Proceedings of the

32nd International Workshop on Statistical Modelling

Volume II

Groningen, Netherlands
3-7 July, 2017

Editors: Marco Grzegorczyk, Giacomo Ceoldo

Proceedings of the 32nd International Workshop on Statistical Modelling
Volume II,
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

---

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 to naturally account for anisotropy and non-stationarity. Space-varying covariate information is accounted for via a semiparametric framework. The 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.

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.

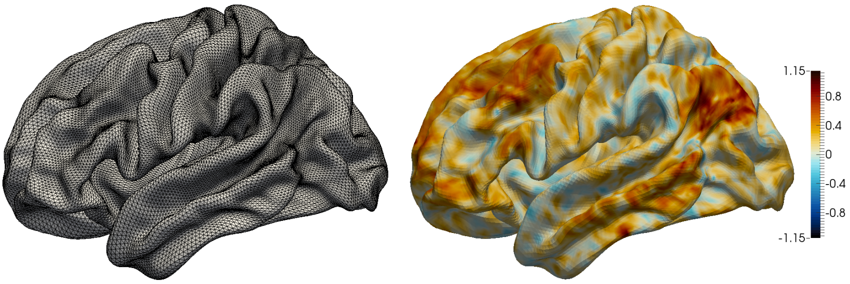


FIGURE 1. Study of high-dimensional neuroimaging signals (data available from The Human Connectome Project Consortium, www.humanconnectomeproject.org). Left: Triangulated surface approximating the left hemisphere of a cerebral cortex. Right: functional connectivity map obtained from fMRI signal. Figure 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 and irrelevant variables is like interpreting point estimates without an associated measure of uncertainty.

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 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 frequent occurrence of interval censoring, this interval censoring is often treated

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.

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

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

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

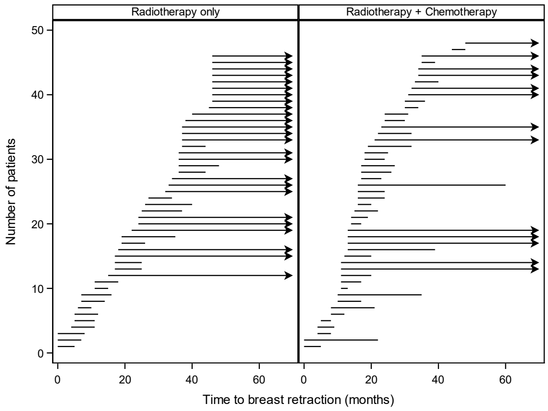


FIGURE 1. Breast cancer study. Observed intervals in months for time to breast retraction of early breast cancer patients per treatment group.

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 parametric 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 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

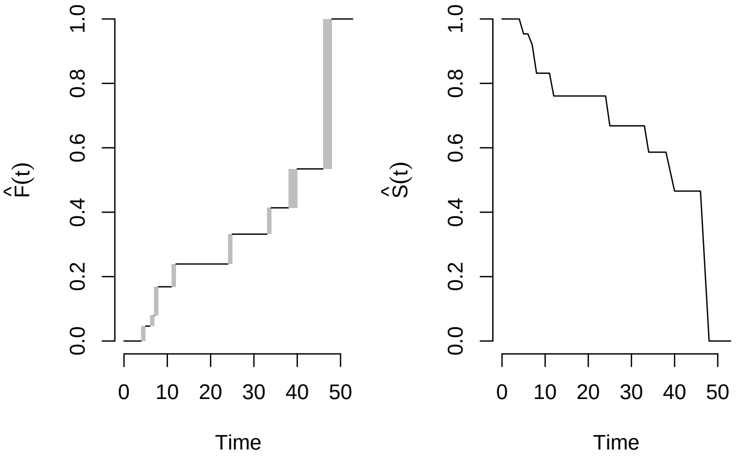


FIGURE 2. NPMLE of the cumulative distribution function (left panel) and NPMLE of the survival function with the additional assumption of a piecewise linear survival curve (right panel).

(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 $(l = 1, \dots, L)$. 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 $T_1, \dots, T_n$ are assumed to satisfy

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

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 $S(t \mid \boldsymbol{X}) = S_0\{\exp(-\boldsymbol{X}^\top \boldsymbol{\beta})\, t\}$.

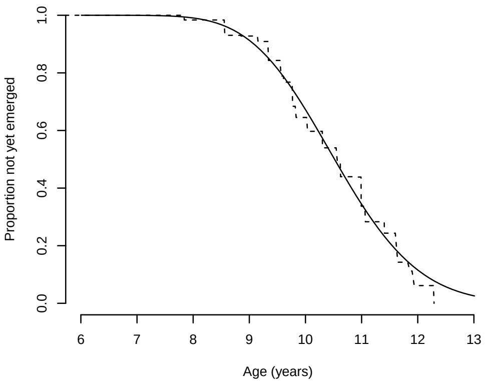


FIGURE 3. Signal Tandmobiel® study (boys). Estimated survival function compared to NPMLE (dashed line) for the time to emergence of tooth 44 estimated using the penalized Gaussian mixture with R package **smoothSurv**.

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 non-homogeneous 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

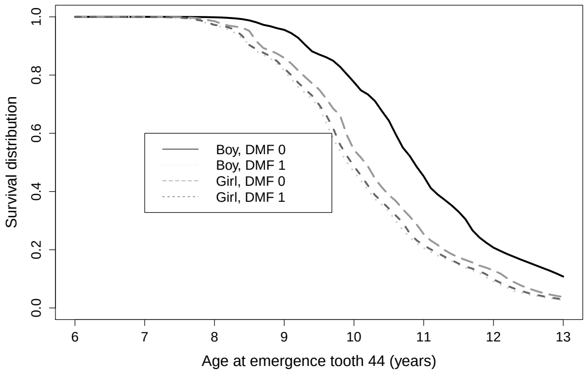


FIGURE 4. Signal Tandmobiel® study. Survival functions for emergence of permanent tooth 44 in gender by DMF groups based on a semiparametric Bayesian model with MDP priors, obtained with R function **DPsurvint**.

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.

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** (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.

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 $(T_{i1}, \dots, T_{in_i})^\top$ 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 $(i = 1, \dots, n;, l = 1, \dots, n_i)$.

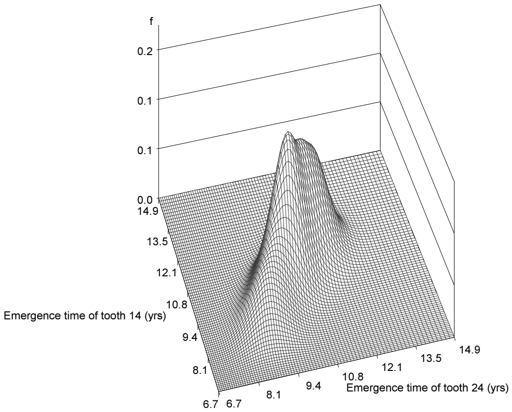


FIGURE 5. Signal Tandmobiel® study. Density of penalized normal mixture model for emergence of permanent teeth 14 and 24 obtained from SAS using macro **%smooth**.

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 DPpackage. *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 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.

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

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

$$\mathbf{F} = \mathbf{E}\mathbf{W}^T,$$

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 $\mathbf{D} = \mathbf{Y}\mathbf{X}^{-1}$. 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

$$\mathbf{P} = (\mathbf{W}\mathbf{C})^{+}\mathbf{T}\mathbf{W}\mathbf{C}$$

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 illustrates this incremental training regime. The left and center panels show the predictions based on $\mathbf{A}$ and $\mathbf{P}$ 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.

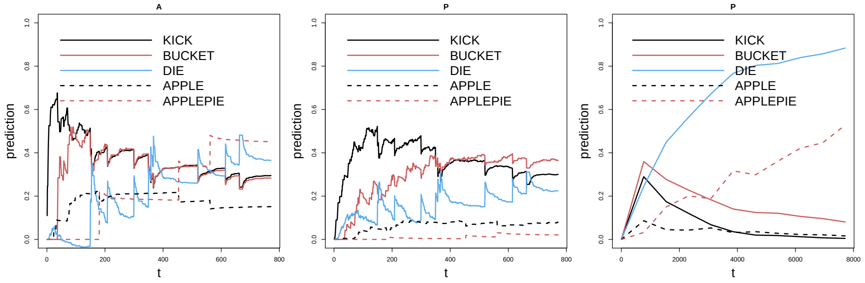


FIGURE 1. Prediction strengths for selected lexomes in the learning events of sentences 8 and 9 in Table 1, using incremented coupled Rescorla-Wagner and Widrow-Hoff. Left and center panels: frequencies as in the table; right panel: frequencies increased tenfold.

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, $\mathbf{W}$ (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.

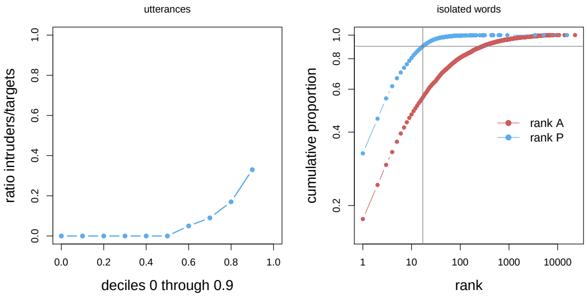


FIGURE 2. Left: Quantiles of the ratio of intruders (false positives) to targets (correct identifications), full utterances. Right: Rank and corresponding cumulative proportion based on **A** (red) and **P** (blue), isolated words.

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 **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.

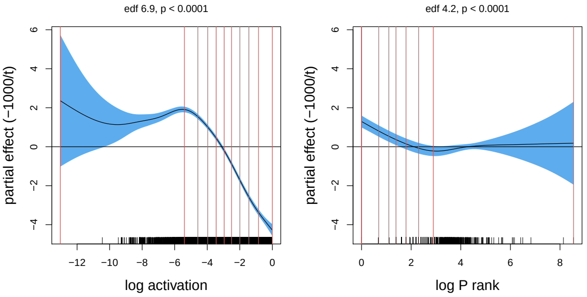


FIGURE 3. Partial effects in a GAM fitted to power-transformed ($-1000t-1$) reaction times. Left: log activation; Right: log prediction rank. Vertical lines denote deciles. The 90% decile of log prediction rank is at rank 18 (red lines indicate deciles). Regression analyses were carried out with GAMs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

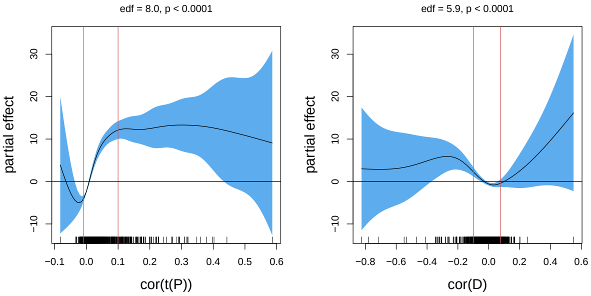


FIGURE 4. Partial effects of the correlations of row vectors of **P** (left) and column vectors of **D** as predictors of human similarity ratings for 2,369 word pairs. Red vertical lines indicate 5% and 95% percentiles. Regression analyses were carried out with GAMs (Wood, 2006).

The column vectors of **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 $\mathbf{W}$ and $\mathbf{D}$ can be very large, most of the weights are close to zero. E.g., for $\mathbf{W}$, 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 $\mathbf{W}$ 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 backpropagation networks in that there is direct error injection twice, once for $\mathbf{W}$ using the Rescorla-Wagner equations, and once for $\mathbf{D}$, 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 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

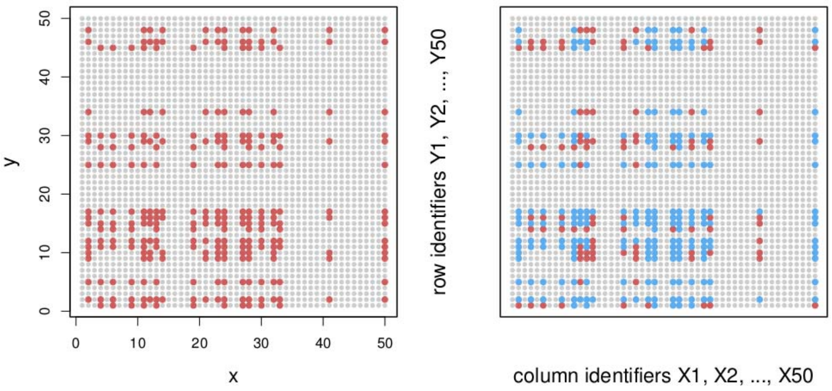


FIGURE 5. A non-linearly separable classification problem with a majority class in gray (2240) and a minority class in red (260). Left: data points in a Cartesian grid ($x = 1, 2, \ldots, 50; y = 1, 2, \ldots, 50$). Right: rerepresentation with row and column identifiers as cues for a Rescorla-Wagner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

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

Predict NBA 2016-2017 Regular Season MVP Winner

Mason Chen[1]

[1] Mason Chen Consulting, San Jose, USA

E-mail for correspondence: mason.chen.training@gmail.com

Abstract: This project is to build a statistical model to predict who will win the 2017 NBA Most Valuable Player (MVP) Award. The “MVP Index” has been derived from combining each player’s Z statistics with equal weight as a ”uniform” model. The team has further derived the ”weighted” model by adding the weight factor which was calculated based on the dispersion/separation between the top two MVP winners and the remaining players not in top five. Based on the Power Model, team can improve the Accuracy Index to 70% at power=3. Our final model would suggest that Westbrook and Harden shall own the Co-MVP Awards.

Keywords: Descriptive Statistics; Z Transformation; Power Model; Discriminant Cluster; Data Mining.

1 Problem Statement and Project Objective

In major professional sports, the coach and team management are looking for ways to win more games. Sports statistical modeling analytics (A. Maymin, 2012; M. Oh, 2015) is becoming a critical approach to uncover the winning patterns hidden in sports data collected during each game played. The objective of this paper is to build a statistical model to predict the NBA 2016-2017 regular season most valuable player (MVP).

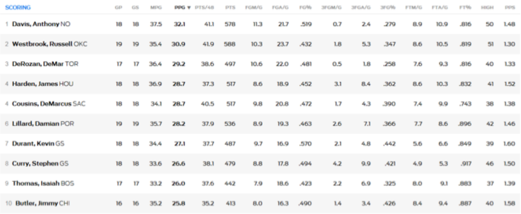
2 Data Collection, MVP Index and Model Accuracy Index

Section 2 has three subsections: (1) Data Collection, (2) Derive MVP Index to judge player performance, (3) Derive Model Accuracy Index to assess model performance.

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

2.1 Data Collection

In order to predict the 2016-2017 season MVP, team has collected three raw data of the 2015-2016 season from the public Sports domain: (1) top 50 player statistics in Figure 1 (NBA Player Statistics, 2015-2016), (2) team win%, and (3) historical MVP winners.

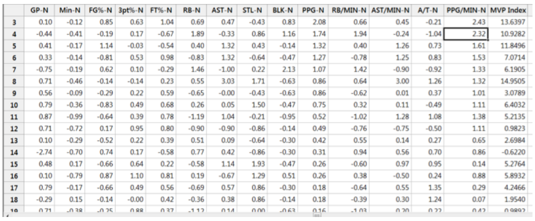


SCORING	GP	GS	MPG	PPG ▼	PTS/48	PTS	FGM/G	FGA/G	FG%	3FGM/G	3FGA/G	3FG%	FTM/G	FTA/G	FT%	HIGH	PPS
1 Davis, Anthony NO	18	18	37.5	32.1	41.1	578	11.3	21.7	.519	0.7	2.4	.279	8.9	10.9	.816	50	1.48
2 Westbrook, Russell OKC	19	19	35.4	30.9	41.9	588	10.3	23.7	.432	1.8	5.3	.347	8.6	10.5	.819	51	1.30
3 DeRozan, DeMar TOR	17	17	36.4	29.2	38.6	497	10.6	22.0	.481	0.5	1.8	.258	7.6	9.3	.816	40	1.33
4 Harden, James HOU	18	18	36.9	28.7	37.3	517	8.6	18.9	.452	3.1	8.4	.362	8.6	10.3	.832	41	1.52
4 Cousins, DeMarcus SAC	18	18	34.1	28.7	40.5	517	9.8	20.8	.472	1.7	4.3	.390	7.4	9.9	.743	38	1.38
6 Lillard, Damian POR	19	19	35.7	28.2	37.9	536	8.9	19.3	.463	2.6	7.1	.366	7.7	8.6	.896	42	1.46
7 Durant, Kevin GS	18	18	34.4	27.1	37.7	487	9.7	16.9	.570	2.1	4.8	.442	5.6	6.6	.849	39	1.60
8 Curry, Stephen GS	18	18	33.6	26.6	38.1	479	8.8	17.8	.494	4.2	9.9	.421	4.9	5.3	.917	46	1.50
9 Thomas, Isaiah BOS	17	17	33.2	26.0	37.6	442	7.9	18.6	.423	2.2	6.9	.325	8.0	9.1	.883	37	1.39
10 Butler, Jimmy CHI	16	16	35.2	25.8	35.2	413	8.0	16.3	.490	1.4	3.4	.426	8.4	9.4	.887	40	1.58

FIGURE 1. Top 50 Player Statistics.

2.2 Derive MVP Index

Before building the model, the player statistics have been standardized to the Z scale in order to remove any mean and standard deviation effect. This Z transformation can eliminate any statistics bias or domination. The Z scale will also analyze each player's performance as compared to the other top 50 NBA players in the same 2015-2016 season. The "MVP Index" has been derived from summing each player's Z statistics as shown in Figure 2.



