_\theta^2, \sigma_\eta^2) d\theta d\eta\Big) \tag{11}$$

Exact solution to this interval is infeasible but it can be approximated following Laplace:

$$\Phi_{\text{marg}} = -\frac{p_\theta}{2}\log(\sigma_\theta^2) - \sum_{m=1}^{t}\frac{p_m}{2}\log(\sigma_m^2) + \ell_p(\hat{\theta}, \hat{\eta}; \sigma_\theta^2, \sigma_\eta^2) - \frac{1}{2}\log\Big|-Q(\hat{\theta}, \hat{\eta}; \sigma_\theta^2, \sigma_\eta^2)\Big| \tag{12}$$

where $Q(\theta, \eta; \sigma_\theta^2, \sigma_\eta^2) = \nabla^2_{\theta\eta}\ell_p(\theta, \eta; \sigma_\theta^2, \sigma_\eta^2)$. Solving

$$\frac{\partial \Phi_{\text{marg}}}{\partial \sigma_\theta^2} = 0 \tag{13}$$

$$\frac{\partial \Phi_{\text{marg}}}{\partial \sigma_{\eta_m}^2} = 0 \tag{14}$$

for σ_θ^2 and $\sigma_{\eta_m}^2$ we obtain the desired smoothing parameters estimates:

$$\sigma_\theta^2 = \frac{\theta^T R_\theta \theta}{p_\theta - \mathrm{tr}\Big((G - Q_\theta - \sum Q_m)^{-1} Q_\theta}\Big) \tag{15}$$

$$\sigma_m^2 = \frac{\eta_m^T R_m \eta_m}{p_m - \mathrm{tr}\Big((G - Q_\theta - \sum Q_m)^{-1} Q_m}\Big) \tag{16}$$

3 Example

Metastatic breast cancer is the most common cause of cancer death among Australian women. The ANZ0001 trial is a randomized trial with three chemotherapy

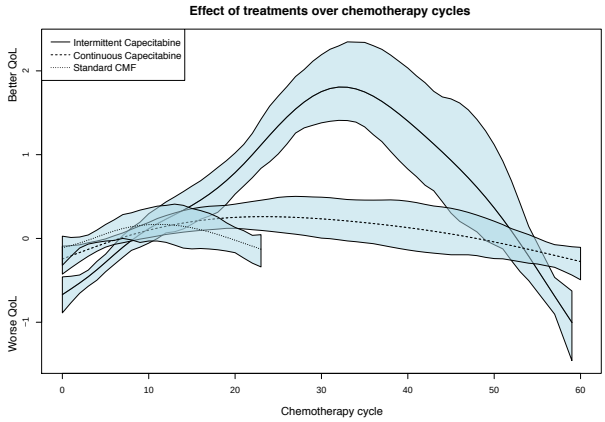


FIGURE 1. Effects of treatments over chemotherapy cycles

treatment arms ($n = 292$ patients with complete quality of life measurements) concluded in 2005 (Stockler et al., 2011). QoL is assessed at each chemotherapy treatment cycle on a LASA scale. The treatments Intermittent Capecitabine (IC) and Continuous Capecitabine (CC) are compared with the standard combination treatment CMF, each with its own protocol. There is no maximum duration of treatment, but it is interrupted on disease progression, or when patient intolerance or unacceptable toxicity are recorded. In this analysis we aim to verify which treatment has a better impact on QoL, and in particular how this impact changes over chemotherapy cycle. The research questions can be answered by modelling the overall QoL v_{ij} for patient i at chemotherapy cycle j as:

$$\log \frac{\gamma_{ij}}{1 - \gamma_{ij}} = g(v_{ij}) + x_i\beta + s(j|t_i)$$

where x_i is the age of patient i, $s(j|t_i)$ is a smooth term that depends on cycle number j and the treatment arm t_i. In Figure 1, the effects of the three treatments over chemotherapy cycles are shown. The CC treatment has no clear advantage over the CMF treatment in terms of QoL, but can be sustained for longer periods,

while the IC treatment has a positive effect on the QoL, with a peak after about 30 cycles (higher values mean higher quality of life).

All the analyses have been performed in the R Statistical Software (R Core Team, 2016) using the package ordinalCont (Manuguerra and Heller, 2016). The data set used in the analysis is included in the package.

References

Heller, G.Z., Manuguerra, M., and Chow, R. (2016). How to analyze the Visual Analogue Scale: Myths, truths and clinical relevance. *Scandinavian Journal of Pain*, **13**, 67 – 75.

Ma, J. (2010) Positively constrained multiplicative iterative algorithm for maximum penalized likelihood tomographic reconstruction. *IEEE trans. on Nucl. Sci.*, **57**(1), 181 – 192.

Manuguerra, M. and Heller, G.Z. (2010). Ordinal Regression Models for Continuous Scales. *The International Journal of Biostatistics*, **6**(1).

Manuguerra, M., and Heller, G.Z. (2016). ordinalCont: Ordinal Regression Analysis for Continuous Scales. *R package version 1.0.2*, URL https://cran.r-project.org/web/packages/ordinalCont.

R Core Team (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Ramsey, J.O. (1988) Monotone Regression Splines in Action. *Statistical Science*, **3**(4), 425 – 441.

Stockler, Martin R et al. (2011). Capecitabine versus classical cyclophosphamide, methotrexate, and fluorouracil as first-line chemotherapy for advanced breast cancer. *J of Clinical Oncology*,**29**, 34, 4498 – 4504.

Modeling agreement on continuous recordings in the presence of a binary scale

Sophie Vanbelle[1], Emmanuel Lesaffre[2]

[1] Department of Methodology and Statistics, CAPHRI, Maastricht University, The Netherlands
[2] Leuven Biostatistics and Statistical Bioinformatics Centre, Belgium

E-mail for correspondence: sophie.vanbelle@maastrichtuniversity.nl

Abstract: Recording variables continuously over time is increasingly frequent in medical and behavioral sciences. Indeed, the use of video and portable devices permits to study individuals in their natural environment and provide real-time assessments. The measurement instruments developed for ambulatory use and the coding schemes used to analyze video recordings have to show reliability and validity before daily practice implementation. Studies using continuous records are usually characterized by a large amount of data presenting serial correlation and a small sample of subjects. We developed a method to derive appropriate reliability and agreement indexes to analyse this type of data in the presence of a covariate structure (including time as component) when records are made on a binary scale. The method is illustrated on a validation study for a new single-unit activity monitor (CAM) in patients with chronic organ failure.

Keywords: Bayesian methods, transient event, time-event sequential data, reliability, e-Health.

1 Introduction

Continuous recordings, defined as second by second or even closer records in time, are frequent in medical, behavioral and social sciences. Indeed, complex behaviors can be quantified by observers on video recordings or data can be collected with the use of handled portable electronic data-entry and storage devices (e.g. e-Health). In both cases, the reliability and the validity of the coding scheme/measurement instrument have to be assessed before using the continuous recordings in daily practice.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

For example, the following study was designed to validate a new single-unit activity monitor (CAM) in patients with chronic organ failure (Annegarn et al., 2011). This new monitor was developed to implement daily physical activity as an outcome measure of cardiopulmonary rehabilitation. In the study, the activity (non-weight bearing posture (NWBP), weight-bearing posture (WBP) or dynamic activity (DA)) of 10 patients in rehabilitation was recorded during one hour of daily routine by the CAM. In the present paper, we restrict attention on the distinction between non-weighted bearing postures (NWBP) and the two other activities (WPB+DA). The CAM was wore at two different places of the body simultaneously for comparative purposes, namely on the leg and on the trunk. The patients were also videotaped by a researcher during the one hour time period. The video, considered as criterion standard, was then analyzed second by second by a researcher blinded to the values obtained by the CAM. The aim of the study was to determine (1) the agreement level between the video assessments and the CAM, wore on the leg or on the trunk and (2) the body place providing the highest agreement level with the video.

The existing population-based and unit-specific models developed to determine agreement in the presence of a covariate structure and repeated measurements have to be ruled out to analyze the present study. Population-based models are not practicable. The large amount of repeated measurements (3600 in the CAM study) and the small number of subjects involved (10 in the CAM study) do not permit to obtain stable parameter estimates. In the same way, the only unit-specific approach permitting to directly model agreement according to predictors in the presence of repeated measurements (Vanbelle and Lesaffre, 2015) is not manageable because of the large amount of observations. We therefore developed a partial-Bayesian approach, based on this latter work, to model agreement obtained on continuous recordings over time in the presence of a covariate structure.

2 Method

Instead of considering each time point separately, we group them into small time intervals. The cumulative distribution of the outcome over one interval (i.e. number of time points under NWBP) is binomial given three assumptions: (1) the intervals have fixed length, (2) the probability of being engaged in NWBP is constant within each time interval and (3) the observations within a time interval are independent. The first assumption holds by definition of the intervals. We assume that the second assumption holds because the time intervals are small. The third assumption does not hold as observations close in time are very likely to be correlated. To account for this correlation, we assumed that the cumulative distribution of the outcome within each interval follows a beta-binomial distribution. We make the same assumption for the cumulative distribution of agreement between the CAM and video assessments within each interval.

Then, based earlier work, we consider the likelihood formed by the cumulative distribution of the probability to observe NWBP with the CAM and the video and the cumulative distribution of the agreement. Models relating these cumulative distributions to covariates (namely time, body place) are then fitted jointly with a partial-Bayesian approach using Markov chain Monte Carlo (MCMC). The MCMC calculations were performed using JAGS. Vague independent priors

were considered for all model parameters. The dependency between the probabilities over the one hour observation period was taken into account using random effects. Marginal posterior probability distributions were obtained by averaging over these random effects. The length of the time intervals was fixed to 30 seconds.

3 Results

The probability to be engaged in a NWBP is displayed in Figure 1 over the one hour observation period. It can be remarked that the probability of being in NWBP varies over time and differs markedly between CAM-trunk on one hand and CAM-Leg and video on the other hand.

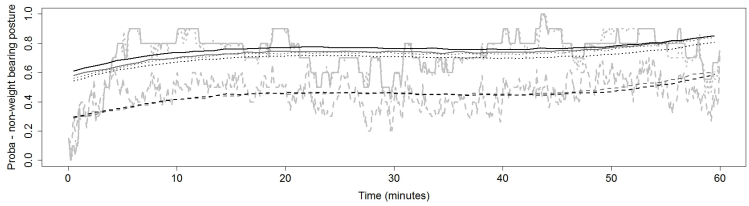


FIGURE 1. CAM study. Probability of being in a NWBP in the sample (light gray). Data were summarized every 5 seconds for the clarity of the graph. Posterior marginal probability of being in a NWBP using a binomial (gray) and beta-binomial (black) model. The plain line corresponds to the video, the dotted line to the CAM-leg and the dashed line to the CAM-trunk

The evolution of the agreement level between the CAM and the video, quantified through Cohen's kappa coefficient, is depicted in Figure 2 over the one hour observation period. Two important results have to be noted. First, the agreement level between the CAM and the video decreases over time. This is an indication of a possible inappropriate observation period. The observation period was either too short to obtain stable agreement estimates or too long for the researcher to avoid fatigue effects when rating the video recordings. Secondly, the agreement level is higher between the CAM-leg and the video than between the CAM-trunk and the video. The new CAM device should therefore best be placed on the leg than on the trunk to detect non-weight bearing postures.

4 Discussion

We developed a method to model the agreement in the presence of continuous recordings and a covariate structure. To this end, we summarized the data into small time intervals. The difference between the results obtained with the binomial and the beta-binomial models supports the need to take the dependency between observations within time intervals into account. Taking the average probability over a time interval corresponds to temporal aggregation, a kind of smoothing technique used in time series analysis. We should be careful about

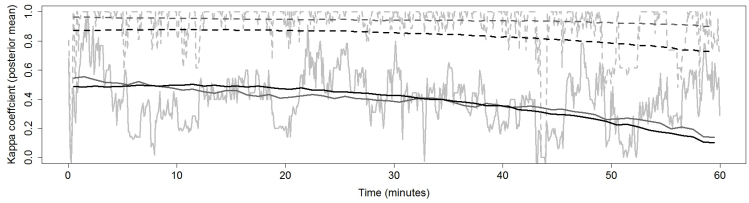


FIGURE 2. CAM study. Agreement (Cohen's kappa coefficient) on NWBP in the sample (light gray). Data were summarized every 5 seconds for the clarity of the graph. Posterior marginal mean for the agreement level between CAM-Leg and video (dashed line) and between CAM-trunk and video (plain line) using a binomial (gray) and beta-binomial (black) models.

the choice of the interval length not to hide important patterns in the evolution of the outcome and the agreement over time. The dependency between time intervals is currently taken into account using a random intercept in the models. However, the use of autoregressive models is currently under study.

Acknowledgments: This research is part of project 451-13-002 funded by the Netherlands Organisation for Scientific Research.

References

Annegarn, J., Spruit, M.A., Uszko-Lencer, N.H.M.K., Vanbelle, S., Savelberg, H.H.C.M., Schols, A.M.W.J., Wouters, E.F.M. and Meijer, K. (2011). Objective Physical Activity Assessment in Patients With Chronic Organ Failure: A Validation Study of a New Single-Unit Activity Monitor. *Archives of Physical Medicine and Rehabilitation*, **92**, 1852 – 1857.e1.

Vanbelle, S. and Lesaffre, E. (2015). Modeling agreement on categorical scales in the presence of random scorers. *Biostatistics*, **17**, 79 – 93.

Bayesian Inference for Mixed Effects Multinomial Logit Models

Helga Wagner[1], Sylvia Frühwirth-Schnatter[2]

[1] Johannes Kepler University Linz, Austria
[2] Vienna University of Economics and Business, Austria

E-mail for correspondence: helga.wagner@jku.at

Abstract: We consider Bayesian inference for very flexible multinomial logit models with choice and unit specific covariates where we allow for global as well as category specific fixed and random effects. To achieve model identification we use a sum-to-zero constraint on the category specific fixed and random coefficients. In contrast to the conventional identification strategy where one of the outcome categories is chosen as baseline and the corresponding category specific parameters are fixed to zero, the sum-to-zero constraint allows for a symmetric prior of the category specific effects of all outcome categories. By augmenting the model with latent utilities posterior inference using MCMC methods is straightforward. In our empirical analysis we use panel data from EU-SILC in Austria 2004 – 2007 to analyse labor supply of households. Inference is particularly challenging as due to the fact that data are collected using a rotating panel design, panels are very short with a maximum of 4 observations per household.

Keywords: random utility model, sum-to-zero identification, discrete choice model, longitudinal data, EU-SILC

1 Mixed Effects Multinomial Logit Model

We consider modelling of longitudinal discrete choice data using covariate information. Covariates can be either choice or subject specific and we allow for choice-specific effects of subject-specific covariates. Additionally, to model both heterogeneity between subjects as well as dependence within subjects we include a subject-specific random intercept and allow for random effects of choice specific covariates. We assume that each subject i, $i = 1, \ldots, N$ chooses one of $m + 1$ unordered alternatives $\{0, \ldots, m\}$ at occasion $t = 1, \ldots, T_i$ and denote this choice by Y_{it}. The probability that alternative k is chosen by subject i at occasion t is

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

then modelled by a multinomial logit model as

$$\mathrm{P}(y_{it} = k) = \frac{\exp(\mathbf{x}_{itk}^f \boldsymbol{\alpha} + \mathbf{x}_{it}\boldsymbol{\beta}_k + \mathbf{x}_{itk}^r \boldsymbol{\zeta}_i + \xi_{ik})}{\sum_{l=0}^{m} \exp(\mathbf{x}_{itl}^f \boldsymbol{\alpha} + \mathbf{x}_{it}\boldsymbol{\beta}_l + \mathbf{x}_{itl}^r \boldsymbol{\zeta}_i + \xi_{il})} \quad k = 0, \ldots, m \tag{1}$$

Here $\mathbf{x}_{it}$ denotes the vector of subject specific covariates with choice specific effects $\boldsymbol{\beta}_k$; $\boldsymbol{\alpha}$ is the vector of fixed effects of the choice specific covariates $\mathbf{x}_{itk}^f$ and $\boldsymbol{\zeta}_i$ the vector of random effects of the choice specific covariates $\mathbf{x}_{itk}^r$. Finally ξ_{ik} is the category specific random intercept of subject i.

As the same choice probabilities would result for choice specific effects $\boldsymbol{\beta}_k + \tilde{\boldsymbol{\beta}}$ and subject specific random intercepts $(\xi_{ik} + \tilde{\xi})$,

$$\mathrm{P}(y_{it} = k) = \frac{\exp(\mathbf{x}_{itk}^f \boldsymbol{\alpha} + \mathbf{x}_{it}(\boldsymbol{\beta}_k + \tilde{\boldsymbol{\beta}}) + \mathbf{x}_{itk}^r \boldsymbol{\zeta}_i + (\xi_{ik} + \tilde{\xi}))}{\sum_{l=0}^{m} \exp(\mathbf{x}_{itl}^f \boldsymbol{\alpha} + \mathbf{x}_{it}(\boldsymbol{\beta}_l + \tilde{\boldsymbol{\beta}}) + \mathbf{x}_{itl}^r \boldsymbol{\zeta}_i + (\xi_{il} + \tilde{\xi}))},$$

the model is not identified. To achieve identification the usual strategy is to choose one of the categories, e.g. category 0 as reference category and set the corresponding category specific effects to zero, i.e.

$$\boldsymbol{\beta}_0 = \mathbf{0} \qquad \text{and} \qquad \xi_{i0} = 0.$$

However this identification strategy causes problems in the interpretation of the estimated subject- and category-specific random intercepts ξ_{ik} . Though no random intercept is explicitely included for the reference category 0, this is implicitely the case as the probability that subject i chooses category 0 depends on all other category-specific random intercepts $\xi_{ik}, k = 1, \ldots, m$ due to the fact that

$$\sum_{k=0}^{m} \mathrm{P}(y_{it} = k) = 1.$$

An alternative identification strategy, which is symmetric with respect to all categories, is to impose a sum-to zero constraint on the category-specific effects (Burgette and Hahn, 2013),

$$\sum_{k=0}^{m} \boldsymbol{\beta}_k = \mathbf{0} \qquad \text{and} \qquad \sum_{k=0}^{m} \xi_{ik} = 0$$

In a Bayesian approach this constraint can be imposed by specifying a reduced rank Normal prior distribution on the corresponding effects, where the sum to zero constraint holds a priori with probability 1.

2 Prior Distributions

To specify the prior on the fixed effects, we define the vector of the category specific regression coeffcients for each covariate j as $\boldsymbol{\beta}^j = (\beta_{0,j}, \ldots, \beta_{m,j})^T$ and let $\mathbf{e}_{m+1} = (1, 1 \ldots, 1)^T$. We assume that the vectors $\boldsymbol{\beta}^j$ are a priori independent for $j = 1, \ldots, d$. The sum-to-zero constraint on $\boldsymbol{\beta}^j$ given as

$$\sum_{k=0}^{m} \beta_{k,j} = \mathbf{e}_{m+1}^T \boldsymbol{\beta}^j = 0 \tag{2}$$

is imposed by specifying a reduced rank Normal prior

$$\boldsymbol{\beta}^j \sim \mathcal{N}(\mathbf{b}_j, \delta_j^\beta \mathbf{C}_m),$$

where the mean vector $\mathbf{b}_j$ is restricted to sum to zero

$$\mathbf{e}_{m+1}^T \mathbf{b}_j = 0.$$

δ_j^β is a scale factor and the matrix $\mathbf{C}_m$ has a compound symmetry structure

$$\mathbf{C}_m = \frac{m+1}{m}\mathbf{I}_{m+1} - \frac{1}{m}\mathbf{e}_{m+1}\mathbf{e}_{m+1}^T = \begin{pmatrix} 1 & -\frac{1}{m} & \dots & -\frac{1}{m} \\ -\frac{1}{m} & 1 & \dots & -\frac{1}{m} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{m} & -\frac{1}{m} & \dots & 1 \end{pmatrix}.$$

With this specification the sum to zero constraint (2) holds a priori with probability 1, as $\mathrm{E}(\mathbf{e}_{m+1}^T \cdot \boldsymbol{\beta}^j) = \mathbf{e}_{m+1}^T \cdot \mathbf{b}_j = \mathbf{0}$ and

$$\mathrm{V}(\mathbf{e}_{m+1}^T \cdot \boldsymbol{\beta}^j) = \delta_j^\beta \mathbf{e}_{\mathbf{m+1}}^{\mathbf{T}} \mathbf{C}_\mathbf{m} \mathbf{e}_{\mathbf{m+1}} = \mathbf{O}.$$

For the subject-specific vectors of random intercept $\boldsymbol{\xi}_i = (\xi_{i0}, \xi_{i1}, \dots, \xi_{im})^T$ we assume prior independence across subjects and specify the prior hierarchically as

$$\begin{aligned} \boldsymbol{\xi}_i &\sim \mathcal{N}(\boldsymbol{\gamma}, \delta^\xi \mathbf{C}_m) \\ \boldsymbol{\gamma} &\sim \mathcal{N}(\mathbf{0}, \delta^\gamma \mathbf{C}_m) \\ \delta^\xi &\sim \mathcal{G}^{-1}(\nu, Q) \end{aligned}$$

Note, that the specification of this reduced rank prior yields a sparse parameterisation of the covariance matrix of the category specific random intercepts where only one parameter, the scale parameter δ^ξ has to be estimated from the data.

Finally, for the global fixed and random effects we specify standard multivariate Normal priors.

3 Posterior Inference

Posterior inference is performed by MCMC methods, where we use the representation of the model in terms of random latent utilities, first introduced by McFadden (1974). Let u_{itk} denote the latent utility of category k for subject i at timepoint t

$$u_{itk} = \mathbf{x}_{itk}^f \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta}_k + \mathbf{x}_{itk}^r \boldsymbol{\zeta}_i + \xi_{ik} + \epsilon_{itk}, \qquad k = 0, \dots, m$$

where ϵ_{itk} are i.i.d extreme value distributed random variables. If y_{it} is defined as the category with maximum utility

$$y_{it} = k \Leftrightarrow u_{itk} = \max_l u_{itl}$$

the marginal distribution of y_{it} is given by the multinomial logit model in equation (1). MCMC iterates between sampling the latent utilities and the effects in the multinomial logit model, for which we use the data augmented independence MH sampler with a multivariate proposal density described in Fussl et al. (2013).

4 Application

The goal of our empirical analysis is to provide a microsimulation model for labour supply of couples with children in Austria, which allows to evaluate the consequences of interventions like changes in tax system, social security contributions or social transfers (e.g. parental leave benefits) on labour supply of the households. We use a unitary model with discrete choices, i.e. we assume that choices on labour supply are made on the unit level of the household.

For the analysis we use data from EU-SILC (European Survey on Income and Living Conditions) from 2003 – 2006. EU-SILC has a four year rotating panel design and provides longitudinal microdata on income, poverty, housing, labor and living conditions. The data for our analysis comprises 4218 observations of 2289 households which are observed in at least one but not more than four years. Figure 1 shows the labour supply (in working hours) of men and women.

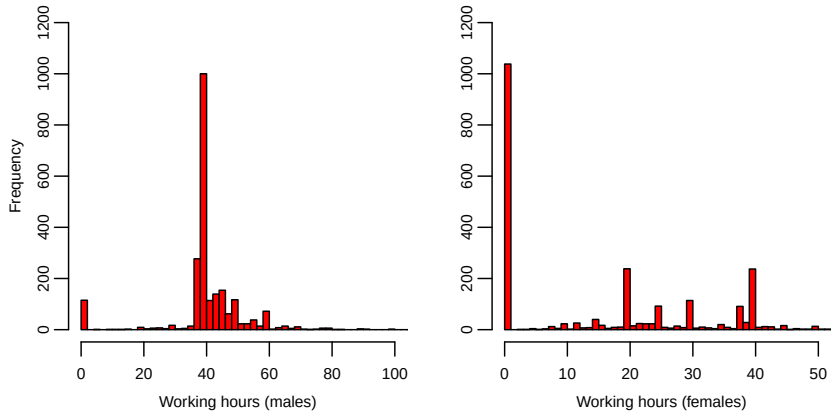


FIGURE 1. Working hours of males (left) and females (right) in couple households with children (in the first panel year).

We exclude non-working men from the analysis and – as the vast majority of men is working full-time – define three choices with respect to the labour supply of women as *not working* (N), *working part-time* (P), and *working full time* (F). A challenging feature of the data is their structure as an unbalanced panel with 1079 households observed only at one time point and only 167 households observed in all four panel time points. Additionally only one choice is observed for the majority of households (1936 of 2289), whereas all three possible labour supply choices were observed in only 12 households. For the three possible labour supply choices we fit a multinomial logit model where we include household income and household leisure time, both measured on the log-scale, as choice-specific covariates with fixed effects. Additionally we include the following global covariates: indicators for age of the youngest child (categorized; baseline: 0-3; 3-5, 6-10; over 10) and the number of further children in each of these age groups and model their effects as category-specific. To account for heterogeneity of households with respect to their labour supply choice we include a household specific random intercept.

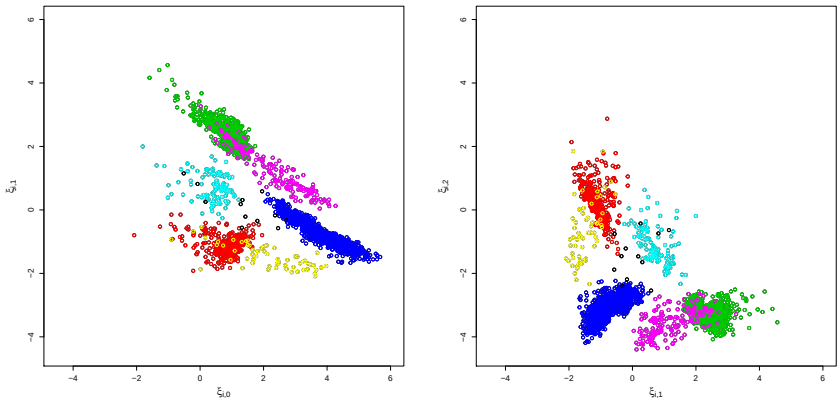


FIGURE 2. Random effects for groups defined by observed working status of mothers (not working (N)/working part-time (P)/ working full-time (F); blue: N; green: P; red: F; magenta:N/P; yellow: N/F; cyan: P/F; black: N/P/F)

As expected, the global effect of household income is positive, thus increasing the utility of part-time as well as full time work , and the effect of leisure time is negative. The utilities for both part-time as well as full-time work of mothers increase with the age of the youngest child and decrease with the number of children. Heterogeneity among mothers is considerable with an estimated random intercept variance of 4.8. Figure 2 shows bivariate plots of the random intercepts, in different colors for clusters of households defined by the observed labour supply, e.g. for households observed only with the mother non-working (shown in blue) the mean random intercept is positive for *not working* (N), roughly zero for *part-time working* (P) and negative for *full-time working* (F).

Predicted household choices, which are determined as the category with the highest mean posterior probability, are compared to the observed choices in the sample in Table 1. Results indicate good insample predictive performance of the model with 92.2% choices predicted correctly.

TABLE 1. Comparison of predicted to observed choices in the sample

	actual choice			
predicted choice	non-working	part-time	full-time	Sum
non-working	**1711**	106	46	1863
part-time	71	**1436**	66	1573
full-time	12	31	**739**	781
Sum	1794	1573	851	4218

References

Burgette, L. and Hahn, P. R. (2013). *A symmetric prior for multinomial probit models*. Technical Report.

Fussl, A. et al. (2013). Efficient MCMC for Binomial Logit Models. *ACM Transactions on Modeling and Computer Simulation*. **23**, Article 3, 121.

Frühwirth-Schnatter S. and Frühwirth R. (2010). Data augmentation and MCMC for binary and multinomial logit models. In: *Kneib T. and Tutz G. (Eds..), Statistical Modelling and Regression Structures Festschrift in Honour of Ludwig Fahrmeir*. Heidelberg: Physica-Verlag, 111-132.

Mc Fadden, D. (1981). Econometric models of probabilistic choice. In: *C. F. Manski and D. McFadden (Eds.), Structural Analysis of Discrete Data with Econometric Applications*, Cambridge MA: MIT Press, 198 – 272.

Simple Effect Measures for Interpreting Models for Ordinal Categorical Data

Alan Agresti[1], Claudia Tarantola[2]

[1] University of Florida, USA
[2] University of Pavia, Italy

E-mail for correspondence: aa@stat.ufl.edu

Abstract: We survey effect measures for models for ordinal categorical data that can be simpler to interpret that the model parameters. For describing the effect of an explanatory variable while adjusting for other explanatory variables, we present probability-based measures, including a measure of relative size and effect measures based on an instantaneous rate of change. We also survey summary measures of predictive power that are analogs of R-squared and multiple correlation measures for quantitative response variables. We suggest new measures of effect and of predictive power. In a longer companion paper available at www.stat.ufl.edu/~aa/articles/agresti_tarantola.pdf, we illustrate the measures for an example and provide R code for calculating them.

Keywords: Cumulative link models; Cumulative logits; Marginal effects; Proportional odds; R-squared

1 Introduction

Popular models for ordinal categorical response variables are generalized linear models that employ non-linear link functions. As a consequence, the model effect parameters relate to measures, such as odds ratios and probits, that are not easily understood by non-statisticians. This article surveys simpler ways to interpret the effects of explanatory variables and to summarize the model's predictive power.

Section 2 presents alternative ways to summmarize the effect of an explanatory variable. These include simple comparisons of the probability of extreme-response outcomes at extreme values of an explanatory variable, measures of instantaneous effect on the extreme-response probabilities, and measures for comparing groups that result directly from latent variable models that induce the standard ordinal models. Section 3 surveys measures of predictive power, including measures that

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

resemble those for ordinary linear models, possibly estimated for a latent variable model. We also propose a multiple correlation measure that generalizes the Spearman correlation. A companion paper (Agresti and Tarantola 2017) available at `www.stat.ufl.edu/~aa/articles/agresti_tarantola.pdf` illustrates existing and newly-proposed measures with an example, gives further details and references, and provides `R` code for the analyses.

2 Ordinal Effect Measures for Individual Explanatory Variables

For an ordinal response variable y with c categories, we consider models in which the explanatory variables may be a mixture of quantitative and categorical variables. We denote explanatory variable values by $\boldsymbol{x} = (x_1, \ldots, x_p)^T$. To discuss ways of summarizing effects for a categorical explanatory variable, we refer also to a separate indicator variable z that distinguishes between two groups. The most popular models are special cases of the *cumulative link model*

$$\text{link}[P(y \leq j)] = \alpha_j - \beta z - \beta_1 x_1 - \cdots - \beta_p x_p \ , \quad j = 1, \ldots, c-1, \tag{1}$$

for link functions such as the logit and probit. The nonlinear link function naturally produces effects on the link scale. For example, for cumulative logit models with this proportional odds structure, β is the difference between logits of cumulative probabilities for the two groups, and $-\beta_1$ is the change in the cumulative logit per each 1-unit increase in x_1, adjusting for the other explanatory variables. This leads to odds ratios as natural effect measures. For instance, $\exp(-\beta_1)$ is a multiplicative effect of each 1-unit increase in x_1 on the cumulative odds of response $\leq j$ vs. $> j$, for each j.