	GP-N	Min-N	FG%-N	3pt%-N	FT%-N	RB-N	AST-N	STL-N	BLK-N	PPG-N	RB/MIN-N	AST/MIN-N	A/T-N	PPG/MIN-N	MVP Index
3	0.10	-0.12	0.85	0.63	1.04	0.69	0.47	-0.43	0.83	2.08	0.66	0.45	-0.21	2.43	13.6397
4	-0.44	-0.41	-0.19	0.17	-0.67	1.89	-0.33	0.86	1.16	1.74	1.94	-0.24	-1.04	2.32	10.9282
5	0.41	-0.17	1.14	-0.03	-0.54	0.40	1.32	0.43	-0.14	1.32	0.40	1.26	0.73	1.61	11.8496
6	0.33	-0.14	-0.81	0.53	0.98	-0.83	1.32	-0.64	-0.47	1.27	-0.78	1.25	0.83	1.53	7.0714
7	-0.75	-0.19	0.62	0.10	-0.29	1.46	-1.00	0.22	2.13	1.07	1.42	-0.90	-0.92	1.33	6.1905
8	0.71	-0.46	-0.14	-0.14	0.23	0.55	3.03	1.71	-0.63	0.86	0.64	3.00	1.26	1.32	14.9505
9	0.56	-0.09	-0.29	0.22	0.59	-0.65	-0.00	-0.43	-0.63	0.86	-0.62	0.01	0.37	1.01	3.0789
10	0.79	-0.36	-0.83	0.49	0.68	0.26	0.05	1.50	-0.47	0.75	0.32	0.11	-0.49	1.11	6.4032
11	0.87	-0.99	-0.64	0.39	0.78	-1.19	1.04	-0.21	-0.95	0.52	-1.02	1.28	1.08	1.38	5.2135
12	0.71	-0.72	0.17	0.95	0.80	-0.90	-0.90	-0.86	-0.14	0.49	-0.76	-0.75	-0.50	1.11	0.9823
13	0.10	-0.29	-0.52	0.22	0.39	0.51	0.09	-0.64	-0.30	0.42	0.55	0.14	0.27	0.65	2.6984
14	-2.74	-0.70	0.74	0.17	-0.58	0.77	0.42	-0.86	-0.30	0.31	0.94	0.56	0.70	0.86	-0.6220
15	0.48	0.17	-0.66	0.64	0.22	-0.58	1.14	1.93	-0.47	0.26	-0.60	0.97	0.95	0.14	5.2764
16	0.10	-0.79	0.87	1.10	0.81	0.19	-0.67	1.29	0.51	0.26	0.38	-0.50	0.24	0.88	5.8932
17	0.79	-0.17	-0.66	0.49	0.56	-0.69	0.57	0.86	-0.30	0.18	-0.64	0.55	1.35	0.29	4.2466
18	-0.29	0.15	-0.14	-0.00	0.42	-0.36	0.38	0.86	-0.14	0.18	-0.39	0.30	1.24	0.07	1.9540
19	0.71	-0.38	-0.25	0.88	0.37	-1.12	0.14	0.00	-0.63	0.16	-1.03	0.20	0.22	0.42	0.9892

FIGURE 2. Z Transformation and MVP Index.

2.3 Derive Model Accuracy Index

To evaluate the model accuracy, the author has derived "Accuracy Index" of predicting the top five MVP players. MVP index of each top player was sorted and compared to the actual top 5 MVP winners in that season.

3 Build Statistical Modeling Algorithms

Based on the above analysis, author has derived and compared: (1) Uniform, (2) Weighted, (3) Power, and (4) Discriminant Clustering Model.

3.1 Uniform Model

In the Uniform model, the “MVP Index” has been derived from combining each player’s Z statistics with equal weight. The “Uniform” model can predict the top five MVP winners at 47% accuracy.

3.2 Weighted Model

The “Weighted” model added the weight factor which reflects player Z scale statistics categories are more important. To determine the weight level, the descriptive dispersion statistics between the top two MVP winners and the remaining players not in top 5 was calculated for each Z scale statistics. The “Weighted” model has further improved the Accuracy Index from 47% to 52% shown in Figure 3. The weighted MVP index formula is shown below:

$$Weighted\, MVP\, Index = 1.03 *' GP - N' + 0.73 *' Min - N' + 1.71 *' FG\% - N' + 0.61 *' 3pt\% - N' + 0.67 *' FT\% - N' + 1.17 *' RB - N' + 1.66 *' AST - N' + 1.5 *' STL - N' + 1.16 *' BLK - N' + 2.7 *' PPG - N' + 0.89 *' RB/MIN - N' + 1.3 *' AST/MIN - N' + 0.96 *' A/T - N' + 3.13 *' PPG/MIN - N'$$

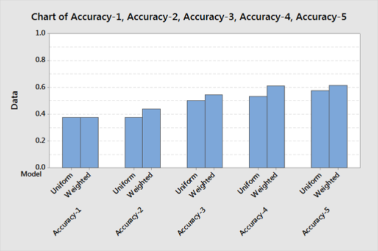


FIGURE 3. Model Accuracy: Uniform model vs. Weighted model.

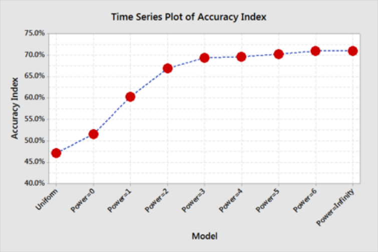


FIGURE 4. Model Accuracy: compare all three models.

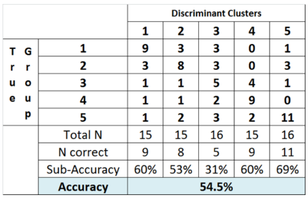
3.3 Power Model

To further optimize the accuracy, author has added the ”Team Winning” factor. Most historical MVP winners were from the teams with best or better regular season records. Author has assessed the team winning factor based on the

“Power” model from power= 0 (equivalent to the Weighted Model), 1 to 6 to power= infinity (pick MVP from the best Team). The weighted MVP Index will be multiplied by the power of the team winning% in the Power model. Based on the Power Model, team can improve the Accuracy Index to 70% at Power=3. The Power model has indicated the importance of team winning % on MVP selection process. There is little benefit but more over-fit risk to further increase the power level beyond 3 shown in Figure 4.

3.4 Data Mining Discriminant Model

Data Mining Discriminant analysis (J. Garcia, 2013) has been utilized to identify basketball performance indicators in regular season and playoff games. The Discriminant model accuracy is at 54.5% not better than Power model shown in Figure 5.



		Discriminant Clusters				
		1	2	3	4	5
True Group	1	9	3	3	0	1
	2	3	8	3	0	3
	3	1	1	5	4	1
	4	1	1	2	9	0
	5	1	2	3	2	11
	Total N	15	15	16	15	16
	N correct	9	8	5	9	11
	Sub-Accuracy	60%	53%	31%	60%	69%
	Accuracy	**54.5%**				

FIGURE 5. Data Mining Discriminant Model.

4 Conclusions

Power model has shown the better prediction capability which has indicated the importance of team winning performance on the MVP selection process. Power=3 model will be utilized to predict 2017 MVP winner on the first day of each month starting from Dec 2016,end in April 2017. The final model prediction will suggest Co-MVP Awards, and presented in IWSM.

References

Maymin, A., Maymin, P., and Shen, E. (2012). NBA chemistry: Positive and negative synergies in basketball. *Proceedings MIT Sloan Sports Analytics Conference.*

Oh, M., Keshri, S., and Iyengar, G. (2015). Graphical model for basketball match simulation. *Proceedings MIT Sloan Sports Analytics Conference.*

Garcia, J., Ibanez, S., Santos, Leite, N., and Sampaio, J. (2015). *Journal of Human Kinetics*, **36**, 2013, 161 – 168 Section II Sports Training.

Detect Exam Cheating Pattern on Multiple-Choice Exams

Mason Chen[1], Charles Chen[2]

[1] Mason Chen Consulting, United States
[2] Morrill Learning Center, United States

E-mail for correspondence: `mason.chen.training@gmail.com`

Abstract: This project will demonstrate on how to apply Data Mining Algorithm on detecting the assessment exam cheating pattern in Schools. In order to detect the cheating pattern, JMP multivariate statistics was used to determine whether there was any association pattern among the students from the same exam table. Hierarchical Clustering and Dendrogram Tree were used to identify the grouping affinity behavior related to exam cheating pattern. This cheating case study can also be used to study the students answering patterns on any particular subject to help instructors on designing their future curriculum based on the Data Mining Patterns.

Keywords: Data Mining; Heat Map; Clustering Analysis; Dendrogram Tree; Principle Component Analysis.

1 Problem Statement and Project Objective

For each instructor, designing an effective assessment exam is a critical job (Kevin Y. etc.). Its much easier to grade any exam with questions of multiple choices than with comprehensive questions. In this case study, there were 75 students sat in 25 different small tables (3 students per table) in a very limited space. Instructor has modified the original exam into three different orders (versions A, B, C). Three students from the same table will take different versions. However, students were still smart enough to synchronize the question sequence quickly. The objective of this paper is to find a data mining algorithm in order to detect any cheating pattern from the same table.

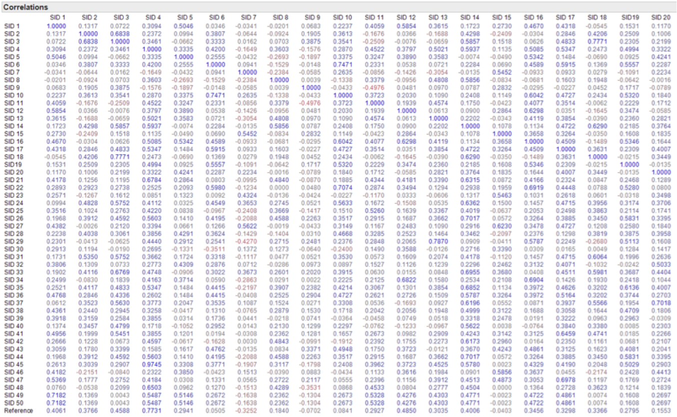
This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

2 Data Collection and Multivariate Correlation Analysis

The raw data includes each students ID, Exam Version, Answers, and Table Number. In order to reduce the computing time, the bottom 25 scorers have been excluded from the analysis in this paper. There is unlikely that we can find any cheating pattern from these bottom scorers.

2.1 Multivariate Correlation Analysis

JMP 12 Multivariate Correlation Analysis was used to study any correlation between Top 50 Students in Figure 1. From the Multivariate Correlation Analysis, there are $Combination(50, 2) = 1,225$ correlation coefficients between any two students. This massive correlation table is a good start to visualize any correlation pattern, but not effectively to draw any systematic pattern conclusion.



Correlations

	SID 1	SID 2	SID 3	SID 4	SID 5	SID 6	SID 7	SID 8	SID 9	SID 10	SID 11	SID 12	SID 13	SID 14	SID 15	SID 16	SID 17	SID 18	SID19	SID 20
SID 1	1.0000	0.1317	0.0722	0.3094	0.5046	0.0346	-0.0341	-0.0201	0.0683	0.2237	0.4059	0.5854	0.3615	0.1723	0.2730	0.4670	0.4318	-0.0545	0.1531	0.1170
SID 2	0.1317	1.0000	0.6838	0.2372	0.0994	0.3807	-0.0644	-0.0924	0.1905	0.3613	-0.1676	0.0366	-0.1688	0.4298	-0.2409	-0.0304	0.2846	0.4206	0.2509	0.1006
SID 3	0.0722	0.6838	1.0000	0.3461	-0.0662	0.3333	0.0162	0.0703	0.3875	0.3541	-0.2509	-0.0076	-0.0659	0.5857	0.1518	0.0626	0.4833	0.7771	0.2305	0.2199
SID 4	0.3094	0.2372	0.3461	1.0000	0.3335	0.4200	-0.1649	0.3603	-0.1576	0.2870	0.4522	0.3797	0.5021	0.5937	0.1135	0.5085	0.5347	0.2473	0.4994	0.3322
SID 5	0.5046	0.0994	-0.0662	0.3335	1.0000	0.2555	-0.0432	-0.2693	-0.1897	0.3375	0.3247	0.3890	0.3583	-0.0074	-0.0490	0.5342	0.1484	-0.0690	0.0925	0.4241
SID 6	0.0346	0.3807	0.3333	0.4200	0.2555	1.0000	0.0941	-0.1529	-0.0148	0.7471	0.2331	0.0538	0.0721	0.2284	0.0690	0.4589	0.5915	0.1369	0.5557	0.2287
SID 7	-0.0341	-0.0644	0.0162	-0.1649	-0.0432	0.0941	1.0000	-0.2384	-0.0585	0.2635	-0.0856	-0.1426	-0.3054	-0.0135	0.5452	-0.0933	0.0933	0.0279	-0.1091	0.2234
SID 8	-0.0201	-0.0924	0.0703	0.3603	-0.2693	-0.1529	-0.2384	1.0000	0.0039	-0.1338	0.3379	-0.0956	0.4808	0.5856	-0.0834	-0.0681	0.1603	0.1948	-0.0642	-0.0016
SID 9	0.0683	0.1905	0.3875	-0.1576	-0.1897	-0.0148	-0.0585	0.0039	1.0000	-0.0433	-0.4976	0.0481	0.0970	0.0787	0.2832	-0.0295	-0.0227	0.0452	0.1717	-0.0789
SID 10	0.2237	0.3613	0.3541	0.2870	0.3375	0.7471	0.2635	-0.1338	-0.0433	1.0000	0.3723	0.2030	0.1090	0.2408	0.1149	0.6042	0.4727	0.2434	0.5320	0.1840
SID 11	0.4059	-0.1676	-0.2509	0.4522	0.3247	0.2331	-0.0856	0.3379	-0.4976	0.3723	1.0000	0.1939	0.4574	0.1750	-0.0423	0.4077	0.3514	-0.0062	0.2229	0.1712
SID 12	0.5854	0.0366	-0.0076	0.3797	0.3890	0.0538	-0.1426	-0.0956	0.0481	0.2030	0.1939	1.0000	0.0613	0.0900	0.2864	0.6298	0.0351	-0.1645	0.3474	-0.0585
SID 13	0.3615	-0.1688	-0.0659	0.5021	0.3583	0.0721	-0.3054	0.4808	0.0970	0.1090	0.4574	0.0613	1.0000	0.2202	-0.0343	0.4119	0.3854	-0.0390	0.2360	0.2821
SID 14	0.1723	0.4298	0.5857	0.5937	-0.0074	0.2284	-0.0135	0.5856	0.0787	0.2408	0.1750	0.0900	0.2202	1.0000	0.1078	0.1134	0.4722	0.6290	0.2185	0.3764
SID 15	0.2730	-0.2409	0.1518	0.1135	-0.0490	0.0690	0.5452	-0.0834	0.2832	0.1149	-0.0423	0.2864	-0.0343	0.1078	1.0000	0.3658	0.3264	-0.0350	0.1608	0.1835
SID 16	0.4670	-0.0304	0.0626	0.5085	0.5342	0.4589	-0.0933	-0.0681	-0.0295	0.6042	0.4077	0.6298	0.4119	0.1134	0.3658	1.0000	0.4509	-0.1489	0.5346	0.1644
SID 17	0.4318	0.2846	0.4833	0.5347	0.1484	0.5915	0.0933	0.1603	-0.0227	0.4727	0.3514	0.0351	0.3854	0.4722	0.3264	0.4509	1.0000	0.3631	0.2309	0.4007
SID 18	-0.0545	0.4206	0.7771	0.2473	-0.0690	0.1369	0.0279	0.1948	0.0452	0.2434	-0.0062	-0.1645	-0.0390	0.6290	-0.0350	-0.1489	0.3631	1.0000	-0.0215	0.3449
SID19	0.1531	0.2509	0.2305	0.4994	0.0925	0.5557	-0.1091	-0.0642	0.1717	0.5320	0.2229	0.3474	0.2360	0.2185	0.1608	0.5346	0.2309	-0.0215	1.0000	-0.0135
SID 20	0.1170	0.1006	0.2199	0.3322	0.4241	0.2287	0.2234	-0.0016	-0.0789	0.1840	0.1712	-0.0585	0.2821	0.3764	0.1835	0.1644	0.4007	0.3449	-0.0135	1.0000
SID 21	0.4178	0.1256	0.1195	0.6784	0.2864	0.0803	-0.0995	0.4840	-0.0870	0.1885	0.4344	0.4181	0.3390	0.6315	0.0872	0.4166	0.2324	0.0847	0.2468	0.1289
SID 22	0.2893	0.2923	0.2738	0.2525	0.2093	0.5980	-0.1234	0.0000	0.0480	0.7074	0.2874	0.3494	0.1294	0.2938	0.1959	0.6919	0.4448	0.0788	0.5280	0.0800
SID 23	0.2571	-0.1267	0.1612	0.0851	0.1323	0.0092	0.4324	-0.0136	-0.0424	-0.0227	-0.1170	0.0333	-0.0606	0.1317	0.5463	0.1031	0.2618	0.0601	-0.0318	0.3498
SID 24	0.0994	0.4828	0.5752	0.4112	0.0325	0.4549	0.3653	0.2745	0.0521	0.5633	0.1672	-0.1508	0.0535	0.6362	0.1500	0.1457	0.4715	0.3956	0.3174	0.3706
SID 25	0.3516	0.1024	0.2763	0.4220	0.0838	-0.0967	-0.2408	0.3669	-0.1417	0.1510	0.5260	0.1639	0.3367	0.4019	-0.0637	0.2053	0.2498	0.3863	0.2114	0.1741
SID 26	0.1968	0.3912	0.4592	0.5603	0.1410	0.4195	-0.2088	0.4588	0.2263	0.3517	0.2915	0.1687	0.3662	0.7017	0.0572	0.3264	0.3885	0.3450	0.5831	0.3395
SID 27	0.4382	-0.0026	0.2120	0.3394	0.0661	0.1266	0.5622	-0.0019	-0.0433	0.3149	0.1167	0.2483	0.1090	0.2916	0.6230	0.3478	0.4727	0.1208	0.2580	0.1840
SID 28	0.2238	0.4038	0.3061	0.3856	0.4291	0.3624	-0.1429	-0.1404	0.0310	0.4668	0.3285	0.2523	0.1464	0.3462	-0.2097	0.2376	0.1298	0.3819	0.3875	0.3958
SID 29	0.2301	-0.0413	-0.0625	0.4440	0.2912	0.2541	-0.4270	0.2715	0.2481	0.2376	0.2848	0.2065	0.7870	0.0909	-0.0411	0.5787	0.2249	-0.2680	0.5113	0.1608
SID 30	0.2913	0.1194	-0.0190	0.2695	-0.1331	-0.3511	0.1372	0.1273	-0.0640	-0.2400	0.1490	0.3588	-0.0126	0.2716	0.3390	0.0309	0.0165	0.0049	0.1284	0.1417
SID 31	0.1731	0.5350	0.5752	0.3662	0.1724	0.3318	-0.1117	0.0477	0.0521	0.3530	0.0573	0.1609	0.2074	0.4178	-0.1120	0.1457	0.4715	0.6064	0.1996	0.2636
SID 32	0.3806	0.1309	0.0733	0.2773	0.4309	0.2876	0.0712	-0.0286	0.0973	0.0897	0.1527	0.1126	0.1239	0.2296	0.2462	0.3132	0.4071	-0.1032	-0.0242	0.5033
SID 33	0.1902	0.4116	0.6769	0.4748	-0.0906	0.3022	0.3673	0.2601	0.2020	0.3915	0.0630	0.0155	0.0848	0.6955	0.3680	0.0408	0.4511	0.5981	0.3687	0.4404
SID 34	0.2499	-0.0830	0.1839	0.4163	0.3714	0.0590	-0.2863	0.0291	0.0022	0.2225	0.2125	0.6822	0.1580	0.2534	0.2108	0.6904	0.1426	0.1930	0.2418	0.1044
SID 35	0.2521	0.4117	0.4833	0.5347	0.1484	0.4415	-0.2197	0.3907	0.2382	0.4214	0.3067	0.1301	0.3854	0.6852	0.1134	0.3972	0.4626	0.3202	0.6136	0.4007
SID 36	0.4768	0.2846	0.4336	0.2602	0.1484	0.4415	-0.0408	0.2525	0.2904	0.4727	0.2621	0.2726	0.1509	0.5787	0.3264	0.3972	0.5164	0.3202	0.3744	0.2703
SID 37	0.0612	0.3523	0.5630	0.3773	0.2047	0.3535	0.1087	0.1524	0.0271	0.3308	0.0536	-0.1693	0.0927	0.6196	0.0552	0.0871	0.3937	0.5566	0.1954	0.7018
SID 38	0.4361	0.2440	0.2945	0.3258	-0.0417	0.1310	-0.0765	0.2879	0.1530	0.1718	0.2042	0.2056	0.1948	0.4999	0.3122	0.1688	0.3058	0.1644	0.4709	0.1806
SID 39	0.3918	0.3159	0.2584	0.3855	0.0314	0.1736	0.0441	-0.0218	0.0741	-0.0364	-0.0458	0.0749	0.0518	0.3318	0.2478	0.0191	0.3222	0.0963	0.2963	-0.0309
SID 40	0.1374	0.3457	0.4799	0.1718	-0.1052	0.2952	0.0143	0.2130	0.1299	0.2297	-0.0762	-0.1233	-0.0967	0.5622	0.0038	-0.0764	0.3840	0.3380	0.0085	0.2303
SID 41	0.4956	0.1999	0.5451	0.3855	0.1201	0.0194	-0.0308	0.2362	0.1281	0.1657	0.2673	0.0982	0.2909	0.4243	0.3142	0.3125	0.6459	0.4741	0.0185	0.2266
SID 42	0.2666	0.1228	0.0673	0.4597	-0.0617	-0.1628	0.0030	0.4843	-0.0991	-0.1912	0.2392	0.1755	0.2273	0.6173	0.2960	0.0164	0.2350	0.1161	0.0681	0.2107
SID 43	0.3059	0.1780	0.3399	0.1585	0.1617	0.4762	-0.0135	0.0834	0.3371	0.4948	0.1750	0.3723	-0.0121	0.3670	0.4243	0.4861	0.3125	0.1623	0.4081	0.2041
SID 44	0.1968	0.3912	0.4592	0.5603	0.1410	0.4195	-0.2088	0.4588	0.2263	0.3517	0.2915	0.1687	0.3662	0.7017	0.0572	0.3264	0.3885	0.3450	0.5831	0.3395
SID 45	0.2613	0.3039	0.2907	0.9745	0.3308	0.3771	-0.1907	0.3117	-0.1798	0.2408	0.3962	0.3723	0.4525	0.5780	0.0023	0.4329	0.4190	0.2048	0.5029	0.2903
SID 46	0.4182	-0.2151	-0.0840	0.2322	0.3850	-0.0423	0.1513	-0.0390	0.0883	-0.0434	0.1133	0.3616	0.1984	0.0901	0.5856	0.3637	0.0455	-0.2174	0.2428	0.4413
SID 47	0.5369	0.1777	0.2752	0.4184	0.0308	0.1331	0.0565	0.2722	0.2117	0.0555	0.2396	0.1156	0.3912	0.4513	0.4873	0.3053	0.6978	0.1197	0.1769	0.2724
SID 48	0.0760	-0.0538	0.2099	0.6503	0.0962	0.1270	-0.1513	0.4289	-0.3531	0.0868	0.4533	0.0804	0.2777	0.4504	0.0000	0.1364	0.2728	0.3623	0.1214	0.1839
SID 49	0.7182	0.1369	0.0043	0.5487	0.5146	0.2672	-0.1638	0.2362	-0.1304	0.2673	0.5328	0.4276	0.4303	0.4771	-0.0023	0.4722	0.4861	0.0074	0.1608	0.2697
SID 50	0.7182	0.1369	0.0043	0.5487	0.5146	0.2672	-0.1638	0.2362	-0.1304	0.2673	0.5328	0.4276	0.4303	0.4771	-0.0023	0.4722	0.4861	0.0074	0.1608	0.2697
Reference	0.4061	0.3766	0.4588	0.7731	0.2941	0.0505	-0.3252	0.1840	-0.0702	0.0841	0.2927	0.4850	0.3035	0.4006	-0.0403	0.3456	0.3298	0.3366	0.2795	0.1553

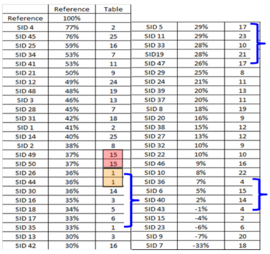
FIGURE 1. Multivariate Correlation Analysis.

2.2 Sort Students Score

To further detect any cheating pattern, authors have sorted students score (reference column) from top to bottom and list Table information in Figure 2. Its clear that, on some tables, some students (No.1, No.15, No.17, and No.4) have observed same score or similar score.

3 Build Data Mining Algorithms

Authors are looking for Data Mining Algorithms including JMP 12 Hierarchical Clustering Dendrogram, Heat Map, and Principle Component Analysis to detect the cheating pattern with high confidence.



	Reference	Table			
Reference	100%				
SID 4	77%	2	SID 5	29%	17
SID 45	76%	25	SID 11	29%	23
SID 25	59%	16	SID 33	28%	10
SID 34	53%	7	SID19	28%	21
SID 41	53%	11	SID 47	26%	17
SID 21	50%	9	SID 29	25%	8
SID 12	49%	24	SID 24	21%	11
SID 48	48%	19	SID 39	20%	13
SID 3	46%	13	SID 37	20%	11
SID 28	45%	7	SID 8	18%	19
SID 31	42%	18	SID 20	16%	9
SID 1	41%	2	SID 38	15%	12
SID 14	40%	25	SID 27	13%	12
SID 2	38%	8	SID 32	10%	9
SID 49	37%	15	SID 22	10%	10
SID 50	37%	15	SID 46	9%	16
SID 26	36%	1	SID 10	8%	22
SID 44	36%	1	SID 36	7%	4
SID 30	36%	14	SID 6	5%	15
SID 16	35%	3	SID 40	2%	14
SID 18	34%	5	SID 43	-1%	4
SID 17	33%	6	SID 15	-4%	2
SID 35	33%	1	SID 23	-6%	6
SID 13	30%	3	SID 9	-7%	20
SID 42	30%	16	SID 7	-33%	18

FIGURE 2. Sort Score and Exam Table Information.

3.1 Hierarchical Clustering Dendrogram Analysis

Hierarchical Clustering analysis (Michael A.) is utilized to analyze the cheating pattern. In Figure 3 Dendrogram Tree, JMP software will calculate the answer affinity among all 1,225 pairs and group the first pair at the strongest affinity (similar pattern). JMP software then find the next affinity pair until all done as shown in Dendrogram. In Figure 4, clustering history has shown 4 out of 5 pairings sat on the same exam table. Hierarchical clustering analysis can defend the cheating pattern at high confidence based on the low pairing distance separated from the remaining pairing distance shown in Figure 4. Students from Exam Tables 1, 4, and 15 have been identified having cheating pattern in the exam.

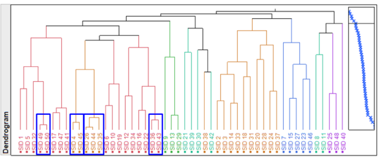
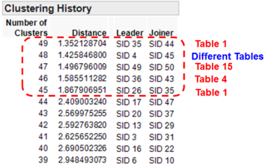


FIGURE 3. Hierarchical Dendrogram Tree.



Clustering History

Number of Clusters	Distance	Leader	Joiner	
49	1.352128704	SID 35	SID 44	Table 1
48	1.425846800	SID 4	SID 45	Different Tables
47	1.496796009	SID 49	SID 50	Table 15
46	1.585511282	SID 36	SID 43	Table 4
45	1.867906951	SID 26	SID 35	Table 1
44	2.409003240	SID 17	SID 47	
43	2.569975255	SID 20	SID 37	
42	2.592763820	SID 13	SID 29	
41	2.625652250	SID 3	SID 31	
40	2.690502326	SID 16	SID 22	
39	2.948493073	SID 6	SID 10	

FIGURE 4. Clustering History.

3.2 Heat Map Analysis

Instead of visualizing the cheating in Figure 3 Correlation Table, Heat Map in Figure 5 below was conducted to visualize the cheating pattern among the student identified in previous Dendrogram. Its very clear from the Heap Map, SID (26, 35, 44), SID (36, 43) and SID (49, 50) have similar heap map color pattern.

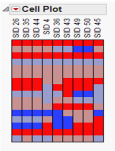


FIGURE 5. Heat Map Analysis.

3.3 Principle Component Analysis

Authors also conducted JMP 12 Principle Component Analysis (Soren H.) shown in Figure 6 based Matrix Eigenvalue and Eigenvector algorithm to derive the two strongest principle components in a linear combination of all the answering variable dimensions. Its very clear that SID (26, 35, and 44), SID (36, 43) and SID (49, 50) are assigned in the same region based on two principle components (in X-Y).

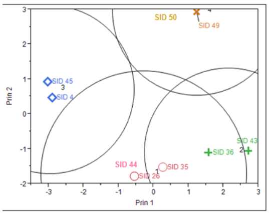


FIGURE 6. Principle Component Analysis.

4 Conclusions

Authors have utilized various Data Mining Algorithms to detect the exam cheating patterns. The same concept and algorithm can be applied to any other applications to uncover the hidden patterns such as Sports Analytics, Customer Relational Management, or Biostatistics.

References

Michael R. Anderberg (1973). *Cluster analysis for applications*, Academic Press, New York, ISBN 0120576503.

Soren Hojsgaard (2011). Multivariate analysis Principal component analysis (PCA) *Statistics and Decision Theory Research Unit.* 1 – 58.

Simulation-based evaluation of the impact of singletons on the fixed effects in a linear mixed model

R. Bruyndonckx[12], N. Hens[13], M. Aerts[1]

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

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

Abstract: Data that are collected in medical sciences often have a hierarchical structure. Regardless of sparseness caused by the presence of a large number of small units or a small number of large units, linear mixed models are used to account for within-unit correlation. Using a simulation study, we assess the impact of an increasing proportion of singletons (i.e. one subunit in a unit) at the highest or lowest level of the hierarchy on the fixed effects in a linear mixed model. Additionally, we assess whether, when high proportions of singletons are present, performance improves by removing or grouping singletons, splitting singletons at the highest level or ignoring the dependency within units at the lowest level. We show that, in the presence of singletons at the lowest level, the model is quite stable. Ignoring clustering and dropping the singletons come with biased standard error estimates. Grouping the singletons does not improve the model's performance. In the presence of singletons at the highest level, the model is unstable. Grouping, dropping and splitting the singletons either decrease type I error rate or increase power, while worsening the other. The likelihood ratio test and Wald test perform poorly. The performance of the permutation test however is superior to that of the F test. In conclusion, the linear mixed model is stable in the presence of singletons at the lowest level, but care should be taken in the presence of singletons at the highest level. In that case, the permutation test rather than the F test should be used to assess the significance of included fixed effects.

Keywords: F test, hierarchical data, permutation test, sparseness

1 Introduction

Data that are collected in e.g. medical sciences often have a hierarchical structure, with units at a lower level nested within units at a higher level. Most multi-level settings consist of a small number of units that tend to be quite large. An example is seen in the Ceftriaxone data, which contain information on doses (expressed in mg/kg/day) of ceftriaxone prescribed to hospitalized children. Here, some geographical regions contain one country, while all but one country contain more than ten children. The opposite, where a large number of units tend to be quite small, occurs less often but is omnipresent in specific fields such as family research. An example can be seen in the Ceftriaxone data, where 47.6% of included departments contain only one child. Units containing only one subunit (e.g. regions containing only one country or departments containing only one child) are referred to as singletons. Regardless of sparseness caused by high proportions of singletons, hierarchical data are generally analysed with linear mixed models.

In this study, we are interested in the impact of different proportions of singletons on the fixed effects in a linear mixed model. We will use simulation studies to assess the impact of an increasing proportion of singletons at the highest level (i.e. countries within regions) and at the lowest level (i.e. children within departments). We will assess whether, when high proportions of singletons are present, the model's performance improves by applying some frequently used techniques to cope with singletons: ignoring the dependency within units, removing singletons, splitting singletons (at highest level) and grouping singletons in an artificial unit.

2 Methods

2.1 singletons at the lowest level of the hierarchy

A simulation study, focussing on a two-level setting with 350 children divided over 50 departments and including variables both at the level of the child (age and reason to treat) and at the level of the department (size), was set up. We used a low (0.15), realistic (0.27, from the Ceftriaxone data) and high (0.64) intracluster correlation coefficient (ICC). The proportion of singletons ranged from 0 to 95% (in steps of 5%), with the number of singletons rounded upwards (e.g. 5% singletons implies 3 departments with one child) and the remaining children divided equally amongst the remaining departments. For each of the 60 scenarios, 1000 datasets were simulated.

We fitted a linear mixed model, including fixed effects for department size, reason for treatment and age, and a random effect for department, to all simulated datasets. Stability of fixed effects estimates and standard errors was evaluated using the relative difference between the average parameter estimate (or standard error) and the true parameter estimate (or standard error) (RDM (or RDE)). Additionally, we considered the rejection rate of the F tests for age, reason for treatment and department size. Because some scenarios contained a very low (or very high) proportion of singletons, we considered ignoring within-cluster correlation, dropping the singletons and grouping them in an artificial unit.

2.2 Singletons at the highest level of the hierarchy

A simulation study, focussing on a three-level setting with 240 children divided over $5 - 40$ countries in five regions, was set up. Data were simulated under the null hypothesis (H_0), assuming that region has no effect on dose, as well as under a specific alternative hypothesis (H_A). For each of the 18 scenarios, 1000 datasets were simulated.

We fitted a linear mixed model, including a fixed effect for region and a random effect for country, to all simulated datasets. The performance of the F test for region was assessed using the rejection rate, referred to as type I error rate (under H_0) or corrected power (under H_A). Because one can think of several ways to eliminate data sparseness, we considered dropping the singletons, grouping them in an artificial region and splitting them into two artificial countries. For one sparse scenario (Scenario 3), we studied the likelihood ratio test, Wald test and permutation test as alternatives to the F test.

3 Results

3.1 Singletons at the lowest level of the hierarchy

The RDM fluctuated regardless of the proportion of singletons. The RDE was stable throughout the simulation study. The rejection rate for effects at the level of the child increased slightly with an increasing proportion of singletons and ICC, while the rejection rate for the effect at the level of the department decreased slightly with an increasing proportion of singletons and ICC.

The RDM for both levels was not affected by ignoring clustering and dropping or grouping singletons. When ignoring clustering, the RDE was overestimated and the F test rejection rate underestimated, with the respective over- and underestimation being worse with higher ICC. When ignoring clustering, the RDE at the level of the department was underestimated while the F test rejection rate was overestimated. When dropping singletons, the RDE increased with an increasing proportion of singletons and ICC, while the F test rejection rate decreased. When grouping the singletons into an artificial department, the RDE decreased slightly with an increasing proportion of singletons and ICC, while the F test rejection rate increased slightly

3.2 singletons at the highest level of the hierarchy

In the absence of singletons, the Type I error rate was low, while the corrected power was high, indicating that the F test performs well. In the presence of singletons, type I error rate was high and power low, indicating that the performance of the F test is inadequate, and there is a need for an alternative approach.

When dropping or regrouping the singletons, type I error rate and corrected power decreased. When splitting the singletons, type I error rate and corrected power increased. The performance of the Wald test and likelihood ratio test, investigated for a scenario with a small number of countries per region (Scenario 3), was worse than that of the F test. Performance of the permutation test however was superior to that of the F test.

4 Conclusions

The linear mixed model appears to be stable enough to handle high proportions of singletons at the lowest level of the hierarchy, even when the intracluster correlation is high. Alternatives which are frequently used, such as ignoring clustering and removing or grouping the singletons, should be avoided as they provide biased standard error estimates for the fixed effects.