Such effect measures are not easy to interpret by scientists who need to understand the effects in more real-world terms. In addition, with nonlinear link functions, effects often behave in a way that is counterintuitive to those who are mainly familiar with ordinary linear models. In logistic regression models, for example, if an explanatory variable is added to the model that is uncorrelated with x_1, the partial effect of x_1 is typically different than in the model without the other explanatory variable, whereas it would be identical in an ordinary linear model fitted by least squares. Probability-based effects summarized in this paper are easier to understand and are typically more stable.

2.1 Extreme-category range-based probability summaries

In practice with ordinal responses, the highest and lowest response categories often represent a noteworthy state, such as the *best* or *worst* outcome. It is informative to report how $P(y = 1)$ and $P(y = c)$ change as explanatory variables change. As any explanatory variable increases, cumulative link models that contain solely main effects imply monotonicity in these probabilities. A simple summary of the effect of an explanatory variable reports the way that estimates of $P(y = 1)$ and $P(y = c)$ change over the range of values of that explanatory variable, when other explanatory variables are set at particular values such as their means.

2.2 Marginal effect measures

A second type of simple summary for ordinal responses considers the rate of change in $P(y = 1)$ and $P(y = c)$, as a function of the explanatory variable. We express the model (1) as

$$F^{-1}[P(y \leq j)] = \alpha_j - \beta z - \boldsymbol{x}^T\boldsymbol{\beta}, \quad j = 1, \ldots, c-1, \tag{2}$$

where F^{-1} is the inverse of a standard cdf, $\boldsymbol{x}$ is a column vector of explanatory variable values (excluding z), and $\boldsymbol{\beta}$ is a column vector of parameters for $\boldsymbol{x}$. Let $f(y) = \partial F(y)/\partial y$.

For a quantitative explanatory variable x_k, the rate of change in $P(y = 1)$ at a particular value of x_k, when other explanatory variables are fixed at certain values $\boldsymbol{x}^*$, is $\partial P(y = 1|\boldsymbol{x} = \boldsymbol{x}^*)/\partial x_k$. Some books, such as Long and Freese (2014), refer to such an instantaneous effect as a *marginal effect*. For the cumulative link model, the marginal effect of x_k is $-f(\alpha_j - \beta z - \boldsymbol{x}^T\boldsymbol{\beta})\beta_k$ on $P(y = 1)$ and $f(\alpha_j - \beta z - \boldsymbol{x}^T\boldsymbol{\beta})\beta_k$ on $P(y = c)$. For the n sample observations, we can find the marginal effect of x_k at each of the n observed values of the explanatory variables, and then average them. This is called an *average marginal effect* (AME). For categorical explanatory variables, we can find the difference between $P(y = 1)$ (and/or $P(y = c)$) when $z = 1$ and when $z = 0$, for the n sample observations on $\boldsymbol{x}$, and average the obtained values.

For binary data, Mood (2010) pointed out that such measures have behavior reminiscent of effects in ordinary linear models. For example, they are roughly stable when we add an explanatory variable to the model that is uncorrelated with the variable for which we are describing the effect.

2.3 A probability summary for ordered comparison of groups

We next present an alternative way to summarize the effect of a categorical explanatory variable on an ordinal response y, suggested by Agresti and Kateri (2017). We discuss this in the context of comparing two groups ($z = 0$ and $z = 1$). We regard the ordinal response as crude measurement of an underlying continuous latent variable y^* that is the response variable in an ordinary linear model. Anderson and Philips (1981) showed that the cumulative link model (2) is implied by a model in which a latent response has conditional distribution with standard cdf given by the inverse of the link function. Let y_1^* and y_2^* denote independent underlying latent variables for y, representing the underlying distributions when $z = 1$ and when $z = 0$ respectively. At a particular setting $\boldsymbol{x}$ for other explanatory variables, $P(y_1^* > y_2^*; \boldsymbol{x})$ is a summary measure of relative size. This measure is most meaningful when the groups are stochastically ordered.

The normal latent variable model with $y^* \sim N(\beta z + \beta_1 x_1 + \cdots + \beta_p x_p, 1)$ implies model (2) with probit link. For this model,

$$P(y_1^* > y_2^*; \boldsymbol{x}) = P\left[\frac{(y_1^* - y_2^*) - \beta}{\sqrt{2}} > \frac{-\beta}{\sqrt{2}}\right] = \Phi\left(\frac{\beta}{\sqrt{2}}\right). \tag{3}$$

This is true regardless of the $\boldsymbol{x}$ value, so we denote it by $P(y_1^* > y_2^*)$. For the logit link, Agresti and Kateri (2017) showed that

$$P(y_1^* > y_2^*) \approx \frac{\exp(\beta/\sqrt{2})}{[1 + \exp(\beta/\sqrt{2})]}, \tag{4}$$

for the β coefficient of z in the cumulative logit model. For a log-log link, which is relevant when we expect underlying latent variables to have extreme-value distributions, Agresti and Kateri noted that

$$P(y_1^* > y_2^*) = \frac{\exp(\beta)}{[1 + \exp(\beta)]},$$

for the β coefficient of z in the cumulative link model with log-log link. Ordinary confidence intervals for the β model parameter induce confidence intervals for the stochastic superiority measure.

Agresti and Kateri suggested that practitioners can more easily interpret $P(y_1^* > y_2^*)$ than parameters such as odds ratios and differences in probits that naturally result in cumulative link models. They also proposed related measures for the observed y scale that need not relate to latent variables.

3 Summary Measures of Predictive Power

Next we discuss ways to summarize how well we can predict y using the fit of the chosen ordinal model. Measures of predictive power can be useful for comparing different models, such as to see whether it helps substantively to add an interaction term.

3.1 Concordance index

Consider all pairs of observations that have different outcomes on y. The *concordance index* estimates the probability that the predictions and the outcomes are concordant, that is, that the observation with the larger y-value also has a stochastically higher set of model-fitted probabilities. Appealing features of the concordance index are its simple structure and its generality of potential application. Because it utilizes ranking information only, however, it cannot distinguish between different link functions or linear predictors that yield the same stochastic orderings.

3.2 R-squared type measures

An alternative approach to summarizing predictive power adapts standard measures for quantitative response variables. A way to construct such a measure without assigning arbitrary scores to the categories of y is to estimate R^2 for the linear model for an underlying latent response variable. McKelvey and Zavoina (1975) suggested this measure for the cumulative probit model, for which the underlying latent variable model is the ordinary normal linear model. Let y_i^* denote the value of the latent variable for subject i. The R^2 measure for the latent variable model has the usual proportional reduction in variation form

$$R^2 = \frac{\sum_i (y_i^* - \bar{y}^*)^2 - \sum_i (y_i^* - \hat{y}_i^*)^2}{\sum_i (y_i^* - \bar{y}^*)^2} = \frac{\sum_i (\hat{y}_i^* - \bar{y}^*)^2}{\sum_i (y_i^* - \bar{y}^*)^2},$$

the estimated variance of $\hat{y}^*$ divided by the estimated variance of y^*. After fitting a cumulative link model we can estimate the variance of $\hat{y}^*$ by the variance of

the linear predictor $\hat{y}^* = \hat{\beta}z + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ without the intercept. We cannot observe the latent variable or its sample variance, but we can estimate that variance by the estimated variance of $\hat{y}^*$ plus the variance of the latent variable distribution, which is 1 for the probit link and $\pi^2/3 = 3.29$ for the logit link (i.e., standard logistic distribution).

An alternative proportional-reduction-in-variability approach uses a likelihood-based measure. Denote the maximized log-likelihood values by L_M for the working model fit and L_0 for the null model (i.e., containing only intercept terms). The *pseudo R-squared* measure is

$$1 - \frac{L_M}{L_0}.$$

A weakness of such a measure is that the scale is not the same as for y. Interpreting the numerical value is difficult, other than in a comparative sense for different models.

3.3 Multiple correlation measures

Some statisticians prefer correlation measures over related R^2 measures, because of the appeal of working on the original scale and its proportionality to the effect size. For example, for the ordinary linear model, for fixed marginal standard deviations, doubling the slope also doubles the correlation.

We could estimate the multiple correlation for an underlying latent variable model, such as by taking the square root of the McKelvey and Zavoina (1975) R^2 measure. Another way to form a multiple correlation measure that connects with models for cumulative probabilities uses as scores the average cumulative proportions for the marginal distribution of y. For sample proportions $\{p_j\}$ in that marginal distribution, the average cumulative proportion in category j is

$$v_j = \sum_{k=1}^{j-1} p_k + \left(\frac{1}{2}\right) p_j, \quad j = 1, 2, \ldots, c.$$

Such scores, which are linearly related to the midranks, are sometimes referred to as *ridits*. With such created scores, one could construct the correlation for the n sample observations between the score for the observed outcome category for a subject and the estimated mean score generated by the model-fitted probability values for the subject. With ridit scores, this is a multiple correlation version of the Spearman correlation. In future research, it would be of interest to study properties of such a measure, including bias reduction in estimating the population analog.

References

Agresti, A., and Kateri, K. (2017). Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics*, to appear.

Agresti, A., and Tarantola, C. (2017). Simple effect measures for interpreting models for ordinal categorical data. Submitted for publication.

Anderson, J. A., and Philips, P. R. (1981) Regression, discrimination, and measurement models for ordered categorical variables. *Applied Statistics*, **30**, 22–31.

Long, J.S. and Freese, J. (2014). *Regression Models for Categorical Dependent Variables Using Stata*, 3rd ed. College Station, TX: Stata Press.

McKelvey, R.D. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *J. Math. Sociol.* **4**: 103–120.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *Eur. Sociol. Rev.* **26**: 67–82.

The truth about the effective dimension

Paul H. C. Eilers[1]

[1] Erasmus MC, Rotterdam, the Netherlands

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

Abstract: The hat matrix of a model is valuable. Its trace gives the effective dimension. In mixed models partial effective dimensions can be defined that make variance estimation easy and reliable. Efron (2004) casts doubt on the trace of the matrix and advocates an alternative definition of the effective dimension, based on a covariance formula. Unfortunately he uses the robust *lowess* smoother, thereby blurring the issue.

Keywords: Hat matrix; Penalty; Mixed model; Robustness

1 Introduction

In a linear model we can write $\hat{y} = Hy$, where y is an observed vector and H is the so-called hat matrix (Ye, 1998). It depends only on the explanatory variables and penalties, if present. The diagonal of H provides the effective (model) dimension: $ED = \sum_i h_{ii}$. The situation seems clear cut, but one can find several publications that cast doubt on this definition, e.g Efron (2004) and Janson et al. (2015). Here I want to show that H deserves its place. Also I show that in mixed models we can define partial hat matrices and partial effective dimensions. The latter play a central role in the estimation of variances, as already pointed out by Harville (1972).

Efron uses the robust *lowess* smoother, which has a redescending influence function, effectively making the model strongly non-linear and generating worrying results. A linear smoother, like *locfit* or P-splines eliminates all issues.

The effective dimension is a property of a model, when fitted to data, not of the data themselves. I illustrate that with the Whittaker smoother, varying the order of the differences in the penalty.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

2 Theory and examples

2.1 The hat matrix

Consider a mixed model with fixed and random components: $y = X\beta + Zc + e$, with $c \sim N(0, G)$ and $e \sim N(0, R)$. It is well known that one finds estimates for β and c by solving the system

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} \hat{X}'R^{-1}y \\ \hat{Z}'R^{-1}y \end{bmatrix}. \tag{1}$$

In many applications one can assume that $R = \sigma^2 I_1$ and $G = \tau^2 I_2$, with unknown σ and τ, and proper sizes of the identity matrices I_1 and I_2. Then one can write

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda I \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{c} \end{bmatrix} = S \begin{bmatrix} \hat{\beta} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} \hat{X}'y \\ \hat{Z}'y \end{bmatrix} = \begin{bmatrix} X' \\ Z' \end{bmatrix} y, \tag{2}$$

with $\lambda = \sigma^2/\tau^2$. Because $\hat{y} = [X\ Z][\hat{\beta}'\hat{c}]'$, the hat matrix is

$$H = \begin{bmatrix} X & Z \end{bmatrix} S^{-1} \begin{bmatrix} X' \\ Z' \end{bmatrix}. \tag{3}$$

Another useful matrix is

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} = S^{-1} \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix}. \tag{4}$$

It is easy to show that $\mathrm{tr}(H) = \mathrm{tr}(K) = \mathrm{tr}(K_{11}) + \mathrm{tr}(K_{22})$. Hence the effective dimension $\omega = \mathrm{tr}(H)$ can be written as the sum of two components, one for the fixed and one for the random part.

Generalization is straightforward. If we have two random components: $y = X\beta + Zc + \breve{Z}\breve{c}$, with $\breve{c} \sim N(0, \breve{\tau}^2 I)$, we have the blocks K_{11}, K_{22} and K_{33} on the diagonal of K. Their traces add up to the effective dimension.

Harville (1977) presents an algorithm for variance estimation that uses traces of sub-matrices of a matrix $I - T$. It turns out that $I - T$ is equal to the sub-matrix of K that corresponds to the random components. The algorithm uses the fact that $c'c/\mathrm{tr}(K_{22}) = \tau^2$ and $\breve{c}'\breve{c}/\mathrm{tr}(K_{33}) = \breve{\tau}^2$. Also $\sigma^2 = (y - \hat{y})'(y - \hat{y})/(n - ED)$, where n is the number of observations. These relations hold for the true values of σ^2, τ^2 and $\tilde{\tau}^2$, but it immediately suggests an iterative algorithm. Practice has shown that iterating with these expressions, updating the variances, re-estimating the coefficients and traces, is reliable and relatively fast. Harville's algorithm predates Schall's (1991) work. It has been used successfully by Rodriguez-Alvarez et al. (2016) for modeling agricultural field trials with up to ten unknown variances. As reported by Velazco and others (2017), it never failed on a large set of agricultural experiments.

Harville's formula for T is quite complicated, because he starts by explicitly eliminating the fixed part of the mixed model. Everybody else does it too. This is an intriguing element of the mixed model folklore, as if the REML paradigm, correcting for fixed effects, should be explicitly visible to be believed. It is a complication we can easily avoid, using the matrix K.

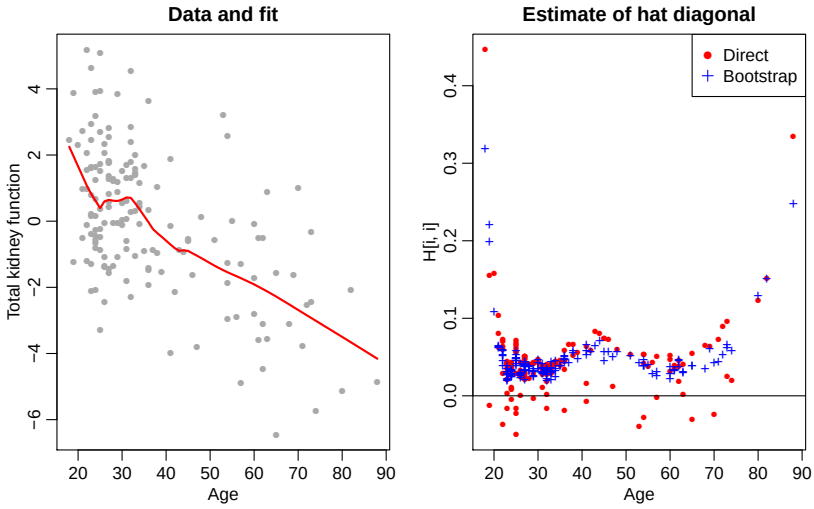


FIGURE 1. Smoothing of kidney data with *lowess*. Left: data and *lowess* fit. Right: estimates of $\partial\hat{y}_i/\partial y_i$, obtained by data perturbation (red dots) and by bootstrapping (blue crosses).

2.2 The covariance alternative

Ye (1998) not only used the diagonal of the hat matrix as the effective dimension, but also an equivalence: $\partial\hat{y}_i/\partial y_i = \text{cov}(\hat{y}_i, y_i)/\sigma^2$. Remarkably, it seems to be more popular than the trace of H. Efron (2004) casts doubt on the hat matrix by showing a scatterplot of data on kidney function and applying the *lowess* smoother. He also presents a plot of $\partial\hat{y}_i/\partial y_i$. It shows a rather wild pattern and even a number of negative values. Efron shows how to get alternative estimates, using the above equivalence and bootstrapping. The covariance in the formula cannot be computed from the data directly, but by simulating pseudo-data vectors with mean $\hat{y}$ and independent errors with standard deviation $\hat{\sigma}$ and smoothing each of them.

In Figure 1 I have reproduced the data, the *lowess* fit, and the corresponding estimates of $\partial\hat{y}_i/\partial y_i$. In the paper Efron does not mention how he obtained them, but the example also appears in the recent book (Efron and Hastie, 2016), where numerical differentiation is mentioned. Following this approach, I made a small change (0.01) to each observation in turn and re-computed the fit.

These surprising results for $\partial\hat{y}_i/\partial y_i$ are caused by the non-linearity of the smoother. Figure 2 influence function of *lowess* and its derivative. The latter is negative for moderately large residuals. Their influence is reduced so aggressively that for the more extreme observations an increase of the size of the residual moves the smooth curve in the opposite direction.

If Efron would have used *locfit* instead of *lowess*, he would have obtained non-controversial estimates for the hat diagonal. P-splines would have been even more attractive, because the hat matrix (or the matrix K) is easily computed. Writing P-splines with a second order difference penalty as a mixed model (Currie and

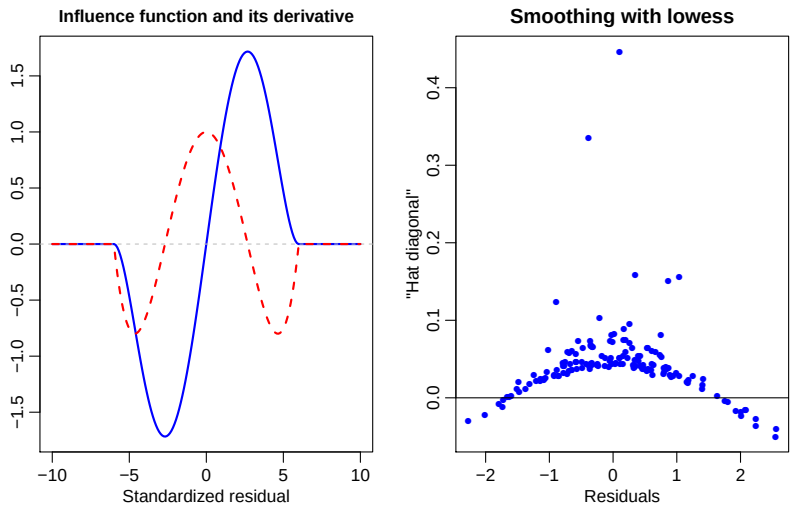


FIGURE 2. Left panel: the *lowess* smoother has a re-descending influence function, with a derivative that is negative for moderately large residuals. For standardized residuals with absolute values between 2.7 and 5, a fitted value will move towards the smooth trend when the observation moves away from it. Right panel: estimates of the hat diagonal vs the residuals for *lowess* smoothing of the kidney data.

Durban, 2002) and using the simplified Harville algorithm to estimate variances, we find that very strong smoothing is indicated. This means that effectively we get a straight line fit. The kidney data turn out to be no good candidate for smoothing at all: simple linear regression is sufficient.

2.3 The dimension of the model or the data?

The effective dimension is not a property of the data, but of the model, once penalty parameters have been set. Figure 3 shows simulated data (a sine wave plus independent errors) and the (optimized) fits obtained with the Whittaker smoother: $\hat{y} = (I + \lambda D_d' D_d)^{-1} y$, with D_d the matrix that forms differences of order d. For $d = 1$, the effective dimension is much larger than for $d = 2$ or $d = 3$, while the fits look very similar. The reason is that fist order differences do not allow much flexibility. A further increase of λ would give a smoother fit, but at the price of a larger bias.

These results have practical consequences. It is attractive to specify a smoother in terms of the effective dimension of the fitted model, because a penalty parameter like λ has no intuitive appeal. But a chosen effective dimension might lead to bias if the penalty is not flexible enough.

In the limit, when λ is very large, the Whittaker smoother fits a polynomial of degree $d-1$. The second order penalty can be written as $\lambda \sum_i (z_i - 2\phi z_{i-1} + z_{i-2})^2$, with $\phi = 1$. If we change that to $\phi = \cos(2\pi f/n)$, the limit is a (co)sine with f

periods on the domain from 0 to 1, if covered by n observations. Amplitude and phase are automatically det to values that give the least squares fit to the data. If the data indeed are a cosine wave plus noise, a very large λ would be indicated, with an effective dimension equal to 2 (or equal to 3, if the mean is not zero, and an extra parameter is introduced.) This shows that a good penalty should match the data, a largely unexplored aspect of smoothing.

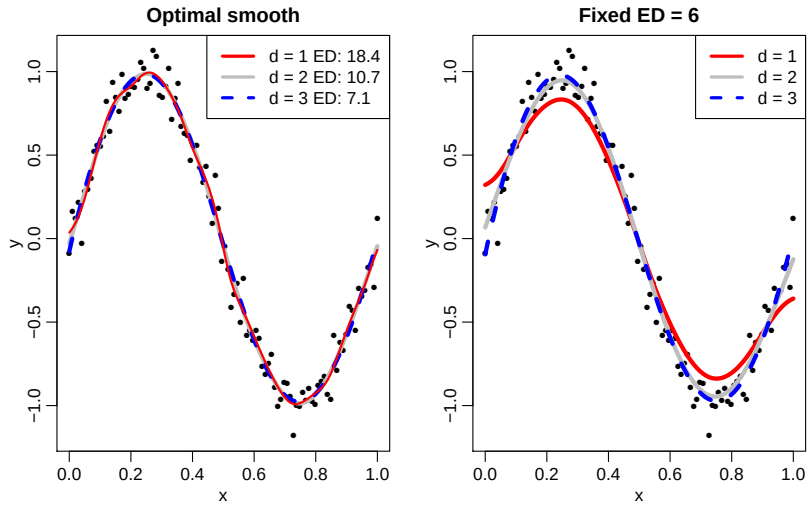


FIGURE 3. Whittaker smoother. Left: optimal smooth for various orders of the differences. Right: fits for a fixed effective dimension.

References

I D Currie, I.D. and Durban, M. (2002) Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, **2**, 333–349.

Efron, B. (2004). The Estimation of Prediction Error. *Journal of the American Statistical Association*, **99**, 619–632.

Efron, B. and Hastie, T (2016) *Computer Age Statistical Inference.* Cambridge University Press.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320-338.

Janson L., Fithian W. and Hastie T.J. (2015) Effective degrees of freedom: a flawed metaphor *Biometrika* **102**, 479–485.

Rodríguez-Álvarez, M. X., Boer, M. P., van Eeuwijk, F. A., and Eilers, P. H. C. (2016). Spatial Models for Field Trials. http://arxiv.org/abs/1607.08255.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719–727.

Velazco, J.G., Rodrguez-lvarez, M.X., Boer, M.P. et al. (2017) Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. *Theoretical and Applied Genetics*, doi:10.1007/s00122-017-2894-4.

Ye, J. (1998). On Measuring and Correcting the Effects of Data Mining and Model Selection. *Journal of the American Statistical Association*, **93**, 120–131.

Score inference in LASSO regression

Giovanna Cilluffo[1,2], Gianluca Sottile[1], Stefania La Grutta[2], Vito M.R. Muggeo[1]

[1] Department of Economical, Business and Statistical Sciences, University of Palermo (Italy),
[2] Institute of Biomedicine and Molecular Immunology A Monroy (IBIM) - National Research Council (CNR) of Palermo (Italy)

E-mail for correspondence: giovanna.cilluffo@unipa.it

Abstract: We discuss the score statistic to carry out inference on the regression coefficients in LASSO regression. The proposed approach relies on the Induced Smoothing framework and leads to results exhibiting good performance in different settings, including the high dimensional one $n < p$. We focus on interval estimation where few proposals have been discussed in literature with unsatisfactory results in some settings. We present results from some simulation experiments and an analysis of the well known prostate cancer dataset.

Keywords: Score Statistics; Induced Smoothing; LASSO; Confidence Intervals.

1 Introduction

The Least Absolutes Shrinkage and Selection Operator (LASSO) represents a very elegant and relatively widespread solution to carry out variable selection and parameter estimation simultaneously (Tibshirani 1996). While point estimation can be performed quite straightforwardly, possible current limitations are computation of standard errors and consequently inferences, namely reliable hypothesis testing and confidence intervals. Cilluffo et al. (2016) present a smooth approximation for the LASSO regression, based on the recent idea of induced smoothing (IS). The IS-LASSO allows to gain reliable standard errors which satisfactorily quantify the estimator variance, even for zero coefficient. In hypothesis testing problems, simulations show good performance of the IS-LASSO Wald statistic when compared with the two competitors covTest (Lockart et al. 2016) and postSel (Lee et al. 2014). However, the regression coefficient estimator is still biased for nonzero coefficient, that prevents the Wald statistic to be used for interval

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

estimation. In this paper we discuss an approach based on Score statistic. The paper is structured as follows: Section 2 describes the basics of the approach and Section 3 reports some simulation evidence. Section 4 is devoted to a real-data analysis, while conclusions are reported in the last section.

2 Methods

For the penalized objective $\frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$, the IS score is

$$\widetilde{U}(\beta) = X^T(y - X\beta) + \lambda\{2\Phi(\beta/v^{1/2}) - 1_p\},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal, v is the main diagonal of the covariance matrix of estimates, V, and a/b means the element-wise ratio of two vectors a and b. V is computed via the sandwich formula, namely $V = \widetilde{H}^{-1}\ \mathcal{I}\ \widetilde{H}^{-1}$ where $\widetilde{H} = \frac{\partial \widetilde{U}}{\partial \beta}$ is the hessian, and $\mathcal{I}$ is the Information matrix.

Let β_1 and β_2 be the interest and nuisance subsets of β, and $\widetilde{U}_j$, $\widetilde{H}_{jk}$ and $\mathcal{I}_{jk}$ $(j, k = 1, 2)$ the corresponding blocks of the Score vector, and of the hessian and Information matrices. It is well know that score inference on β_1 relies on the profiled score

$$\widetilde{U}_{1|2} = \widetilde{U}_1 - \widetilde{H}_{12}\widetilde{H}_{22}^{-1}\widetilde{U}_2, \tag{1}$$

which can be expressed in matrix notation via

$$\widetilde{U}_{1|2} = A\widetilde{U} = [I, -b][\widetilde{U}_1^{\mathrm{T}}, \widetilde{U}_2^{\mathrm{T}}]^{\mathrm{T}}$$

where I is the identity matrix, and $b = \widetilde{H}_{12}\widetilde{H}_{22}^{-1}$. The variance is easily obtained as

$$\mathbb{V}(\widetilde{U}_{1|2}) = A\mathbb{V}(\widetilde{U})A^T \quad \text{with} \quad \mathbb{V}(\widetilde{U}) = \mathcal{I} = \sigma^2(X^TX).$$

Unlike the usual inferential contexts where the regularity conditions are met, $\mathbb{E}[\widetilde{U}] \neq 0$ (at the true parameter value $\beta^0 \neq 0$), and thus the studentized Score statistic to be used for inference takes the form

$$\widetilde{S}_{1|2} = [\widetilde{U}_{1|2} - \mathbb{E}[\widetilde{U}_{1|2}]]^{\mathrm{T}}\mathbb{V}(\widetilde{U}_{1|2})^{-1}[\widetilde{U}_{1|2} - \mathbb{E}[\widetilde{U}_{1|2}]] \xrightarrow{d} \chi^2_{p_1}, \tag{2}$$

where p_1 is the dimension of the interest parameter β_1. We propose to use the studentized Score statistic (2) to carry out hypothesis testing and interval estimation for the regression coefficients in LASSO regression. In particular a $(1-\alpha)$ confidence interval for β_1 is given by

$$\mathrm{CI}_{1-\alpha} = \{\bar{\beta}_1 : \widetilde{S}_{1|2}(\bar{\beta}_1) \leq \chi^2_{p_1;1-\alpha}\}.$$

To illustrate, Figure 1 portrays an example of profile score for two coefficients in a toy dataset.

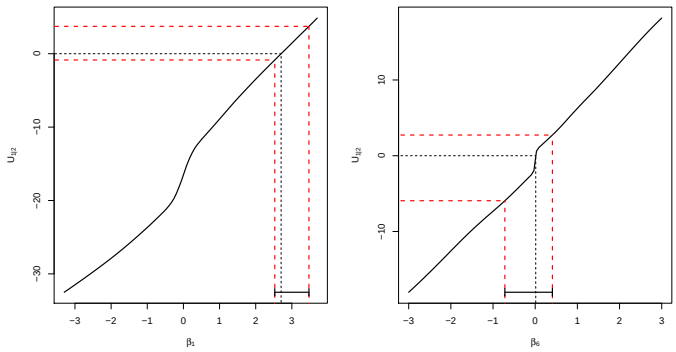


FIGURE 1. Illustrating the profile score with corresponding point estimate and 95% confidence intervals in a simulated dataset. The left and right panels refer to a non-zero and zero coefficient respectively. The functions have been shifted to guarantee $U_{1|2}(\hat{\beta}) = 0$ (thus the dashed horizontal lines do not correspond to quantiles $z_{.025}$ and $z_{.975}$).

3 Simulation Evidence

The proposed Score statistic can be employed for hypothesis testing and interval estimation, but we here focus on interval estimation only. At the best of our knowledge, the only approach currently available to build confidence intervals is via selective inference (pSel) due to Lee et al. (2014). Thus we compare Score-based confidence intervals with respect to those obtained via the pSel. We consider scenarios fulfilling the theoretical condition for guarantee consistency (Zhao and You 2006) of the LASSO, namely: irrepresentable condition; number of nonzero coefficients $\beta_1, \ldots, \beta_k$ (out of p covariates) such that $k \leq n/(2 \log p)$ having magnitudes at least $c\sigma\sqrt{2 \log p}$ for some unspecified numerical constant c (Wainwright 2009). We generate data from a linear regression model $y = X\beta + \epsilon$ with $X \sim N_p(0, I_p)$ and $\epsilon \sim N(0, 1)$ for five different scenarios given by some combinations of sample sizes $n = 50, 100$, and number of covariates $p = 20, 40, 60, 80, 120$ with only 5 non-null coefficients $\beta = (3, -3.1, 4, 3.5, -5, 0, \ldots, 0)^{\mathrm{T}}$. Standard theory of pSel provides inference for a fixed λ, but because it is unknown in practice, we use cross-validation for choosing λ at each simulation run. Table 1 shows results based on 500 and 100 runs, respectively in low and high dimensional settings, wherein the tuning parameter λ has been selected via 5-fold cross validation at each replicate.