The linear mixed model appears to be unstable in the presence of singletons at the highest level of the hierarchy. Frequently used alternatives, such as grouping, splitting or dropping singletons could solve either the problem of a high type I error rate, or the problem of low power, while worsening the other. This forced us to conclude that neither method acts as a solution to the poor performance of the F test. We showed that the performance of the likelihood ratio test and the Wald test was comparable to that of the F test, while the permutation test outperformed the F test, and hence recommend to use the permutation test in the presence of singletons at the highest level of the hierarchy.

References

Garson, D.G. (2013). *Hierarchical linear modeling: guide and applications.* London: Sage publications.

Snijders, T.A.B. and Bosker, R.J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* London: Sage publications.

Verbeke. G and Molenberghs, G. (2009). *Linear mixed models for longitudinal data.* New York: Springer Verlag.

Top-down joint graphical lasso

Eugen Pircalabelu[1], Gerda Claeskens[1], Lourens J. Waldorp[2]

[1] ORSTAT and Leuven Statistics Research Center, KU Leuven, Belgium
[2] Dept. of Psychological Methods, University of Amsterdam, The Netherlands

E-mail for correspondence: eugen.pircalabelu@kuleuven.be

Abstract: We develop a new method to estimate simultaneously multiple graphs and apply it to fMRI data. The method is a 'top-down' method which allows the researcher to zoom repeatedly on edges of the graph obtaining graphs with different levels of detail, in terms of the number of edges and number of nodes. By repeatedly zooming on edges, we estimate graphs with similar structures across coarseness scales, but with different levels of sparsity.

Keywords: FMRI; Graphical models; Joint graphical lasso; Mixed scale data.

1 Introduction and problem description

For $n = 8$ participants the cerebral activity has been measured $T = 240$ times. The researchers used $K = 5$ distinct parcelations of the brain containing different numbers of brain regions of interest (ROIs), obtained in a top-down manner. First 68 ROIs (corresponding to a coarse scale of measurement with anatomically large brain regions) were defined using an anatomical atlas. Subsequently, the large volume regions were further split into several smaller regions and the splitting of the larger regions continued several times. In total five coarseness scales were obtained where the number of ROIs were 68 (for scale 1), 114 (scale 2), 219 (scale 3), 446 (scale 4) and 872 (scale 5). The goal is to estimate brain pathways between the ROIs where each ROI is represented by a node in the graph. We associate with each scale two graphs: a directed and an undirected one, where the graphs at the coarsest scale will contain 68 nodes, while the graphs at the finest scale will contain 872 nodes. Undirected edges indicate a contemporaneous relation between two nodes and directed edges indicate a lagged effect. The sets of edges are unknown and need to be estimated.

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 Graph estimation

Let n be the number of subjects, K the number of scales, p_n^k the number of ROIs at scale k and T the length of the time series (one for each ROI) of brain measurements. The fMRI measurement at scale k for the ith subject, for the jth ROI at time point t is represented by $X_{i,j,t}^k$. The fMRI data at the kth scale for the ith subject is represented by the matrix X_i^k which contains as rows the vectors $X_{i,\cdot,t}^k = (X_{i,1,t}^k, \dots, X_{i,p_n^k,t}^k)^{\mathrm{T}}$ and as columns the vectors $(X_{i,j,1}^k, \dots, X_{i,j,T}^k)^{\mathrm{T}}$. In its rows X_i^k contains measurements for all regions at time t and in its columns a times series for the jth region.

We estimate at each k two graphs: $G^k(E_u^k, V^k)$ having undirected edges only and $H^k(E_d^k, V^k)$ having directed edges only. $V^k = \{1, \dots, p_n^k\}$ denotes the set of all nodes for scale k. E_u^k denotes the set of undirected edges depicted as $a - b$ between a pair of nodes (a, b). Assuming a p_n^k-dimensional Gaussian vector $X_{i,\cdot,t}^k$ with inverse covariance matrix $\Theta \equiv \Sigma^{-1}$, the edge set is defined as $E_u^k = \{(a, b) \in V^k \times V^k | a \neq b \;\&\; \Theta_{a,b}^k \neq 0\}$.

E_d^k represents the set of directed edges of the form $a \to b$. Following Yin and Li (2011) and Abegaz and Wit (2013), for H^k a first order Markov model is used where observations from time t depend only on observations at time $t-1$. We assume that the vector $X_{i,\cdot,t}^k$ conditional on past values, $X_{i,\cdot,t-1}^k$ obeys $X_{i,\cdot,t}^k | X_{i,\cdot,t-1}^k \sim N(\Gamma^k X_{i,\cdot,t-1}^k, \Sigma^k)$. The data generating process depends on scale specific matrices Γ^k and Σ^k. The matrices Γ^k contain the autoregressive coefficients corresponding to lagged effects between the nodes (the ROIs in the example) and are used to define the set of directed edges as $E_d^k = \{(a, b) \in V^k \times V^k | \Gamma_{a,b}^k \neq 0\}$.

We jointly estimate the matrices $\mathcal{A} = \{\Gamma^1, \dots, \Gamma^K, \Theta^1, \dots \Theta^K\}$ by minimizing a top-down penalized negative log-likelihood:

$$\min_{\mathcal{A}} \{ -\sum_{k=1}^{K} (\log\det\Theta^k - \mathrm{trace}(S_{\Gamma^k}\Theta^k)) + \lambda_{n1}(\sum_{a\neq b} |\, \Theta_{a,b}^1 \,| + \sum_{k=2}^{K} \sum_{a\neq b \in U^k} |\, \Theta_{a,b}^k \,|) + \lambda_{n2}(\sum_{a,b} |\, \Gamma_{a,b}^1 \,| + \sum_{k=2}^{K} \sum_{a,b\in D^k} |\, \Gamma_{a,b}^k \,|)\}, \quad (1)$$

such that $\Theta^1, \dots, \Theta^K$ are positive definite and where $S_{\Gamma^k} = 1/(nT) \sum_{i=1}^n \sum_{t=2}^T (X_{i,\cdot,t}^k - \Gamma^k X_{i,\cdot,t-1}^k)(X_{i,\cdot,t}^k - \Gamma^k X_{i,\cdot,t-1}^k)^{\mathrm{T}}$, $U^k = \{(a, b) \in V^k \times V^k | a \neq b \;\&\; \Theta_{a,b}^{k-1} \neq 0\}$ and $D^k = \{(a, b) \in V^k \times V^k | \Gamma_{a,b}^{k-1} \neq 0\}$.

Equation (1) uses ℓ_1 terms which enforce sparsity (see Friedman et al., 2008) for the directed and undirected graphs. The fine scale edges estimated at each scale k depend on the coarser edges estimated at scale $k-1$ which creates a zoom effect on edges in a top-down approach where the structure of the coarser graphs influences the structure of the finer graphs making them similar. Different penalties that enforce sparsity and similarity of graphs can be found in Guo et al. (2011) and Danaher et al. (2014).

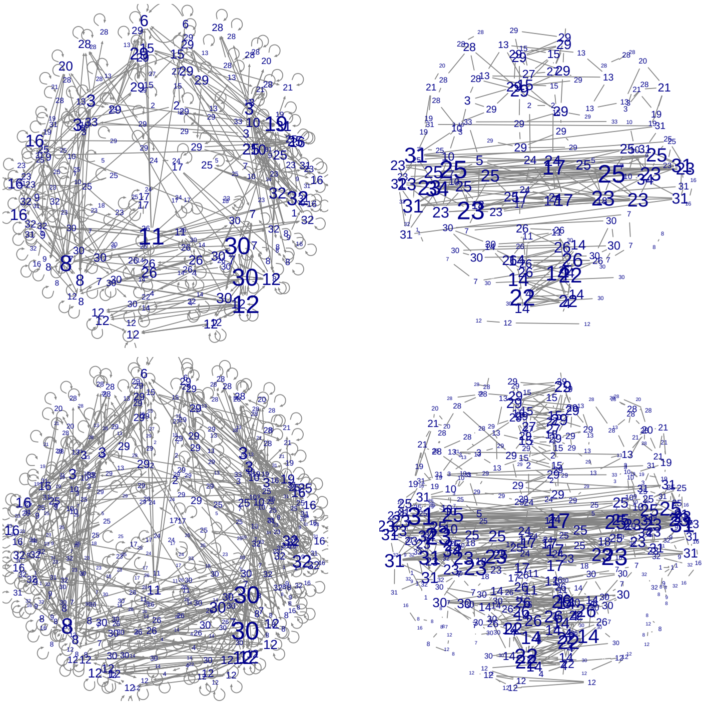


FIGURE 1. Directed (left panels) and undirected (right panels) graphs for scales $k = 3$ (top) and 4 (bottom) estimated using $(\lambda_{n1},\ \lambda_{n2}) = (.4, .2)$

3 fMRI example

Due to space constraints Figure 1 presents only the graphs estimated for scales 3 and 4 from which we conclude that (i) across the scales the graphs are similar to each other (due to the top-down penalty we use) and (ii) the contemporaneous associations between ROIs are more pronounced for regions from opposite sides of the brain, while the lagged associations are more pronounced for regions from the same hemisphere.

Figure 2 shows that the estimated directed and undirected graphs have percentagewise a similar number of edges within the left hemisphere, but for the right hemisphere the directed graphs estimate percentagewise more edges than the undirected ones. Moreover, as we move across the coarseness scales the within hemisphere connections relative to the total connections become more predominant than the between hemisphere connections. This points to a functional connectivity shift in the brain pathways, in the sense that as we increase the splitting of regions this creates partitions that are more homogenous within the hemisphere than across the hemispheres.

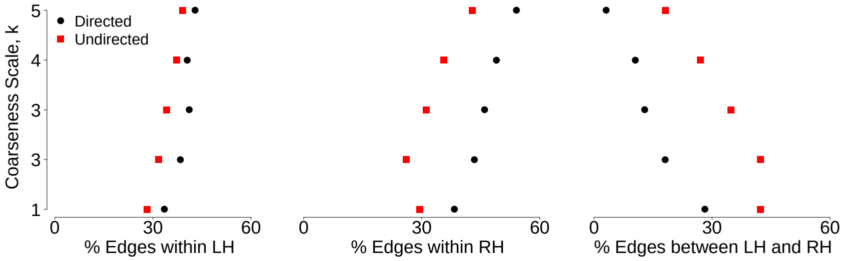


FIGURE 2. Percentages of edges connecting ROIs within the left hemisphere (left), right hemisphere (middle) and from different hemispheres (right).

4 Discussion

The proposed method to jointly estimate sparse graphs at different coarseness scales, starts from the coarsest scale and uses a top-down penalty which constrains the structure of finer scales graphs to depend on the structure of coarser scale graphs. Using a top-down approach reveals how the graphs evolve from having a simpler structure at the coarsest scale to having a more complex structure at the finest scale. The resulting graphs offer the researcher a more complete image about the phenomenon under study.

References

Abegaz, F. and Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, **14(3)**, 586 – 599.

Danaher, P., Wang, P. and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, **76(2)**, 373 – 397.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9(3)**, 432 – 441.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, **98(1)**, 1 – 15.

Pircalabelu, E., Claeskens, G., and Waldorp, L.J. (2016). Mixed scale joint graphical lasso. *Biostatistics*, **17(4)**, 793 – 806.

Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, **5(4)**, 2630 – 2650.

Modelling time series of counts with deflation or inflation of zeros

Marcelo Bourguignon[1]

[1] Universidade Federal do Rio Grande do Norte, Brasil

E-mail for correspondence: m.p.bourguignon@gmail.com

Abstract: In this paper, we introduce a first order non-negative integer valued autoregressive process with zero-modified geometric innovations based on the binomial operator. This new model will enable one to tackle the problem of deflation or inflation of zeros inherent in the analysis of integer-valued time series data. The main properties of the model are derived, such as transition probabilities and zero probability. The methods of conditional maximum likelihood, Yule-Walker and conditional least squares are used for estimating the model parameters. A Monte Carlo experiment is conducted to evaluate the performances of these estimators in finite samples. The proposed model is fitted to time series of number of weekly sales and weekly number of syphilis cases illustrating its capabilities in challenging cases of deflated and inflated count data.

Keywords: INAR(1) process; Zero-modified geometric distribution.

1 Introduction

McKenzie (1985) proposed the INAR(1) process based on the binomial thinning operator, i.e., a sequence $\{X_t\}_{t\in\mathbb{Z}}$ is said to be an INAR(1) process if it admits the representation

$$X_t = \alpha \circ X_{t-1} + \epsilon_t, \tag{1}$$

where $0 \leq \alpha < 1$, $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is a sequence of independent and identically distributed integer-valued random variables, called innovations, with ϵ_t independent of X_{t-k} for all $k \geq 1$, $\mathrm{E}(\epsilon_t) = \mu_\epsilon$ and $\mathrm{Var}(\epsilon_t) = \sigma_\epsilon^2$, and "$\circ$" is binomial thinning operator, defined by

$$\alpha \circ X_{t-1} = \begin{cases} \sum_{j=1}^{X_{t-1}} Y_j, & X_{t-1} > 0; \\ 0, & X_{t-1} = 0, \end{cases}$$

where the so-called counting series $\{Y_j\}_{j\geq 1}$ is a sequence of independent and identically distributed Bernoulli random variables with $\Pr(Y_j = 1) = 1 - \Pr(Y_j = 0) = \alpha$.

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

While processes for integer-valued time series are now abundant, there is a shortage of similar processes when the time series refer to data with deflation or inflation of zeros, i.e., processes for modeling count time series with excess (or deficit) of zeros based on thinning operators were discussed by few authors. Jazi et al. (2012b) introduce a new stationary INAR(1) process with zero inflated Poisson innovations [ZINAR(1)]. Recently, Barreto-Souza (2015) proposed a first-order integer-valued autoregressive process for dealing with count time series with deflation or inflation of zeros [ZMGINAR(1)]. The proposed process has zero-modified geometric marginals. However, the conditional probabilities of this model don't have a simple form and the parameter restrictions aren't liberal.

This paper aims to give a contribution in this direction. The objective of this paper is to propose a new INAR(1) process (1) with zero-modified geometric (ZMG) innovations, denoted by INARZMG(1), based on binomial thinning for modeling nonnegative integer-valued time series with deflation or inflation of zeros.

Let $\{\epsilon_t\}_{t\in\mathbb{Z}}$ be a sequence of discrete i.i.d. random variables following a zero-modified geometric (ZMG) distribution with parameters $\mu > 0$ and $\pi \in (-1/\mu, 1)$. More specifically, we here assume that $\{\epsilon_t\}_{t\in\mathbb{Z}}$ has a probability mass function given by

$$\Pr(\epsilon_t = y) = \begin{cases} \frac{1+\pi\mu}{1+\mu}, & \text{if } y = 0, \\ (1-\pi)\frac{\mu^y}{(1+\mu)^{y+1}}, & \text{if } y = 1, 2, \ldots \end{cases} \tag{2}$$

The process $\{X_t\}_{t\in\mathbb{Z}}$ satisfying Equation (1), with $\{\epsilon_t\}_{t\in\mathbb{Z}} \sim \text{ZMG}(\pi, \mu)$, is Markovian, stationary, and ergodic. Then, the Markov process admits a unique stationary distribution.

The dispersion index, which is the variance-to-mean ratio, will be given by

$$\mathrm{I}_X := \frac{\sigma_X^2}{\mu_X} = 1 + \frac{\mu(1+\pi)}{1+\alpha},$$

it follows that this model presents equidispersion when $\pi = -1$; underdispersion when $\mu \in (0, 1)$ and $\pi \in [-1/\mu, -1)$; and overdispersion when $\pi \in (-1, 1)$.

2 Real data examples

2.1 Modelling deflation of zeros

In the first application, we consider the series of weekly sales of a particular soap product in a supermarket. The proportion of zeros in the series considered is 3.7%. Then, we have evidence that there is deflation of zeros in the number of weekly sales. The time series data and their sample autocorrelation and partial autocorrelation functions are displayed in the Figure 1.

Table 1 provides the estimates of the model parameters and two goodness-of-fit statistics: Akaike information criterion (AIC) and Bayesian information criterion (BIC). From this table, we observe that the proposed model being better.

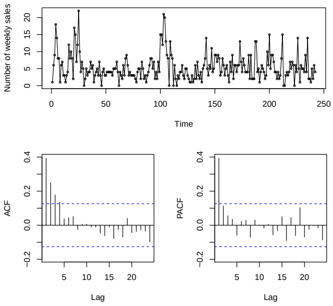


FIGURE 1. Plots of the time series, autocorrelation and partial autocorrelation functions for the number of weekly sales.

TABLE 1. Estimates of the parameters and goodness-of-fit statistics for the first data set.

Model		CML Estimates	AIC	BIC
	$\widehat{\alpha}$	0.3056	1252.88	1263.4
INARZMG(1)	$\widehat{\mu}$	3.2094		
	$\widehat{\pi}$	-0.1826		
	$\widehat{\alpha}$	0.3623	1310.91	1321.4
ZINAR(1)	$\widehat{\lambda}$	4.5972		
	$\widehat{\rho}$	0.2415		
INARG(1)	$\widehat{\alpha}$	0.3712	1260.34	1267.32
	$\widehat{p}$	0.7746		

2.2 Modelling inflation of zeros

For the second application, we consider the weekly number of syphilis cases in the United States from 2007 to 2010 in Vermont state given in ZIM package. The proportion of zeros in the series considered is 96%. Then, we have evidence that there is inflation of zeros in the number of syphilis cases. The series, autocorrelation and partial autocorrelation functions are displayed in Figure 2.

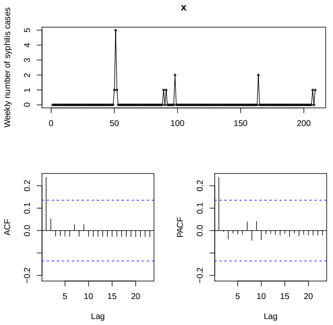


FIGURE 2. Plots of the time series, autocorrelation and partial autocorrelation functions for the number of syphilis cases.

Table 2 gives the CML estimates, AIC and BIC for the fitted models. Since the values of the AIC and BIC are smaller for the INARZMG(1) process compared to those values of the ZINAR(1) and INARG(1) models, the new model seems a competitive model for these data.