Coverage levels referring to the first 10 coefficients are comparable for the nonzero coefficients at the 'simpler' scenario of $n = 50$ and $p = 20$, but otherwise pSel exhibits coverage levels lower than the nominal one, 0.95. Interestingly when the sample size increases to $n = 100$ (keeping fixed the ratio p/n) the coverage levels do not improve. The proposed score-based confidence intervals exhibit good performance in all scenarios.

We also consider scenarios in which the theoretical conditions are not met, namely with same combinations of n and p, but $\sigma = 3$. Table 2 reports results. Coverage

TABLE 1. Coverage levels of 95% CIs from Score and pSel for 10 selected parameters. Results are based on 500 and 100 replicates in low and high dimensional setting, σ is equal to 1 and the optimal λ has been obtained via cross validation.

		coefficents true value									
(n, p)		3	-3.1	4	3.5	-5	0	0	0	0	0
(50,20)	Score	0.95	0.95	0.95	0.95	0.95	0.95	0.96	0.95	0.96	0.94
	pSel	0.94	0.95	0.94	0.95	0.95	1.00	1.00	1.00	1.00	1.00
(50,40)	Score	0.96	0.96	0.96	0.97	0.98	0.95	0.98	0.97	0.98	0.96
	pSel	0.63	0.63	0.62	0.65	0.64	0.87	0.95	0.96	0.94	0.94
(100,80)	Score	0.99	0.98	0.99	0.98	0.97	0.99	0.98	0.98	0.99	0.99
	pSel	0.39	0.39	0.39	0.39	0.39	0.91	0.89	0.89	0.86	0.90
(50,60)	Score	0.98	0.94	0.99	0.94	0.96	0.95	0.95	0.97	0.93	0.95
	pSel	0.76	0.73	0.77	0.75	0.75	1.00	0.99	1.00	1.00	1.00
(100,120)	Score	0.98	0.92	0.97	0.96	0.99	0.96	0.97	0.96	0.95	0.96
	pSel	0.88	0.85	0.88	0.87	0.89	1.00	0.99	0.96	0.99	1.00

levels from pSel are again lower than the nominal level, even in the simple scenario of $n = 50$ and $p = 20$. In high dimensional settings, pSel performs quite bad.

TABLE 2. Coverage levels of 95% CIs from Score and pSel for 10 selected parameters. Results are based on 500 and 100 replicates in low and high dimensional setting, σ is equal to and the optimal λ has been obtained via cross validation.

		true value									
(n, p)		3	-3.1	4	3.5	-5	0	0	0	0	0
(50,20)	Score	0.96	0.97	0.95	0.97	0.97	0.96	0.97	0.97	0.97	0.94
	pSel	0.64	0.64	0.65	0.65	0.64	0.94	0.90	0.92	0.90	0.91
(50,40)	Score	0.96	0.97	0.99	0.98	0.99	0.98	0.98	0.98	0.98	0.98
	pSel	0.49	0.48	0.49	0.49	0.48	0.74	0.86	0.86	0.88	0.87
(100,80)	Score	0.99	0.98	0.98	0.97	0.98	0.97	0.98	0.98	0.99	0.97
	pSel	0.33	0.34	0.34	0.34	0.35	0.86	0.84	0.88	0.83	0.84
(50,60)	Score	0.99	0.99	0.99	0.99	0.99	1.00	0.97	0.98	0.98	0.99
	pSel	0.25	0.24	0.22	0.23	0.25	0.86	0.82	0.86	0.80	0.83
(100,120)	Score	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
	pSel	0.34	0.32	0.33	0.33	0.33	0.92	0.93	0.90	0.90	0.94

4 Application and conclusion

We apply the IS-Lasso to the well-known Prostate Cancer dataset analyzed in Tibshirani (1996). Data refer to $n = 97$ subjects, $p = 8$ covariates and the response variable is the log of prostate specific antigen. Table 3 reports the estimates from ordinary LASSO along with p-values and the 95% confidence intervals returned by pSel and the proposed Score.

Findings are substantially similar between Score and pSel for all but one coefficient. In fact for the variable svi, pSel returns a somewhat 'large' p-value and confidence interval including the zero, while the Score inference provides evidence for a significant effect of such covariate. However for the remaining coefficients with the same findings in terms of significance, the Score based confidence intervals are generally narrower. Results for pSel have been obtained via the package

TABLE 3. Estimates, p-values and 95% confidence intervals for the regression coefficients in the Prostate Cancer dataset. The widths are reported below each CI. Estimates come from ordinary LASSO, p-values and confidence interval come from Score statistic and pSel.

		p-value		95% confidence interval	
covariate	Est	Score	pSel	Score	pSel
lcavol	0.517	0.000	0.000	0.394, 0.726	0.395, 0.717
				0.332	0.322
lweight	0.349	0.008	0.038	0.122,0.820	-0.057, 0.783
				0.698	0.840
age	-0.001	0.118	0.203	-0.039, 0.004	-0.039, 0.026
				0.044	0.065
lbph	0.053	0.078	0.099	-0.012, 0.228	-0.066, 0.222
				0.240	0.288
svi	0.570	0.003	0.077	0.223, 1.141	-0.290, 1.069
				0.918	1.359
lcp	0.000	0.265	*	-0.283, 0.079	*
				0.363	*
gleason	0.000	0.812	*	-0.275, 0.346	*
				0.622	*
pgg45	0.002	0.424	0.166	-0.005, 0.012	-0.004, 0.010
				0.170	0.014

*lcp and gleason variable are not selected by pSel

`selectiveInference` version 1.2.2 (Tibshirani et al., 2017). An R package to perform Score-based inference in LASSO regression will be distributed in due time.

5 Conclusions

The proposed Score statistic seems to be a good inferential tool to build confidence intervals in LASSO regression. The coverage levels of the interval estimators are pretty close to the nominal level in different scenarios, even when the theoretical conditions are not met. While current results are quite promising, further research is necessary to compare our approach with others not yet considered, such as the method proposed by Javanmard and Montanari (2014), which use a desparsifying LASSO to build CI.

References

Cilluffo, G., Fasola, S., La Grutta, S. and Muggeo, V.M.R. (2016) The induced smoothed LASSO. In: *Proceedings of the 31st International Workshop on Statistical Modelling*, Rennes, (France), **1**, 69–74.

Javanmard, A., Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15**, 2869–2909.

Lee, J.D., Sun, D.L., Sun, Y., and Taylor, J. E. (2016) Exact post-selection inference, with application to the lasso. *Annals of Statistics*, **44**, 907–927.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014) A significance test for the lasso. *Annals of statistics*, **42**, 413.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, **58**, 267–288.

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus., J., and Reid, J., (2017) selectiveInference: Tools for Post-Selection Inference. *R package version 1.2.2*, https://CRAN.R-project.org/package=selectiveInference

Wainwright, M.J. (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE Trans. Inform. Theory*, **55**, 2183–2202.

Zhao, P., and Yu, B. (2006) On model selection consistency of Lasso. *Journal of Machine learning research*, **7**, 2541–2563.

Using chain graph models for structural inference with an application to linguistic data

Craig Alexander[1], Jane Stuart-Smith[2], Tereza Neocleous[1], Ludger Evers [1]

[1] School of Mathematics and Statistics, University of Glasgow, UK
[2] School of Critical Studies, University of Glasgow, UK

E-mail for correspondence: c.alexander.1@research.gla.ac.uk

Abstract: Graphical models provide a visualisation of the conditional dependence structure between variables, making them an attractive inference tool. The improved readability makes this an appealing approach to represent complex model output to non-statisticians. In this paper, we introduce a novel approach using graphical models to visualise the output of a mixed effects model with multivariate response with an application to linguistic data.

Keywords: Hierarchical models; Graphical models; Linguistic change.

1 Introduction

In this work, we discuss the use of a chain graph model structure to represent the output from a hierarchical regression model with multivariate response. The chain graph model is inferred in three parts. The dependency structure of the covariates is modelled independently using standard structural inference methods in graphical models. The relationship between the response and explanatory variables and the dependence structure between response variables is jointly inferred using a multivariate Bayesian hierarchical model, where the precision estimates of the residuals and random effects are assumed to conform to an undirected graphical model. We report an application of this model to linguistic data obtained from the Sounds of the City corpus, consisting of Glaswegian speech recordings from the 1970's to the 2000's. From this data, we look to recover the underlying chain graph model detailing which factors affect vowel change.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

2 Chain graph model structure

Implementing a chain graph (Lauritzen & Wermuth, 1989) structure allows the use of directed and undirected edges. Nodes are partitioned into blocks, one for explanatory variables and one for response variables. Edges within blocks are undirected and edges between blocks are directed.

The directed edges between blocks are modelled using a Bayesian hierarchical model. The response is defined as y_{ij}^{l}, which is the j^{th} measurement from the i^{th} group on the l^{th} response. $\boldsymbol{\beta}^{l}$ is the corresponding vector of regression coefficients. The random effects coefficients are denoted by $\mathbf{b}^{l}$.

In vector-matrix notation, the likelihood is defined as:

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}, \mathbf{X}) = \mathcal{N}\left(\boldsymbol{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{b}, \left(\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}^{-1} \otimes \mathbf{I}\right)\right). \tag{1}$$

As conjugate priors have been specified for each parameter where possible, a Gibbs sampler can be used for parameter inference.

The presence of a directed edge in the graphical model corresponds to the value of the respective β_i^j coefficient:

- $\beta_i^j \neq 0 \rightarrow$ an edge is present between variable i and response j.
- $\beta_i^j = 0 \rightarrow$ no edge is present between variable i and response j.

If any interactions are in the model, a factor graph structure is used to represent these, with the interacting explanatory variables connecting to a factor node, which is connected to the relevant response variable.

At each step of the sampler, a new candidate model is proposed by either adding or removing an explanatory variable. By integrating out $\boldsymbol{\beta}$, model evidences are then computed to determine whether we accept or reject the candidate model.

For random effect parameters $\mathbf{b}$ a Gaussian prior is specified of the form:

$$\mathbf{b} \mid \boldsymbol{\Omega}_{\mathbf{b}} \sim \mathcal{N}\left(\mathbf{0}, \left(\boldsymbol{\Omega}_{\mathbf{b}}^{-1} \otimes \mathbf{I}\right)\right). \tag{2}$$

To maintain conjugacy, the random effect and model error precision matrices have G-Wishart (Dobra *et al*, 2011) hyperpriors placed on them:

$$\boldsymbol{\Omega}_{\mathbf{b}}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}} | \mathrm{G} \sim \mathcal{W}_{\mathrm{G}}(\nu_{\mathbf{b}}, \mathbf{S}_{\mathbf{b}}). \tag{3}$$

where G is a defined graph.

The precision estimates from the hierarchical model are taken as input to a zero mean Gaussian graphical model, defined as:

$$\mathcal{M}_G = \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}^{-1}). \tag{4}$$

where $\omega_{i,j} = 0$ corresponds to a missing edge between response i and j.

The normalising constant, $I_{\mathrm{G}}(\nu, \mathbf{S})$, for chordal graphs can be obtained via a closed form solution by factorising it into a product of density functions:

$$I_{\mathrm{G}}(\nu, \mathbf{S}) = \frac{\prod_{i=1}^{d} I_{T_i}(\nu, S_{T_i,T_i})}{\prod_{j=1}^{d-1} I_{S_i}(\nu, S_{S_i,S_i})}. \tag{5}$$

where T_i are cliques and S_i the separators of G. In the case of a non-chordal graph, model selection can be performed by using trans-dimensional MCMC methods such as those discussed in Mohammadi & Wit (2015).

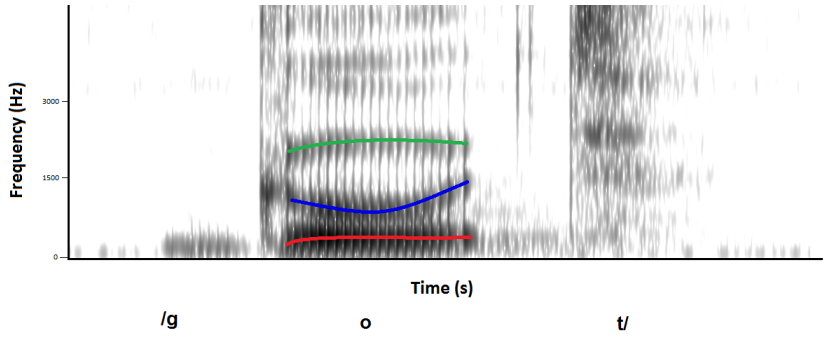


FIGURE 1. Spectrogram of the word *GOAT* spoken by a male Glaswegian speaker. The three coloured lines show the first three formants of the vowel /o/.

3 Application: vowel change in Glasgow

We apply the methodology to data from the Sounds of the City (Stuart-Smith, 2017) project, which is a study modelling vowel change in the Glaswegian dialect over the 20th century. Recordings of speakers over various decades are used and vowel measurements are taken along with phonetic quantities of interest and social and biological factors.

Acoustically, a vowel can be characterised by its main resonances, known as formants (Hz). This is illustrated in Figure 1, with the /o/ vowel formants represented by the three coloured lines. Vowel change is studied in terms of how such frequencies alter over time.

Figure 2 shows the best posterior model selected for the vowel /o/, found in words like GOAT, HOPE, etc, selected with posterior probability 0.414.

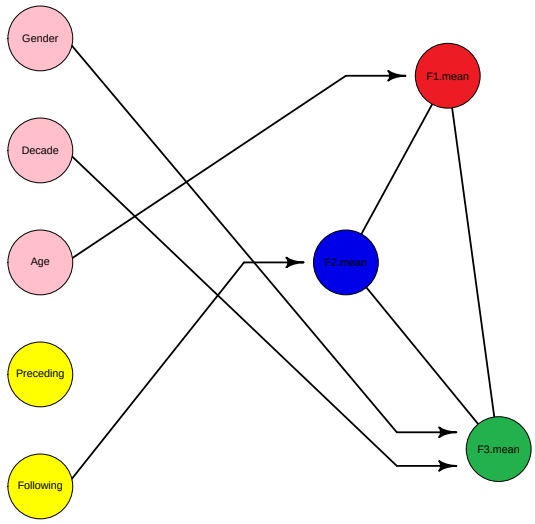


FIGURE 2. Best posterior model selected for the *GOAT* vowel.

From the graphical model, we observe the following significant predictors:

- Age for F1; younger speakers are leading the change in vowel quality. Shifts in F1 relate to raising of the vowel quality, so /o/ becoming more like /u/.
- Shifts in F2 relate to fronting/backing of the vowel quality according to the nature of following consonant.
- Decade and Gender for F3, evidence of vowel change over time and additional differences due to speaker gender.

4 Conclusions & future work

In this work, we have extended beyond previous modelling of sociolinguistic data, by considering multiple formants within the one model, accounting for the high correlation between formants. This approach can also be extended beyond linguistic data, with several academic trials using datasets with similar multivariate and nested features.

In order to promote the usability of our method, we aim to turn it into an online application, allowing users to input their own data.

Acknowledgments: CA is grateful to the Lord Kelvin Adam Smith scholarship for his research studentship.

References

Lauritzen, S.L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, **17**, 31-57.

Dobra, A., *et al* (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *JASA* , **106**.

Stuart-Smith, J., *et al* (2017). *Changing sounds in a changing city. Language and a sense of place*, E. Moore & C. Montgomery, CUP, 38-64.

Mohammadi, A. and Wit, E.C. (2015) Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian Analysis*, **10**, 109-138.

Sample quantiles corresponding to mid p-values for zero–modification tests

Paul Wilson[1], Jochen Einbeck[2]

[1] School of Mathematics and Computer Science/Statistical Cybermetrics Research Group, University of Wolverhampton, WV1 1LY, United Kingdom
[2] Department of Mathematical Sciences, Durham University, DH1 3LE, United Kingdom

E-mail for correspondence: pauljwilson@wlv.ac.uk

Abstract: Wilson and Einbeck (2015, 2016) propose a test for zero-modification relative to a stated model. The basis of the test is that the number of observed zeros follows a Poisson-binomial distribution. The decision to reject, or otherwise, the non zero–modified model is made by either (i) computing the mid p-value corresponding to the number of observed zeros, or (ii) comparing the number of observed zeros to the relevant "traditional" quantile of the appropriate Poisson–binomial distribution. In general either approach will result in the same decision, but occasionally discrepancies may occur. In this paper we investigate the use of mid-distribution quantiles in approach (ii) above, and show that this reduces the possibility of discrepancies.

Keywords: zero-modification, mid p-values, quantiles

1 Introduction

Wilson and Einbeck (2015) proposed a new and intuitive test for zero-modification that uses the observed number of zeros, n_0, in a given sample $y = y_1, y_2, \ldots, y_n$ from count variables Y_i and a set of covariates x_i, $i = 1, 2, \ldots, n$ to establish whether the distributional assumption $Y_i|x_i \sim G(y_i|\mu_i)$ where μ_i is a pre-specified parametric function of the x_i is consistent with N_0, the distribution of the number of zeros under G. This is achieved by referencing the value of n_0 to the appropriate Poisson-Binomial distribution (Chen and Liu, 1997).

To illustrate, consider the case where G is a Poisson model, and thus $p_i = p(0|\mu_i) = e^{-\mu_i}$ and let T_i be a random variable which takes the value 1 if $y_i = 0$ and 0 otherwise. Clearly T_i is a Bernoulli random variable with parameter p_i and

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

thus N_0 may be formulated as the sum over independent Bernoulli experiments $T_1, T_2, \ldots, T_n$.

Based on this simple observation, consider the special case that there are no covariates, that is $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$. In this case, the p_i's are equal also, and so the distribution of N_0 is a binomial distribution $Bin(n, p)$, where $p = e^{-\mu}$, and thus has mean np and variance $np(1-p)$. Based on this distribution, one can immediately compute quantiles corresponding to a given significance level, and use these as critical values for the test; alternatively one may determine the p-value corresponding to n_0, and reject or otherwise the Poisson model based upon this. If the μ_i do depend on covariates, N_0 is the sum of n independent Bernoulli random variables $T_1, T_2, \ldots, T_n$, and hence is a Poisson-Binomial distribution with parameters $p_1, p_2, \ldots, p_n$ and one proceeds by computing quantiles or p-values relative to this distribution, using, for example, the R package `poibin` (Hong, 2013).

Wilson and Einbeck (2016) proposed the use of *mid p-values*

$$\hat{\alpha}_{T,0.5}(t) = P_0[T > t] + 0.5P_0[T = t] = 0.5\left(P_0\left[T \geq t\right] + P_0[T \geq t + 1]\right)$$

which Franck (1986) argues are more appropriate when the test statistic is discrete. Note that if T were continuous, then $P_0[T = t] = 0$ and the mid p-value is equivalent to the "traditional" p-value. It may be shown that the attainment rate of the proposed test when mid p-values are employed is superior to that when traditional p-values are used.

Wilson and Einbeck (2015) utilise the "traditional" quantile $Q(p) = \inf\{t \mid F(t) \geq p\}$ where $F(x) = P(X \leq x)$ is the cumulative distribution function of a random variable X. This may lead to discrepancies. An example, based upon the one-sided version of the test (i.e. we are testing for zero-inflation only), is the following:

1.1 Trajan Data

The data are the number of roots produced by $n = 270$ micropropagated shoots of the columnar apple cultivar Trajan. During the rooting period, all shoots were maintained under identical conditions, but the shoots themselves were cultured on media containing different concentrations of the cytokinin BAP, in growth cabinets with an 8 or 16 hour photoperiod. Full details of the experiment are to be found in Marin (1993). A striking feature of the data is that although almost all shoots produced under the 8 hour photoperiod rooted, only about half of those produced under the 16 hour photoperiod did. Overall $n_0 = 64$ shoots produced zero roots, of which only 2 were from the shorter photoperiod.

These data were analysed by Ridout and Demétrio (1992) and Ridout et al. (1998). If the model of the null hypothesis is a negative binomial (type–II) model, where both the mean and the size parameter are modelled by photoperiod, then a (mid) p-value of 0.0871 for the test of Wilson and Einbeck (2015, 2016) is returned, indicating non-rejection of the negative binomial model at $\alpha = 0.05$. The traditional $5th$ and $95th$ quantiles of the distribution of N_0 are 47 and 66; the interval $[47, 66]$ is referred to as a 90% *fluctuation interval*. As $n_0 = 64$ is interior to this interval we conclude that n_0 *is* consistent with such a model (and

inconsistent with the zero-inflated model) at a level of significance of $\alpha = 0.05$. An 80% fluctuation interval however is $[49, 64]$, and thus based upon this fluctuation interval we would fail to reject the negative-binomial model in favour of the strictly zero-inflated model at a level of significance of 0.10, but we would do so under the "p–value criterion".

2 Quantiles based on mid-distribution functions

Let X be a discrete random variable with distinct values $v_1 < v_2 < \cdots < v_d$, let $P(X = v_i) = p_i$. Ma et al. (2011) recommend the following quantile function for discrete distributions:

$$Q(p) = F^{-1}_{\text{mid}}(p) = \begin{cases} v_1 & \text{if } p < p_1/2 \\ v_k & \text{if } p = \pi_k,\ k = 1, \ldots, d \\ \lambda v_k + (1-\lambda) v_{k+1} & \text{if } p = \lambda\pi_k + (1-\lambda)\pi_{k+1} \\ & 0 < \lambda < 1,\ k = 1, \ldots, d-1 \\ v_d & \text{if } p > \pi_d \end{cases}$$

Where $\pi_k = \sum_{i=1}^{k-1} p_i + p_k/2$, that is, π_k is a lower-tailed mid-p-value.

2.1 Example: Mid Quantiles for a Binomial Distribution

Let $X \sim \text{Bin}(7, 0.35)$, and thus X has pmf and cdf:

x	0	1	2	3	4	5	6	7
$P(X = x)$	0.049	0.185	0.298	0.268	0.144	0.047	0.008	0.001
$P(X \leq x)$	0.049	0.234	0.532	0.800	0.944	0.991	0.999	1.000

and hence the "traditional" 90^{th} quantile of X is 4.

We determine the "mid–quantile" as follows:

$v_5 = 4, v_6 = 5$, $p_5 = 0.144$, $p_6 = 0.047$.

Hence $\pi_4 = 0.800 + 0.144/2 = 0.8720$, $\pi_5 = 0.944 + 0.047/2 = 0.9675$.

Note that $0.9 = 0.707\pi_4 + (1 - 0.707)\pi_5$, hence:

$$Q(0.9) = F^{-1}_{\text{mid}}(0.9) = 0.707 v_4 + (1 - 0.707) v_5 = 3.213$$

2.2 Example: Simulated Poisson Data

The 25 data of Table 1 are a random draw from a random variable W that is believed to follow a Poisson distribution. It is wished to test this belief.

It is estimated, using the adaptive mixture estimator of Wilson and Einbeck (2016), that the mean of W is $\mu = 1.171$, and hence under the null (Poisson) model $P(W = 0) = \exp(-1.171) = 0.310$. Hence the observed number of zeros in

TABLE 1.

0	1	2	3
16	4	4	1

random samples of size 25 drawn from W will be $Bin(25, 0.310)$ distributed. The "traditional" 2.5^{th} and 97.5^{th} quantiles of such a distribution are 7 and 16 respectively, and hence a 95% fluctuation interval for the number of observed zeros under the Poisson distribution is $[7, 16]$ indicating non-rejection of the Poisson model at a level of significance of $\alpha = 0.05$, consistent with the traditional p-value of 0.064, but inconsistent with the mid p-value of 0.045. The 95% fluctuation interval based upon the mid quantiles is however $[6.52, 15.57]$, consistent with the mid p-value. These results are summarised in Table 2.2.

TABLE 2. $n = 25$, H_0:Poisson

$n_0 = 16$	p-value	$95\%FI$
traditional	0.064	$[7, 16]$
mid	0.045	$[6.52, 15.57]$

2.3 Example: Trajan Data Revisited

Here we re-compute the 80% fluctuation interval for the negative binomial model fitted to the Trajan data of Section 1.1 using the mid-distribution quantiles defined above. (Recall, here we are testing for strict zero-inflation, and thus the upper bound of the fluctuation interval serves as a test statistic for a one-sided test). We find that $\pi_{47} = 0.073$ and $\pi_{48} = 0.101$, thus $0.1 = 0.069\pi_{47}+(1-0.069)\pi_{48}$ and hence $Q(0.1) = (0.069 \times 47) + ((1 - 0.069) \times 48) = 47.931$. Similarly $\pi_{63} = 0.882$ and $\pi_{64} = 0.913$, thus $0.9 = 0.419\pi_{63} + (1 - 0.419)\pi_{64}$ and hence $Q(0.9) = (0.419 \times 63) + ((1 - 0.419) \times 64) = 63.581$. Thus, using the mid-distribution quantile we obtain a 80% fluctuation interval of $(47.931, 63.581)$, and hence $n_0 = 64$ is exterior to the confidence interval, and we reject the negative-binomial model in favour of the zero-inflated negative binomial model under both criteria. These results are summarised in Table 2.3.

TABLE 3.

$n_0 = 64$	p-value	90% FI	80% FI
traditional	0.1010	$[47, 66]$	$[49, 64]$
mid	0.0871	$[46.902, 65.679]$	$[47.931, 63.581]$

3 Conclusion

Decisions based upon mid-distribution quantiles as defined above will agree with those based upon mid p-values unless $p < p_1/2$ or $p > \pi_d$. With respect to the test

proposed in Wilson and Einbeck (2015, 2016) these exceptions correspond to the observed data either containing no zeros, or consisting entirely of zeros, and hence the adoption of quantiles based upon mid-distribution functions results in fluctuation intervals that nearly entirely removes discrepancies that may sometimes occur between decisions based upon fluctuation intervals and mid p–values. Given that the power and attainment rates of the test when based upon mid p-values are excellent, such alignment is desirable. The adoption of such quantiles is straightforward. In this paper we only discuss the use of mid-distribution quantiles in relation to the test of Wilson and Einbeck (2015, 2016), but their application to other tests with discrete test statistics is worthy of investigation.

References

Chen, S.X. and Liu, J.S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875–892.

Hong, Y. (2013). poibin: The Poisson Binomial Distribution. R package version 1.2. https://CRAN.R-project.org/package=poibin

Ma, Y., Genton M. and Parzen, E. (2011) Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics* **63**, 227 – 243.

Ridout and Demétrio (1992) Generalized Linear Models for Positive Count Data. *Revista de matematica e estatistica* **10**, 139 – 148.

Ridout, M.S., Demétrio C.B. and Hinde, J. (1998) Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference* **19**, 179 – 192).

Wilson, P. and Einbeck, J. (2015). A simple and intuitive test for number–inflation or number–deflation. In: Wagner, H. and Friedl, H. (Eds). Proc's of the 30th IWSM, Linz, Austria, Vol 2, pages 299–302.

Wilson, P. and Einbeck, J. (2016). On statistical testing and mean parameter estimation for zero–modification in count data regression. In: Dupuy, J. and Josse, J. (Eds). Proc's of the 31st IWSM, Rennes, France, Vol 1, pages 325 – 330.

Two wrongs make a right: addressing underreporting in binary data from multiple sources

Betsabe Blas[1], Scott J. Cook[2], Raymond J. Carroll[3], Samiran Sinha[4]

[1] Department of Statistics, Federal University of Pernambuco, Brazil
[2] Department of Political Science, Texas A&M University, US
[3] Department of Statistics, Texas A&M University and School of Mathematical Sciences, University of Technology Sydney
[4] Department of Statistics, Texas A&M University

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

Abstract: Media-based event data i.e., data comprised from reporting by media outlets are widely used in research in political science. However, events of interest (e.g., strikes, protests, conflict, etc.) are often underreported by these primary and secondary sources, producing incomplete data that risks inconsistency and bias in subsequent analysis. While general strategies exist to help ameliorate this bias, these methods do not make full use of the information often available to researchers. Specifically, much of the event data used in the social sciences is drawn from multiple, overlapping news sources (e.g., Agence France-Presse, Reuters, etc.). Therefore, we propose a novel maximum likelihood estimator that corrects for misclassification in data arising from multiple sources. In the most general formulation of our estimator, researchers can specify separate sets of predictors for the true-event model and each of the misclassification models characterizing whether a source fails to report on an event. As such, researchers are able to accurately test theories on both the causes of and reporting on an event of interest. Simulations evidence that our technique regularly outperforms current strategies that either neglect misclassification, the unique features of the data-generating process, or both. We also illustrate the utility of this method with a model of repression using the Social Conflict in Africa Database.

Keywords: missclasification; binary model; multi-source.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

1 Misclassification

Misclassification is likely to occur with media-based event data, where primary- or secondary-source reports fail to include the occurrence of an actual event. To introduce our method, consider two news outlets, 1 and 2, providing reports, $\mathbf{Y_1}$ and $\mathbf{Y_2}$, on the event of interest $\mathbf{Y_T}$. However, we possess two incomplete reports $\mathbf{Y_1} \neq \mathbf{Y_T}$ and $\mathbf{Y_2} \neq \mathbf{Y_T}$, explained by $\text{pr}(\mathbf{Y_1} = 1|\mathbf{X}, \mathbf{Z_1})$ and $\text{pr}(\mathbf{Y_2} = 1|\mathbf{X}, \mathbf{Z_2})$, where $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are predictors of the (mis-)reporting of an event (e.g., distance to reporting office) by that source, which are otherwise unrelated to $\mathbf{Y_T}$. Following convention in the applied literature, we aggregate these sources to reduce the individual missingness by $\mathbf{Y_{sum}} = \mathbf{1}(\mathbf{Y_1} + \mathbf{Y_2} \geq 1)$. If $\mathbf{Y_{sum}} = \mathbf{Y_T}$, the data are complete and we find that

$$\text{pr}(\mathbf{Y_T} = 1|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) = \text{pr}(\mathbf{Y_T} = 1|\mathbf{X}) = F(\beta^t \mathbf{X}), \tag{1}$$

where $F(\cdot)$ is specified up to the parameter β. We are interested in Eq. 1. However, where $\mathbf{Y_{sum}} \neq \mathbf{Y_T}$, we are unable to simplify as in 1. This means that when observed outcomes are misclassified, fitting Equation 1 will result in biased estimates of $\mathbf{X}$ on $\mathbf{Y_T}$. We make the following assumptions:(a) $\mathbf{Y_1}$ and $\mathbf{Y_2}$ are independent given $(\mathbf{Y_T}, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})$. (b) $\mathbf{Y_{sum}} = 1$ implies $\mathbf{Y_T} = 1$ with probability 1. (c) $\mathbf{Y_T} = 0$ implies that $\mathbf{Y_{sum}} = 0 = \mathbf{Y_1} = \mathbf{Y_2}$ with probability 1.

If we treat $\mathbf{Y_{sum}}$ as the response variable, the problem is related to one studied by Carroll and Pederson (1993), and Hausman et al. (1998). The misclassification probabilities are $\text{pr}(\mathbf{Y_{sum}} = 0|\mathbf{Y_T} = 1, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) = \gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})$, and $\text{pr}(\mathbf{Y_{sum}} = 1|\mathbf{Y_T} = 0, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) = 0$, which follows from Assumption (b). In Hausman et al. (1998) these misclassification probabilities do not depend of covariates, and are instead simply an unknown constant to be estimated. Therefore, we generalize Hausman et al. (1998)'s estimator to allow for misclassification probabilities that are dependent upon the covariates. We have that

$$\begin{aligned}
\text{pr}(\mathbf{Y_{sum}} = 0|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \{1 - \gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})\}\{1 - F(\mathbf{X}, \beta)\} + \gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}); \\
\text{pr}(\mathbf{Y_{sum}} = 1|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \{1 - \gamma(\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})\}F(\mathbf{X}, \beta).
\end{aligned}$$

However, the data are not $(\mathbf{Y_{sum}}, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})$, but $(\mathbf{Y_1}, \mathbf{Y_2}, \mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2})$, that is, we have multiple sources of data. As such, there may be different misclassification rates, which is the fundamental difference between our estimator and existing approaches. Rather than neglect this information, thereby failing to use all of the data, we define

$$\begin{aligned}
\alpha_1(\mathbf{X}, \mathbf{Z_1}) &= \text{pr}(\mathbf{Y_1} = 0|\mathbf{Y_T} = 1, \mathbf{X}, \mathbf{Z_1}); \\
\alpha_2(\mathbf{X}, \mathbf{Z_2}) &= \text{pr}(\mathbf{Y_2} = 0|\mathbf{Y_T} = 1, \mathbf{X}, \mathbf{Z_2}).
\end{aligned}$$

Here by Assumption (b) we have that $\text{pr}(\mathbf{Y_1} = 1|\mathbf{Y_T} = 0, \mathbf{X}, \mathbf{Z_1}) = \text{pr}(\mathbf{Y_2} = 1|\mathbf{Y_T} = 0, \mathbf{X}, \mathbf{Z_2}) = 0$.

Under Assumption (a-c), we have

$$\begin{aligned}
\text{pr}(\mathbf{Y_1} = 0, \mathbf{Z_2} = 0|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= 1 - F(\mathbf{X}, \beta) + \alpha_1(\mathbf{X}, \mathbf{Z_1})\alpha_2(\mathbf{X}, \mathbf{Z_2})F(\mathbf{X}, \beta); \\
\text{pr}(\mathbf{Y_1} = 0, \mathbf{Z_2} = 1|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \alpha_1(\mathbf{X}, \mathbf{Z_1})\{1 - \alpha_2(\mathbf{X}, \mathbf{Z_2})\}F(\mathbf{X}, \beta); \\
\text{pr}(\mathbf{Y_1} = 1, \mathbf{Z_2} = 0|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \{1 - \alpha_1(\mathbf{X}, \mathbf{Z_1})\}\alpha_2(\mathbf{X}, \mathbf{Z_2})F(\mathbf{X}, \beta); \\
\text{pr}(\mathbf{Y_1} = 1, \mathbf{Z_2} = 1|\mathbf{X}, \mathbf{Z_1}, \mathbf{Z_2}) &= \{1 - \alpha_1(\mathbf{X}, \mathbf{Z_1})\}\{1 - \alpha_2(\mathbf{X}, \mathbf{Z_2})\}F(\mathbf{X}, \beta).
\end{aligned}$$

Thus, we can construct the likelihood function.

TABLE 1. Simulation Study Results

	Naive Probit	Hausman Const Pr	Hausman w/ Cov	Multi-Source Const Pr	Multi-Source w/ Cov
		$\beta_0 = -1$			
Bias	0.074	−0.663	−0.023	−0.048	0.003
Std	0.051	0.227	0.106	0.060	0.063
SE	0.055	0.175	0.096	0.067	0.063
MSE	0.008	0.490	0.012	0.006	0.004
CP(%)	77.4	3.7	95.8	91.3	95.8
		$\beta_1 = 1$			
Bias	0.397	−0.379	−0.035	0.287	−0.010
Std	0.055	0.258	0.158	0.092	0.094
SE	0.057	0.218	0.132	0.079	0.095
MSE	0.161	0.210	0.026	0.091	0.009
CP(%)	0.0	56.9	94.0	11.1	96.3

2 Simulations

We consider the following five methods: **Naïve Probit**: $\text{pr}(Y_{\text{sum}} = 1|X) = \Phi(\beta_0 + \beta_1 X)$. **Hausman, Constant Probabilities**: the approach outlined in Hausman et al. (1998), which fits $\text{pr}(Y_{\text{sum}} = 1|X) = \{1 - \text{pr}(Y = 1|Y_T = 1, X)\}\Phi(\beta_0 + \beta_1 X)$. **Hausman with Covariates**: our generalization of Hausman et al. (1998), $\text{pr}(Y_{\text{sum}} = 0|Y_T = 1, X, Z_1, Z_2) = \gamma(X, Z_1, Z_2) = \Phi(\eta_{00} + \eta_{01}X + \eta_{02}Z_1 + \eta_{03}Z_2)$, giving event probabilities $\text{pr}(Y_{\text{sum}} = 1|X, Z_1, Z_2) = \{1 - \gamma(X, Z_1, Z_2)\}\Phi(\beta_0 + \beta_1 X)$. **Multi-Source, Constant Probabilities**: our multi-source method, $\text{pr}(Y_1 = 0|Y_T = 1, X, Z_1) = \Phi(\eta_{10})$ and $\text{pr}(Y_2 = 0|Y_T = 1, X, Z_2) = \Phi(\eta_{20})$. **Multi-Source with Covariates**: our general multi-source method, $\text{pr}(Y_1 = 0|Y_T = 1, X, Z_1) = \Phi(\eta_{10} + \eta_{11}X + \eta_{12}Z_1)$ and $\text{pr}(Y_2 = 0|Y_T = 1, X, Z_2) = \Phi(\eta_{20} + \eta_{21}X + \eta_{22}Z_2)$. We generated $N = 1000$ Monte Carlo trials, with sample size $n = 1000$. X, Z_1 and Z_2 from $N(0, 1)$ distribution. Y_T from Bernoulli$\{F(X, \beta) = \Phi(\beta_0 + X\beta_1)\}$, where Φ denotes the CDF of the standard normal distribution. Next, generate misclassification probabilities by using $\alpha_1 = \Phi(\eta_{10} + \eta_{11}X_i + \eta_{12}Z_{i,1})$ and $\alpha_2 = \Phi(\eta_{20} + \eta_{21}X_i + \eta_{22}Z_{i,2})$, then generate $Y_{i,1} = Y_{i,T}(1 - \text{Bernoulli}(\alpha_1))$ and $Y_{i,2} = Y_{i,T}(1 - \text{Bernoulli}(\alpha_2))$. Given Y_1 and Y_2, generate Y_{sum} using $Y_{\text{sum}} = \mathbf{1}(Y_{i,1} + Y_{i,2} \geq 1)$. We set $\beta_0 = -1$ and $\beta_1 = 1$, $(\eta_{10}, \eta_{11}, \eta_{12}, \eta_{20}, \eta_{21}, \eta_{22}) = (-0.7, 1, 1, -1.4, 1, 1)$.

3 Application

we estimate a model of repression in Africa using the four methods. Our outcome data is taken from SCAD (Salehyan et al. 2012), which generates event data on forty-seven African countries using key word searches of Associated Press (AP) and Agence France-Presse (AFP) news wires.

TABLE 2. Model of Repression in Africa

Model	Naive Probit	Hausman Const Pr	Hausman w/ Cov	Multi-Source Const Pr	Multi-Source w/ Cov
$GDPpc_{t-1}$	0.020	0.020	−0.164	0.022	−0.292
	(0.062)	(0.062)	(0.125)	(0.072)	(0.145)
Pop_{t-1}	0.407	0.407	0.314	0.458	0.330
	(0.053)	(0.053)	(0.085)	(0.063)	(0.095)
$Demo_{t-1}$	−0.655	−0.655	−0.739	−0.757	−0.819
	(0.151)	(0.151)	(0.261)	(0.172)	(0.315)
Constant	−8.011	−8.011	−4.679	−8.568	−3.857
	(1.000)	(0.994)	(1.624)	(1.161)	(2.063)
N	1092	1092	1092	1092	1092

4 Conclusion

Traditionally researchers devote less attention to measurement error in the outcome, however, here we have highlighted the severity of the bias induced by differential misclassification in binary outcomes. Our simulations show that misclassification can produce substantial bias when researchers employ either: i) strategies which assume no misclassification *or* ii) strategies which assume non-differential misclassification. We show how researchers possessing more than one source of data-generating information can achieve this desired result. Specifically, we derive an estimator for applications in which researchers have at least two sources of potentially misclassified data on a single outcome of interest. Under few assumptions, our estimator returns unbiased estimates of the risk probability and allows for source-specific misclassification estimates. We illustrated the utility of this method in a model of state repression in Africa, observing that predictor effects change dramatically when misclassification is ignored.

Acknowledgments: This work was supported by a grant from the National Cancer Institute [U01-CA057030 to R.J.C.]; and a post-doctoral fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico [201192/2015-2 to B.B.].

References

Carroll, R. J., Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 693 – 706.

Hausman, J. A., Abrevaya, J., Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87(2), 239 – 269.

Salehyan, I., Hendrix, C. S., Hamner, J., Case, C., Linebarger, C., Stull, E., Williams, J. (2012). Social conflict in Africa: A new database. *International Interactions*, 38(4), 503 – 511.

Estimation of multivariate distributions for recurrent event data

Luís Meira-Machado[1], Beatriz Sampaio[1]

[1] Centre of Mathematics & Department of Mathematics and Applications, University of Minho, Campus de Azurem, 4800-058 Guimarães, Portugal.

E-mail for correspondence: lmachado@math.uminho.pt

Abstract: In many longitudinal studies information is collected on the times of different kinds of events. Some of these studies involve repeated events, where a subject or sample unit may experience a well-defined event several times along his history. Such events are called recurrent events. In this work we consider the estimation of the marginal and joint distribution functions of two gap times under univariate random right censoring. We also consider the estimation of the bivariate survival function.

Keywords: Censoring; Kaplan-Meier; Nonparametric estimation; Recurrent events; Survival Analysis.

1 Introduction

In many longitudinal studies, subjects can experience recurrent events. This type of data has been frequently observed in medical research, engineering, economy and sociology. In medical research, the recurrent events could be multiple occurrences of hospitalization from a group of patients, multiple recurrence episodes in cancer studies, repeated heart attacks or multiple relapses from remission for leukemia patients. In this work we consider the estimation of the marginal and joint distribution / survival functions of the gap times under univariate random right censoring. These issues have received much attention recently. Among others they were investigated by Lin, Sun and Ying (1999), de Uña-Álvarez and Meira-Machado (2008) or de Uña-Álvarez and Amorim (2011).

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

2 Nonparametric estimators

In the context of recurrent event data, each individual may go through a well-defined event several times along his history. Assume that each study subject can potentially experience K consecutive events at times $T_1 < T_2 < ... < T_K$, which are measured from the start of the follow-up. In this work we are primarily interested in the gap times $Y_1 := T_1$, $Y_2 := T_2 - T_1$, ..., $Y_k := T_k - T_{k-1}$, $k = 2, ..., K$. For simplicity we assume $K = 2$.

Then, (Y_1, Y_2) is a vector of gap times of successive events, which we assume to be observed subjected to (univariate) random right-censoring. Let C be the right-censoring variable, assumed to be independent of (Y_1, Y_2). Because of this, the observed data consists of $(\widetilde{Y}_{1i}, \widetilde{Y}_{2i}, \Delta_{1i}, \Delta_{2i})$, $1 \leq i \leq n$, which are n independent replications of $(\widetilde{Y}_1, \widetilde{Y}_2, \Delta_1, \Delta_2)$, where $\widetilde{Y}_1 = Y_1 \wedge C$, $\Delta_1 = I(Y_1 \leq C)$, $\widetilde{Y}_2 = Y_2 \wedge C_2$, $\Delta_2 = I(Y_2 \leq C_2)$ with $C_2 = (C - Y_1)I(Y_1 \leq C)$ the censoring variable of the second gap time. Here and thereafter, $a \wedge b = min(a, b)$ and $I(\cdot)$ is the indicator function.

Let F_k, $k = 1, 2$ denote the distribution function of the k-th event time T_k. Since T_k and C are independent, the Kaplan-Meier product-limit estimator (Kaplan and Meier, 1958) based on the pairs $(\widetilde{T}_{ki}, \Delta_{ki})$'s, consistently estimates the distribution of the time to the k-th event. Because Y_2 and C_2 will be in general dependent, the estimation of the marginal distribution of the second gap time is not a simple issue. The same applies to the joint distribution function $F_{12}(t_1, t_2) = P(Y_1 \leq t_1, Y_2 \leq t_2)$ and the joint survival function $S_{12}(t_1, t_2) = P(Y_1 > t_1, Y_2 > t_2)$. Some estimators for these quantities will be presented below.

Below we present several different approaches for estimating the bivariate distribution function of (Y_1, Y_2). An estimator based on Inverse Probability of Censoring Weights was first introduced by Lin, Sun and Ying (1999):

$$\widehat{F}_{12}^{\texttt{IPCW}}(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(\widetilde{Y}_{1i} \leq t_1)\Delta_{1i}}{\widehat{G}_1(\widetilde{Y}_{1i})} - \frac{1}{n}\sum_{i=1}^{n} \frac{I(\widetilde{Y}_{1i} \leq t_1, \widetilde{Y}_{2i} > t_2)}{\widehat{G}(\widetilde{Y}_{1i} + t_2)}.$$

where $\widetilde{G}_1$ and $\widetilde{G}$ stand for the Kaplan-Meier estimator (of the censoring distribution) based on the $(\widetilde{Y}_{1i}, 1 - \Delta_{1i})$'s and $(\widetilde{T}_{2i}, 1 - \Delta_{2i})$'s, respectively.

A simple estimator based on the Kaplan-Meier weights was later introduced by de Uña-Álvarez and Meira-Machado (2008). The idea behind their estimator is to weight the data by the Kaplan-Meier weights (W_i) pertaining to the distribution of the total time (in this case, T_2) of the process:

$$\widehat{F}_{12}^{\texttt{KMW}}(t_1, t_2) = \sum_{i=1}^{n} W_i I(\widetilde{Y}_{1i} \leq t_1, \widetilde{Y}_{2i} \leq t_2).$$

A related estimator based on presmoothing ($\widehat{F}_{12}^{\texttt{PKMW}}$) was later proposed by de Uña-Álvarez and Amorim (2011).

Given that $P(Y_1 \leq t_1, Y_2 \leq t_2) = P(Y_2 \leq t_2 \mid Y_1 \leq t_1)P(Y_1 \leq t_1)$ we also consider the landmark estimator (LDM) for which to estimate $P(Y_2 \leq t_2 \mid Y_1 \leq$

t_1) the analysis is restricted to the individuals with an observed first event time less or equal than t_1. This is known as the landmark approach (van Houwelingen et al. 2007). The corresponding estimator (LDM) is given by

$$\widehat{F}_{12}^{\texttt{LDM}}(t_1,t_2)=\sum_{i=1}^{n}W_i^{(t_1]}I(\widetilde{Y}_{2i}\leq t_2)\times\widetilde{F}_1^{KM}(t_1)$$

where F_1^{KM} is the Kaplan-Meier estimator of the distribution of the first time and $W_i^{(t_1]}$ denote the Kaplan-Meier weights of the distribution of T_2 computed from the given sub sample $\left\{i:\widetilde{Y}_1\leq t_1\right\}$.

In this work we also introduce new estimators which are constructed using the cumulative hazard of the total time given a first time but where each observation has been weighted using the information of the first duration. The proposed estimator (WCH - weighted cumulative hazard) is given by $\widehat{F}_{12}^{\texttt{WCH}}(t_1,t_2)=\widehat{P}(Y_1\leq t_1)(1-\widehat{P}(Y_2>t_2\mid Y_1\leq t_1))$ where $\widehat{P}(Y_1\leq t_1)$ is estimated by the Kaplan-Meier estimator of the first event time and $\widehat{P}(Y_2>t_2\mid Y_1\leq t_1)=\prod_{v\leq t_2}(1-\hat{\Lambda}_{Y_2|Y_1\leq t_1}(dv))$,
where

$$\hat{\Lambda}_{Y_2|Y_1\leq t_1}(dv)=\frac{\sum_{i=1}^{n}I(\widetilde{Y}_{1i}\leq t_1,\widetilde{Y}_{2i}=v,\Delta_{2i}=1)/\widehat{G}(\widehat{Y}_{1i}+v)}{\sum_{i=1}^{n}I(\widetilde{Y}_{1i}\leq t_1,\widetilde{Y}_{2i}\geq v,\Delta_{1i}=1)/\widehat{G}(\widehat{Y}_{1i}+v)}.$$

Finally we compare the aforementioned methods with the estimator of the bivariate distribution which is obtained using Nearest Neighbor Estimation (NNE).

Now, we consider the estimation of the bivariate survival function $S(t_1,t_2)=P(Y_1>t_1,Y_2>t_2)$. For this quantity, the estimator constructed using the Kaplan-Meier weights was built assuming the following equality $S(t_1,t_2)=1-P(Y_1\leq t_1)-P(Y_1>t_1,Y_2\leq t_2)$ where the first probability on the right hand side is estimated using the Kaplan-Meier estimator of the first event and the second probability is estimated using Kaplan-Meier weights pertaining to the distribution of the total time (i.e., T_2) in a similar way as introduced above. The weighted cumulative hazard estimator of the bivariate survival function is given by $\widehat{S}_{12}^{\texttt{WCH}}(t_1,t_2)=\widehat{P}(Y_2>t_2\mid Y_1>t_1)(1-\widehat{P}(Y_1\leq t_1))$ where $\widehat{P}(Y_2>t_2\mid Y_1>t_1)$ is obtained using the same ideas given above. This is the Wang and Wells (1998) estimator.

Finally, landmark-based estimators can be introduced to estimate the bivariate survival function. Given that $P(Y_1>t_1,Y_2>t_2)=1-P(Y_2\leq t_2\mid Y_1>t_1)(1-P(Y_1\leq t_1))$ the idea is to estimate $P(Y_2\leq t_2\mid Y_1>t_1)$ by restricting the analysis to the individuals with an observed first event time greater or equal than t_1. The corresponding estimator (LDM) is given by $\widehat{S}_{12}^{\texttt{LDM}}(t_1,t_2)=1-\sum_{i=1}^{n}W_i^{[t_1)}I(\widetilde{Y}_{2i}\leq t_2)\times(1-\widetilde{F}_1^{KM}(t_1))$ where $W_i^{[t_1)}$ denote the Kaplan-Meier weights of the distribution of T_2 computed from the sub sample $\left\{i:\widetilde{Y}_1>t_1\right\}$.

3 Example of Application

Our methodology is motivated by the re-analysis of the German breast cancer data. In this study, patients were followed from the date of breast cancer diagnosis

until censoring or dying from breast cancer. From the total of 686 women, 299 developed a recurrence and 171 died. These data can be viewed as arising from a model with two consecutive events: 'Alive with Recurrence' and 'Dead'. In this section, we present plots for the proposed methods to estimate the bivariate distribution function and bivariate survival function of the two gap times, Y_1 = "Time to recurrence" and Y_2 = "Time from recurrence to death".

Figure 1 reports estimated probabilities for a fixed value of $t_1 = 365$ (days), along time. Plot shown in the left hand side (bivariate d.f) show that all proposed methods behave quite similar something that is not true with regard to the estimation of the bivariate survival function (right hand side).

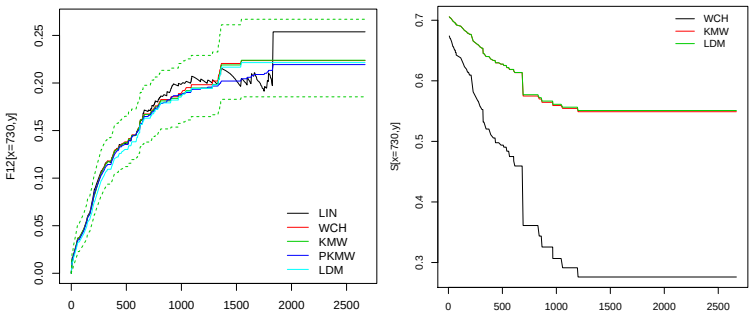


FIGURE 1. Estimates of the bivariate d.f. and bivariate s.f. using the proposed methods. Breast cancer data.

4 Simulation Studies

In this section, we investigate the performance of the proposed estimators through simulations. To simulate the data we consider the bivariate exponential distribution with marginal exponentials with rate parameter 1. This corresponds to the so-called FarlieGumbelMorgenstern copula, where the single parameter controlling for the amount of dependence between the gap times. An independent uniform censoring time C was generated, according to models $Uniform(0, 4)$ and $Uniform(0, 3)$. For each simulated setting we derive the analytic expression of $F_{12}(t_1, t_2)$ and $S_{12}(t_1, t_2)$ for several (t_1, t_2) pairs, corresponding to combinations of the percentiles 20%, 40%, 60%, and 80% of the marginal distributions of the gap times (i.e., 0.2231, 0.5108, 0.9163, 1.6094). Sample sizes $n = 100$, $n = 250$, and $n = 500$ were considered.

Results reveal that the all proposed methods for estimating the bivariate distribution function perform quite well, though the performance of all methods is poorer at the right tail (i.e., larger values of t_1 and t_2) where the censoring effects are stronger. At these points the standard deviation (SD) is in most cases larger. The SD decreases with an increase in the sample size and with a decrease of the censoring percentage. All methods proposed in this work obtain in all settings a negligible bias.

Attained results for the bivariate survival function reveal that the weighted cumulative hazard estimator (WCH) is the recommended approach. This is illustrated

in Figure 2 in which we show the boxplots of the estimates for the bivariate distribution function.

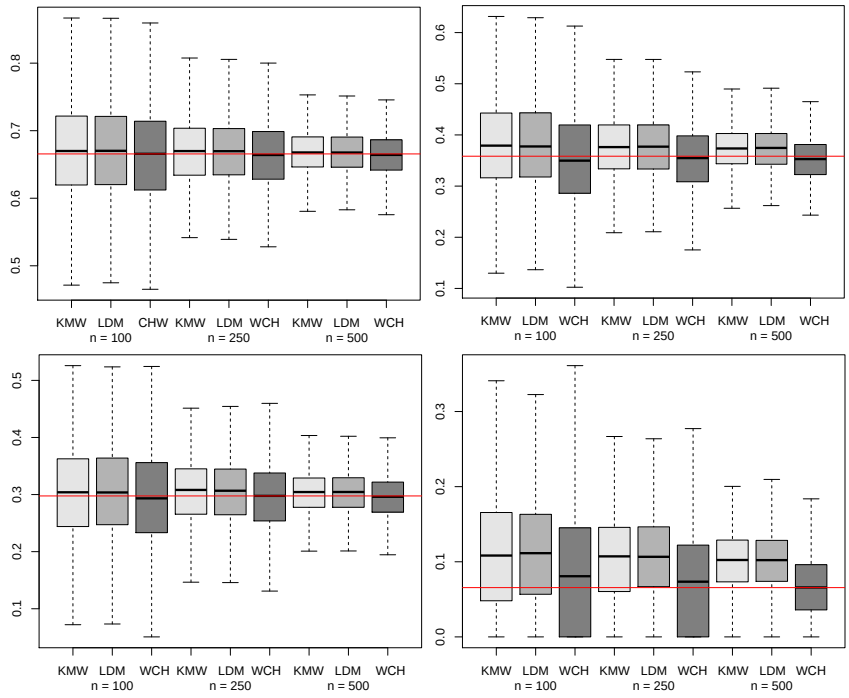


FIGURE 2. Boxplot with estimated probabilities $S_{12}(t_1, t_2)$. On the top results for the pair (0.2231, 0.2231) (left) and (0.2231,0.9163) (right); on the bottom results for the pair (0.9163, 0.5108) (left) and (1.6094,1.6094) (right).

Acknowledgments: This research was financed by Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia", within Project UID/MAT/00013/2013.

References

de Uña-Álvarez J. and Meira-Machado L. (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters*, **78**, 2440 – 2445.

de Uña-Álvarez J and Amorim, A.P. (2011). A semiparametric estimator of the bivariate distribution function for censored gap times. *Biometrical Journal*, **53**, 113 – 127.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53(282)**, 457 – 481.

Lin, D., Sun, W. and Ying, Z. (1999). Nonparametric estimation of the time distributions for serial events with censored data. *Biometrika*, **86(1)**, 59 – 70.

van Houwelingen, H.C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, **34**, 70 – 85.

Wang, M.C. and Wells, M.T. (1998). Nonparametric Estimation of successive duration times under dependent censoring. *Biometrika*, **85**, 561 – 572.

Multi-Parameter Regression and Frailty

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

[1] University of Limerick, Limerick, Ireland
[2] CREST, Ensai, Rennes, France

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

Abstract: We introduce a class of survival models which account for the presence of frailty and in which both scale and shape parameters of the hazard function depend on covariates (we call this "Multi-Parameter Regression"). This general formulation allows us to consider potential origins of time-dependent covariate effects. We apply the models to lung cancer data using the `mpr` package in `R`.

Keywords: frailty; multi-parameter regression; survival; time-dependent effects

1 Introduction

Consider the following basic hazard model

$$h(t) = \lambda\, h_0(t\,;\,\gamma) \tag{1}$$

where $\lambda > 0$ controls the overall size of $h(t)$, and $h_0(t\,;\,\gamma)$ is a non-negative function with shape parameter $\gamma > 0$ characterising its time-evolution. Survival data may exhibit heterogeneity beyond that of (1). Thus, we introduce a latent variable, $U \in [0, \infty)$, representing this heterogeneity and consider the *conditional* hazard function

$$h(t\,|\,U) = U\,\lambda\, h_0(t\,;\,\gamma).$$

Furthermore, we assume that U follows a one-parameter gamma distribution such that $E(U) = 1$ and $Var(U) = \nu$, i.e., the Gamma Frailty (GF) model (cf. Duchateau and Janssen (2007)). Since U is unobservable, so too is the conditional hazard; in fact, we observe the *marginal* hazard,

$$h_m(t) = \frac{\lambda\, h_0(t\,;\,\gamma)}{1 + \nu\,\lambda\, H_0(t\,;\,\gamma)} \tag{2}$$

where $H_0(t) = \int_0^t h_0(u)du$. Note: when $\nu = 0$, we have $h_m(t) = h(t)$.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

2 Multi-Parameter Regression Modelling

Naturally, we wish to consider the effect of covariates on the hazard function. Although it is standard practice to allow *one* distributional parameter to depend on covariates (the "interest" parameter), more flexibility is achieved by allowing *multiple* parameters to depend on covariates simultaneously – we refer to this as Multi-Parameter Regression (MPR) (Burke and MacKenzie, 2016) – and, therefore, we suggest the following:

$$\log \lambda = x^T \beta, \qquad \log \gamma = x^T \alpha,$$

where $x = (1, x_1, \ldots, x_p)^T$ is a vector of covariates, and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ and $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)^T$ are the corresponding vectors of scale and shape regression coefficients. We could go even further and let the frailty variance, ν, depend on covariates, but this is beyond the scope of the current paper. This regression specification leads to the marginal hazard function

$$h_m(t \,|\, x) = \frac{e^{x^T\beta}\, h_0(t\,;\, e^{x^T\alpha})}{1 + \nu\, e^{x^T\beta}\, H_0(t\,;\, e^{x^T\alpha})} \tag{3}$$

which is quite general as it covers a variety of models:

- **PH**: Proportional Hazards model when $\nu = 0$ and $x^T\alpha = \alpha_0$. If $h_0(t)$ is non-parametric, we have Cox's (1972) semi-parametric model.
- **PH-GF**: The Gamma Frailty extension of the PH model when $x^T\alpha = \alpha_0$. See Duchateau and Janssen (2007).
- **MPR**: The Multi-Parameter Regression extension of the PH model when $\nu = 0$. See Burke and MacKenzie (2016) for details.
- **MPR-GF**: The GF extension of the MPR model is given by (3) without constraining any parameters.

Let $x_{(-j)} = (1, x_1, \ldots, x_{j-1}, 0, x_{j+1}, \ldots, x_p)^T$ be the covariate vector with its jth element set to zero so that we may write

$$x^T\beta = x_j\beta_j + x_{(-j)}^T\beta, \qquad x^T\alpha = x_j\alpha_j + x_{(-j)}^T\alpha,$$

i.e, we explicitly express the linear predictor as a term involving x_j and all other terms. This is useful when considering the hazard ratio for x_j:

$$\psi_j(t) = \frac{h_m(t \,|\, x_j = 1)}{h_m(t \,|\, x_j = 0)} = \underbrace{\exp(\beta_j)}_{\text{PH}} \cdot \underbrace{\eta_j(t)}_{\text{MPR}} \cdot \underbrace{\rho_j(t)}_{\text{GF}} \tag{4}$$

where

$$\eta_j(t) = \frac{h_0(t\,;\, e^{\alpha_j + x_{(-j)}^T\alpha})}{h_0(t\,;\, e^{x_{(-j)}^T\alpha})}$$

and

$$\rho_j(t) = \frac{1 + \nu\, e^{x_{(-j)}^T\beta}\, H_0(t\,;\, e^{x_{(-j)}^T\alpha})}{1 + \nu\, e^{\beta_j + x_{(-j)}^T\beta}\, H_0(t\,;\, e^{\alpha_j + x_{(-j)}^T\alpha})}.$$

Thus, the leading term in $\psi_j(t)$ is the familiar PH constant, $\exp(\beta_j)$, the second term, $\eta_j(t)$, appears due to the MPR extension, and the third term, $\rho_j(t)$, appears

due to the GF extension. This general formulation is very useful as it allows us to compare the various models and, in particular, to consider the nature of time-dependent hazard ratios.

The presence of frailty is a standard explanation for observing time-dependent effects in practice (Aalen, 2008, chap. 6). However, frailty on its own is not very flexible in that context: we gain a single additional parameter, ν, which is unlikely to capture the time-dependence arising for each covariate, x_j, $j = 1, \ldots, p$. Furthermore, frailty *imposes* time-dependence on all hazard ratios (i.e., $\rho_j(t)$ always depends on time), and, in the absence of the MPR extension (i.e., setting $x^T\alpha = \alpha_0$), the hazard ratios also *must* converge to one with time.