TABLE 2. Estimates of the parameters and goodness-of-fit statistics for the number of syphilis cases.

Model		CML Estimates	AIC	BIC
INARZMG(1)	$\widehat{\alpha}$	0.0801	96.46	106.49
	$\widehat{\mu}$	0.7239		
	$\widehat{\pi}$	0.9078		
ZINAR(1)	$\widehat{\alpha}$	0.0894	97.80	107.83
	$\widehat{\lambda}$	1.1893		
	$\widehat{\rho}$	0.9444		
INARG(1)	$\widehat{\alpha}$	0.1138	110.32	117.00
	$\widehat{p}$	0.0606		

References

Barreto-Souza(2015) barreto2015 Barreto-Souza, W. (2015). Zero-Modified Geometric INAR(1) Process for Modelling Count Time Series with Deflation or Inflation of Zeros. *Journal of Time Series Analysis*, **36**, 839–852.

Jazi et al.(2012a) jazia Jazi, M. A., Jones, G. and Lai, C. D. (2012a). Integer valued AR(1) with geometric innovations. *Journal of the Iranian Statistical Society*, **11**, 173–190.

Jazi et al.(2012b) jazib Jazi, M. A., Jones, G. and Lai, C. D. (2012b). First-order integer valued AR processes with zero inflated Poisson innovations. *Journal of Time Series Analysis*, **33**, 954–963.

McKenzie(1985) Mc85 McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resour. Bull.*, **21**, 645–650.

An Estimated Parameter of Local Polynomial Regression on Time Series Data

Autcha Araveeporn

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

E-mail for correspondence: kaautcha@hotmail.com

Abstract: The local linear estimation, Nadaraya-Watson estimation, and locally weighted scatter plot smoothing (LOWESS) estimation are the way for estimating unknown parameter of local polynomial regression based on nonparametric regression model, fitting function of independent variable by the weighted least squares is computed for time series data. With the local polynomial regression fitting we can estimate the local linear estimator based on kernel function as a Gausian density function, and the Nadaraya-Watson estimation which is a part of local linear estimation or called a local constant estimation. Furthermore the LOWESS is used a weighted least squares to estimate parameter that modified the kernel function as a tricube kernel. The goal of this article is to show, through application, which three estimators can be used the best for fitting time series data. The Average Mean Square Errors (AMSE) is considered as a criterion to check a bias of fitted estimator. For simulation data, the data is generated by autoregressive process by several coefficients. The Nadaraya-Watson estimation outperforms other methods by showing the minimum of AMSE values.

Keywords: local linear estimation; local polynomial regression; locally weighted scatter plot smoothing; Nadaraya-Watson estimation

1 Introduction

Regression analysis is a statistical tool for the investigation of relationship between independent and dependent variables in term of regression model. An estimating parameter of regression model requires an assumption such as the linearity, statistical independence, homoscedasticity, and normality that the form of the underlying regression analysis. If an inappropriate regression model is used, it is possible to produce misleading conclusions. To overcome the difficulty caused by the restrictive assumption of the regression function, this approach leads to so-called nonparametric regression.

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

Typically, the nonparametric regression methods are overcome this problem, because the fitted model interpolates on the curve of data base on bandwidth values. The local polynomial regression is a class of nonparametric regression method which has a long history in the smoothing of time series.

2 Methods of Parameter Estimation

The simple of nonparametric regression model can be written as

$$y_t = f(x_t) + \varepsilon_t, \qquad t = 1, 2, \ldots, n, \tag{1}$$

where x_t are the independent variable as the time points of time series data, y_t are the dependent variable, and ε_t denote the measurement errors. The Taylor expansion can be approximated by a local polynomial regression of degree p and assumed that $f(x_t)$, as

$$f(x_t) = f(x_0) + (x_t - x_0)f^{(1)}(x_0) + \frac{1}{2}(x_t - x_0)^2 f^{(2)}(x_0) + \ldots, t = 1, 2, \ldots, n.$$

The local polynomial regression estimator is used the weight least-squares to minimize of

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \arg\min \sum_{t=1}^{n} (y_t - f(x_t))^2 \, K_h(x_t - x_0), \tag{2}$$

where $K(\cdot)$ is a kernel function, $h > 0$ is called the bandwidth.
The estimators of local polynomial regression are depended on the process of weight least squares and the bandwidth selection. The following methods are shown the parameter estimation on each method.

2.1 The Local Linear Estimation

For the weighted least squares, the local linear estimator can be approximated by setting $p = 1$, denote for $j = 0, 1$ and 2,

$$S_j(x; h) = \sum_{t=1}^{n} K_h(x_t - x_0)(x_t - x_0)^j,$$

$$T_j(x; h) = \sum_{t=1}^{n} K_h(x_t - x_0)(x_t - x_0)^j y_t.$$

The local linear estimator is

$$\hat{\beta}_0(x; h) = \frac{T_0(x; h)\ S_2(x; h) - T_1(x; h)\ S_1(x; h)}{S_0(x; h)\ S_2(x; h) - S_1^2(x; h)} \tag{3}$$

,and

$$\hat{\beta}_1(x; h) = \frac{T_1(x; h)\ S_0(x; h) - T_0(x; h)\ S_1(x; h)}{S_0(x; h)\ S_2(x; h) - S_1^2(x; h)}. \tag{4}$$

The quality of the local polynomial regression depends on the bandwidth selection method. Ruppert, Sheather, and Wand (1995) studied the ideas of plug-in bandwidth selection of kernel estimators.

2.2 Nadaraya-Watson Estimation

The trend estimator is evaluated by using the Nadaraya-Watson kernel estimator (Nadaraya (1964) and Watson (1964)) and written as

$$\hat{f}(x_t) \quad = \quad \frac{\sum_{t=1}^{n} K_h(x_t - x_0)y_t}{\sum_{t=1}^{n} K_h(x_t - x_0)}, \tag{5}$$

where h is known as the bandwidth parameter. The bandwidth can be chosen using a cross-validation criteria. The kernel functions K can be chosen as the Gaussian density function.

2.3 LOWESS Estimation

Cleveland(1979) introduced an alternative form of the local polynomial regression which is called LOWESS , locally weighted scatter plot smoothing. The basic idea was to start the weighted least square as the kernel function in term of the tricube kernel function, and proposed a nearest neighbour bandwidth by setting $r = nf + 0.5$. For each predictor x_t, let

$$h_k = \mid x_k - x \mid_{(r)},$$

where h_k is the rth order statistic of the sample $\mid x_k - x_1 \mid, \mid x_k - x_2 \mid, \ldots, \mid x_k - x_n \mid$.

3 Application of Simulation Data

The data is generated in term of autoregressive model in order 1 (AR1) following

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

where $\varepsilon_t \sim N(0, 1)$, the coefficient ($\rho$) of AR(1) is defined by $0.1, 0.5, 0.7$, and 0.99, and the sample sizes $n = 100, 200, 300$, and 400. The accuracy of fitting function is consider by the Mean Square Errors (MSE) as follows:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2, \tag{6}$$

where y_t denoted the simulated data and $\hat{y}_t$ denoted the fitting data of 3 estimators. From Table 1, the results show that the average MSE of NW is the minimum values in all cases.

4 Conclusion

We have been discussed the method to estimate parameter of local polynomial regression function. It is concluded that the estimation of NW has indicated a good performance more than LL and LOWESS method. Hence it can be say that NW leads to a nearly enough class of local polynomial regression to estimate adequately fitted time series data. The cross-validation criteria is a good performance for fitting model.

TABLE 1. The average MSE on simulated data of local linear (LL), Nadaraya-Watson (NW), and locally weighted scatter plot smoothing (LOWESS) estimation.

Sample sizes	Method	ρ =0.1	ρ =0.5	ρ = 0.7	ρ = 0.99
n=100	LL	1.0513	0.7593	0.6099	0.4822
	NW	0.5438	0.3604	0.2966	0.2197
	LOWESS	1.2167	1.4810	1.8635	3.3187
n=200	LL	1.0438	0.9299	0.8489	0.6327
	NW	0.5137	0.3494	0.2918	0.2228
	LOWESS	1.1113	1.4033	1.9588	6.5061
n=300	LL	1.0337	1.0671	1.0527	0.7893
	NW	0.5011	0.3501	0.2910	0.2190
	LOWESS	1.0709	1.3906	1.9561	9.9320
n=400	LL	1.0377	1.1317	1.2328	0.9651
	NW	0.5042	0.3468	0.2908	0.2227
	LOWESS	1.0672	1.3736	1.9885	12.3403

References

Cleveland, W.S. (1979). Robust Locally Weight Regression and Smoothing Scatter Plots. *Journal of American Statistics Association*, **74**, 829 – 836.

Nadaraya, E.A. (1964). On Estimating Regression. *Theory of Probability and Its Application*, **10**, 186 – 196.

Ruppert, D, Sheather, S. J.,Wand M. P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. Journal of the American Statistical Association, **90**, 1257 – 1270.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā, Series A*, **26**, 359 – 372.

New links for binary regressions: An application to the coca cultivation in Peru

Artur J. Lemonte[1], Jorge L. Bazán[2]

[1] Universidade Federal do Rio Grande no Norte, Brazil
[2] Universidade de São Paulo, Brazil

E-mail for correspondence: arturlemonte@gmail.com

Abstract: Binary response data arise naturally in applications. In general, the well-known logistic and probit regression models form the basis for analysing binary data in practice. These regression models make use of symmetric link functions (logit and probit links). However, many authors have emphasized the need of asymmetric links in modeling binary response data. In this paper, we consider a broad class of parametric link functions that contains as special cases both symmetric as well as asymmetric links. Furthermore, this class of links is quite flexible and simple, and may be an interesting alternative to the usual regression models for binary data. We consider a frequentist approach to perform inferences, and the maximum likelihood method is employed to estimate the model parameters. We also propose residuals for the link models to assess departures from model assumptions as well as to detect outlying observations. Additionally, the local influence method is discussed, and the normal curvatures for studying local influence are derived under two specific perturbation schemes. Finally, an application to the coca leaf cultivation in Peru is considered to show the usefulness of the proposed link models in practice.

Keywords: Binary response model; Maximum likelihood estimation; Parametric link function; Symmetric distributions.

1 New link functions

In the generalized linear model setup, we have that

$$\mu_i = F(\eta_i), \quad i = 1, \ldots, n, \tag{1}$$

where $F(\cdot)$ is a CDF and its inverse $F^{-1}(\cdot)$ is typically called link function. The link function is symmetric when $F(\cdot)$ is a CDF of a symmetric distribution.

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

1.1 Power symmetric and reciprocal power symmetric distributions

We have the following definition.

Definition 1 *A univariate random variable T is said to have a power distribution with location, scale and shape parameters given by $\xi \in \mathbb{R}$, $\phi > 0$ and $\lambda > 0$, respectively, if its CDF has the form*

$$F_{\mathrm{p}}(t) = G\left(\frac{t-\xi}{\phi}\right)^{\lambda}, \quad t \in \mathbb{R},$$

where $G(\cdot)$ is any absolute continuous CDF. The standard power distribution arises when $\xi = 0$ and $\phi = 1$, and its CDF is $F_{\mathrm{p}}(z) = G(z)^{\lambda}$, for $z \in \mathbb{R}$.

In the above construction, the function $G(\cdot)$ is refereed to as the baseline CDF. We can also define the (standard) reciprocal power distribution by considering $F_{\mathrm{rp}}(z) = 1 - G(-z)^{\lambda}$, for $z \in \mathbb{R}$. From the above definition, it is evident that we can introduce several new class of distributions. In this paper, we will assume that the baseline CDF $G(\cdot)$ belongs to the symmetric family of distributions in its standardized form (Fang et al., 1990). In this case we can express $F_{\mathrm{rp}}(z) = 1 - [1 - G(z)]^{\lambda}$. Therefore, we have the standard power symmetric (PS) distribution as well as the standard reciprocal power symmetric (RPS) distribution.
The PS distributions are skewed to the right if $\lambda > 1$ and to the left if $0 < \lambda < 1$, and the RPS distributions are skewed to the left if $\lambda > 1$ and to the right if $0 < \lambda < 1$. The additional shape parameter λ introduces a great degree of skewness to the distribution. In fact, let us consider the definition of skewness measure in terms of the mode for the PS and RPS distributions; that is, if $Z \sim \mathrm{PS}(\lambda)$, then $\gamma_{\mathrm{p}} = 1 - 2F_{\mathrm{p}}(M) = 1 - 2G(M)^{\lambda}$, where M is the mode of Z; and if $Z \sim \mathrm{RPS}(\lambda)$, it follows that $\gamma_{\mathrm{rp}} = 1 - 2F_{\mathrm{rp}}(N) = 2G(-N)^{\lambda} - 1 = -\gamma_{\mathrm{p}}$, where $N = -M$ is the mode of Z. Clearly, $-1 \leq \gamma_{\mathrm{p}}, \gamma_{\mathrm{rp}} \leq 1$. Additionally, note that small values of λ yield values of γ_{p} and γ_{rp} near -1 and 1, respectively. On the other hand, γ_{p} and γ_{rp} are close to 1 and -1, respectively, for high values of λ. In short, $\gamma_{\mathrm{p}} \to -1(1)$ as $\lambda \to 0^{+}(\infty)$, whereas $\gamma_{\mathrm{rp}} \to -1(1)$ as $\lambda \to \infty(0^{+})$. Values of this skewness measure near $1(-1)$ will indicate extreme right (left) skewness. Therefore, we can have a high degree of skewness (positive as well as negative) for the PS and RPS distributions depending on the values of the additional shape parameter λ. Thus, asymmetric link functions constructed from the PS (RPS) distributions can be very skewed. Finally, we have that $\gamma_{\mathrm{p}} = \gamma_{\mathrm{rp}} = 0$ when $\lambda = 1$, as expected.

1.2 The PS and RPS link models

The first class of asymmetric links uses $F_{\mathrm{p}}(\eta_i)$ (i.e. the PS distributions) for $F(\cdot)$ in (1), that is,

$$\mu_i = F_{\mathrm{p}}(\eta_i) = G(\eta_i)^{\lambda}, \quad i = 1, \ldots, n. \tag{2}$$

When $F(\cdot)$ in (1) involves $F_{\mathrm{rp}}(\eta_i)$ (i.e. the RPS distributions) we have the second class of asymmetric links given by

$$\mu_i = F_{\mathrm{rp}}(\eta_i) = 1 - G(-\eta_i)^{\lambda}, \quad i = 1, \ldots, n. \tag{3}$$

The additional parameter $\lambda > 0$ characterizes the skewness of the link functions associated with models (2) and (3). For $\lambda = 1$ the models (2) and (3) are the same and, in addition, the link function is symmetric. These models include commonly used symmetric links such as the probit and logit links. They also include the symmetric robit (Student-t) link as a special case. Notice we are considering wide classes of link functions for binary response data, which contains several symmetric (and asymmetric) links as special cases.

As expected, the probability μ_i approaches zero at the same rate as it approaches one when $\lambda = 1$, since in this case the link function is symmetric. When $0 < \lambda < 1$ ($\lambda > 1$), the probability μ_i approaches one (zero) at a faster rate than it approaches zero (one) for the PS models. On the other hand, the probability μ_i approaches one (zero) at a faster rate than it approaches zero (one) when $\lambda > 1$ ($0 < \lambda < 1$) for the RPS models. In short, the additional parameter λ can produce a considerable degree of skewness to the links, since it controls the "tail" behaviour of the links.

Finally, the parameter λ may be named as "acceleration" parameter for the PS link models, in the sense that it will accelerate the point (value) of the linear predictor at which the slope power of the curve becomes greatest. In contrast, this parameter may be named as "deceleration" parameter for the RPS link models, once now we have an opposing behavior.

2 Application to real data

We have a study on eradication of the coca cultivation in Peru, and we consider the following regression models for modeling these data:

$$\mu_i = F_{\mathrm{p}}(\eta_i) = G(\eta_i)^{\lambda}, \quad \mu_i = F_{\mathrm{rp}}(\eta_i) = 1 - G(-\eta_i)^{\lambda},$$

where $\mu_i = \Pr(\mathrm{Erad} = 1)$, and

$$\eta_i = \beta_1 + \beta_2 \mathrm{Pcult}_i + \beta_3 \mathrm{Cpar}_i + \beta_4 \mathrm{Pcco}_i + \beta_5 \mathrm{Polev}_i,$$

for $i = 1, \ldots, 1947$. We shall consider several link functions to model these binary data, which are based on the Normal, Logistic, Laplace, Cauchy, Student-t and PE distributions. The degrees of freedom $\nu > 0$ in the Student-t and PS(RPS)-Student-t link models as well as the value of $k \in (-1, 1]$ in the PE and PS(RPS)-PE link models were obtained by using the profile log-likelihood procedure We also include the complementary log-log (cLogLog) and log-log (LogLog) link functions given by $\log[-\log(1-\mu_i)] = \eta_i$ and $-\log[-\log(\mu_i)] = \eta_i$, respectively, for the sake of comparison.

To compare the different regression models, we consider selection criteria on the candidate models; that is, the value of the log-likelihood function evaluated at the ML estimates ($\widehat{\ell}$), and the AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) and HQIC (Hannan-Quinn Information Criterion) criteria (see Table 1). According to the AIC, BIC and HQIC criteria, the PS- and RPS-Student-t regression models outperform all regression models and therefore should be preferred, indicating that these regression models may be suitable to describe these binary data.

References

Fang, K.T., Kotz, S., Ng, K.W. (1990). *Symmetric Multivariate and Related Distributions.* Chapman and Hall, London.

TABLE 1. AIC, BIC, HQIC and $\widehat{\ell}$ for the fitted models.