Hazard ratios are constant in the basic PH model since $h_0(t\,;\,\gamma)$ is common to all individuals. Thus, compared with frailty, a more fundamental explanation for the appearance time-dependent hazard ratios is perhaps simply that hazard functions take on different shapes for different individuals, i.e., γ depends on covariates. In contrast to the GF extension, the MPR extension produces an additional parameter *per* covariate, α_j, $j = 1, \ldots, p$, which allows flexibility in modelling the hazard ratios for each of these covariates separately. Moreover, when $\alpha_j = 0$, we see that $\eta_j(t) = 1$ and, therefore, in the absence of frailty (i.e., setting $\nu = 0$), the hazard ratio reduces to the usual PH constant, i.e., the MPR extension does *not* impose time-dependence on all hazard ratios.

The MPR and GF extensions offer alternative explanations for non-PH effects. However, we may contemplate the general MPR-GF model and whether or not both extensions are required simultaneously in practice.

3 Analysis of Lung Cancer Data

We applied the above models to a lung cancer dataset collected in Northern Ireland between October 1991 and September 1992 (see Burke and MacKenzie, 2016) using the `mpr` package in `R` (Burke, 2016). We consider only single-factor treatment models here and specialise to the case where $h_0(t) = \gamma\, t^{\gamma-1}$ (Weibull baseline).

A brief summary of the models appears in Table. 1 where, interestingly, the most

TABLE 1. Summary of Treatment Models.

	PH	PH-GF	MPR	MPR-GF
$\hat{\nu}$	—	0.91	—	0.74
AIC	57.7	25.3	20.4	0.0
BIC	33.9	6.3	15.6	0.0

complex model (MPR-GF) minimises both AIC and BIC. Thus, it appears that both extensions (MPR and GF) *can* be supported simultaneously in practice; the improved fit is evident when we compare the fitted models to the Kaplan-Meier curves (not shown). The hazard ratios arising from the various models (not shown) clearly display the flexibility achieved by modelling the shape of the hazard function. Note that the estimated frailty variance, $\hat{\nu}$, is smaller in the case

of the MPR-GF which we may expect as more variation is explained under this model compared with the simpler PH-GF model.

4 Discussion

The class of models considered here allows us to compare various approaches. From a practical perspective, frailty as an explanation of time-dependent effects is somewhat contrived, whereas the MPR approach directly extends the PH model more fundamentally without latent assumptions. However, in practice, combining MPR and frailty is useful: in our application, the MPR-GF model is best in terms of AIC and BIC. Thus, the GF extension of the PH model fails to absorb all unexplained variation (e.g., due to covariate-dependent hazard shape) but, equally, unexplained variation still exists beyond the MPR model.

Acknowledgments: The first author would like to thank the Irish Research Council (www.research.ie) for supporting this work.

References

Aalen, O. O. and Borgan, Ø. and Gjessing, H. K. (2008). *Survival and event history: A process point of view*. Springer.

Burke, K. (2016). mpr: Multi-Parameter Regression (MPR). R package version 1.0.4 (http://cran.r-project.org/package=mpr).

Burke, K. and MacKenzie, G. (2016). Multi-parameter regression survival modeling: An alternative to proportional hazards. *Biometrics*, DOI: 10.1111/biom.12625.

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

Duchateau, L. and Janssen, P. (2007). *The Frailty Model*. Springer.

Comparison of Semiparametric Copula and Frailty Models for Clustered Survival Data

Il Do Ha[1], Jong-Min Kim[2]

[1] Department of Statistics, Pukyong National University, Busan, South Korea
[2] Division of Science and Mathematics, University of Minnesota-Morris, USA

E-mail for correspondence: idha1353@pknu.ac.kr

Abstract: In clustered survival data, dependence among individual survival times within a cluster has been usually described using copula models and frailty models. In this paper we compare the two modelling approaches under semiparametric setting. We also investigate relationships between copula-based likelihood and frailty-based marginal likelihood, and propose corresponding profile likelihood-based efficient procedures. The proposed method is demonstrated using a well-known CGD recurrent infection data set with different cluster sizes.

Keywords: Copula model; Frailty model; Marginal likelihood; Profile likelihood.

1 Introduction

A dependence among individual survival times within a cluster can be modelled using frailty or copula. Here, the frailty is an unobserved random effect acting multiplicatively on the individual hazard rate (Vaupel et al., 1979) and has been widely used for modelling the dependence among clustered survival data in biomedical studies (Duchateau and Janssen, 2008; Ha, Jeong and Lee, 2017). The copula is a convenient way to study the dependence between random variables; following Sklar's (1959) theorem, a copula expresses the joint distribution of random variables as a function of marginal distributions of each variable (Joe, 1997). In particular, copula is useful for modelling a dependency in finance risks (Kim and Jung, 2016).

In this paper we consider clustered survival data where the size of each cluster may be different. Let $T_{i1}, \ldots, T_{in_i}$ be survival times (time-to-events) for the jth ($j = 1, \ldots, n_i$) observation of the ith ($i = 1, 2, \ldots, q$) cluster. The survival data may be correlated because T_{ij}'s are observed on the same cluster. Here q is the number of clusters and n_i is the number of individuals in cluster i (i.e. cluster size). Let

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

$x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ be the p covariates corresponding to T_{ij}. The observable random variables are $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I(T_{ij} \leq C_{ij})$, where C_{ij} is the censoring time. We assume the non-informative and independence for censoring. In this paper we compare two popular models (i.e. semi-parametric marginal copula and conditional frailty models) using profile likelihood approaches. The proposed method is illustrated using a practical CGD example (Fleming and Harrington, 1991).

2 Likelihood-based procedure for copula models

Under the Archimedean copula family, the joint survival function of $T_{i1}, \ldots, T_{in_i}$ for cluster i given x_{ij} can be expressed as

$$S(t_{i1}, \ldots, t_{in_i}|x_{ij}) = \varphi_\theta[\varphi_\theta^{-1}\{S_1(t_{i1}|x_{i1})\} + \cdots + \varphi_\theta^{-1}\{S_{n_i}(t_{in_i}|x_{in_i})\}],$$

where $S_j(t_{ij}|x_{ij})$ is a marginal survival function for T_{ij} given x_{ij} $(j = 1, 2, \ldots, n_i)$. Here, the generator φ_θ of Archimedean copula can also be expressed as a Laplace transform of a positive distribution function $G_\theta(y)$; e.g., for the Clayton copula $\varphi_\theta(s) = (1+\theta s)^{-1/\theta}$ for $\theta \geq 0$. Thus, the joint survival function above for cluster i can be rewritten as

$$S(t_{i1}, \ldots, t_{in_i}|x_{ij}) = \int \exp\left[-y \sum_{j=1}^{n_i} \varphi_\theta^{-1}\{S_j(t_{ij}|x_{ij})\}\right] dG(y). \tag{1}$$

Here we assume the marginal survival function is obtained from the Cox's proportional hazards (PH) model:

$$\lambda_{ij}(t|x_{ij}) = \lambda_0(t) \exp(x_{ij}^T \beta), \tag{2}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard and $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of regression parameters corresponding to covariates x_{ij}. Since φ_θ is the Laplace transform of G_θ, following Prenen et al. (2017) and the derivative of (1), the Clayton copula-based log-likelihood for all q clusters has a closed form:

$$\begin{aligned}
\ell_c &= \sum_{i=1}^{q}\left[\sum_{j=1}^{n_i} \delta_{ij}\left\{\log f_{ij} - \log \varphi_\theta'\{\varphi_\theta^{-1}(S_{ij})\}\right\} + \log \varphi_\theta^{(d_i)}\left\{\sum_{j=1}^{n_i} \varphi_\theta^{-1}(S_{ij})\right\}\right] \\
&= \sum_{ij} \delta_{ij}\{\log \lambda_{ij} + \theta \Lambda_{ij}\} - \sum_{i=1}^{q}\left[(d_i + \theta^{-1}) \log(1 + S_{i+}^*) - \sum_{l=0}^{d_i - 1} \log(1 + l\theta)\right],
\end{aligned}$$

where $d_i = \sum_{j=1}^{n_i} \delta_{ij}$, $S_{ij} = S_j(y_{ij}|x_{ij})$, $f_{ij} = f_j(y_{ij}|x_{ij})$ and $S_{i+}^* = \sum_{j=1}^{n_i}(S_{ij}^{-\theta} - 1)$. Here $\lambda_{ij} = \log \lambda_0(y_{ij}) + x_{ij}^T \beta$ and $\Lambda_{ij} = \Lambda_0(y_{ij}) \exp(x_{ij}^T \beta)$ with the baseline cumulative hazard $\Lambda_0(t)$ are from the marginal Cox model (2). For the estimation of copula models with the the model (2), we present the two-stage semi-parametric estimation procedure (Shih and Louis, 1995). Since the functional form of $\lambda_0(t)$ in (2) is unknown, we consider a step function for $\Lambda_0(t)$ (Breslow, 1972). That is, in the first stage, β is estimated by maximizing the Breslow partial likelihood $(\ell_p(\beta))$ and Λ_0 is estimated by the Breslow's estimator. In the second stage, θ is estimated by maximizing a profile likelihood based on the estimates ($\hat{\beta}$ and $\hat{\Lambda}_0$) obtained from the first stage,

$$\ell_c(\theta) = \ell_c|_{\beta=\hat{\beta}, \Lambda_0=\hat{\Lambda}_0}.$$

In this paper, we also propose a likelihood ratio test (LRT) for testing the absence of association parameter θ, $H_0 : \theta = 0$. However, care is necessary because such a null hypothesis is on the boundary of the parameter space ($\theta \geq 0$). Thus, the standard chi-square distribution can not be applied. The null distribution for the LRT statistic follows an asymptotic chi-square mixture distribution, i.e. a mixture of χ^2_0 and χ^2_1 with equal weights of 0.5, (Stram and Lee, 1994). Thus, it is calculated as

$$LR_c = -2\{\ell_p(\hat{\beta}) - \ell_c(\theta)\},$$

leading that the p-value is calculated from $p = \frac{1}{2}P(\chi^2_1 > \mathrm{LR}_c)$. Here $\ell_p(\hat{\beta})$ is the partial log-likelihood evaluated at $\beta = \hat{\beta}$ under $H_0 : \theta = 0$.

3 Likelihood-based procedure for frailty models

Let U_i be a shared frailty (random effect) of the ith cluster. The semi-parametric frailty model (Duchateau and Janssen, 2008) is described as follows. The conditional hazard function of T_{ij} given x_{ij} and $U_i = u_i$ takes the form of

$$\lambda_{ij}(t|x_{ij}, u_i) = \lambda_0(t)\exp(x_{ij}^T\beta)u_i, \tag{3}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard. Popular distributions for the frailty U_i are gamma and log-normal; for the gamma frailty model $E(U_i) = 1$ and $\mathrm{var}(U_i) = \theta$. Following Nielsen et al. (1992), the marginal joint survival function is given by

$$S_m(t_{i1}, \ldots, t_{in_i}|x_{ij}) = \int \exp\left[-u_i \sum_{j=1}^{n_i} \varphi_\theta^{-1}\{S_{jm}(t_{ij}|x_{ij})\}\right] dG_\theta(u_i), \tag{4}$$

where $S_{jm}(t_{ij}|x_{ij}) = \int S(t_{ij}|x_{ij}, u_i)dG_\theta(u_i)$ with distribution function $G_\theta(\cdot)$ of U_i. As mentioned by Goethals (2008) and Prenen et al. (2017), the two joint survival functions (1) and (4) are similar in that both joint functions take the form of the same copula, but $S_j(t_{ij}|x_{ij}) \neq S_{jm}(t_{ij}|x_{ij})$, which gives a major difference between both models. We find that the marginal log-likelihood for all q clusters under frailty models (3) is obtained by by replacing f_{ij} and S_{ij} in ℓ_c by the corresponding functions f^m_{ij} and S^m_{ij}; under gamma frailty model (3), it has also a closed form:

$$\begin{aligned}
\ell_m &= \sum_{i=1}^{n}\left[\sum_{j=1}^{n_i}\delta_{ij}\left\{\log f^m_{ij} - \log\varphi'_\theta\{\varphi_\theta^{-1}(S^m_{ij})\}\right\} + \log\varphi_\theta^{(d_i)}\left\{\sum_{j=1}^{n_i}\varphi_\theta^{-1}(S^m_{ij})\right\}\right]. \\
&= \sum_{ij}\delta_{ij}(\log\lambda^F_{ij}) - \sum_{i=1}^{q}\left[(d_i + \theta^{-1})\log(1 + \theta\Lambda^F_{i+}) - \sum_{l=0}^{d_i-1}\log(1 + l\theta)\right],
\end{aligned}$$

where $\lambda^F_{ij} = \log\lambda_0(y_{ij}) + x^T_{ij}\beta$ and $\Lambda^F_{i+} = \sum_{j=1}^{n_i}\Lambda_0(y_{ij})\exp(x^T_{ij}\beta)$ are from the frailty models (3). Following Ha et al. (2010), we propose the use of a simple grid search method to implement the likelihood method. In the inner loop, given θ, (β, Λ_0) are obtained by solving the estimating equations $\partial\ell_m/\partial(\beta, \Lambda_0) = 0$ via the Newton-Raphson method. In the outer loop, given (β, Λ_0), the profile likelihood $\ell_m(\theta) = \ell_m|_{\beta=\tilde{\beta}, \Lambda_0=\tilde{\Lambda}_0}$ is maximized for θ. Here $\tilde{\beta}$ and $\tilde{\Lambda}_0$ are the estimates obtained from ℓ_m. We can again use the LRT for testing the absence of θ in the frailty models, which is equivalent to test $u_i = 0$ for all i (i.e. the absence of frailty effect). Here, the corresponding LRT is given by $LR_m = -2\{\ell_p(\hat{\beta}) - \ell_m(\theta)\}$.

TABLE 1. Results of fitting copula and frailty models for the CGD data.

	Treatment	Association	Profile likelihood
Model	$\hat{\beta}$ (SE)	$\hat{\theta}$ (p-value)	$\ell(\hat{\theta})$
Clayton copula	$-1.086(0.268)$	$0.804(< .001)$	-347.86
Gamma frailty	$-1.138(0.345)$	$1.383(< .0001)$	-345.84

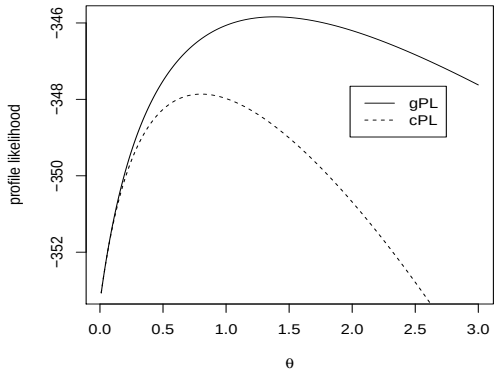


FIGURE 1. Profile likelihoods (PL) for association parameter θ in the CGD data; gPL, $\ell_m(\theta)$ under gamma frailty model; cPL, $\ell_c(\theta)$ under Clayton copula model.

4 An illustrated example

Fleming and Harrington (1991) provided a clustered survival data set on a placebo-controlled randomized trial of gamma interferon (γIFN) in chronic granulomatous disease (CGD). Here 128 patients had recurrent infections with $n_i = 1 \sim 8$, and we consider a main covariate, i.e. Treatment (0 = placebo, 1 = γ-IFN). Table 1 shows both models give similar significance, but different estimates for θ (see also Figure 1) as compared to $\hat{\beta}$.

Acknowledgments: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2015R1D1A3A01015663).

References

Duchateau, L. and Janssen, P. (2008). *The frailty model*. Springer.

Ha, I. D., Noh, M. and Lee, Y. (2010). Bias reduction of likelihood estimators in semi-parametric frailty models. *Scandinavian Journal of Statistics*, **37**, 307–320.

Ha, I. D., Jeong, J.-H. and Lee, Y. (2017). *Statistical modelling of survival data with random effects: h-likelihood approach*. Springer, in press.

Kim, J.M. and Jung, H. (2016). Linear time-varying regression with CopulaDC-CGARCH models for volatility. *Economics Letters*, **145**, 262-265.

Prenen, L., Braekers, R. and Duchateau, L. (2017). Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *JRSS, B*, in press.

Conditional inference survival trees for nonstandard data

Wei Fu[1], Jeffrey S. Simonoff[1]

[1] New York University, USA

E-mail for correspondence: jsimonof@stern.nyu.edu

Abstract: Tree methods (recursive partitioning) are a popular class of nonparametric methods for analyzing data. One extension of the basic tree methodology is the survival tree, which applies recursive partitioning to censored survival data. This has mainly been designed for right-censored data. We discuss application of the conditional inference survival tree method to two important but less standard data types, left-truncated and right-censored (LTRC) data and interval-censored data. Further, we show that LTRC trees can be used to analyze survival data with time-varying covariates, essentially building a time-varying covariates survival tree. Implementation of the methods is easy, and simulations and real data analysis results show that the proposed methods work well from both a predictive point of view and in uncovering tree structure in the underlying survival process.

Keywords: Interval-censored data; Left-truncated and right-censored data; Survival tree.

1 Introduction

Right-censored data are often studied using a (semi-)parametric model such as the Cox proportional hazards (PH) model or accelerated failure time model. However, the parametric assumptions imposed by these models are often either not met or unrealistic in practice, and more flexible nonparametric methods are desired. One such method is the survival tree.

Various authors have proposed tree methods for right-censored data. Hothorn *et al.* (2006) (hereafter HHZ), in particular, implemented a survival tree using the log-rank test as the splitting method. They embedded the survival tree algorithm into a large framework of conditional inference trees, which has the desirable property of selecting the splitting variable in an unbiased way (an unbiased tree has

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

the property that when there is no relationship between the response and any covariates all covariates have the same probability of being the split variable). Like other proposed methods, this method is designed for the most basic setup of survival outcome – right-censored data with time-independent covariates. However, other types of survival data such as left-truncated and right-censored (LTRC) data, interval-censored (IC) data, and survival data with time-varying covariates arise commonly in practice.

In this paper, we discuss extension of the conditional inference survival tree method to LTRC and IC data. The resulting methods are easy to implement in practice. Through data reformulation, the LTRC survival tree method can then also be used to fit survival data with time-varying covariates.

2 Conditional inference trees for survival data

The conditional inference tree algorithm of HHZ is based on the idea of separating the two steps of choosing the variable for splitting and then choosing the split point of that variable. The splitting variable is chosen based on conditional distributions that are constructed assuming that the response and covariates are independent. After the splitting variable is selected, the split point can be determined by any criterion. The association of Y and a covariate X_j is measured by linear statistics of the form

$$T_j(L_n, w) = vec\left(\sum_{i=1}^{n} w_i g_i(X_{ji}) h\left(Y_i\right)^T\right) \in \mathbb{R}^{p_j q}$$

(equation 3.1 in HHZ), where $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ is a nonrandom transformation of covariate X_j and $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$ is the influence function of the response Y. For a univariate numeric response Y, the choice of influence function is the identity, i.e. $h\left(Y_i\right) = Y_i$.

2.1 Log-rank score for right-censored data

HHZ discussed the construction of a conditional inference survival tree for right-censored data. Subjects can be represented as a triple $(t_i, \delta_i, \mathbf{x}_i)$, $i = 1, 2, ..., n$, where t_i is the observed event time or censored time for the ith subject, $\delta_i = 1$ if t_i is the event time and $\delta_i = 0$ if t_i is the censored time and $\mathbf{x}_i$ is the covariate vector for the ith subject. We assume that censoring is noninformative given $\mathbf{x}_i$. The response variable for the ith subject is $Y_i = (t_i, \delta_i)$. The influence function for such a bivariate response is the so-called log-rank score. The main function of the log-rank score is to assign a univariate value U_i (scalar) to the bivariate response $Y_i = (t_i, \delta_i)$, so the algorithm can then execute in the same way as in the univariate numeric response case.

2.2 Log-rank score for LTRC and IC data

Pan (1998) extended the rank invariant tests of Peto and Peto (1972) to left-truncated and interval-censored data, with the log-rank score being

$$U_i = \frac{\hat{S}(l_i)\log\hat{S}(l_i) - \hat{S}(r_i)\log\hat{S}(r_i)}{\hat{S}(l_i) - \hat{S}(r_i)} - \log\hat{S}(L_i).$$

Here, L_i is the left-truncation time and (l_i, r_i) is the interval in which the true event time lies. The log-rank scores for LTRC and IC data can be derived from this score equation as special cases, which are then used in the general conditional inference tree framework. The log-rank score for an LTRC observation (L_i, R_i, δ_i) is

$$U_i = 1 + \log \hat{S}(R_i) - \log \hat{S}(L_i) \quad \text{if} \quad \delta_i = 1$$

and

$$U_i = \log \hat{S}(R_i) - \log \hat{S}(L_i) \quad \text{if} \quad \delta_i = 0.$$

Note that $\hat{S}$ is the nonparametric maximum likelihood estimator (NPMLE) of the survival function, constructed based on the product-limit estimator, i.e. Kaplan-Meier (KM) estimator. Details are given in Fu and Simonoff (2017a), and implementation of the method is provided in the `R` package `LTRCtrees`, which is available at CRAN.

A crude but common approach to IC data is to impute the midpoint or endpoint of the censoring interval as the actual survival time, but it is known that this can lead to bias and incorrect inferences (Lindsey and Ryan, 1998). The appropriate conditional inference tree, however, is easily defined using the log-rank score for an IC observation, which is given by

$$U_i = \frac{\hat{S}(L_i) \log \hat{S}(L_i) - \hat{S}(R_i) \log \hat{S}(R_i)}{\hat{S}(L_i) - \hat{S}(R_i)},$$

where L_i and R_i are the lower and upper boundaries of the censoring interval for the ith observation, respectively. Once again $\hat{S}$ is the NPMLE of the survival function, here constructed using the EM-algorithm as proposed by Turnbull (1976). If the event time is observed, so $L_i = R_i$ and the interval (L_i, R_i) reduces to a point, the corresponding log-rank score is $U_i = 1 + \log \hat{S}(L_i)$. More details are given in Fu and Simonoff (2017b), and an `R` function that implements the method is provided in the `R` package `LTRCtrees`.

3 Properties of the trees

Simulations show that when the true structure is a tree, the LTRC and IC trees perform significantly better than the Cox PH model, regardless of sample size, censoring distribution and left-truncation rate, as would be expected. Further, when the true structure is a complex nonlinear relationship the trees can outperform the Cox PH model, and never perform significantly worse than the Cox PH model, demonstrating that the trees have more robust performance than does the Cox PH model. The IC trees are uniformly better than building a tree based on imputing the midpoint or endpoint of the censoring interval as the actual survival time.

4 Real data examples

4.1 LTRC data

The assay of serum free light chain data in the `R` package `survival` is used as a data example. An analysis could use age as a covariate and time from enrollment

in the study as the response, but as noted by Klein and Moeschberger (2003), age should actually be used as a left-truncation point, since the real response of interest should be the subject's life length, not the time from enrollment in the study.

Figure 1 gives conditional inference trees with the actual death/censoring time as response. The tree that accounts for the left truncation (left panel) identifies the top FLC decile (FLC=10) as the most important covariate of overall survival, the same split used in the original analysis based on subject area knowledge. In contrast, the tree that ignores the left-truncation (right panel) only uncovers the effect of FLC for males.

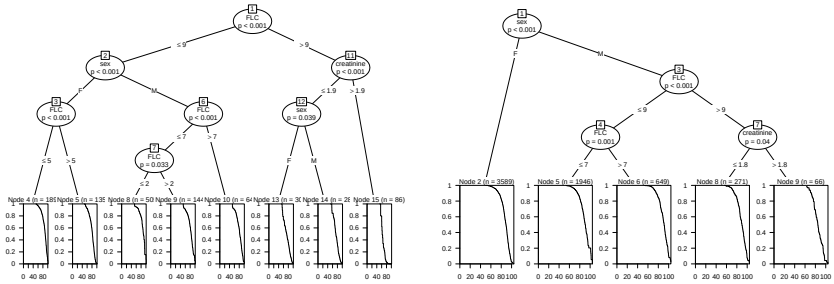


FIGURE 1. Survival trees for the serum FLC data, accounting for the left-truncation (left) and ignoring it (right), respectively.

4.2 IC data

The Signal Tandmobiel® study is a longitudinal prospective oral health study that was conducted in the Flanders region of Belgium for which data are provided as the `tandmob2` data set in the `R` package `bayesSurv`. The response variable is the time to emergence of tooth 24, which is interval-censored since emergence is only identified at scheduled dentist visits.

Figure 2 gives the IC tree for the emergence time of the tooth. Decay and gender are associated with earlier emergence time. More orthodontic removal of deciduous teeth is associated with earlier emergence time, which could reflect a reverse causality effect. There are province effects for some children, apparently due to misclassification effects of certain examiners.

5 Adapting LTRC trees to time-varying covariate data

Fu and Simonoff (2017a) also described how LTRC trees can be used to construct trees for time-varying covariate data. The general strategy is to first split each subject into several "pseudo-subjects," inside which covariates are constant. So, for example, if a covariate changes at time points t_1 and t_2 with the event time T, the subject is split into three pseudo-subjects over the intervals $(0, t_1)$, $[t_1, t_2)$, and $[t_2, T]$. This results in three created LTRC pseudo-subjects with time-independent

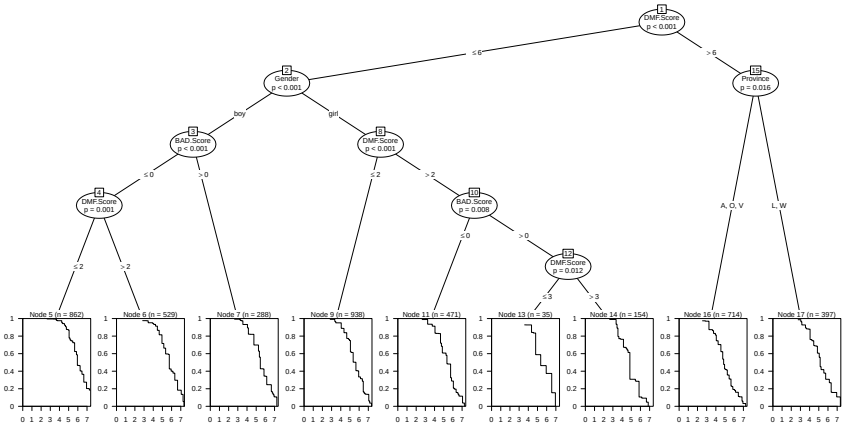


FIGURE 2. IC tree for the emergence time in years after the age of 5 of tooth 24 (the permanent upper left first premolar).

covariates, and the LTRC tree algorithm is then applied on all such constructed pseudo-subjects to fit a tree. Simulations show that this constructed tree has properties similar to those of the (underlying) LTRC tree.

The PBC data in the `R` package **`survival`** are used as an example. The data were collected at entry and at yearly intervals on 45 variables. Figure 3 gives conditional inference trees for these data. The tree in the left panel is based on only baseline covariate values, while the one is the right panel uses the follow-up data, in which all of the covariates except age become time-varying covariates. The trees are quite noticeably different, with different variables appearing and the top-level split variables being different.

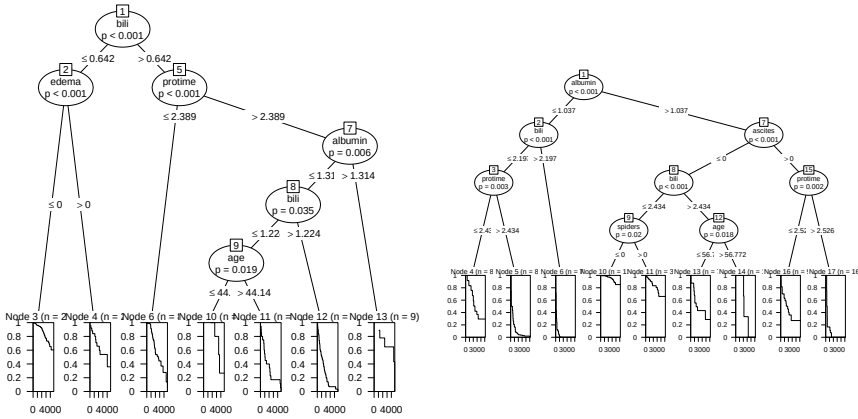


FIGURE 3. Survival trees for PBC data based on baseline covariates (left) and time-varying covariates (right), respectively.

References

Fu, W. and Simonoff, J.S. (2017a). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*, **18**, 352 – 369.

Fu, W. and Simonoff, J.S. (2017b). Survival trees for interval-censored survival data. *arXiv:1702.07763*, `https://arxiv.org/abs/1702.07763`.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, **15**, 651 – 674.

Klein, J.P. and Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.*, Springer, New York.

Lindsey, J.C. and Ryan, L.M. (1998). Methods for interval-censored data. *Statistics in Medicine*, **17**, 219 – 238.

Pan, W. (1998). Rank invariant tests with left truncated and interval censored data. *Journal of Statistical Computation and Simulation*, **61**, 163 – 174.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A*, **135**, 185 – 207.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290 – 295.

Bayesian Joint Modelling of Distributional Regression

Elisabeth Waldmann[1], Nadja Klein[2], David Taylor-Robinson[3]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] Georg-August-Universität Göttingen, Germany
[3] University of Liverpool, England

E-mail for correspondence: elisabeth.waldmann@fau.de

Abstract: When modelling repeated measurement and time to event data simultaneously, analysis is often based on combining random effects regression on the mean with survival analysis in a so called joint model. The assumptions made by this model, however, are not necessarily met, when analysing real life data sets. Especially the assumption of equidispersion is too strict in many cases. In simple longitudinal setups this issue is often dealt with distributional regression, which introduces separate predictors for all parameters of the distribution. This work aims at combining the setup of a Bayesian distributional model with a survival model, linking both, the conditional mean and conditional variance to the risk of event and thus developing a new approach towards distributional joint modelling.

Keywords: Joint Modelling; Mixed Models; Time-to-event data; Bayesian Inference.

1 Introduction

Amongst the many statistical methods that have been developed for analysing data from longitudinal studies, the term *joint modelling* refers to the statistical modelling of data in which each subject provides data on two qualitatively different kinds of outcome variable: a time-sequence of repeated measurements; and a (possibly right-censored) time-to-event variable. In most cases the model used for the longitudinal data is a mean regression model, which very often lacks in flexibility. Distributional regression gives a more complete overview over the dependent variable. It has recently been implemented thoroughly in a Bayesian framework by Klein et al (2014). Instead of only modelling the expectation of a conditional distribution further parameters of the distribution are assigned pre-

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

dictors. In case of the Gaussian distribution the variance parameter σ^2 is modeled by a (potentially additive) predictor η_{σ^2}. Data-sets with heteroscedastic characteristics are modeled more accurately when using distributional regression. The latter is often the case in longitudinal studies and the structure of the variance potentially holds information on the pattern of the time events analysed in the survival analysis. Our interest in this model class was triggered when analysing lung function decline in cystic fibrosis patients from a Danish cystic fibrosis registry. It has been shown that the onset of certain pulmonary infections is associated with an acceleration of the loss of lung function, when using the onset of infection as a covariate in a longitudinal model (Qvist et al. 2015). The onset of infection however, could be seen as a process influenced by the same covariates as the lung function decline itself and hence should be modeled as a related yet separate process. A further interest lies in better understanding how the variability in lung function is related to lung function decline and other outcomes in cytic fibrosis. This contribution hence deals with including this distributional regression feature into joint modelling and developing a Monte Carlo Markov Chain (MCMC) algorithm.

2 Distributional Regression in Joint Modelling

2.1 Joint Modelling

There are many different approaches in combining time-to-event outcomes with longitudinal data in joint models. The type we are going to refer to here, is a model as suggested by Faucett et al (1996) which is based on the following likelihoods:

$$\begin{aligned} f(Y_{ij}|\eta_{ij,\cdot},\sigma) &\propto \frac{1}{\sigma^{m_i}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{m_i}\left(y_{ij}-(\eta_{\mathrm{l}ij}+\eta_{\mathrm{ls}ij})\right)^2\right) \\ f(s_i,\delta_i|Y_i) &= \lambda_0(s_i)^{\delta_i}\exp\left(\eta_{\mathrm{s}i}+\alpha\eta_{\mathrm{ls}ij}\right) \\ &\quad \cdot \exp\left(\int_0^{s_i}\lambda_0(u)\exp\left(\eta_{\mathrm{s}i}+\alpha\eta_{\mathrm{ls}ij}\right)du\right), \end{aligned}$$

where Y_{ij} is the j-th outcome for individual i. The event time/censoring time is denoted by s_i and δ_i refers to the censoring indicator. The predictors for the two parts of the model are composed of so called sub-predictors $\eta_{\cdot ij}$ refer to longitudinal, survival and shared sub-predictors (l, s and ls). In case of the longitudinal process, the composed predictor is $\eta_{\mathrm{l}ij}+\eta_{\mathrm{ls}ij}$, in case of the time-to-event process the predictor is composed of $\eta_{\mathrm{s}ij}+\alpha\eta_{\mathrm{ls}ij}$, where α is the so called association parameter, measuring the connection between the two parts of the model. Faucett et all (1996) suggest a MCMC algorithm to estimate the corresponding parameters.

2.2 Distributional Regression

Distributional regression models – also known as generalized additive models for location, shape and scale (GAMLSS) – are based on not only measuring the

impact on the mean but on more parameters of the conditional distribution. In case of the Gaussian distribution the second parameter which is of interest is the variance. The classical likelihood for the Gaussian distribution is hence extended by a predictor η_{σ^2}. In order to ensure positivity of the variance, an exponential link function is used, such that $\hat{\sigma}^2 = \exp(\hat{\eta}_{\sigma^2})$, with η_{σ^2} being a additive predictor including a variety of different types of effects. Klein et al (2014) suggested an MCMC algorithm using iteratively weighted least squares proposal densities.

2.3 Distributional Regression in Joint Modelling

Combining the two above explained concepts leads to an extended set of sub-predictors, now including sub-predictors specifically designed for quantifying the impact of the covariates on the variation in the data and the connection of them to the survival probabilities. Thus the set of longitudinal sub-predictors is $\eta_{\mathrm{l},\mu}(x_{\mathrm{l}ij,\mu})$,$\eta_{\mathrm{l},\sigma}(x_{\mathrm{l}ij,\sigma})$ and the set of shared sub-predictors: $\eta_{\mathrm{ls},\mu}(x_{\mathrm{ls}ij,\mu})$,$\eta_{\mathrm{ls},\sigma}(x_{\mathrm{ls}ij,\sigma})$. The survival sub-predictor does not have to be extended. Furthermore the model is completed by a variance specific association parameter α_{σ^2}. The extended likelihoods are hence:

$$
\begin{aligned}
f\left(Y_i|\eta_{\cdot,i,\cdot}\right) \quad &\propto \quad \frac{1}{\exp\left(\eta_{\sigma^2,i,\mathrm{l}}+,\eta_{\sigma^2,i,\mathrm{ls}}\right)} \\
&\quad \cdot \exp\left(-\frac{1}{2\exp\left(\eta_{\sigma^2,i,\mathrm{l}}+\eta_{\sigma^2,i,\mathrm{ls}}\right)}\sum_{j=1}^{n_i}\left(y_{ij}-\left(\eta_{\mu,i,\mathrm{l}}+\eta_{\mu,i,\mathrm{ls}}\right)\right)^2\right) \\
f\left(s_i,\delta_i|\eta_{\cdot,i,\cdot}\right) \quad &= \quad \lambda_0(s_i)^{\delta_i}\exp\left(\delta_i\left(\alpha_\mu\eta_{\mu,i,\mathrm{ls}}+\alpha_{\sigma^2}\eta_{\sigma^2,i,\mathrm{ls}}+\eta_{\mu,i,\mathrm{s}}+\eta_{\sigma^2,i,\mathrm{s}}\right)\right) \\
&\quad \cdot \exp\int_0^{s_i}\lambda_0(u)\exp\left(\alpha_\mu\eta_{\mu,i,\mathrm{ls}}+\alpha_{\sigma^2}\eta_{\sigma^2,i,\mathrm{ls}}+\eta_{\mu,i,\mathrm{s}}+\eta_{\sigma^2,i,\mathrm{s}}\right)du
\end{aligned}
$$

The MCMC algorithm was constructed as a mixture of the one done by Faucett et al (1996) and Klein et al (2014).

3 Simulations

In order to evaluate the algorithm we first simulated a typical setting, similar to the simulation study in Waldmann et al (2017) were done. The sub-predictor were constructed as linear functions. The shared mean sub-predictor included a fixed and random time-effect, as well as a random intercept. We simulated 300 individuals of which each had a maximum of 5 observations. The baseline hazard was chosen to be constant and we varied the association parameters. We did not include a survival sub-predictor. Results for one simulation setup are shown in Figure 1. The black lines indicate the true value, the boxplots display the means of 100 simulation runs. The results show that our algorithm captures the simulated effects very closely.

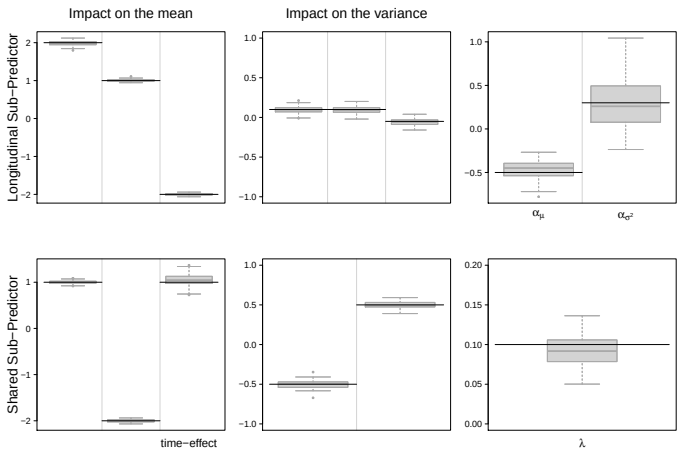


FIGURE 1. Results for the regression parameters from all four sub-predictors as well as association parameters and fixed baseline hazard. The solid black lines show the true values, the boxplots the MCMC sample means of 100 repetitions of the same model.

4 Lung Function Decline in Cystic Fibrosis Patients

In order to establish if our model is applicable to real life data, we choose a subset of a data set taken from the Danish cystic fibrosis registry. After choosing the patients that have at least two observations, before the infection the data set contained a total of 6268 of 489 patients of which 53 were infected with pseudomonas aeruginosa (PA) in the course of the study. The model calculated aims at explaining both, the progression of lung function, measured in Forced Expiratory Volume in 1 second (FEV1) and the risk of onset of PA. The covariates chosen for the model were preselected based on previous studies (Waldmann et al (2017)). As covariates for both longitudinal sub-predictor height and weight of the patient as well as three binary covariates indicating, if the patient had one of three different additional lung infections were chosen. For the shared sub-predictors we considered time and sex for the mean predictor and pancreatic insufficiency and year of birth for the variance. The boxes for the MCMC samples of the covariates are displayed in the first two rows of Figure 2. Boxes of variables that do not include the value 0 in the 95% interval of the sample (i.e. which are considered to have a significant influence) are black, the rest is gray. The results for the mean regression is similar to previous findings: while patients with higher weight have better lung function, height, additional infections and diabetes have a negative impact. The variance is not influenced by weight and the additional infection, whereas height and diabetes have a negative influence on the variance, i.e. the lung function varies less for taller patients and for patients having diabetes. The second row has to be interpreted taking into account the association parameters which are displayed in the last row. The association with the mean shared sub-predictor is negative, such that we can conclude that the time impact, which is negative for the lung function itself is positive for the risk of being infected. The

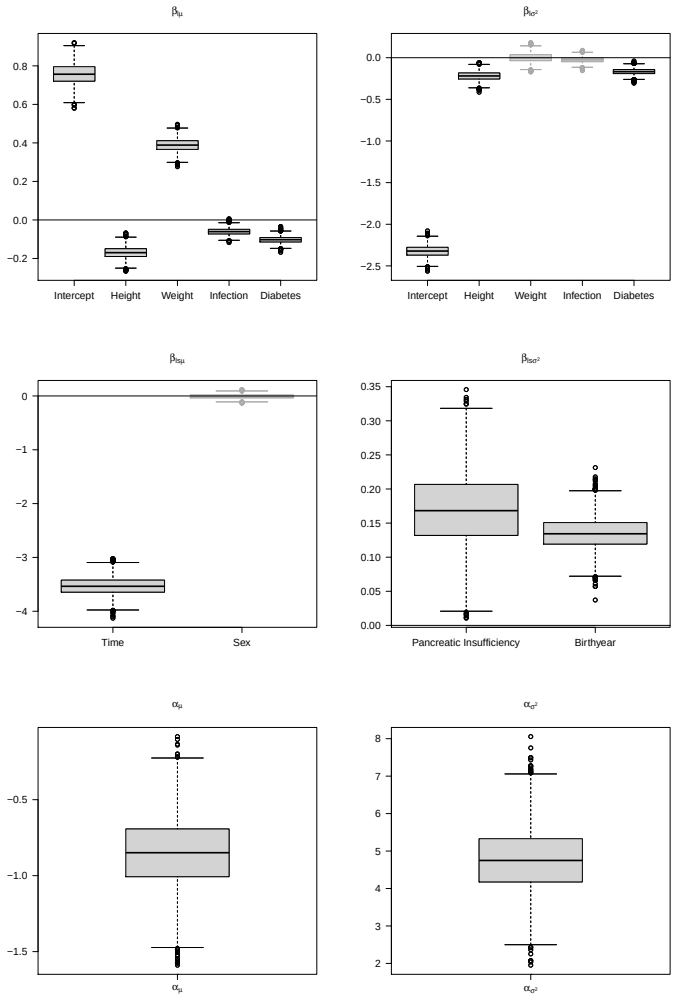


FIGURE 2. Results from a model for lung function decline and risk of lung infection in cystic fibrosis patients. The upper four plots contain the boxplots of the MCMC samples for the linear model parameters (black if 95% of the MCMC sample do not include zero, gray otherwise). The two plots in the last row show the boxplots of the MCMC samples for the association parameters.

association parameter for the variance shared sub-predictor is positive. Hence the increase in variance for higher years of birth and pancreatic insufficiency increases the risk of infection.

5 Conclusion

In this work, we show how to combine distributional regression and joint modelling in order to extract more information about data generating processes in longitudinal and time-to-event data. The results for the cystic fibrosis study can be seen as a first hint in the direction of answering the question of the impact of varying lung function on susceptibility for lung infections rather than the lung function level itself. To answer this question with more ultimate certainty, however, a more refined model has to be constructed.

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., Pressler T. (2016). Infection and lung function decline; comparing the harmful effects of cystic fibrosis pathogen. *Journal of Cystic Fibrosis* **15(3), 380–385.**

Klein N., Kneib T., Klasen S., Lang S. (2014). Bayesian Structured Additive Distributional Regression for Multivariate Responses. *Journal of the Royal Statistical Society: Series C*, **64**, 569 – 591.

Waldmann E., Taylor-Robinson D., Klein N., Kneib T., Pressler T., Schmid M., Mayr A.(2017): Boosting Joint Models for Longitudinal and Time-to-Event Data. *Biometrical Journal.* accepted.

Incorporating cyclical effects and time-varying covariates in models for single-source capture-recapture data

Thomas Husken[1], Maarten Cruyff[1], Peter van der Heijden[12]

[1] Utrecht University, the Netherlands
[2] University of Southampton, England

E-mail for correspondence: t.f.husken@uu.nl

Abstract: The objective of capture-recapture analysis is to estimate the size of an elusive population, for which the zero-truncated Poisson model is a basic model. We extend this model to the more general recurrent events model to include cyclical effects and time-varying covariates. An application to police data on victims of domestic violence provides strong evidence for the presence of weekly and seasonal cyclical effects on the rate of police reports.

Keywords: single-source capture-recapture; cyclical effects; time-varying covariates; recurrent events model; zero-truncation.

1 Introduction

The zero-truncated Poisson model is a well-established model for the analysis of single-source capture-recapture data. Such data typically arise when each observation of a member of an elusive population is recorded in a registration file. Counting the number of records for each individual population member yields a zero-truncated count distribution, because population members with a zero count are not in the register. Under the assumption that the counts follow a Poisson distribution, an estimate of the Poisson parameter can be obtained that in turn can be used to estimate the frequency of the zero count. Relevant covariates can be included to model individual differences in Poisson parameters, this leads to the zero-truncated Poisson regression model (TPR) (see Cruyff & van der Heijden, 2013; van der Heijden et al., 2003).

Like any type of events data, single-source capture-recapture data can exhibit seasonal or cyclical patterns. For example, a homeless person may be more likely

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

to stay in a homeless shelter during winter than in the summer and a problematic drug user may have a higher probability of being admitted to the hospital in the weekend than during the week. However, the TPR is unable to incorporate these types of cyclical effects.

In this paper we present a method that allows for the inclusion of cyclical effects in single-source capture-recapture data. In this method, the TPR is extended to the zero-truncated recurrent events model (TREM), which includes a time dimension. This dimension makes the model more general than the TPR, since it allows for the inclusion of time-varying covariates and cyclical effects. The resulting model can accommodate a wide variety of effects: time-invariant, cyclical, time-varying, and interactions thereof.

2 Method

The TREM is an extension of the TPR that allows for the modelling of time-varying covariates and cyclical effects in single-source capture-recapture data. The likelihood of the TREM is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{y_i} \lambda_{ij} \left(\frac{e^{-\Lambda_i(\tau)}}{1 - e^{-\Lambda_i(\tau)}} \right) \right\}, \tag{1}$$

where $\Lambda_i = \sum_{t=1}^{\tau} \lambda_{it}$, and $\ln \lambda_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_p x_{itp}$ (Cook & Lawless, 2007, p. 273-278). Here, x_{itp} specifies the value covariate p takes at time point t for person i. Additionally, y_i is the total number of captures over the observation period for individual i. Note that this allows for time-varying covariates since x_{itp} may vary over time.

Cyclical effects are modelled by adding a cosine term to the linear predictor of the TREM:

$$\ln \lambda_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_p x_{itp} + \alpha \cos\left(\frac{2\pi}{k} t - \theta \right), \tag{2}$$

where α is the amplitude and θ the horizontal shift. The period k is a constant and determines how often the cyclical component peaks. The cosine term in Equation (2) is non-linear and therefore difficult to estimate, but can be rewritten to a linear function. The most common method is using a trigonometric identity to parametrise the cyclical effect as

$$\alpha \cos\left(\frac{2\pi}{k} t - \theta \right) = \beta_{\cos} \cos\left(\frac{2\pi}{k} t \right) + \beta_{\sin} \sin\left(\frac{2\pi}{k} t \right), \tag{3}$$

from Cryer & Chan (1994, Ch. 3, p. 34). The final expression can be easily included in the linear predictor of the recurrent events model, where $\cos(\frac{2\pi}{k} t)$ and $\sin(\frac{2\pi}{k} t)$ are entered as covariates. The interpretation of the two cyclical regression coefficients is not very intuitive, but they can be transformed back in terms of α and θ:

$$\begin{aligned} \alpha &= \sqrt{\beta_{\cos}^2 + \beta_{\sin}^2}, \\ \theta &= \arctan2(\beta_{\sin}, \beta_{\cos}), \end{aligned} \tag{4}$$

which is also known as a polar transformation. The linear parametrisation of Equation (3) has the advantage that interaction effects with time-invariant covariates can be included almost in the classical manner. This provides the option to include main effects of time-invariant covariates, cyclical effects, and interactions between these two simultaneously.

Parameter estimates of the TREM are obtained by optimizing the loglikelihood given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{y_i} \ln \lambda_{ij} - \sum_{i=1}^{n} \Lambda_i - \sum_{i=1}^{n} \ln[1 - e^{-\Lambda_i}]. \tag{5}$$

Analytical closed form expressions for the score function and Hessian of the TREM can be derived from the loglikelihood (see Hu & Lawless, 1996). These expressions are then used to set up a Newton-Raphson algorithm. Standard errors are obtained through the observed information matrix.

Given the parameter estimates $\hat{\boldsymbol{\beta}}$ of the TREM, the Horvitz-Thompson population size estimate is obtained as

$$\hat{N} = \sum_{i=1}^{n} \frac{I_i}{1 - e^{-\hat{\Lambda}_i}}, \tag{6}$$

where $I_i = 1$ if case i is observed in the sample and $I_i = 0$ otherwise. The denominator $1 - e^{-\hat{\Lambda}_i}$ is the probability that a population member is observed in the sample. The variance of $\hat{N}$ is calculated through the Delta method, as presented in van der Heijden et al. (2003).

3 Application to domestic violence data

The application is a data set of domestic violence victims from the Netherlands in the period 2004 - 2006. The response of interest is the number of times a police report was filed for domestic violence for a certain individual. Although information on perpetrators of domestic violence is also available, we do not focus on that group in this paper. Hence, our population of interest is defined as victims of domestic violence. There are a total of 56,575 observed victims of domestic violence in the period 2004 - 2006. These data are made available by the Dutch national police.

The variables gender and age are available as subject-specific covariates. Gender is included as a time-invariant covariate. Age is modelled with time-varying linear and quadratic contrasts, meaning that an individual can move from one age group to another during the observation period. The age categories are: 0-17, 18-29, 30-39, 40-49, 50+.

Cyclical effects with periods 366 and 7 are included, representing seasonal and weekly effects, respectively. Interaction effects between the cyclical week effect and the linear and quadratic age effects allow each age group to have a different cyclical week effect. Furthermore, a linear effect of time is added in the final model to allow for an increase or decrease of capture probabilities over the time period of three years. Finally, an interaction effect of the cyclical season effect and the linear time effect is included so that the cyclical seasonal is allowed to vary over time.

TABLE 1. Regression coefficients and point and interval estimate of the population size for the TREM model fit to the domestic violence data

Variable	Coding	$\hat{\boldsymbol{\beta}}$	**SE**	
Intercept		-8.63	0.03	***
Gender	(male = 0, female = 1)	0.63	0.03	***
Age	Linear	0.26	0.03	***
	Quadratic	-0.50	0.02	***
$\cos_{366}$		-0.06	0.01	***
$\sin_{366}$		-0.02	0.01	**
$\cos_{7}$		0.05	0.01	***
$\sin_{7}$		-0.04	0.01	***
Age (Linear)*$\cos_{7}$		-0.03	0.01	*
Age (Linear)*$\sin_{7}$		0.03	0.01	
Age (Quadratic)*$\cos_{7}$		0.12	0.01	***
Age (Quadratic)*$\sin_{7}$		0.01	0.01	
Time		0.34	0.01	***
Time*$\cos_{366}$		0.01	0.01	
Time*$\sin_{366}$		0.16	0.01	***
$\hat{N}$		211,155		
95%-CI		206,460 - 215,848		

Note: *** $= p < 0.001$, ** $= p < 0.01$, * $= p < 0.05$.

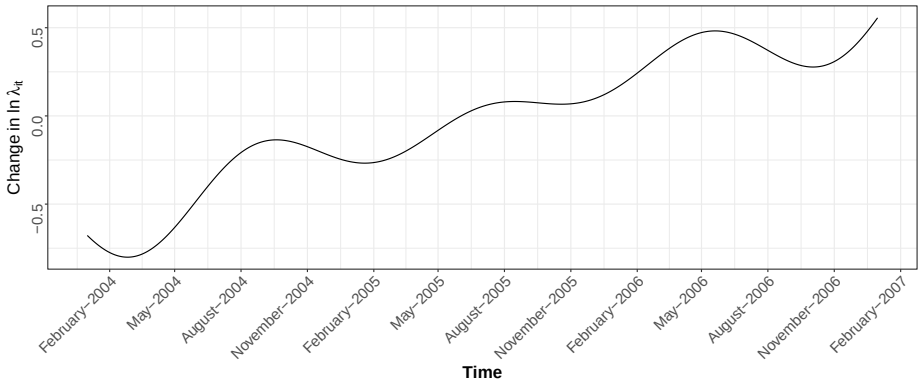


FIGURE 1. Fitted general cyclical trend for the domestic violence application

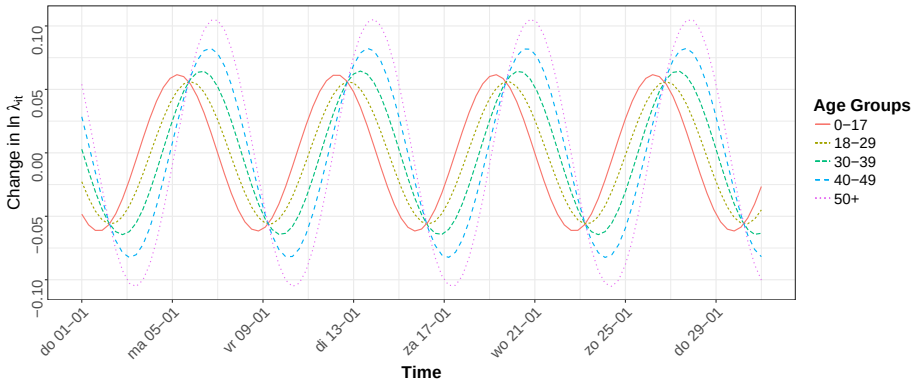


FIGURE 2. Fitted cyclical week effect and its interaction with age for the domestic violence application for the month January 2004

Table 1 shows the regression coefficients and point and interval estimate of the population size for the model fitted to the data. The effects of gender and age are significant. Women are more likely to be mentioned as a victim in a police report for domestic violence than men, and there is a quadratic effect of age.

Both the cyclical main effects are significant, indicating the presence of seasonal and weekly variation in capture probabilities. Additionally, the interaction effects of age (linear and quadratic) with the cosine terms of the cyclical week effect are significant, so that the cyclical week effect is different for each age group.

A positive effect of time is found, indicating that the capture probabilities increase over the course of the observation period. One of the two interaction components of time with the cyclical season effect is also significant, so that the cyclical season effect is different over the three years. The population size estimate of domestic violence victims in the time period 2004 - 2006 is 211,155 (95%-CI: 206,460 - 215,848).

The general cyclical trend of the fitted model (omitting cyclical week effects) is presented in Figure 1. This trend consists of three components: the cyclical season main effect, the linear effect of time, and the interaction between the two. In general, we can say that the cyclical season effect is stronger in 2004 and 2006 than in 2005, and that the capture probabilities increase over the course of the observation period. Additionally, the cyclical season effect in 2006 peaks in May, while in the 2004 and 2005 the effect peaks in September.

In Figure 2, the cyclical week effect and the interaction of this effect with age is presented. These effects are plotted for the month January in 2004, and repeat throughout the length of the observation window. The age group 30-39 is the reference group in this analysis, represented by the green curve. The strength of the cyclical week effect is lowest for this reference group. The cyclical week effect is strongest for the 50+ age group, while the strength of the cyclical effects of the other groups is somewhere in between. For all age groups, the cyclical week effect peaks after the weekend. The groups 40-49 and 50+ peak on Wednesday, and the other groups on Monday (18-29) and Tuesday (0-17 and 30-39).

References

Cook, R.J. & Lawless, J.F. (2007). *The statistical analysis of recurrent events*. Springer Science & Business Media.

Cruyff, M.J. & van der Heijden, P.G.M. (2013). *Sensitivity Analysis and Calibration of Population Size Estimates Obtained with the Zero-Truncated Poisson Regression Model*. Statistical Modelling, 14(5): 361-373

Cryer, J.D. & Chan, K. (2008). *Time series analysis*. Princeton university press.

Hu, X. J. & Lawless, J. F. (1996). *Estimation of rate and mean functions from truncated recurrent event data.* Journal of the American Statistical Association, 91(433):300-310.

van der Heijden, P.G.M., Bustami, R., Cruyff, M.J., Engbersen, G. & van Houwelingen, H.C. (2003). *Point and interval estimation of the population size using the truncated Poisson regression model*. Statistical Modelling, 3(4): 305-322.

Statistical modelling of cell movement

Diana Giurghita[1], Dirk Husmeier[1]

[1] University of Glasgow, United Kingdom

E-mail for correspondence: d.giurghita.1@research.gla.ac.uk

Abstract: In this paper we demonstrate an application of the unscented Kalman filter in the context of cell movement, using a model defined in terms of stochastic differential equations (SDEs).

Keywords: Chemotaxis, Cancer, Stochastic Differential Equations, Unscented Kalman Filter.

1 Introduction

Many important biological processes, such as wound healing, tissue development and cancer cell invasion, are based on the collective movement of cells. One of the main mechanisms for directed cell movement is chemotaxis, where cells follow chemical gradients (chemoattractants) present in their environment. These gradients might arise from the presence of a local source of chemoattractant or due to local depletion of the chemical in the environment (Tweedy et al. 2016). An example of the former scenario is the migration of breast tumour cells that respond to the epidermal growth factor released by macrophages. In an attempt to acquire a deeper understanding of the mechanisms behind cell movement many population based models have been formulated using partial differential equations, with very few of them attempting to fit these models to actual data.

In this paper, we propose a model that describes the movement of any individual cell being driven by an external resource gradient using SDEs of the form:

$$dX_t = \sigma dB_t^X, \quad dY_t = \frac{\alpha\beta \exp[-\beta(Y_t - \gamma t)]}{\{1 + \exp[-\beta(Y_t - \gamma t)]\}^2} dt + \sigma dB_t^Y \tag{1}$$

Equations (1) describe the evolution in time of the x and y coordinates of a cell in 2D space. σdB_t^X and σdB_t^X are Brownian motion terms which in this model represent the intrinsic randomness in a cell's movement. The coordinate in the y direction has a drift term that is described by three parameters: α - the amplitude

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

of the resource gradient, β - the steepness of the gradient and γ which indicates how fast the gradient changes over time. The strength of the random component in the cell movement equations is indicated by the diffusion coefficient σ.

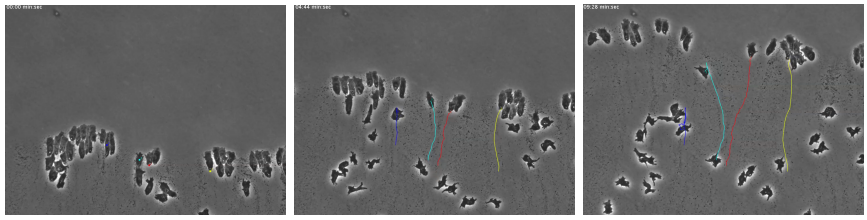


FIGURE 1. Three frames from the video recording *Dictyostelium* cells movement.

In this paper, we present our approach to fitting this model to cell movement data using a non-linear Bayesian filter. We provide some insight into the particularities of this model using simulated data and we discuss the results of this analysis from a real data set, describing the movement of *Dictyostelium* cells (see Figure 1).

2 Methods

Inference in non-linear dynamical systems poses numerous challenges due to the stochastic nature of the data, intractable likelihoods and unidentifiable parameters. Recent developments have tackled this problem using likelihood-free methods (sequential Monte Carlo ABC) or computational methods (particle Markov Chain Monte Carlo) (Golightly & Wilkinson, 2011), however these can become too computationally expensive as the number of time points or parameters increases. The unscented Kalman filter (UKF) is an online Bayesian filtering method that can easily be scaled up to higher dimensions (Julier & Uhlmann, 1997). Intuitively, the UKF starts from the initial distribution of the state vector, drawn from a multivariate normal distribution, which is then iterated through a prediction and updating step for each measurement available using the transition and observation models.

We introduce the UKF by referring to a general state-space model:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}, \boldsymbol{\epsilon}_t), \quad \mathbf{y}_t = \mathbf{g}(\mathbf{x}_t, \boldsymbol{\nu}_t) \tag{2}$$

where $\mathbf{x}_t$ represents the vector of the hidden states, $\mathbf{y}_t$ are the measurements, $\boldsymbol{\epsilon}_t$ is the process noise at time t, $\boldsymbol{\nu}_t$ is the observation noise at time t and the functions $\mathbf{f}$ and $\mathbf{g}$ represent the transition and, respectively, the observation models. The model parameters $\boldsymbol{\theta}$ can be included as dynamical variables in the hidden states vector $\mathbf{x}_t$, which means they will be estimated at every time point along with the observed system states (Sitz et al., 2002). The advantage of the method comes from the fact that the probability distribution of the predictor step: $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ and the probability distribution of the updating step $p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1})$ can be obtained in closed form using properties of the Gaussian distribution (Julier & Uhlmann, 1997):

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \approx \mathcal{N}(\mathbf{x}_t|\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Sigma}}_t) \tag{3}$$

$$p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1}) \approx \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \tag{4}$$

where $\bar{\boldsymbol{\mu}}_t$ and $\bar{\boldsymbol{\Sigma}}_t$ are the prediction mean and covariance at time t and $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ are the update mean and covariance at time t (see Julier & Uhlmann, 1997 for full derivations). Therefore, the algorithm essentially updates the mean and covariance of the Gaussian distribution of the state vector at each iteration. The approximation of the Gaussian distribution is made using the unscented transform, which consists of a set of deterministically chosen sigma-points that are passed through the non-linear function and weighted to obtain the mean and covariance of the Gaussian. The unscented transformation is used twice for each iteration of the algorithm: in the prediction and respectively in the update step.