Link model	$\widehat{\ell}$	AIC	BIC	HQIC
Normal	-1223.03	2456.05	2483.92	2466.30
Logistic	-1222.85	2455.71	2483.58	2465.96
Laplace	-1223.56	2457.12	2484.99	2467.37
Cauchy	-1223.30	2456.61	2484.48	2466.85
Student-$t\,(\nu = 2.36)$	-1222.73	2455.47	2483.34	2465.71
PE $(k = 0.42)$	-1223.28	2456.55	2484.42	2466.80
PS-Normal	-1222.95	2457.90	2491.35	2470.20
PS-Logistic	-1222.44	2456.87	2490.32	2469.17
PS-Laplace	-1220.75	2453.50	2486.94	2465.79
PS-Cauchy	-1220.66	2453.33	2486.77	2465.62
PS-Student-$t\,(\nu = 0.25)$	-1218.22	2448.45	2481.89	2460.74
PS-PE $(k = -0.2)$	-1219.81	2451.62	2485.06	2463.92
RPS-Normal	-1222.85	2457.70	2491.15	2470.00
RPS-Logistic	-1222.65	2457.31	2490.75	2469.60
RPS-Laplace	-1221.09	2454.18	2487.62	2466.47
RPS-Cauchy	-1221.30	2454.60	2488.04	2466.89
RPS-Student-$t\,(\nu = 0.19)$	-1218.20	2448.40	2481.84	2460.70
RPS-PE $(k = -0.34)$	-1219.85	2451.69	2485.14	2463.99
cLogLog	-1223.10	2456.20	2484.07	2466.45
LogLog	-1224.50	2459.00	2486.87	2469.25

Wavelet regression and generalized linear models with continuous distributions

Aluísio Pinheiro[1], Eufrásio Lima Neto[2]

[1] Universidade Estadual de Campinas, Campinas, Brazil
[2] Universidade Federal da Paraíba, João Pessoa, Brazil

E-mail for correspondence: apinheiro.unicamp@gmail.com

Abstract: We discuss in this paper the use of wavelet regression methods in generalized linear models setups. The use of such a non-parametric link function would be interesting in practice for several reasons. The first one is that researchers in different areas are used to the employment of the paradigm of generalized linear models. The second one is that researchers are more interested in the set of explanatory variables which adequately predicts the responses than on the specific form on which this relationship happens. The correct specification of the link function is paramount to a good fit. The wavelets can be employed in three ways: (i) by helping choosing the link function from a pre-specified set of fundtions; (ii) by providing a fully non-parametric estimation of the mean; and (iii) by providing a non-parametric link function.

Keywords: Non-parametric link function; non-linear regression;semi-parametric regression

1 Introduction

The generalized linear models (GLMs) represent one of the most important developments in statistical theory over the past several decades[6]. A GLM is characterized by three terms. The first is the random component with the response variable belonging to the exponential family of distributions. The second term is the systematic component represented by a linear predictor that includes the explanatory variables. Finally, the third term is the link function which connects the linear predictor to the response variable mean.

An important step in the GLM model employment is the choice of the link function. This function represents the mathematical structure or the regression equation that relates the random component to the linear predictor in this class of models. Link misspecification can lead to several problems on a GLM application, such as bias in the regression parameters and in the mean response estimates

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, 3]. A methodology which finds an appropriate link function for a GLM is still an open problem. Techniques have been proposed to evaluate if a predetermined link function is adequate for a fitted GLM model [1, 4, 7]. Most of the current techniques are straightforward adaptations from linear models' techniques.
Wavelets have been developed in functional analysis as bases for $L_2(\mathbf{R})$, as well as some of its subspaces. These classes of functions contain a large number of diverse elements, which makes them suitable for broad theoretical and numerical applications. For instance, they form unconditional bases for some large functional classes, which leads to optimal estimators and tests [8].

2 The problem

A Multi-Resolution Analysis (MRA) in $L_2(R)$ is a nested sequence of closed subspaces, $\{V_j\}_{j\in Z}$ with four basic properties: (i) **Hierarchy** - $V_j \subset V_{j+1} \subset L_2(R)$ $\forall j \in Z$; (ii) **Dense Union and Trivial Intersection** - $\bigcup_{j\in Z} V_j = L_2(R)$ and $\bigcap_{j\in Z} V_j = \{0\}$; (iii) **Self-Similarity** - $m(2^j t) \in V_j \Leftrightarrow m(t) \in V_0$ $\forall j \in Z$; and (iv) **Natural Basis** - $\exists \phi \in V_0$ so that $T^k\phi(t) = \phi(t-k) \forall k \in Z$ spans V_0, i.e., $V_0 = \left\{m \in L_2(R) \mid f(t) = \sum_{k\in Z} c_k\phi(t-k)\right\}$ for some appropriate sequence $\{c_k\}_{k\in Z}$, and $\{\phi(\cdot - k), k \in Z\}$ is called an orthonormal basis of V_0. By mirror filtering another function, ψ, can be built, so that any function $m \in L_2(R)$ can be written in L_2-sense as:

$$m(t) = \sum_{j\in Z}\sum_{k\in Z}\beta_{j,k}\psi_{j,k}(t) = \sum_{k\in Z}\alpha_{j_0,k}\phi_{j_0,k}(t) + \sum_{j\geq j_0\in Z}\sum_{k\in Z}\beta_{j,k}\psi_{j,k}(t),$$

for an arbitrary j_0, where $\phi_j(t) = 2^{j/2}\phi(2^j t - k)\quad j \in Z\quad k \in Z$ and $\{\psi_{j,k}(t) = 2^{\frac{j}{2}}\psi(2^j t - k), j \in Z, k \in Z\}$[5].
Let $Y = \{y_1, \ldots, y_n\}$ be a set of observations that represents a random sample of the response variable Y. Suppose a set of explanatory variables $X_1, \ldots, X_p$ and

$$\eta = g(\mu) = \mathbf{X}\beta, \tag{1}$$

where $\mathbf{X}$ is the design matrix formed by the observed values of the explanatory variables $X_1, X_2, \ldots, X_p$, β is vector of parameters, η is the vector of linear predictors, μ is the vector of means of Y, i.e., with $\eta=(\eta_1, \ldots, \eta_n)^T, \mu = (\mu_1, \ldots, \mu_n)^T$ and $\beta=(\beta_0, \ldots, \beta_p)^T$. $g(\mu)$ is the link function, which connects the response variable mean with the explanatory variables. If Y is continuous, a few functions available for a **GLM** are: the *identity, logarithmic, inverse, power*, among others.
We propose wavelets in three ways. The first procedure identifies the most appropriate link function, g, for a GLM from a set of pre-specified link functions. The second is a purely non-parametric wavelet regression without any paradigm related to GLM for the estimation of the mean function m of Y. Finally the third setup employs a wavelet estimation of the link function g within the GLM model. The proposals are illustrated by simulation studies and by their application on three data sets. The data sets present problems on insurance, measurements on babies and SONAR signal analysis. The Monte Carlo simulation studies with different sample sizes, random components, true link functions and criteria selection are considered and the overall and relative performances of each proposal are recorded.

3 Simulation Studies

We perform an experimental study to evaluate the performance of the wavelet proposals. The artificial data sets consider a predefined relationship between the response variable Y and the linear predictor $\eta = X\beta$, where X represents a set of explanatory variables. The synthetic data sets are generated according to 30 different configurations, taking into account 3 sample sizes (128, 256, 512), 3 probability distributions for the response variable Y (Gaussian, gamma, inverse Gaussian) and 4 link functions (identity, logarithm, inverse, $1/\mu^2$). Note that the $1/\mu^2$ link function was considered only in the inverse Gaussian model. We considered one explanatory variable X, uniformly distributed in $(0.5, 1.5)]$. A Monte Carlo simulation was considered for each configuration with 1000 replications.

At each time, we generate an artificial data set according to a predefined GLM. The WM and other eligible GLMs are fitted to this data set. We then compare the performances of the three procedures. On the first wavelet proposal, we calculate the frequency of correct classification (choice) of link functions. This was never smaller than 85% in our simulation studies. The link functions *log* and *inverse* exhibited a better true classification rate when compared with *identity* link function. The WM demonstrated a better performance in the random components gamma and inverse Gaussian, when compared with the Gaussian distribution. These results highlight that the WM can be used for choosing an appropriate link function when the response variable presents an asymmetric distribution and a nonlinear relationship between the variables.

For the non-parametric and semi-parametric proposals, performance of the parametric correct link function is superior, specially for smaller samples. This, however, depends on the initial correct link function specification, which is not expected in many situations. Moreover, non-linear and segmented linkages are true in many real-life situations. In these latter cases the non-parametric and the semi-parametric are superior to the parametric, with a clear advantage for the semi-parametric in estimation and prediction precision.

4 Concluding remarks

We propose the use of wavelets in generalized linear model setups. The percentage of true link function classification is higher than 85% but for one configuration. Moreover, this degree of accuracy is based on very parsimonious wavelet representation. If the true link function is compared to the non-parametric and semi-parametric approaches, it will be superior to them, as expected. However, the advantage the fully parametric method has depends on the right choice of the link function which is not necessarily true. Moreover, for many interesting applications the true nature of the link function is not parsimoniously represented by any set of link functions. For these more realistic situations wavelets may be the only efficient choice.

We believe that wavelets may be useful on GLM setups for the purposes here specified but further research will yield other ways in which the combined use of wavelet representation and generalized linear models will provide the practitioner with a very reliable and efficient tool.

Bibliography

[1] Cole, M. J. and McDonald, J. W. (1989). Bootstrap goodness-of-link testing in generalized linear models. *Statistical Modelling*, **57**, 84–94.

[2] Czado, C. and Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, **33**, 213–231.

[3] Czado, C. and Raftery, A. E. (2006). Choosing the link function and accounting for link uncertainty in generalized linear models using Bayes factors. *Statistical Papers*, **47** (3), 419–442.

[4] Hinkley, D.V. (1985). Transformations diagnostic for linear models. *Biometrika*, **72**, 487–496.

[5] Morettin, P.A., Pinheiro, A. and Vidakovic, B. (2016). *Wavelets in Functional Data Analysis.* Springer, New York, *in press.*

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

[7] Pregibon D. A. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society C*, **29** (1), 15–23.

[8] Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* John Wiley & Sons.

Bivariate residual plots with simulation polygons

Rafael A. Moral[1], John Hinde[2], Clarice G. B. Demétrio[1]

[1] ESALQ/USP, Brazil
[2] NUI Galway, Ireland

E-mail for correspondence: `rafael.moral@usp.br`

Abstract: When using univariate models, goodness-of-fit can be assessed through many different methods, including graphical tools such as half-normal plots with a simulation envelope. This is straightforward due to the notion of ordering of a univariate sample, which can readily reveal possible outliers. In the bivariate case, however, it is often difficult to detect extreme points and verify whether a sample of residuals is a reasonable realisation from a fitted model. We propose a new framework, implemented as the `bivrp` R package, available on the Comprehensive R Archive Network. Our framework uses the same principles of the simulation envelope in a half-normal plot, but as a simulation polygon for each point in a bivariate sample.

Keywords: `bivrp` package; Bivariate models; Goodness-of-fit; Graphical methods.

1 Introduction

For both univariate and multivariate models, it is possible to obtain summary goodness-of-fit statistics to compare model fits, such as information criteria (e.g. Akaike and Bayesian) and adjusted R-squared values. However, while these statistics are useful for comparing models fitted to the same data, they cannot be used to inform on the suitability of a particular model to the data. In this case, diagnostic analyses play a significant role. For univariate models, a possible alternative is the use of half-normal plots with simulation envelopes (Moral et al., *in press*). Here, we extend this approach to bivariate models. Our proposal makes it possible for one to graphically examine whether the observed data is a plausible realisation of the fitted model.

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

2 Methodology

Let $\mathbf{y}_i = (y_{1i}, y_{2i})^{\mathrm{T}}$ be a vector of bivariate responses to which a model is fitted. Also, let $\mathbf{r}_i = (r_{1i}, r_{2i})^{\mathrm{T}}$ be the vector of bivariate model diagnostics (e.g. residuals, Cook's distances, leverages). Then, a very basic and straightforward way of ordering the bivariate diagnostics is by calculating the angle

$$\alpha_i = \tan^{-1}\left(\frac{r_{2i}}{r_{1i}}\right)$$

they form with the origin (see Figure 1(a)), to obtain the ordered diagnostics $\mathbf{r}_{(i)} = (r_{1(i)}, r_{2(i)})^{\mathrm{T}}$. Here, the ordering is done not to define extremes, but to aid in the simulation process.

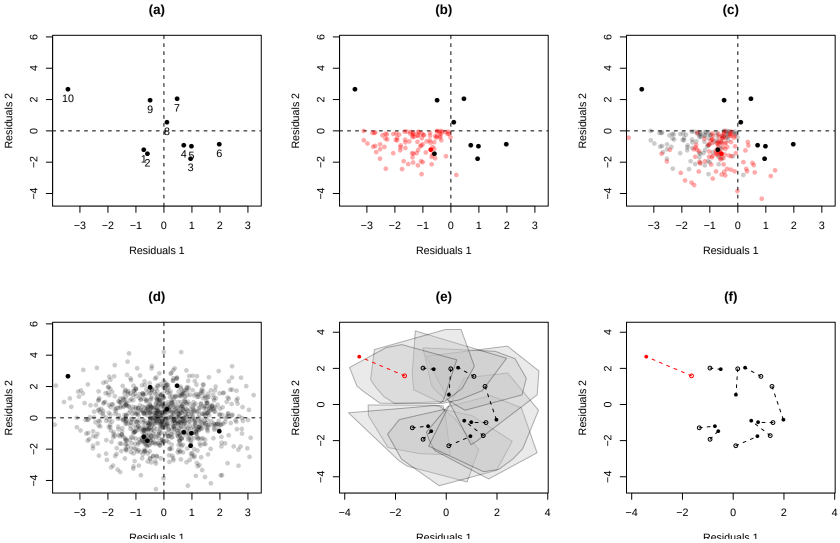


FIGURE 1. The process of constructing the bivariate residual plot with simulation polygons .

We then simulate 99 bivariate response variables $\mathbf{y}_i^s = (y_{1i}^s, y_{2i}^s)^{\mathrm{T}}, s = 1, \ldots, 99$ from the fitted model, using the same model matrices, error distribution and fitted parameters, and refit the same model to each simulated sample, obtaining the same type of model diagnostics $\mathbf{r}_{(i)}^s = (r_{1(i)}^s, r_{2(i)}^s)^{\mathrm{T}}$, ordered by the angles they form with the origin. We have, for each bivariate diagnostic $\mathbf{r}_{(i)}$, 99 simulated bivariate diagnostics $\mathbf{r}_{(i)}^s$ (see Figures 1(b) and (c)), forming the whole cloud of simulated diagnostics (see Figure 1(d)). We then obtain the convex hulls of each set of the s sets of points and obtain a reduced polygon whose area is 95% of the original convex hull's area, forming the simulated polygon (see Figure 1(e)). The points are then connected to the centroids of their respective simulated polygons and, if they lie outside the polygons, they are drawn in red (see Figure 1(e)). For the final display, the polygons are erased so as to ease visualization (see Figure 1(f)).

3 Example and discussion

We illustrate our approach with a simple simulated example with a moderate sample size $n = 80$.. Let Y_{1i} and $Y_{2i}, i = 1, \ldots, 80$, be two normally distributed correlated random variables. We may write

$$\mathbf{Y}_i = \begin{bmatrix} Y_{1i} \\ Y_{2i} \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mu_{1i} \\ \mu_{2i} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right), \tag{1}$$

which yields, marginally,

$$Y_{ji} \quad \sim \quad N(\mu_{ji}, \sigma_j^2), \quad j = 1, 2, \tag{2}$$

with $\text{Cov}(Y_{1i}, Y_{2i}) = \sigma_{12}$. Hence, the marginal distributions remain unaltered regardless of σ_{12}, which when equal to zero implies independence between the random variables.

We supposed a simple linear regression of the form $\mu_{ji} = \beta_{j0} + \beta_{j1} x_i$ between both variables and a covariate x_i, which ranges from 1 to 10 in equal steps of approximately 0.114, giving 80 distinct x_i values. We fixed the true parameter values as $\beta_{10} = 2, \beta_{11} = 0.4, \beta_{20} = \beta_{21} = 0.2, \sigma_1^2 = 2, \sigma_2^2 = 3$, and a negative covariance $\sigma_{12} = -1.7$.

We simulated the data in R (R Core Team, 2016) using the function `rmvnorm` from the package `mvtnorm` (Genz and Bretz, 2009) with the seed fixed as 2016 to allow for reproducibility. We then fitted the bivariate model using the BFGS algorithm implemented in the function `optim` twice, first estimating the covariance σ_{12} and then the second time fixing $\sigma_{12} = 0$, which is the same as fitting two separate linear regressions. The parameter estimates were virtually the same for both approaches, with the obvious exception of the covariance estimate. The contribution to the likelihood when estimating the covariance was 43.33 on 1 degree of freedom, which is highly significant, giving evidence that σ_{12} is different from zero, as expected.

We then proceeded to produce the bivariate residual plot with simulation polygons for both model fits. The raw residuals may be defined as

$$r_{ji}^o = y_{ji} - \hat{\mu}_{ji}, \quad \hat{\mu}_{ji} = \hat{\beta}_{j0} + \hat{\beta}_{j1} x_i. \tag{3}$$

We observe that fitting the models independently for each variable yields a poor model fit, with many points lying outside of the simulation polygons when compared with the joint model (see Figure 2). We observe that because of the negative correlation between the variables the expected two-dimensional pattern of the scatterplot of residuals is actually an ellipse (Figure 2(b)) instead of a circle (Figure 2(a)). Hence, even though when inspecting each variable marginally it might seem that the behaviour of the residuals is as expected by the 99 simulations, this is not true when looking at the bivariate nature of these residuals. Here, our proposed approach worked well in identifying these bivariate patterns and was appropriate to aid in model selection and in identifying possible outliers. We do not rule out other forms of diagnostic checking, either analytical or graphical. Our approach represents a potentially helpful framework for assessing goodness-of-fit for bivariate models.

Acknowledgments: Special Thanks to FAPESP (proc no. 2014/12903-8 and 2014/03310-3).

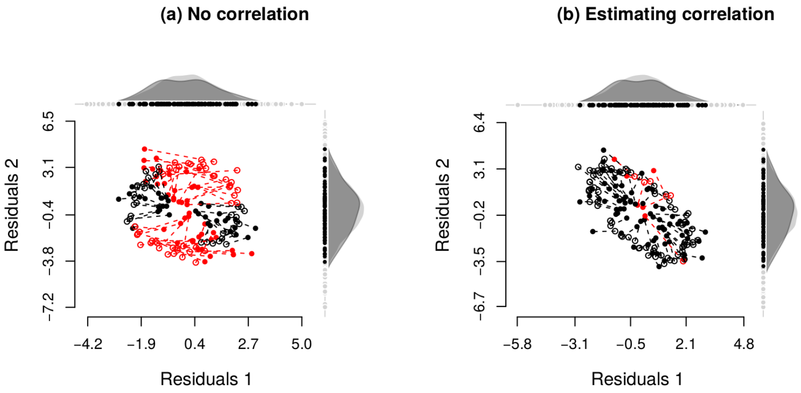


FIGURE 2. Diagnostic plots for the bivariate normal simulated data using raw residuals, while assuming there was no covariance between the two random variables (a) or estimating this covariance with a joint model (b). Filled red points are outside of their respective simulated polygon; there are (a) 45 and (b) 6 points outside of the simulated polygons out of 80 total points.

References

Genz, A. and Bretz, F. (2009). *Computation of multivariate normal and t probabilities* Berlin: Springer.

Moral, R.A., Hinde, J. and Demétrio, C.G.B. (*in press*). Half-normal plots and overdispersed models in R: the `hnp` package *Journal of Statistical Software.*

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Gradient test for variance component models

Antonio Hermes Marques da Silva Junior[12]

[1] Durham University, Durham, UK
[2] Universidade Federal do Rio Grande do Norte, Natal, Brazil

E-mail for correspondence: a.hermes.marques-da-silva-jun@durham.ac.uk

Abstract: Variance component models are an extension for the generalised linear models for group-wise data. The gradient test is an asymptotic likelihood-based test with chi-square as reference distribution and its statistic does not make use of any matrix operation. We motivate this work by a real data application. We propose a bootstrap simulation for checking the accuracy of the chi-square approximation for the gradient test in comparison to the likelihood test. We also show numerical confidence regions based on the inversion of the test.

Keywords: asymptotic test; random effects; random intercepts

1 Motivation data and the model

We take the data from an experiment given by Markussen (2017). It is of interest to investigate how the continuous measurement of redness of pork meat after slaughter is affected by the storage (in light or darkness), by the time (1, 4 or 6 days) and by the breed (old and new, 10 pigs each). Six chops were taken from each pig and allocated according to the scheme shown on Table 1. This gives $2 \times 10 \times 6 = 120$ samples of pork chops in total. Given that the response variable is strictly positive, we consider that the redness measurements y_{ijk} of a given replicate corresponding to the ith breed, the jth storage and the kth time are independently distributed as $\mathcal{IG}(\mu_{ijk}|Z, \phi)$ with means $\mu_{ijk}|Z$ and a fixed dispersion ϕ. We also assume the linear predictor is linked to $\mu_{ijk}|Z$ as

TABLE 1. Factor allocation [source: Markussen (2017)].

Storage	1 days	4 days	6 days
Dark	chop 1	chop 2	chop 3
Light	chop 4	chop 5	chop 6

$$\mu_{ijk}|Z = Z + \alpha_i + \tau_j + \beta_k \quad i = 1, 2, \quad j = 1, 2, \quad k = 1, 2, 3 \tag{1}$$

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

where Z is a random intercept representing the base level for each pig, $\alpha_1 = \tau_1 = \beta_1 = 0$ and (1) is one of the configurations of the variance component model defined in Aitkin et al (2009). Because Z is an unknown random variable, the EM approach in conjunction with the maximum likelihood method can be applied for parameter estimation. We assume that the distribution of Z is unspecified for all the model adjustments and for estimation purposes we used the Nonparametric maximum likelihood (NPML) implementation of Einbeck and Hinde (2006).

2 The gradient test

Consider including in (1) the interaction between storage and time, i.e. testing the null hypothesis $\mathcal{H}_0 : ((\tau\beta)_{22}, (\tau\beta)_{23})^\top = 0$. Let ℓ be the total log-likelihood and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$ the vector of fixed effects parameters where $\boldsymbol{\theta}_1 = ((\tau\beta)_{22}, (\tau\beta)_{23})^\top$ is our vector of parameters of interest and $\boldsymbol{\theta}_2$ is a vector of nuissance parameters. The unrestricted MLE for $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^\top, \hat{\boldsymbol{\theta}}_2^\top)^\top$ and the restricted to the null hypothesis is $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1^{0\top}, \tilde{\boldsymbol{\theta}}_2^\top)^\top$, where $\boldsymbol{\theta}_1^{0\top}$ is an arbitrary vector (which, in our application is equal to 0). From now on the top accents $\wedge$ and $\sim$ represent the MLE unrestricted and restricted to the null hypothesis, respectively. Let $\mathcal{U} = \partial\ell/\partial\boldsymbol{\theta} = (\mathcal{U}_1^\top, \mathcal{U}_2^\top)^\top$ the respective partitioned score vector. Terrell (2002) proposed the gradient statistic for testing $\mathcal{H}_0$ denoted as $\xi_{\mathcal{T}} = \tilde{\mathcal{U}_1}^\top(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^0)$. Note that $\xi_{\mathcal{T}}$ does not have any matrix computation in its formula which turns to be its main advantage. In theory, the reference distribution for $\xi_{\mathcal{T}}$ is χ_q^2 where q denotes the dimension of $\boldsymbol{\theta}_1$. Because of that, $\xi_{\mathcal{T}}$ is comparable to the $\xi_{\mathcal{LR}}$, the likelihood ratio statistic. The gradient test statistic formula for generalised linear models has been defined by Marques da Silva Junior et al (2016) and can be extended for variance component models. Table 2 shows the estimates for the test statistics, the chi-squared p values and the equivalent bootstrap version.

TABLE 2. Likelihood ratio and gradient statistics for the null hypothesis. The p values were computed using the chi-square distribution with two degrees of freedom and p values* uses empirical bootstrap to reproduce the reference distributions.

	likelihood ratio	gradient
Statistic	8.794883	10.25232
p value	0.01230879	0.005939328
p value*	0.01880188	0.00730073

3 Bootstrap and confidence intervals

The main purpose of the bootstrap experiment here is to verify how accurate is the chi-square approximation for the test statistic. We propose therefore a bootstrap in two levels taking the model under null hypothesis as true. In the first level we resample the estimated random intercepts obeying the respective estimated probabilities and in the second level we generate responses given the new intercepts. Then the model in (1) is fitted and both likelihood ratio and gradient statistics are computed. We replicate the procedure 9999 times and the results can be seen in Figure 1. We can produce confidence regions for $\boldsymbol{\theta}_1$ inverting

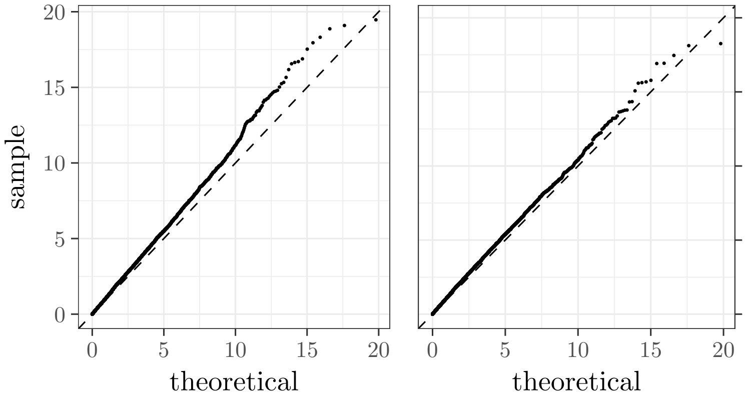


FIGURE 1. Bootstrap samples of the likelihood ratio statistic (left) and gradient statistic (right) compared to the theoretical χ^2_2 for the test with hypothesis $\mathcal{H}_0 : (\tau\beta)_{22} = (\tau\beta)_{23} = 0$.

the gradient test however there is no analytic procedure so far. Numerically, we took a grid of two sequences of 51 values for each $(\tau\beta)_{jk}$ in the interval $(\widehat{\tau\beta})_{jk} \pm 3\widehat{\text{se}}((\widehat{\tau\beta})_{jk})$. Then, we fit the model in (1) with $(\tau\beta)_{jk}$ as offset for each position of the grid and compute $\xi_{\mathcal{T}}$ for $\mathcal{H}_0$. Therefore, the region consists on the values of $\boldsymbol{\theta}_1$ that satisfy $\xi_{\mathcal{T}} < \chi^2_2(1-\alpha)$, where $1-\alpha$ is the confidence level. The same procedure has been done for $\xi_{\mathcal{LR}}$. The Figure 2 shows the contour maps for the 90% confidence regions.

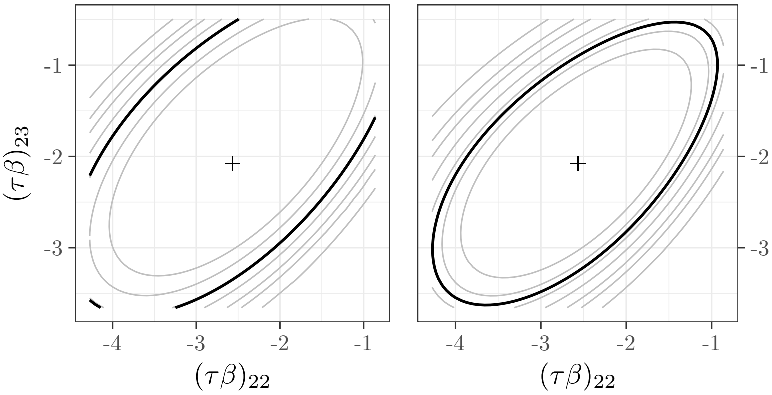


FIGURE 2. 90% confidence regions in black for $(\tau\beta)_{22}$ and $(\tau\beta)_{23}$ based on the numerical inversion of the likelihood ratio test (left) and the gradient test (right).

4 Discussion

The gradient test is an alternative to the classic likelihood ratio test for variable selection on variance component models. Also, the gradient statistic is of simple computation as it does not require matrix operations in its formula. According to our bootstrap simulations, the empirical distribution of the gradient statistic has better approximation to the chi-square than the likelihood ratio statistic. The confidence region using the inverse of the gradient test has smaller area than the one which uses the likelihood ratio test for our real data example. In summary, our results show that the gradient test is preferable for variance component models.

Acknowledgments: We wish to thank J. Einbeck for the discussions and helpful comments. We gratefully acknowledge grant n^o 9622/13-6 from CAPES foundation through Brazil's Science Without Borders Program.

References

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

Einbeck, J. and Hinde, J. (2006). A note on NPML estimation for exponential family regression models with unspecified dispersion parameter. *Austrian Journal of Statistics* **35** 233 – 243.

Markussen, B. (2017). *Lecture notes for Statistical methods for the Biosciences.* Copenhagen: Department of Mathematical Sciences, University of Copenhagen.

Marques da Silva Junior, A.H., Einbeck, J. and Craig, P.S. (2016). Gradient test for generalised linear models with random effects. In J.-F. Dupuy and J. Josse, eds., *Proceedings of the 31st International Workshop on Statistical Modelling*, vol. 1. 213 – 218.

Terrell, G.R. (2002). The gradient statistic. *Computing Science and Statistics*, **34** 206 – 215.

Constrained Mortality Forecast

Carlo G. Camarda[1]

[1] Institut National d'Études Démographiques, Paris, France

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

Abstract: Forecasting mortality can be achieved by parametric models, Lee-Carter variants and smoothing approaches. All of these methods either impose rigid modeling structures or produce implausible outcomes. In this paper, we propose a novel approach that combines well established smoothing model as P-splines and demographic prior information. Specifically, we constrain future smooth mortality patterns to lay within a range of reasonable age-profiles and time-trends, both computed from observed patterns. We enforce these shape constraints by an asymmetric penalty approach on forecast mortality. We illustrate the proposed approach on England & Wales males.

Keywords: Mortality forecast; Smoothing; Demographic constraints; Age-Time patterns; Asymmetric penalty.

1 Introduction

Modelling and forecasting mortality is crucial in epidemiology and population studies as well as in the insurance and pensions industry. In the last few decades, several methodologies have been suggested.
Currie et al. (2004) proposed a model based on two-dimensional P-splines to smooth mortality over age and time. Despite P-splines outperforming all competitors in modeling mortality, this approach suffers from all the issues that a purely data-driven approach reveals when employed for forecasting purposes. Forecast mortality simply follows estimated trends in a slavish manner and mortality structure over age and time is not fully considered in the forecast values. Unreasonable trends from a demographic perspective could consequently emerge. This paper aims to enhance two-dimensional P-splines, incorporating demographic knowledge to allow for abetter performance in forecasting mortality trends.

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

2 Data and model

Suppose that we have mortality data, deaths, and exposures to the risk of death, arranged in two matrices, $\boldsymbol{D} = (d_{ij})$ and $\boldsymbol{E} = (e_{ij})$, each $m \times n_1$, whose rows and columns are classified by age at death, $\boldsymbol{a}$, $m \times 1$, and year of death, $\boldsymbol{y}_1$, $n_1 \times 1$, respectively.

We assume that the number of deaths d_{ij} at age i in year j is Poisson distributed with mean $\mu_{ij}\, e_{ij}$. The value of μ_{ij} is commonly named force of mortality and its estimation is the object of all mortality model. Forecasting mortality aims to reconstruct trends in μ_{ij} for n_2 future years, $\boldsymbol{y}_2, n_2 \times 1$.

In the following we will illustrate the method for England & Wales, males, ages 0-100, years 1960-2013, aiming to forecast up to 2050 (HMD, 2017).

By arranging data as a column vector, that is, $\boldsymbol{d} = \texttt{vec}(\boldsymbol{D})$ and $\boldsymbol{e} = \texttt{vec}(\boldsymbol{E})$, we model our Poisson death counts as follows: $\boldsymbol{\eta} = \ln(E(\boldsymbol{d})) = \ln(\boldsymbol{e}) + \boldsymbol{B}\,\boldsymbol{\alpha}$, where $\boldsymbol{B}$ is the regression matrix computed as a Kronecker product of B-spline bases over the two dimensions: $\boldsymbol{B} = \boldsymbol{B}_{y_1} \otimes \boldsymbol{B}_a$. The coefficients vector $\boldsymbol{\alpha}$ are estimated by a penalised version of the iteratively reweighted least squares (IRWLS) algorithm. In the original paper by Currie et al. (2004), forecasting is treated as a missing value problem, and data and B-spline bases are augmented as follows:

$$\breve{\boldsymbol{E}} = [\boldsymbol{E} : \boldsymbol{E}_2]\,, \qquad \breve{\boldsymbol{D}} = [\boldsymbol{D} : \boldsymbol{D}_2]\,, \qquad \breve{\boldsymbol{B}} = [\boldsymbol{B}_{y_1} : \boldsymbol{B}_{y_2}] \otimes \boldsymbol{B}_a\,,$$

where $\boldsymbol{D}_2$ and $\boldsymbol{E}_2$ are filled with arbitrary future values. If we define a weight matrix $\boldsymbol{V}$:

$$\boldsymbol{V} = \mathrm{diag}(\mathbf{1}_{(m \times n_1)} : \mathbf{0}_{m \times n_2})\,,$$

we can adapted the penalised IRWLS algorithm as follows

$$(\breve{\boldsymbol{B}}^T \boldsymbol{V} \tilde{\boldsymbol{W}} \breve{\boldsymbol{B}} + \boldsymbol{P})\tilde{\boldsymbol{\alpha}} = \breve{\boldsymbol{B}}^T \boldsymbol{V} \tilde{\boldsymbol{W}} \tilde{\boldsymbol{z}}\,, \tag{1}$$

where a difference penalty $\boldsymbol{P}$ enforces smoothness behaviour of mortality over both age and years.

Outcomes from this approach are portrayed as dashed lines in Figure 2. It is clear that, given the optimal amount of smoothness selected by BIC, a plain extrapolation of past trend via the penalised coefficients is not satisfactory.

The aim of this paper is to incorporate prior demographic knowledge by enforcing future mortality to follow data-driven age-profiles and reasonable changes over time. Operating at the shape level, we first compute the derivatives of the fitted linear predictor over the observed years. This can be done by a linear combination of a modified version of the B-splines over the two domains and the estimated coefficients. In formulas:

$$\frac{\partial}{\partial \boldsymbol{a}}\boldsymbol{\eta} = (\boldsymbol{B}_{y_1} \otimes \boldsymbol{C}_a)\,\hat{\boldsymbol{\alpha}} = \boldsymbol{D}_a^{y_1}\,\hat{\boldsymbol{\alpha}}, \qquad \frac{\partial}{\partial \boldsymbol{y}_1}\boldsymbol{\eta} = (\boldsymbol{C}_{y_1} \otimes \boldsymbol{B}_a)\,\hat{\boldsymbol{\alpha}} = \boldsymbol{D}_{y_1}^{a}\,\hat{\boldsymbol{\alpha}}\,,$$

where, for instance, the matrix $\boldsymbol{C}_a$ is given by

$$\boldsymbol{C}_a = \frac{1}{h}\left[B_k^{q-1}(a) - B_{k-1}^{q-1}(a)\right]$$

with h, q and k being the knot-distance, degree and positions of $\boldsymbol{B}_a$.