3 Simulation results

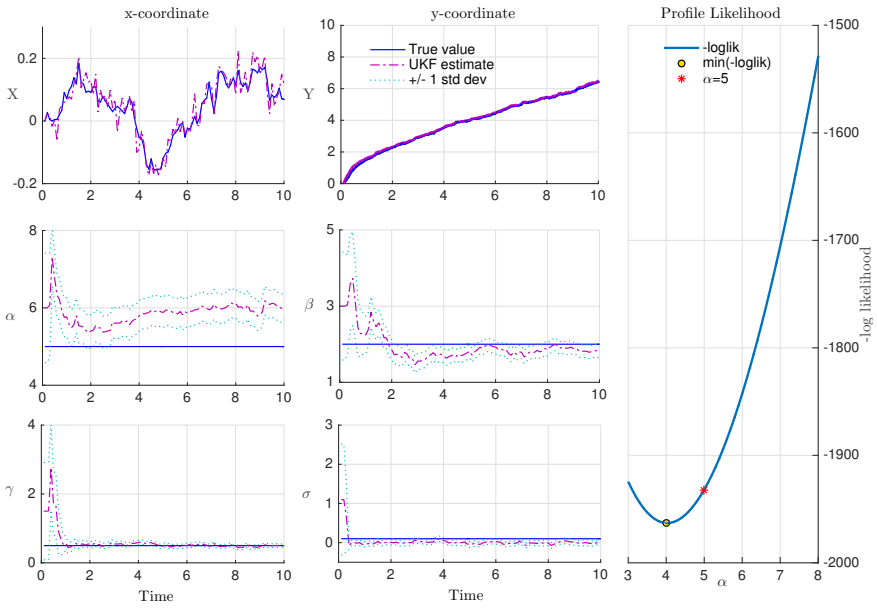


FIGURE 2. Simulation results: UKF tracking of cell coordinates: x and y, and parameters: $\alpha, \beta, \gamma, \sigma$ for time interval $[0, 10]$. Parameter estimates include ± 1 standard error bounds. On the right hand side, negative log profile likelihood plot for α, obtained by fixing the other three parameters at their true values.

We apply the Euler-Maruyama discretisation to bring the system in Equation (1) into the standard state-space model described in Section 2:

$$X_t = X_{t-1} + \sigma \Delta B_t^X \tag{5}$$

$$Y_t = Y_{t-1} + \frac{\alpha\beta \exp[-\beta(Y_{t-1} - \gamma t)]}{\{1 + \exp[-\beta(Y_{t-1} - \gamma t)]\}^2} dt + \sigma \Delta B_t^Y \tag{6}$$

Here ΔB_t^X and ΔB_t^Y are just sums of random normal increments between time $t-1$ and t. Equations (5) and (6) thus define the transition function $\mathbf{f}$ from Equation 2. In this scenario, we assume the process is observed with a small amount of Gaussian noise $\nu_t \sim \mathcal{N}(0, 0.1^2)$, so $\mathbf{g}$ from Equation 2 is just the identity function.

We then fit the UKF to a synthetic data set using the following parameters: $\alpha = 5, \beta = 2, \gamma = 0.5, \sigma = 0.1$. The results summarised in Figure 2 indicate good agreement between the estimated UKF path and the true cell path. The UKF also provides good estimates for the parameters: $\hat{\beta} = 1.88, \hat{\gamma} = 0.51, \hat{\sigma} = 0.04$ with relatively small standard errors: $0.17, 0.83, 0.06$, except for $\hat{\alpha}$ where the estimates indicate a more substantial deviation from the true parameter (bias: 0.44 and standard error is 0.35).

A potential source of bias as the one observed in Figure 2 can be investigated by looking at the likelihood i.e.: marginal likelihood with respect to the hidden states. In order to do that, we first derive the probability of the observed system at time t conditional on the state of the system at time $t-1$ by integrating out the latent variable $\mathbf{x}_t$:

$$p(\mathbf{y}_t|\mathbf{y}_{t-1}) = \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{t-1})d\mathbf{x}_t \tag{7}$$

$$= \int \mathcal{N}(\mathbf{y}_t|\mathbf{x}_t, \mathbf{R}_t)\mathcal{N}(\mathbf{x}_t|\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Sigma}}_t)d\mathbf{x}_t \tag{8}$$

Using the Gaussian convolution integral results (Bishop, 2006) we simplify (8) to $p(\mathbf{y}_t|\mathbf{y}_{t-1}) = \mathcal{N}(\mathbf{y}_t|\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Sigma}}_t + \mathbf{R}_t)$, where $\bar{\boldsymbol{\mu}}_t$ and $\bar{\boldsymbol{\Sigma}}_t$ are the predicted mean and covariance at time t, and $\mathbf{R}_t$ is the measurement noise covariance matrix at time t. The log likelihood is then:

$$\mathcal{L} = \log \prod_t p(\mathbf{y}_t|\mathbf{y}_{t-1}) = \sum_t \log \mathcal{N}(\mathbf{y}_t|\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Sigma}}_t + \mathbf{R}_t) \tag{9}$$

$$\propto \sum_t \{\log \det(2\pi\boldsymbol{\Sigma}_t) + 0.5(\mathbf{y}_t - \bar{\boldsymbol{\mu}}_t)^T \boldsymbol{\Sigma}_t^{-1}(\mathbf{y}_t - \bar{\boldsymbol{\mu}}_t)\}, \tag{10}$$

Where $\bar{\boldsymbol{\Sigma}}_t + \mathbf{R}_t = \boldsymbol{\Sigma}_t$. We evaluate the marginal log likelihood in (10) by considering a grid of values for each parameter in the model and fitting the UKF with each parameter combination. The results summarising the profile likelihood for the α parameter in Figure 2 can be used to calculate the Cramer-Rao lower bound, which provides an indication of the intrinsic uncertainty specific to the problem. In this case, the minimum standard deviation attainable by an estimator of α is 0.14. Considering the standard error obtained from the UKF estimation for α is 0.35, this then indicates that the estimated value of the parameter is reasonably close to the true value.

4 Real data application

Dictyostelium cells are widely used in experiments as proxies for understanding the mechanisms of human disease because of their similarities to important human cells (leukocytes and cancer cells) in terms of biology and response to chemotaxis (Tweedy et al., 2016). The data consists of two cell paths corresponding to Dictyostelium cells locations extracted from a time series of high

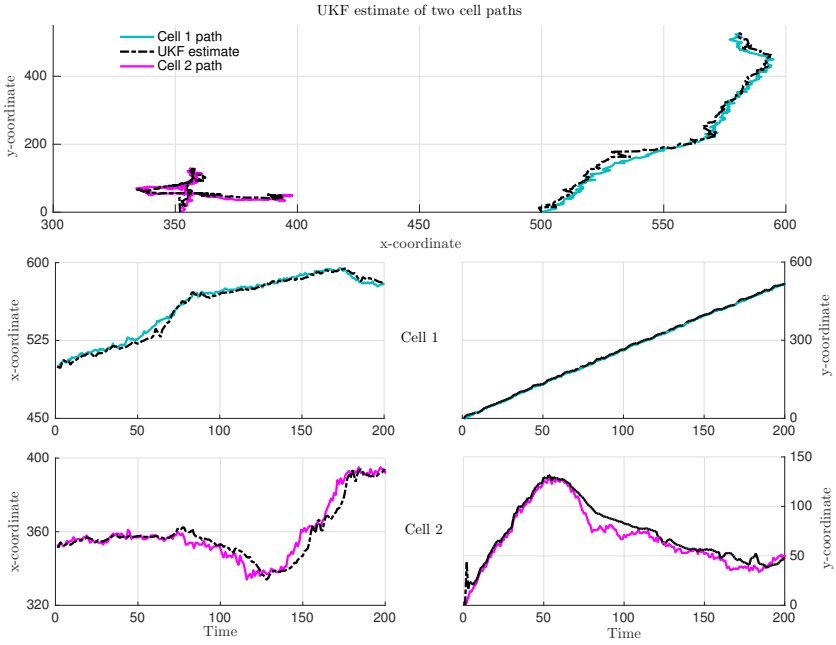


FIGURE 3. UKF tracking of two cells with different movement patterns.

resolution microscopy images (see Figure 1). We emphasise that the main interest for biological applications is the inference of the parameters. However, since the true parameters for the real data are unknown, we use the tracking of the cell trajectories as a proxy for assessing the accuracy of inference (see Figure 3). As can be seen from the left panel of Figure 3, we have picked two cells with very different behaviour (one dominated by drift, the other dominated by diffusion). In both cases, the path reconstructed with the UKF is very accurate.

5 Conclusions and future work

In this paper, we demonstrate the application of the UKF, a Bayesian filtering technique that adequately trades off accuracy versus computational efficiency, to a real-world problem potentially relevant to cancer research: the movement of *Dictyostelium* cells, which has not been tackled at individual cell level before. Our results indicate that the UKF can be successfully used for parameter inference and tracking cells displaying various movement patterns. Future work will extend this work by applying the UKF to a population of cells. Additionally, we plan to fit models describing alternative movement mechanisms, such as the self-induced gradient model described by Tweedy et al. (2016) and employ model selection criteria to choose the best model.

Acknowledgments: The research described in this article is part of the research programme of SoftMech, the Centre for multiscale soft tissue mechanics with application to heart & cancer, funded by the Engineering and Physical Sciences Research Council (EPSRC) of the UK, grant reference number EP/N014642/1.

References

Bishop, C. (2006) Pattern Recognition and Machine Learning. *Springer*

Golightly, A. & Wilkinson, D. J. (2011) Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface focus*, **1(6)**, 807 – 820

Julier, S. J. & Uhlmann, J. K. (1997) A New Extension of the Kalman Filter to Nonlinear Systems. *AeroSense'97*, 182 – 193

Sitz, A., Schwarz, U., Kurths, J. & Voss, H. U. (2002) Estimation of parameters and unobserved components for nonlinear systems from noisy time series *Phys. Rev. E*, **66(1)**, 016 – 210

Tweedy L., Knecht D.A., Mackay G.M. & Insall R.H. (2016) Self-Generated Chemoattractant Gradients: Attractant Depletion Extends the Range and Robustness of Chemotaxis. *PLoS Biol*, **14(3)**

Regularisation of Generalised Linear Mixed Models with autoregressive random effect

Jocelyn Chauvet[1], Catherine Trottier[2], Xavier Bry[1]

[1] University of Montpellier, France
[2] University of Paul-Valéry - Montpellier 3, France

E-mail for correspondence: jocelyn.chauvet@umontpellier.fr

Abstract: We address regularised versions of the Expectation-Maximisation (EM) algorithm for Generalised Linear Mixed Models (GLMM) in the context of panel data (measured on several individuals at different time-points). A random response y is modelled by a GLMM, using a set X of explanatory variables and two random effects. The first one introduces the dependence within individuals on which data is repeatedly collected while the second one embodies the serially correlated time-specific effect shared by all the individuals. Variables in X are assumed many and redundant, so that regression demands regularisation. In this context, we first propose a L_2-penalised EM algorithm, and then a supervised component-based regularised EM algorithm as an alternative.

Keywords: Regularised EM algorithm; Generalised Linear Mixed Model; Autoregressive random effect; Panel data analysis.

1 Introduction

One of the main purposes of panel data analysis is to account for the dependence induced by repeatedly measuring an outcome on each individual over time. Besides, due to the fact that it is nowadays increasingly possible to collect large amounts of data, the potentially high level of correlation among explanatory variables should be taken into account. To this end, ridge-, lasso- and component-based regularisations have recently been highlighted.

In the Linear Mixed Models (LMM) framework, Eliot et al. (2011) proposed to extend the classical ridge regression to longitudinal biomarker data. They suggested a variant of the EM algorithm to maximise a ridge-penalised likelihood. This variant includes a new step to find the best shrinkage parameter - in the Generalised Cross-Validation (GCV) sense - at each iteration.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

With a view towards variable selection, Schelldorfer et al. (2014) proposed a L_1-penalised algorithm for fitting a high-dimensional Generalised Linear Mixed Models (GLMM), using Laplace approximation and an efficient coordinate gradient descent.

In the GLM framework, in order both to regularise the linear predictor and to facilitate its interpretation, Bry et al. (2013) developed a PLS-type method – Supervised Component-based Generalised Linear Regression (SCGLR) – which yields explanatory components. Chauvet et al. (2016) extended SCGLR to GLMM by using an adaptation of Schall's algorithm (Schall (1991)).

To the best of our knowledge, the random effects in the previous strategies are assumed normally distributed with independent levels. However, in the panel data framework, the question naturally arises of the autocorrelation of the time-specific random effect. Consequently, our objective is twofold: on the one hand, to extend the Mixed Ridge Regression of Eliot et al. (2011) to the GLMMs with an AR(1) random effect; and on the other hand, to present the main ideas of a new version of SCGLR which handles the high dimensional case.

2 Model hypotheses

In this section, we recall the main hypotheses of the GLMM framework and we introduce the random effect distributions. For the sake of simplicity, we consider balanced panel data with N individuals, each of them observed at the same T time-points. We denote by $n = N \times T$ the total number of observations. Let X be the $n \times p$ fixed effects design matrix, and U the $n \times q$ random effects design matrix. Let also Y be the n-dimensional random response vector, β the p-dimensional vector of fixed effects, and ξ the q-dimensional vector of random effects. We observe a realisation y of Y, but ξ is not observed. We conventionally assume that:

(i) the $Y_i \,|\, \xi,\ i \in \{1, \ldots, n\}$ are independent and their distribution belongs to the exponential family;

(ii) the conditional mean $\mu_i = \mathbb{E}(Y_i \,|\, \xi)$ depends on β and ξ through the link function g and the linear predictor $\eta_i = x_i^{\mathrm{T}}\beta + u_i^{\mathrm{T}}\xi$, with $\eta_i = g(\mu_i)$.

Less conventionally, we consider two random effects ξ_1 and ξ_2 with different roles and distributions:

(i) ξ_1 is the individual-specific random effect. Assuming individuals are independent, we suppose:

$$\xi_1 \sim \mathcal{N}_N\left(0,\, \sigma_1^2 I_N\right),$$

with σ_1^2 the unknown "individual" variance component.

(ii) ξ_2 is the serially correlated time-specific effect common to all the individuals, which can be viewed as some latent phenomenon not measured in the explanatory variables. As these effects tend to persist over time, we model them with a stationary order 1 autoregressive process (AR(1)), i.e. for each $t \in \{1, \ldots, T-1\}$,

$$\begin{aligned} \xi_{2,t+1} &= \rho\xi_{2,t} + \nu_t, \\ \nu_t &\overset{\text{iid}}{\sim} \mathcal{N}\left(0,\, \sigma_2^2\right), \end{aligned}$$

where ρ is the unknown parameter of the AR(1) and σ_2^2 the unknown "temporal" variance component. Such time-specific effects arise naturally for instance in an economic context (e.g. all companies share a common economic climate which tend to persist over time), or in biology (e.g. the ecological environment is often too complex to be directly observed through the explanatory variables).

Finally, ξ_1 and ξ_2 are assumed independent. Denoting $\xi = (\xi_1^{\mathrm{T}}, \xi_2^{\mathrm{T}})^{\mathrm{T}}$, $U_1 = I_N \otimes \mathbf{1}_T$, $U_2 = \mathbf{1}_N \otimes I_T$ and $U = [U_1 \,|U_2]$, linear predictor η can be matricially written:

$$\eta = X\beta + U\xi.$$

3 Methods

Owing to the GLMM dependence structure, the Fisher scoring algorithm was adapted by Schall (1991). We, in turn, adapt Schall's algorithm by introducing a regularised EM at each step in order to take into account the high level of correlation in X and the unconventional random effects distributions. Two steps appear in our method: the linearisation step and the estimation step.

Linearisation step. For each $i \in \{1, \ldots, n\}$, a classic order 1 linearisation of y_i around μ_i is given by: $g(y_i) \simeq z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$. Matricially, this approximation provides a working variable z entering the following linearised model

$$\mathcal{M}: \quad z = X\beta + U\xi + e,$$

with $\mathrm{Var}(e \,|\, \xi) = \mathrm{Diag}\left(\left[g'(\mu_i)\right]^2 \mathrm{Var}(Y_i \,|\, \xi)\right)_{i=1,\ldots,n} = \Gamma.$

Estimation step. Instead of solving Henderson's system associated with $\mathcal{M}$ seen as a LMM (as proposed by Schall (1991)), we rather propose a regularised EM step. We suggest an adaptation of the L_2-penalised EM algorithm of Eliot et al. (2011) for low dimensional data ($p < n$), and a supervised component-based regularised EM algorithm for the high dimensional case ($p \gg n$), because then, interpretable dimension reduction is needed.

3.1 The low dimensional case

Our estimation step is based on Green (1990), who popularised the use of the EM algorithm for penalised likelihood estimation, and Golub et al. (1979), who encouraged the use of the GCV for efficiently choosing the ridge parameter λ. However, contrary to the homoskedastic LMM considered in Eliot et al. (2011), $\mathcal{M}$ contains heteroskedastic errors. We will then opt for the modified GCV criterion suggested by Andrews (1991), p. 372.

Denoting $\theta = \left(\beta, \sigma_1^2, \sigma_2^2, \rho\right)$, we present the current iteration of our L_2-penalised EM algorithm for GLMM with AR(1) random effect.

(1) Linearisation step. Set:

$$\mathcal{M}^{[t]}: z^{[t]} = X\beta + U\xi + e, \text{ with } \mathrm{Var}(e \,|\, \xi) = \Gamma^{[t]}.$$

(2) Estimation step.

(2.a) Denoting L the complete log-likelihood of the linearised model, define the associated complete penalised log-likelihood L_{pen} by:

$$L_{\text{pen}}(\theta; z, \xi) = L(\theta; z, \xi) - \frac{\lambda}{2}\beta^{\mathrm{T}}\beta.$$

(2.b) Denoting $\widehat{z}^{[t]}$ the fitted values and $S_\lambda^{[t]}$ the "hat-matrix" satisfying the equality $\widehat{z}^{[t]} = S_\lambda^{[t]} z^{[t]}$, set:

$$\lambda^{[t]} \longleftarrow \arg\min_{\lambda} \left\{ \text{GCV}(\lambda) = \frac{n^{-1} \left\| z^{[t]} - S_\lambda^{[t]} z^{[t]} \right\|^2_{\Gamma^{[t]-1}}}{\left[1 - n^{-1}\text{tr}\left(S_\lambda^{[t]}\right)\right]^2} \right\}.$$

(2.c) EM step. Set:

$$\mathcal{Q}_{\text{pen}}\left(\theta, \theta^{[t]}\right) = \mathbb{E}_{\xi|z}\left[L_{\text{pen}}(\theta; z^{[t]}, \xi) \,|\, \theta^{[t]}, \lambda^{[t]}\right],$$

$$\theta^{[t+1]} \leftarrow \arg\max_{\theta} \mathcal{Q}_{\text{pen}}\left(\theta, \theta^{[t]}\right).$$

(3) Updating step. Set $\xi^{[t+1]} = \mathbb{E}_{\xi|z}\left(\xi \,|\, \theta^{[t+1]}\right)$, and update working variable $z^{[t+1]}$ and variance-covariance matrix $\Gamma^{[t+1]}$.

Steps **(1)-(3)** are repeated until stability of parameters β, σ_1^2, σ_2^2 and ρ is reached.

3.2 The high dimensional case

In the $p \gg n$ case, we need to decompose the linear predictor on a small number of interpretable dimensions. To that end, we propose to iteratively maximise a component-based regularised $\mathcal{Q}-$function.

Let $C = XU$ be the set of principal components of X with non-zero eigenvalues and $f = Cw$ the component we currently seek. Let also ϕ denote a structural relevance (SR) criterion (see Bry and Verron (2015)):

$$\phi(w) = \left(\sum_{j=1}^{p}\left[\text{cor}^2\left(x^j, f\right)\right]^l\right)^{\frac{1}{l}}, \quad l \geqslant 1.$$

$s \in [0, 1]$ being a parameter tuning the relative importance of the SR with respect to L, the $\mathcal{Q}-$function would then be:

$$\mathcal{Q}_{\text{reg}}\left(\theta, \theta^{[t]}\right) = \mathbb{E}_{\xi|z}\left[L_{\text{reg}}(\theta; z, \xi) \,|\, \theta^{[t]}\right], \text{ with}$$

$$L_{\text{reg}}(\theta; z, \xi) = (1 - s)L(\theta; z, \xi) + s\phi(w).$$

Parameters s and l are tuned by cross-validation and higher rank components are computed like rank 1 component, after adding extra orthogonality constraints.

4 Numerical results

In order to evaluate the performance of our L_2-penalised EM algorithm, we conducted simulation studies in the canonical Poisson case. We present some graphical diagnoses in FIGURE 1, which aim at answering three questions: (1) Is the convergence assured? (2) How good are the estimations? (3) Are they sensitive to the value of ρ? The answers to these questions is given in the figure's caption.

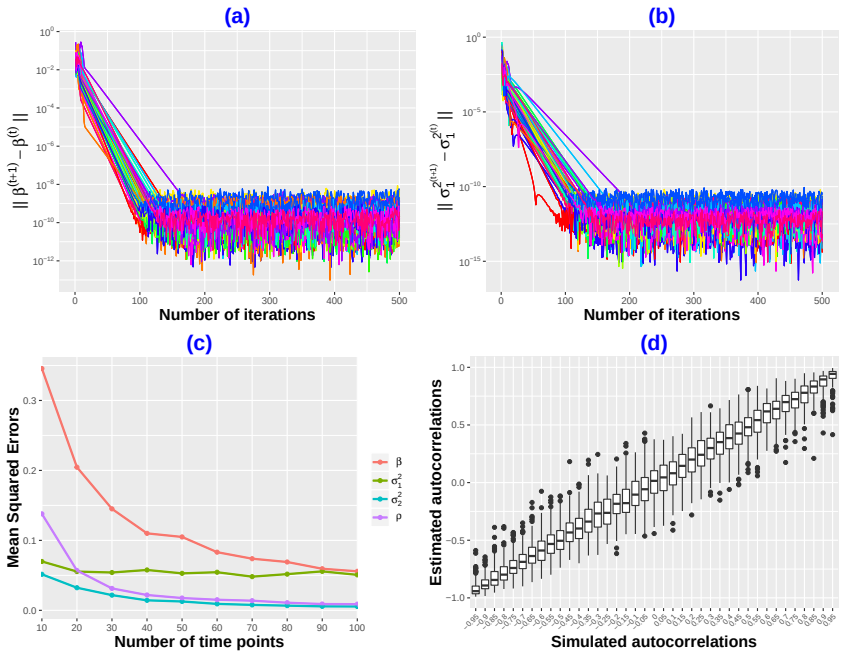


FIGURE 1. Graphical diagnoses relative to the L_2-penalised EM algorithm. **(a),(b):** *40 trajectories of the L_2-convergence criterion for parameters β and σ_1^2 (A similar behaviour is observed for parameters σ_2^2 and ρ). About a hundred iterations is necessary to achieve convergence.* **(c):** *MSEs of parameters $\beta, \sigma_1^2, \sigma_2^2$ and ρ on simulated data where $N = 10$ and $T \in \{10, 20, \ldots, 100\}$. As expected, MSEs of β, σ_2^2 and ρ decrease towards zero. In contrast, since N is fixed, the MSE of σ_1^2 is constant.* **(d):** *Boxplots of estimated ρ according to real value.*

References

Andrews, D.W. (1991). Asymptotic optimality of generalized CL, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, **47**, 359 – 377.

Bry, X., Trottier, C., Verron, T., and Mortier, F. (2013). Supervised component generalized linear regression using a pls-extension of the fisher scoring algorithm. *Journal of Multivariate Analysis*, **119**, 47 – 60.

Bry, X. and Verron, T. (2015). THEME: THEmatic model exploration through multiple co-structure maximization. *Journal of Chemometrics*, **29**, 637 – 647.

Chauvet, J., Trottier, C., Bry, X., and Mortier, F. (2016). Extension to mixed models of the Supervised Component-based Generalised Linear Regression. *COMPSTAT: Proceedings in Computational Statistics*.

Eliot, M., Ferguson, J., Reilly, M.P., and Foulkes, A.S. (2011). Ridge Regression for Longitudinal Biomarker Data. *The International Journal of Biostatistics*, **7**, 1, Article 37.

Golub, G.H., Heath, M., and Wahba, G. (1979). Generalized cross- validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215 – 223.

Green, P.J. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society, Series B*, **52**, 443 – 452.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719 – 727.

Schelldorfer, J., Meier, L., and Bühlmann, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using l_1- penalization. *Journal of Computational and Graphical Statistics*, **23**, 460 – 477.

Optimal Number of Clusters Based on Inter Cluster Elements Mapping

K. M. Matawie[1], A. Mehar[1], A. Maeder[2]

[1] Western Sydney University, Australia
[2] Western Sydney University and Flinders University, Australia

E-mail for correspondence: k.matawie@westernsydney.edu.au

Abstract: To estimated the optimal number of clusters and evaluate the associated quality of the formed clusters is one of the major issue in cluster analysis. In this paper, we present and extend the MMM index to find the optimal number of clusters and their quality. The new index determines the combined mapped elements information from related clusters by comparing the occurrence of common elements across the sets of clusters from successive k number of clusters. This requires comparing the k resultant clusters in each set with respect to 'forward' and 'backward' mapping of common elements for adjacent and non-adjacent clusters (all possible distant) at all possible k. This method will also provide indicators for the similarity and overlapped (dissimilarity) of mapped elements. The optimal or best estimated number of clusters and their quality will be decided using the combination of maximum average similarity and minimum average overlap measures. The evaluation and performance of this index is illustrated and tested using real dataset.

Keywords: k-means, clustering, cluster quality, forward and backward mapping, similarity and overlap

1 Introduction

Unsupervised clustering is a data analysis technique which has no a priori information available to determine the intrinsic structure in the dataset. Due to this lack of a priori knowledge, and the consequential need to infer or discover structure in the data, most clustering algorithms tried to find similar objects within a dataset according to their characteristics: this is usually accomplished by using some well-established distance measures (e.g. Euclidean, Maximum, Manhattan). The labelling of objects using distance measures allows the dataset to be parti-

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

tioned into clusters, where characteristics of objects in the same cluster are more similar than they are to objects in other clusters. This approach has been utilized successfully in various types of problems across many fields. As described in Ng and Han (1994) clustering algorithms are classified into two major types partitional and hierarchical clustering. We will be focusing here on partitional clustering algorithms and in particular the k-means algorithm. The main challenge in using the k-means algorithm is the optimal (best) k value (number of clusters) giving well-structured or quality clusters. To determine the best k value often requires a large number of steps in a given dataset, when particularly iterative or exhaustive search is applied. For this reason a number of techniques have been developed based on use of heuristics such as to minimise the distance from centroid of a cluster to each object in the cluster, or to maximise the distance between clusters (intra and inter cluster distances), or to use an average scatter of clusters and total separation between clusters. They are to some extend limited in effectiveness by the assumptions made about what constitutes the best clustering, this is also called the unsupervised machine learning. The MMM approach Mehar at al.(2010,2013) and Matawie et al(2015) allows this optimisation to be undertaken for k clusters using knowledge of the clustering solutions obtained from k to k+1 'forward' and k+1 to k 'backward' for only adjacent distances in consecutive k situation. In this paper extended MMM will be presented with better evaluation criterion based on the forward and backward changes in cluster membership over the range of all successive k values adjacent and non-adjacent clusters. In this work it is assumed that minimum k is 2 ($k_{min} = 2$) , and maximum K is 16 ($k_{max} \leq 16$) and r=1 to 14 , limiting k and therefore to the most practical finite range. The details and development of the enhanced MMM is explained in details in section 2. The forward and backward inter cluster elements mapping when k=2 with all possible k+r , where r=1,2,.....,14 , is given in Figure-1 below with bold and light arrowheads indicating the forward and backward directions respectively.

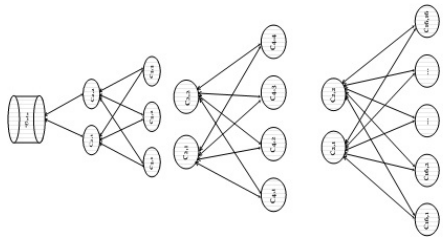


FIGURE 1. Forward and backward mapping of elements when k=2 for different k+r distant, and r=1,2,...., K-k.

2 MMM Index Extended

In this section we will extend the approach of developing the MMM index described in Mehar at al.(2010) by including information from adjacent and non-adjacent (more distant) mapping of the common elements between different clusters. To develop a practical implementation of this approach we need to consider

different resultant of k number of clusters from the k-means algorithm, over a range of successive k values always starting with minimum number of clusters $k_{min} = 2$ to maximum $k_{max} = K$, where K = N-1 and N= total number of elements (objects or observations) in the dataset D. Following are notations and details used to define and describe the approach. Assume we have dataset D containing N number of observations (elements). Each observation may have number of variables that may be used to determine information about the relationship between them. Using a k-means clustering algorithm on D with a certain set of parameters to control the behaviour of the k-means algorithm such as choosing variables, number of variables, number of clusters (centres), number of maximum iterations, number of initial random starting sets of seeds and algorithm to obtain the required partitioning of D into k clusters, k $\leq K$. Suppose at any k number of clusters, we define the first set of clusters to be $C_{(k)i}$ where k $\in$ 2,3,.....,K and i =1,2,.....,k. We define another set of clusters $C_{(k+r)j}$, where r $\in$ 1,2,.....,K-k and j = 1,2,....,k+r. As k get larger and closer to K the K-k set of clusters will have fewer elements, per cluster, than the preceding set of clusters until reaching one element in each K cluster (when K = N-1). Let us define $m_{(k,k+r)ij}$ as the number of common elements in the forward inter cluster mapping from the source cluster $C_{(k)i}$ to the target cluster $C_{(k+r)j}$ i.e. the number of elements of $C_{(k)i} \cap C_{(k+r)j}$. These forward mapped number of elements are used to construct forward mapped inter cluster matrices $[M_{(k,k+r)ij}]$ for i=1,2,....,k and j = 1,2,....,k+r, i.e. the mapped elements from a particular cluster $C_{(k)i}$ at k to all different sets of target clusters $C_{(k+r)j}$, j=1,2,...., k+r. Similarly but in reverse, the backward inter cluster mapping $m_{(k+r,k)ji}$ is defined to be the numbers of common elements from the target to source cluster, and the backward mapped inter cluster matrices is defined to be $[M_{(k+r,k)ji}]$. These backward mapped inter cluster matrices are also defined as transpose of the M forward inter cluster mapped matrices. These M matrices are rectangular in size (k,(k+r)) and ((k+r),k) respectively. We now define the forward inter cluster proportion of elements as $[P_{(k,k+r)ij}]$ matrices from a source cluster $C_{(k)i}$ to all the target sets of clusters $C_{(k+r)j}$. Each row (vector) in P matrix is summed up to 1 and the proportion matrix can be computed as

$$P_{(k,k+r)ij} = \frac{m_{(k,k+r)ij}}{\sum_{j=1}^{k+r} m_{(k,k+r)ij}} \tag{1}$$

Equation (1) simply shows the inter cluster proportion of the elements mapped from the source cluster $C_{(k)i}$ to the target cluster $C_{(k+r)j}$. Similarly, we can obtain the backward inter clusters proportion $[P_{(k+r,k)ji}]$ matrices. The inner product of the forward and backward proportion matrices is computed and constructed as it will determine the mutual similarity between clusters. This inner product will result in a matrix of (k,k) combined mapped proportion matrices called $O_{(k,k)}$ for different k+r distances(see equation (2) below). These combined mapped proportion matrices O provide the combined similarity at each entry of the diagonal, while the off diagonal entries are the dissimilarity or overlap proportions of the clusters.

$$[O_{(k,k)}] = [P_{(k,k+r)ij}][P_{(k+r,k)ji}] \tag{2}$$

It is important to note, that the inner product of the forward and backward are different due to the cardinality difference as ((k, k+r),(k+r, k)) is not equal to

((k+r, k),(k, k+r)), also the backward to forward would be less informative than forward to backward at a given k value.

Finally, we define the combined mapped elements matrices as $[Q_{(k,k)}]$ to be calculated from the inner product of matrix [**k**] by matrix $[O(k,k)]$ for each k and k+r distance, where [**k**] is a (k,k) diagonal matrix with diagonal entries determined by the size or number element of each cluster from k-means algorithm at k. The diagonal entries of $[Q_{(k,k)}]$ matrix $q_{(ii)}$ at each k with different k+r distances are the number of elements belong or remained in the same cluster (within cluster) while the off diagonal $q_{(ij)}$ entries are the number of elements belong or moved from other clusters (representing the overlap at k).

2.1 Computing Cluster Similarity and Overlap

In this section similarity, overlap, average similarity and average overlap are computed from Q, the combined mapped elements matrix. We define the similarity as the trace value of the Q matrix at each k that is mapped to different k+r distances. Equations (3) and (4) below shows the similarity and overlap of the Q matrices respectively;

$$TraceQ_{(k,k+r)} = \sum_{i=1}^{k} q_{ii} \tag{3}$$

$$Overlap_{(k,k+r)} = N - TraceQ_{(k,k+r)} \tag{4}$$

In addition, the traces of each k to different k+r mapped distance would be used to define and compute the average similarity and overlap of k to k+r, equation (4) and (5) below calculate these averages.

$$AverageTrace_k = \mu_k = \frac{\sum_{r=1}^{K-k} Trace_{(k,k+r)}}{K-k} \tag{5}$$

$$AverageOverlap_k = N - \mu_k \tag{6}$$

where, k = 2,3,.....,K and r = 1,2,3,....K-k.

The best (optimal) estimated number of clusters K* can be found to be the maximum average of the traces from equation (4) above;

$$K^* = Max(AverageTrace_k) \tag{7}$$

Equation 7 is considered as a criteria for the best estimated number of cluster, however, there are more details needed to cover the situations when the best estimated number of clusters are fully or partially separated.

3 Real Data Example

To illustrate the proposed method we will use a well-known Ruspini real dataset with four clusters that is known in advance. This particular dataset is also widely

used in the literature to illuminate and evaluate clustering number and structure by many research papers, and was first used and analysed by Ruspini (1970) to investigate fuzzy clustering. This data is a two dimensional numerical dataset that includes 75 elements, Figure 3(a and b) shows the scatter plot of this dataset and the k-means clusters when k=4 respectively. Forward and backward mapping proportion and combined matrices at different k and for all k+r distances (up to K=16) are calculated and their Trace, AverageTrace and number of overlapped elements at each k are also calculated and presented in Figure 3 (c and d). The average trace at each k is given by the solid black line in Figure 3 (c) showing that the average trace was maximum at k= 2, 3 and 4 and was equal to the total number of the element (N=75). This indicates there is a continuous potential to split into new and fully separated clusters up until k=4, and as k increased and moved away from 4 the trace averages start decreasing and fluctuating but always below the value at k=4. The overlap calculations given in figure 3(d) showed zero overlap up until k=4 and it increased after k=4 by which we strongly nominate k=4 as the best estimated number of clusters.

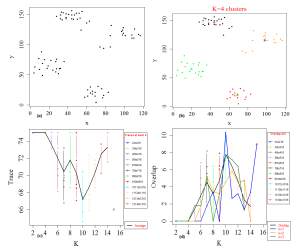


FIGURE 2. Plot (a) Rusipini data scatter plot, (b) k-means clusters, (c and d)the calculations of trace and overlap at different k to k+1.

4 Conclusion

A good choice of K is an important feature for building meaningful homogenous and well separated clusters when applying the k-means clustering algorithm for a dataset. In this study we extended and enhanced the MMM index approach based on the results from a k-means clustering algorithm, and demonstrated its effectiveness to explore the clustering structure for finding a best (correct) value of K. This approach aimed to maximize the mapping similarity based on the forward and backward inter cluster elements mapping that shows to be more descriptive, informative and analytical by which the best and stable set of clusters are reached. Development of this approach is still in progress and it will cover its advantage and effectiveness when applied on variety of simulated and real datasets including the evaluation and comparison with other existing validation indexes.

References

Matawie, K., Mehar, M.A. and Maeder, A. (2015). An approach to determine clusters overlap for k-means clustering. *International Workshop on Statistical Modelling, Linz- Austria*, 163 – 166.

Mehar, A., Maeder, A., Matawie, K. and Ginige, A. (2010). *Blended Clustering for Health Data Mining.* Springer Berlin Heidelberg, 130 – 137.

Mehar, A.M., Matawie, K. and Maeder, A. (2013). Determining an optimal value of K in K-means clustering. *IEEE International Conference on Bioinformatics and Biomedicine, Shanghai- china*, 51 – 55.

Ng, R.T. and Han, J. (1994). Efficient and Effective Clustering Methods for Spatial Data Mining. *Proceedings of 20th International Conference on Very Large Data Bases (VLDB)*, 144 – 155.

Ruspini, E.H. (1970). Numerical methods for fuzzy clustering. *Information Sciences*,**2(3)**, 319 – 350.

Nowcasting infectious disease outbreaks using constrained P-spline smoothing

Jan van de Kassteele[1], Paul Eilers[2], Jacco Wallinga[13]

[1] National Institute for Public Health and the Environment, the Netherlands
[2] Erasmus Medical Center, the Netherlands
[3] Leiden University Medical Center, the Netherlands

E-mail for correspondence: jan.van.de.kassteele@rivm.nl

Abstract: During an infectious disease outbreak it is crucial to have timely information on epidemic trends. However, the reporting of new cases is subject to delay. The real-time assessment of the expected number of cases based on partial information is called nowcasting. We present a nowcasting model based on two dimensional P-spline smoothing with additional constraints. Our aim is to predict the number of occurred-but-not-yet-reported cases. We force the underlying time-varying delay distribution to be unimodal, and to be zero at a predefined maximum delay. We illustrate our method on a large measles outbreak in the Netherlands. We show that even with very limited information our model is able to predict the number of occurred-but-not-yet-reported cases very well.

Keywords: Smoothing; Constrained P-splines; Asymmetric penalty; Infectious disease outbreaks

1 Introduction

During an infectious disease outbreak the National Institute for Public Health and the Environment - RIVM has the responsibility to real-time monitor the number of cases, in order to inform relevant health authorities. However, there is a delay between the first day of symptoms onset and the time that the case is registered. A consequence is that the epidemic curve of reported cases drops to zero at the present day.

The assessment of the current situation based on imperfect or partial information is called nowcasting (Donker et al., 2011; Höhle and an der Heiden, 2014). When the delay distribution is known, it is possible to estimate the number of new cases in real-time, e.g. by dividing the number of reported cases by the fraction of

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

reported cases. In practice however, it is very difficult to obtain stable estimates, especially when the number and fraction of reported cases is low, or even zero. Typically, no cases are reported during the weekend.

Statistical modeling techniques can make an improvement here. The number of reported cases can be tabulated by calendar time and delay. If we set a maximum delay (where all cases have been reported), we get the so-called reporting trapezium. The aim is to predict the number of occurred-but-not-yet-reported cases outside the trapezium, while accounting for a time dependent delay distribution. We model the expected number of reported cases in the reporting trapezium by a smooth function of calendar time and delay using two dimensional P-splines. However, to safely extrapolate the surface outside the reporting trapezium, especially at the beginning of the outbreak, additional constraints are required. First, the surface is expected to be unimodal in the delay direction. Second, it should go to zero at the maximum delay. Our approach is fast; one run takes about 30 seconds.

2 Methods

The number of reported cases by calendar time and delay is a stochastic variable $\mathbf{Y}$ and is assumed to follow a Negative Binomial distribution with reporting intensity $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y})$ and overdispersion parameter θ. The log-intensity is modeled by a linear predictor $\boldsymbol{\eta} = \log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$. To separate the calender time and delay trend surface from weekday effects, we write model matrix $\mathbf{X}$ as a partitioned matrix $[\mathbf{X}_s|\mathbf{X}_w]$ and coefficients $\boldsymbol{\beta}$ as a partitioned vector $[\boldsymbol{\beta}_s|\boldsymbol{\beta}_w]$. Here $\mathbf{X}_s = \mathbf{X}_d \otimes \mathbf{X}_t$, where $\mathbf{X}_t$ and $\mathbf{X}_d$ are $n_t \times k_t$ and $n_d \times k_d$ B-spline basis matrices for marginal variables calender time t and delay d respectively. $\mathbf{X}_w$ is a $n_t n_d \times k_w$ matrix with weekday expanded to a set of dummy variables. Hence, weekday effects are expressed as deviations from the trend surface.

Smoothness of the trend surface $\mathbf{X}_s\boldsymbol{\beta}_s$ is achieved by quadratic penalization of the coefficients. The penalized log-likelihood function becomes

$$\ell^* = \ell(\boldsymbol{\beta}_s, \boldsymbol{\beta}_w, \theta|\mathbf{y}) - \tfrac{1}{2}\lambda_t\boldsymbol{\beta}_s'\mathbf{D}_t'\mathbf{D}_t\boldsymbol{\beta}_s - \tfrac{1}{2}\lambda_d\boldsymbol{\beta}_s'\mathbf{D}_d'\mathbf{D}_d\boldsymbol{\beta}_s.$$

We assume that outside the reporting trapezium the delay distribution is constant and the tails of it are linear (on a log-scale), so we have $\mathbf{D}_t = \mathbf{I}_{k_d} \otimes \mathbf{D}_{(1)}$ and $\mathbf{D}_d = \mathbf{D}_{(2)} \otimes \mathbf{I}_{k_t}$, which are first and second difference operator matrices in respectively time and delay direction. λ_t and λ_d are unknown smoothing parameters.

An unimodal function estimate in the delay direction can be achieved by a priori only allowing negative values for the second order differences between the coefficients in the delay direction (Eilers, 2005). The modified penalized log-likelihood function then becomes

$$\ell^{**} = \ell^* - \tfrac{1}{2}\kappa_u\boldsymbol{\beta}_s'\mathbf{D}_u'\mathbf{V}_u\mathbf{D}_u\boldsymbol{\beta}_s$$

where $\mathbf{D}_u = \mathbf{D}_{(2)} \otimes \mathbf{I}_{k_t}$. Typically $\kappa_u = 10^6$ and $\mathbf{V}_u = \mathrm{diag}(\mathbf{v}_u)$ is a matrix with asymmetric weights

$$\mathbf{v}_u = \begin{cases} 1 & \text{if} \quad \mathbf{D}_u\boldsymbol{\beta}_s \geqslant 0 \\ 0 & \text{if} \quad \mathbf{D}_u\boldsymbol{\beta}_s < 0 \end{cases}.$$

In order to ensure a reporting intensity near zero at the maximum delay, we set an additional constraint $\mathbf{X}_s\boldsymbol{\beta}_s < \mathbf{g}$ at the boundary, where $\mathbf{g}$ is a vector with negative numbers, say -5. The modified penalized log-likelihood function then becomes

$$\ell^{***} = \ell^{**} - \tfrac{1}{2}\kappa_b(\mathbf{X}_s\boldsymbol{\beta}_s - \mathbf{g})'\mathbf{V}_b(\mathbf{X}_s\boldsymbol{\beta}_s - \mathbf{g}),$$

where $\mathbf{V}_b = \text{diag}(\mathbf{bv}_b)$ is a matrix with asymmetric weights

$$\mathbf{v}_b = \begin{cases} 1 & \text{if} \quad \mathbf{X}_s\boldsymbol{\beta}_s \geqslant \mathbf{g} \\ 0 & \text{if} \quad \mathbf{X}_s\boldsymbol{\beta}_s < \mathbf{g} \end{cases},$$

and $\mathbf{b}$ is a vector with elements equal to 1 at the locations where the boundary constraint is applied, and 0 otherwise. Typically $\kappa_b = 10^6$.

Finally, two additional constraints are applied. To obtain stable estimates of the weekday effects, a ridge penalty $-\frac{1}{2}\kappa_w\boldsymbol{\beta}_w'\boldsymbol{\beta}_w$ with $\kappa_w = 1$ is added to log-likelihood function. To make estimation of the trend surface numerically stable, a small ridge penalty $-\frac{1}{2}\kappa_s\boldsymbol{\beta}_s'\boldsymbol{\beta}_s$ with $\kappa_s = 10^{-3}$ is added.

Given λ_t, λ_d and θ, the coefficients $\boldsymbol{\beta}$ are found by penalized iterative weighted least squares, by repeatedly solving the system

$$(\mathbf{X}'\mathbf{W}\mathbf{X} + \mathbf{P})^{-1}\begin{pmatrix}\boldsymbol{\beta}_s \\ \boldsymbol{\beta}_w\end{pmatrix} = \mathbf{X}'\mathbf{W}\mathbf{z} + \begin{pmatrix}\kappa_b\mathbf{X}'\mathbf{V}_b\mathbf{g} \\ \mathbf{0}_{k_w}\end{pmatrix},$$

where $\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is the working variable and weight matrix $\mathbf{W} = \text{diag}(\mathbf{rw})$. $\mathbf{r}$ is a vector with elements equal to 1 if the element lies within the reporting trapezium, and 0 otherwise and $\mathbf{w} = \boldsymbol{\mu}^2/(\boldsymbol{\mu} + \frac{\boldsymbol{\mu}^2}{\theta})$ is a weight vector, corresponding to the Negative Binomial distribution.

The corresponding penalty matrix $\mathbf{P}$ is given by

$$\begin{aligned}\mathbf{P} = \text{blockdiag}(&\lambda_t\mathbf{D}_t'\mathbf{D}_t + \lambda_d\mathbf{D}_d'\mathbf{D}_d + \\ &\kappa_u\mathbf{D}_u'\mathbf{V}_u\mathbf{D}_u + \kappa_b\mathbf{X}_s'\mathbf{V}_b\mathbf{X}_s + \kappa_s\mathbf{I}_{k_tk_d}, \kappa_w\mathbf{I}_{k_w}).\end{aligned}$$

The overdispersion parameter θ is found maximizing the log-likelihood given the current estimates of $\boldsymbol{\beta}$ within the IWLS algorithm. Smoothing parameters λ_t and λ_d are found by minimizing the BIC.

The nowcast is then achieved as follows. First generate 1000 Monte Carlo samples of $\boldsymbol{\beta}$ and θ. Given $\boldsymbol{\mu} = \exp(\mathbf{X}\boldsymbol{\beta})$ and θ, draw 1000 samples from the Negative Binomial distribution for each combination of calendar time and delay outside the reporting trapezium. By summarizing the already reported number of cases and the predicted number of cases over the delays, we obtain the predictive distribution by date. The time dependent delay distribution is found by conditioning the trend surface $\exp(\mathbf{X}_s\boldsymbol{\beta}_s)$, in a $n_t \times n_d$ matrix, on its column sums.

3 Application: nowcasting a measles outbreak

During May 2013 - March 2014, the Netherlands was affected by a large measles outbreak. Figure 1 shows the situation halfway the outbreak on October 1, 2013. Although zero cases have been reported on October 1 (dark bars), we are still able to predict the number of occurred-but-not-yet-reported cases (light bars). As more information is available in the past, the prediction interval gets smaller.

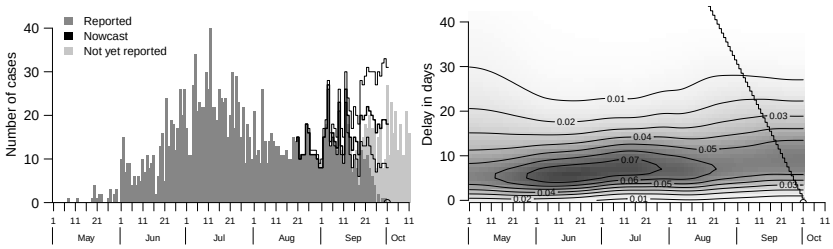


FIGURE 1. Left panel: Nowcast of the measles outbreak on October 1, 2013. The black line is the nowcast (incl. 95% prediction interval) to a maximum delay of six weeks back. Right panel: Time dependent delay distribution.

The right panel shows the estimate of the time dependent delay distribution. The triangle on the right has been obtained by extrapolation. The highest reporting probabilities occur with a delay between five and nine days.

4 Conclusions

We have presented a nowcasting model based on constrained P-splines. The model takes care of a time dependent delay distribution, right truncation and day-of-the-week effects. Without the additional constraints, stable extrapolation outside the reporting trapezium is almost impossible. Even with very limited information, the model is able to predict the number of occurred-but-not-yet-reported cases very well on a daily resolution.

References

Donker T., van Boven, M., . . . , and Wallinga, J. (2011). Nowcasting pandemic influenza A-H1N1 2009 hospitalizations in the Netherlands. *Eur. J. Epidemiol.*, **26** (3), 195 – 201.

Eilers, P.H.C. (2005). Unimodal smoothing. *J. Chemometrics*, **19** (5-7), 317 – 328.

Höhle, M. and an der Heiden, M. (2014). Bayesian nowcasting during the STEC O104:H4 outbreak in Germany. *Biometrics*, **70** (4), 993 – 1002.

A general statistical framework to study the diversification of species

Francisco Richter[1], Rampal Etienne[1], Ernst Wit[1]

[1] University of Groningen, The Netherlands

E-mail for correspondence: f.richter@rug.nl

Abstract: The mechanisms that control the diversification of species are poorly understood. Sophisticated diversification models have been developed, but they have been developed on a case-by-case basis and no general method to study the combined effect of ecological factors exists.

Such a general method has remained elusive for several reasons. Firstly, evolutionary processes have extremely complex dynamics. Secondly, decay and fossilization degrade crucial evidence useful for phylogenetic analyses. Thirdly, diversification processes have many potential explanatory variables, which increases the dimensionality of the models enormously.

To overcome these issues, we propose a general diversification model expressing the evolutionary species diversification dynamics as a combination of two generalized linear models. The fact that we typically only have data on currently existing species can be described as a missing data problem and we developed an MCEM-type algorithm for it.

We show that our method performs well for cases where an exact solution is available, and discuss potential future usage of our approach.

Keywords: Diversification models; GLM; EM.

1 Introduction

Biodiversity, the term used to describe the wide variety of species on Earth, is declining at enormous rates. To conserve biodiversity, we must understand the mechanisms how it comes about and how it is maintained, in assemblages of species, so-called ecological communities.

Figure 1 shows different sources of information regarding evolutionary processes. It is our aim to incorporate the sources of high-dimensional data under a unified statistical framework in order to overcome the main challenges that evolutionary biologists currently face. Particularly, the lack of information of extinct species and the huge complexity of current stochastic differential diversification models

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

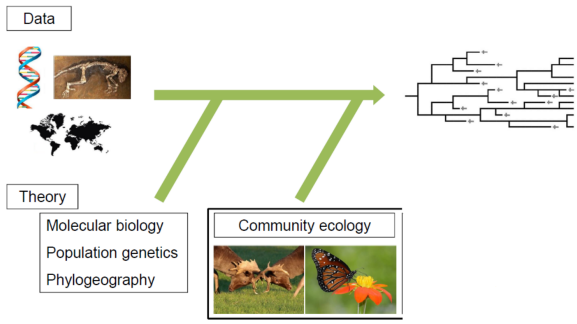


FIGURE 1. Inserting feedback with spatial ecological interactions into the phylogenetic data we are able to infer diversifications dynamics.

are bottlenecks for a proper inference on a general scenario. The framework we propose is shown to be an efficient and general method with the potential to provide practical solutions for a large number of open questions in evolutionary biology and ecology.

2 Methodology

A *phylogenetic time-tree* is defined as a graph (V, E) with the following properties:

1. It is undirected and acyclic, i.e, a tree,
2. All nodes have degrees 1,2 or 3, depending on the biological meaning of the node. The tree is binary,
3. It has a time-dimension.

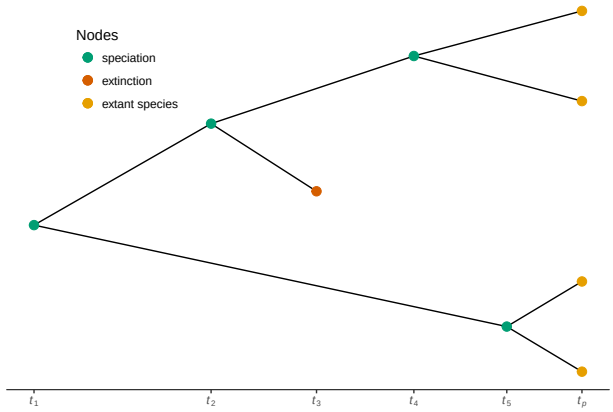


FIGURE 2. Phylogenetic time-tree.

If we consider the space of trees $\mathcal{T}$, extending the ideas of Gavryushkin (Gavryushkin et. al. 2016), we define the space of phylogenetic trees as

$$S = \{(rt(T), \tau) | T \in \mathcal{T}, \tau \in \mathbb{R}^{\mathbb{N}}\}$$

where $rt(T)$ is the ranked topology of the tree T and τ is the vector containing the waiting times between nodes (speciations or extinctions).

Assuming that the evolutionary mechanics follows a birth-death process (Nee et. al. 1994), the Markov nature of the dynamical system means that the likelihood is exactly the product of the conditional densities (Reynolds 1973), in other words, a multiplication of exponentially distributed waiting times and a multinomial event selection variables

$$L(\theta|Y) = \prod_{i=1}^{N} \sigma_i(\theta) e^{-\sigma_i(\theta)\tau_i} \frac{\rho_i(\theta)}{\sigma_i(\theta)} = \prod_{i=1}^{N} e^{-\sigma_i(\theta)\tau_i} \rho_i(\theta), \tag{1}$$

where $\sigma_i(\theta)$ and $\rho_i(\theta)$ are linear functions of the speciation and extinction rates $\lambda(\theta)$ and $\mu(\theta)$, which in turn are monotone functions of many potential explanatory (ecological) variables.

We have shown that for complete phylogenetic trees this approach is equivalent to current diversification models. However, assuming knowledge on the complete phylogenies is not realistic because information of extinct species is almost never available. Figure 2 shows the missing information on phylogenetic trees, where most of the times we only see the tree on the right with extant species only. To overcome that issue we implement an EM algorithm considering extinct species as missing data.

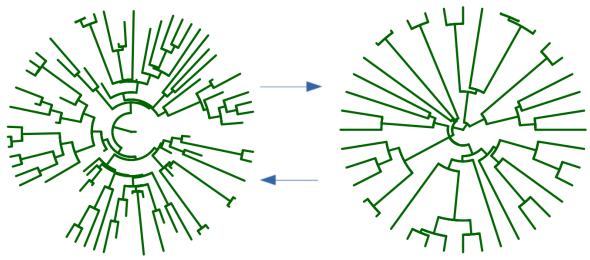


FIGURE 3. Phylogenetic trees where we can visualize the loss of information on reconstructed trees. At the left we have a tree with all extinct species whereas the right plot shows the same tree with only observable species.

Thus, if we define D as the observed phylogenetic tree and D^+ the extinct-extant species tree, we iteratively performs

E-step: Compute $Q(\theta|\theta^*) = E_{\theta^*}[logP(D^+|\theta)|D]$,

M-step: Choose θ to be the value of $\theta \in \Omega$ which maximizes $Q(\theta|\theta^*)$

However, the expectation

$$E_{\theta^*}[logP(D^+|\theta)|D] = \int_{\mathcal{X}(y)} \log f_{D^+}(x;\theta) f_{D^+|D}(x|D,\theta^*) dx$$

has not a close form due to the complexity of the time-tree space. Thus, an approximation via Monte-Carlo method is needed

$$E_{\theta^*}[logP(D^+|\theta)|D] \approx \frac{1}{N}\sum_{i=1}^{N} logP(D_i^+|\theta)$$

for sampling sets of complete-trees conditioned to the observed tree. However, the marginal distribution needed for sampling is in fact a complex combination of non-homogeneous Poisson processes and an approximated sampling method is needed. On figure 4 we can see the estimations for the Diversity-dependence (DD) model described on next section. Because the simulated sampled trees are approximations in the sampling space and not completely unbiased, we incorporate an importance sampling correction when needed.

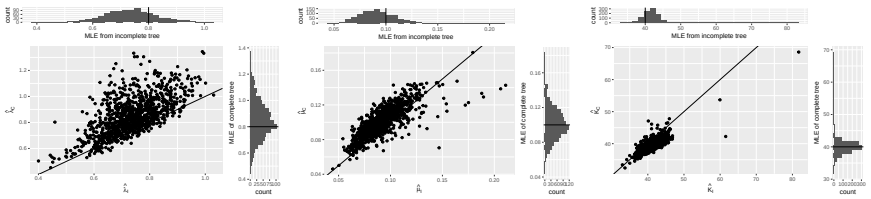


FIGURE 4. Parameter estimations for the DD model. Every point represents a pylogenetic tree, the y-axes shows the MLE of the complete tree and the x-axes shows the MLE of a set of reconstructed trees using the reconstruction algorithm . From the histograms we can see that the estimations are precise, however while the complete tree estimations are unbiased, the incomplete estimations are a bit biased. To correct that we implement include importance weights on the final estimations.

Application: Diversity-dependence model

The existence of upper limits to diversity has been discussed extensively in the last decades. Etienne et. al. (2012) propose a diversity-dependence model as a standard model for macro-evolutionary dynamics, with diversification rates defined by

$$\lambda_i = \lambda_0 - (\lambda_0 - \mu_0)\frac{n_i}{K}, \qquad \mu_i = \mu_0$$

where n corresponds to the number of species, and μ_0, λ_0, K are parameters of interest. μ_0 and λ_0 are initial extinction and speciation rates and K is the so-called carrying capacity. ,

If we define $\sigma_i = \sum_{j=1}^{N} \lambda_0 - (\lambda_0 - \mu_0)\frac{n_i}{K} + \mu_0 = n_i(\lambda_0 + \mu_0) - n_i^2\beta_0$ for $\beta_0 = \left(\frac{\lambda_0 - \mu_0}{K}\right)$, and $\rho_i = E_i(\lambda_0 - n_i\beta_0) + (1 - E_i)\mu_0$. n_i is defined as the number of species at time t_i, E is a binary vector containing the topology of the tree. For equation 1, the log-likelihood function of the phylogenetic tree is

$$l(\lambda_0, \mu_0, K|Y) = -\sum_{i}^{m} n_i t_i [\lambda_0(1 - \frac{n_i}{K}) + \mu_0(1 + \frac{n_i}{K})]$$
$$+ \sum_{i}^{m} \log\left(\lambda_0 I_{x_i}(1 - \frac{n_i}{K}) + \mu_0(I_{x_i}\frac{n_i}{K} + I_{1-x_i})\right)$$

We incorporate the Diversity-dependence model within our framework with satisfactory results, reproducing the evolutionary dynamics and estimations efficiently using a general approach, easily extensible to many more variables and biological scenarios. In figure 5 we can see the expected number of species versus time for two clades: Dendroica and Foraminifera. Here we can see the influence on the so-called carrying capacity on their evolutionary dynamics.

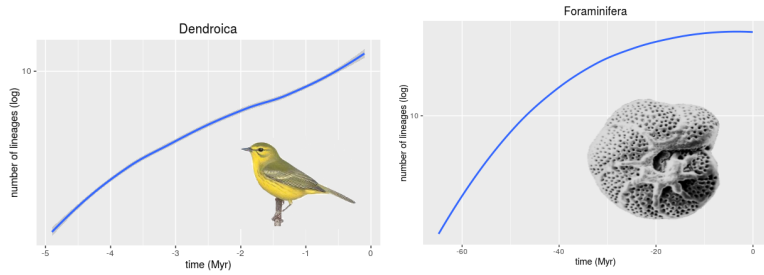


FIGURE 5. Expectation of lineages through time for the obtained parameters under the proposed framework. Here we can observe two clades with different responses to diversification. While Foraminifera shows an upper limit on diversification, Dendroica shows opposite behavior.

3 Conclusions and further work

In this manuscript we propose an alternative and general method to the usual diversification models. This framework is a methodological innovation with many potential applications, including been able to study the effects of climate, species interaction, protracted speciation on evolutionary dynamics, among many others. Further work will be the implementation of a differential geometric path finding method (Augugliaro et al., 2013) in order to deal efficiently with high-dimensional covariates.

Acknowledgments: This work is part of the research programme *Mathematics of planet Earth* with project number 657.014.005, which is financed by the Netherlands Organisation for Scientific Research (NWO).

References

Augugliaro, L., Mineo, A. M., and Wit, E. C. (2013). *Differential geometric least angle regression: a differential geometric approach to sparse generalized lin-*

ear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(3):471498.

Etienne, R. S. et. al. (2011).*Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record.* Proceedings of the Royal Society of London B: Biological Sciences, page rspb20111439.

Etienne, R. S. and Rosindell, J. (2012). *Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification.*, Systematic Biology, 61(2):204213.

Gavryushkin, Alex, and Alexei J. Drummond. (2016). *The space of ultrametric phylogenetic trees.*, Journal of theoretical biology 403 (2016): 197-208.

Nee, S., May, R. M., and Harvey, P. H. (1994). *The reconstructed evolutionary process.*, Philosophical Transactions of the Royal Society of London B: Biological Sciences, 344(1309):305311.

Reynolds, John F. (1973) *On estimating the parameters of a birth-death process.* Australian & New Zealand Journal of Statistics 15.1 (1973): 35-43.

Index

We are very grateful to the following organisations for sponsoring the 32nd IWSM 2017:

- **The Statistical Modelling Society**

- **Rijksuniversiteit Groningen (RUG)**

- **Johann Bernoulli Institute, RUG**

- **Netherlands Organisation for Scientific Research (NWO)**

  , (NWO sponsors our Open Access Session)

- **STAR clusterronde, NWO**

NWO Star Cluster

- **Koninklijke Nederlandse Akademie van Wetenschappen**

- **Toyota Motor Corporation**

to be continued

- **C.R. Rao Stichting, Groningen** C.R. Rao Stichting
- **Center for Language and Cognition, RUG** CLCG Center for Language and Cognition Groningen
- **Behavioural and Cognitive Neurosciences, RUG** bcn
- **R Studio** R Studio
- **Statistical Analysis Software (SAS)** sas THE POWER TO KNOW.
- **The welcome reception is sponsored by:**
  - **The Province of Groningen** provincie groningen
  - **The Muncipality of Groningen** Gemeente Groningen
  - **The Rijksuniversiteit Groningen (RUG)** university of groningen