Figure 1 presents the boxplots of the derivatives of the mortality patterns for each age over both ages and years. The idea is to incorporate this information

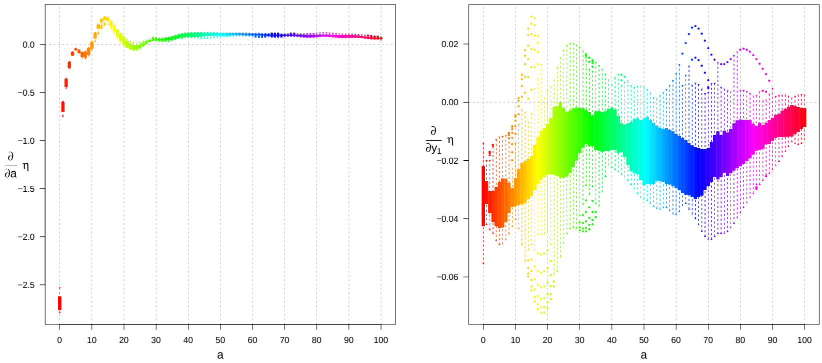


FIGURE 1. Boxplots for each age of the derivatives with respect to ages (left) and years (right). England & Wales, males, ages 0-100 years 1960-2013.

in the model for constraining future mortality to lay within a range plausible age-patterns and time-trends.

Due to space restriction, we describe the model for the age-pattern. Let us denote by $\boldsymbol{\delta}_L^a$ and $\boldsymbol{\delta}_U^a$ the 99% confidence intervals of the derivatives with respect to ages, and by $\boldsymbol{g}_L^a = \mathbf{1}_{n_1+n_2} \otimes \boldsymbol{\delta}_L^a$ the repetition of these values over both dimensions. We can enforce our shape constraints by adding two asymmetric penalties within the system described in (1) (Bollaerts et al., 2006):

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

As example, the penalty terms for the lower bounds are given by

$$\begin{aligned} \boldsymbol{P}_L^a &= \kappa \, \boldsymbol{D}_a^{y_1+y_2\,T} \, \mathtt{diag}(\boldsymbol{s} \cdot \boldsymbol{v}_L^a) \, \boldsymbol{D}_a^{y_1+y_2} \\ \boldsymbol{p}_L^a &= \kappa \, \boldsymbol{D}_a^{y_1+y_2\,T} \, \mathtt{diag}(\boldsymbol{s} \cdot \boldsymbol{v}_L^a) \, \boldsymbol{g}_L^a \end{aligned}$$

where

$$\boldsymbol{v}_L^a = \begin{cases} 0 & \text{if} \quad \boldsymbol{D}_a^{y_1+y_2} \tilde{\boldsymbol{\alpha}} \geqslant \boldsymbol{g}_L^a \\ 1 & \text{if} \quad \boldsymbol{D}_a^{y_1+y_2} \tilde{\boldsymbol{\alpha}} < \boldsymbol{g}_L^a \end{cases} ,$$

and $\boldsymbol{s}$ is a 0/1 vector equal to 1 when the constraint is to be applied (future years). The size of κ regulates how strictly the constraints are enforced. In this paper, we chose $\kappa = 10^4$. Similar reasoning is applied for the rate-of-change over years.

3 Results

Figure 2 presents the outcomes of the proposed model. For comparison we add the outcomes from a plain P-spline approach and a smooth Lee-Carter variant (Delwarde et al., 2007). Our proposed approach outperforms Lee-Carter model in fitting observed data and it provides reasonable outcomes in future years.

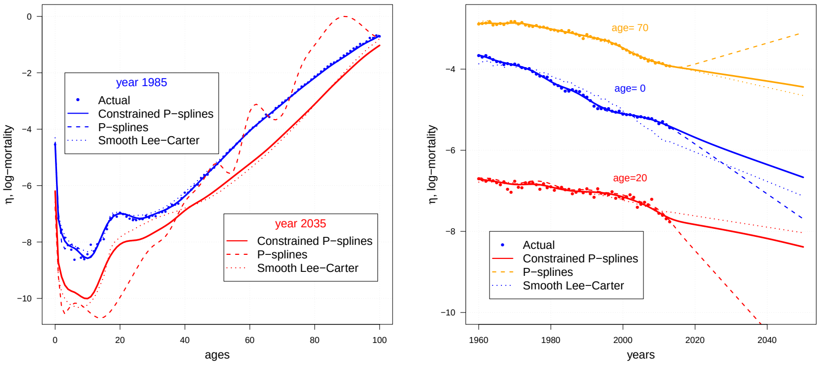


FIGURE 2. Actual, modelled and forecast death rates (log-scale) over ages for selected years (left) and over years for selected ages (right) by the proposed model as well as by plain P-spline approach and a smooth Lee-Carter variant. England & Wales, males, ages 0-100 years 1960-2013, forecast up to 2050.

References

Bollaerts, K., Eilers, P. H. C., and van Mechelen, I. (2006). Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, **59**, 451-469.

Currie, I. D., Durbán, M. and Eilers, P. H. C. (2004). Smoothing and Forecasting Mortality Rates. *Statistical Modelling*, **4**, 279-298.

Delwarde, A., Denuit, M., and and Eilers, P. H. C. (2007). Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: A penalized log-likelihood approach. *Statistical Modelling*, **7**, 29-48.

Human Mortality Database (2016). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at `www.mortality.org`. Data downloaded on January 2017.

Prior Specifications to Handle Monotone Likelihood in the Cox Regression Model

Enrico A. Colosimo[1], Frederico M. Almeida[1], Vincius D. Mayrink[1]

[1] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

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

Abstract: The phenomenon of monotone likelihood is observed in the fitting process of a Cox model when the likelihood converges to a finite value while at least one parameter estimate diverges to infinity. A penalty solution suggested by Heinze and Schemper (2001) leads to finite parameter estimates by means of penalized maximum likelihood estimation. In this paper, we explore other penalties for the partial likelihood function in the flavor of Bayesian prior distributions. This work was motivated by a real situation involving a melanoma skin data set.

Keywords: Firth correction, MCMC, Partial likelihood, Survival analysis.

1 Introduction

The proportional hazards model (PHM) is probably one of the most important statistical methods for the analysis of censored data. When fitting the PHM to some data sets, one may observe a phenomenon known as monotone likelihood or separation (Bryson and Johnson, 1981). The monotone likelihood tends to occur associated to one category of a categorical covariate. In this sense, the larger the number of dichotomous regressors included in the model, the higher is the chance of monotone likelihood.

This paper was motivated by the analysis of a melanoma data set (Cherobin, et al., 2017). Melanoma is a neoplasm that shows high mortality when diagnosed in advanced stages. The PHM analysis was used to assess factors associated with time until metastasis occurrence. However, presence of mitosis, one of most important covariate, had lack of metastasis in those tumors without mitosis in the histologic exam; this means the occurrence of monotone likelihood.

A solution suggested by Heinze and Schemper (2001) is based on the procedure of Firth (1993). This method produces finite parameter estimates by means of penalized maximum likelihood estimation. Penalization is a very general method

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

of stabilizing estimates, which has both frequentist and Bayesian rationales. Firth method is a well known example of penalty which can be derived from a Jeffreys type of prior in Bayesian inference. However, this approach has some drawbacks, especially biased estimators and high standard errors (Greenland and Mansournia, 2015).
The goal of this paper is to propose and compare others penalties for the partial likelihood function in the flavor of prior distributions in the Bayesian context.

2 Notation and Cox Regression Model

The Cox regression model uses the exponential formulation for the hazard function:

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

in which $\lambda_0(t)$ is a baseline hazard function (an unknown non-negative function of time), $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters (to be estimated), and $\mathbf{x} = (x_1, x_2, \ldots, x_p)^{\mathrm{T}}$ is a covariate vector.
The estimation of coefficients $\boldsymbol{\beta}$ in Eq. (1) is based on the partial log-likelihood function:

$$l(\boldsymbol{\beta}) = \mathrm{logL}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \log\left(\sum_{j\in R_{(t_i)}} \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)\right)\right], \tag{2}$$

in which $R_{(t_i)} = \{k:\ t_k \geq t_i\}$ is the risk set at time t_i, δ_i is the failure indicator ($\delta_i = 1$ means failure and $\delta_i = 0$ means censoring), and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ corresponds to the covariate row vector for the i-th individual. The Maximum Partial Likelihood Estimator (MPLE; $\widehat{\boldsymbol{\beta}}$) of $\boldsymbol{\beta}$ is obtained by maximizing Eq. (2).
The Firth method for bias reduction estimates $\boldsymbol{\beta}$ as the maximum of the $l(\boldsymbol{\beta}) = \mathrm{logL}(\boldsymbol{\beta})$ penalized by $r(\boldsymbol{\beta}) = \log|I(\beta)|^{-1}$. That is,

$$l^*(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + (1/2)\log|I(\boldsymbol{\beta})|.$$

The augmenting term $1/2\log|I(\boldsymbol{\beta})|$ is the log of a Jeffreys prior, apart from a constant, and thus the maximizer of $l^*(\boldsymbol{\beta})$ is the posterior mode given this prior. Other penalties structures have been proposed recently by Greenland and Mansournia (2015). In this paper, the main interest is to evaluate the properties of the Jeffreys prior and propose other priors in the monotone partial likelihood situation.

3 Prior Distributions and Bayesian Approach

Partial likelihood modeling version based on Eq. (2) can be used for a Bayesian analysis. The Bayesian inference using the Cox's partial likelihood is the topic of Sinha *et al.* (2003). According to the authors the regression coefficients can be well estimated in this situation.
The main focus of our study is to investigate the impact of different prior specifications for the coefficients associated with the covariates included in the Cox regression model used to fit a data set configured with the monotone likelihood scenario; the following cases are investigated:

- $\beta_i \sim \mathrm{N}(m, v)$, with mean m variance v, for $i = 1, \ldots, p$. Typically, one may choose $m = 0$ and $v > 0$ small/large to explore informative/vague initial uncertainty.
- $\beta_i \sim \text{Log-F}(l_1/2, l_2/2)$ with l_1 and l_2 being the degrees of freedom of the original F distribution. Higher variability is associated with small values of l_1 and l_2.

The joint posterior distribution of $\boldsymbol{\beta}$, given the observed time points and covariates, does not have a closed form, i.e., a proper probability density cannot be identified via Bayes' rule in this case. Markov Chain Monte Carlo (MCMC) algorithm called Gibbs Sampler methods are required to sample from this unknown target distribution.

4 Melanoma Data Revisited

The prognostics factors under consideration are: (i) Gender β_2 (*Female*); (ii) Histological Type (HT) with the following levels β_{31} (*Nodular*), β_{32} (*Acral lentiginous*) and the extensive malign+superficial spreading level assumed as reference; (iii) Breslow index (CB) with levels β_{41} ($1-4mm$), β_{42} ($> 4mm$) and ($< 1mm$) assumed as the reference level; (iv) Ulceration β_5 (*Yes*). The fifth factor, β_1 (mitotic rate or mitosis), is the one associated with the monotone likelihood issue.

	Standard approach		Bayesian approach				
	Cox	Firth	$N(0,1)$	$N(0,5)$	log-F$(1,1)$	log-F$(2,2)$	log-F$(9,9)$
$\hat{\beta}_1$	18.52	2.2214	1.4033	2.7724	3.6909	2.4310	1.0801
$\hat{\beta}_2$	-0.651	-0.6414	-0.589	-0.6327	-0.6298	-0.6203	-0.531
$\hat{\beta}_{31}$	0.984	0.9578	1.0376	1.0770	1.0487	1.0465	0.9940
$\hat{\beta}_{32}$	0.236	0.2978	0.1721	0.2332	0.2152	0.2114	0.1488
$\hat{\beta}_{41}$	0.110	1.0356	0.7312	1.0612	1.0602	0.9221	0.5263
$\hat{\beta}_{42}$	0.184	1.7598	1.4229	1.7823	1.7983	1.6486	1.2067
$\hat{\beta}_5$	0.699	0.6848	0.7925	0 .7594	0.7332	0.7656	0.7844

TABLE 1. Parameter estimates for the Melanoma data set.

Tables 1 and 2 present the main results. As expected, the estimates for the parameters not related to the monotone likelihood issue are very similar for all cases. The most reliable prior distributions, $N(0, 5)$ and log-F$(1, 1)$ (based on simulation studies, not presented), provide estimates larger than those of Firth correction, and standard errors slightly smaller than the one for the standard approach. Indeed, Mitosis is not an important factor according to the Firth method ($z = 2.2214/1.5003$) and it is closely significant when using the Bayesian approach with prior $N(0, 5)$.

References

Bryson, M. C., Johnson, M. E. (1991). The Incidence of Monotone Likelihood in the Cox Model. *Technometrics*, **23**, 381–383.

	Standard approach		Bayesian approach				
	Cox	Firth	$N(0,1)$	$N(0,5)$	log-F$(1,1)$	log-F$(2,2)$	log-F$(9,9)$
$\hat{\sigma}_1$	5131.0	1.5003	0.6956	1.4272	2.3215	1.4229	0.5599
$\hat{\sigma}_2$	0.356	0.3554	0.3342	0.3537	0.3484	0.3479	0.3149
$\hat{\sigma}_{31}$	0.449	0.4511	0.3872	0.4535	0.4436	0.4405	0.3667
$\hat{\sigma}_{32}$	0.634	0.6183	0.5334	0.6292	0.6071	0.5871	0.4515
$\hat{\sigma}_{41}$	0.692	0.6657	0.4875	0.6625	0.6135	0.5894	0.4021
$\hat{\sigma}_{42}$	0.734	0.7147	0.5301	0.6953	0.6505	0.6318	0.4329
$\hat{\sigma}_5$	0.385	0.3872	0.3574	0.3910	0.3859	0.3756	0.3390

TABLE 2. Standard error estimates for the Melanoma data set.

Cherobin, A. C. F., Wainstein, A. J., Colosimo, E. A., Goulart, E. A. (2017). Prognostic factors for metastatic melanoma *An. Bras. Derm.* (in press).

Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, **80**, 27–38.

Greenland, S., Mansournia, M. A. (2015). Penalization, Bias Reduction and Default Priors in Logistic and Related Categorical and Survival Regressions. *Statistics in Medicine*, **34**, 23, 3133–3143.

Heinze, G., Schemper, M. (2001). A Solution to the Problem of Monotone Likelihood. *Biometrics*, **57**, 1, 114–119.

Sinha, D., Ibrahim, J. G., Chen, M. H. (2003). A Bayesian Justification of Cox's Partial Likelihood. *Biometrika*, **90**, 3, 629–641.

Spatially Balanced Sampling

Jennifer Brown[1], Blair Robertson[1], Trent McDonald[2]

[1] University of Canterbury, New Zealand
[2] Western Ecosystems Technology Inc, USA

E-mail for correspondence: Jennifer.Brown@canterbury.ac.nz

Abstract: Spatially balanced sampling is an emerging area in statistical sampling. These designs are popular because they are one way to ensure the selected sample has spatial coverage over the entire survey area. This feature of spatial coverage aids in the resultant sample being representative of the population of interest.

One of the first and the most commonly used spatially balanced design is called GRTS (Generalized Random Tessellation Stratified sampling) where sample effort is spread evenly over the target region. The term spread evenly in this context means having coverage of survey effort over the region. The coverage from GRTS has a stochastic component rather than a fixed interval, regularly spaced coverage as in a systematic sampling design.

We have extended the idea of GRTS to a new design called Balances Acceptance Sampling (BAS). The BAS design allows surveys to be balanced in dimensions higher then two (n - dimensional space). Until now, most designs have considered balance in 2-D geographic space. With BAS we can achieve balance in 3-D space, or in higher dimensions. In some applications these dimensions can be features other than the spatial measures of geographic location, and the design allows aspects such as time for repeat surveys to be incorporated into sample balance.

Keywords: Sample; Environmental Sampling; Spatial Sampling.

1 Introduction

Spatial sampling is a broad category referring to designs that incorporate spatial reference in site selection. The purpose of sampling is usually to estimate a population parameter, such as the mean or total of some characteristic of interest. For example, in environmental applications surveys may be to estimate the density or abundance of a plant species of interest. In social science applications the survey may be to estimate household income within a city. Interest in spatial sampling has increased because of the availability of georeferenced data, computation capacity, and the overarching need for surveys to be cost-efficient.

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

The gold standard of survey designs is simple random sampling (Cochran 1977), where selection of individual units in the sample is at random. In the above examples, individual units could be a quadrat or plot within which plants are counted, or, individual households from which household income is recorded. A spatial sample would appear very similar to a simple random sample except that the selection process would use information on the spatial location of the sample units (Wang et al. 2012).

A feature of many natural and social systems is that characteristics observed in one area are likely to be similar to the ones observed in an adjacent area. For example, the counts of plants in one plot will be similar to the counts in an adjacent plot (Grafström et al. 2012, Stevens and Olsen 2004). This is because of the underlying biotic and abiotic processes that drive the species distribution. In human populations individual neighbourhoods tend to have more similar household incomes than are observed among neighbourhoods, because of socio-economic factors. One way to view this is that there is limited new information provided from a sample unit that is spatially adjacent to another unit that has already been measured. This motivated development of survey designs where the sample is evenly, or near to evenly, spread over the study area. A design that generates samples that are well-spread over the population is called a spatially balanced survey design.

There are many different methods of spatially balanced survey designs (Wang et al. 2012). The design that stimulated the most interest in spatially balanced sampling, and began the use of this phrase, is Generalized Random Tessellation Stratified sampling (GRTS), developed by Stevens and Olsen (2003). In this design an invertible mapping technique is used to transform two-dimensional space into one-dimensional space. Then, a systematic sample is selected along the linear representation. Sampling location geo-references are generated from selecting points at regular intervals in this one-dimensional space (Brewer and Hanif 1983). The one-dimensional space is then mapped back to the two-dimensional original space. By maintaining the spatial properties of the original units, the resultant sample is spatially balanced, with neither no one area being over-represented with high sample intensity nor under-represented with low sample intensity.

2 Other spatially balanced designs

Local Pivotal Method (LPM) is a design for spatially balanced sampling (Grafström et al. 2012). The method is based on a method introduced by Deville and Till (1998). The algorithm involves sequentially updating each point's inclusion probability. Starting with N points, neighbouring points compete to be included in the sample. The winning point of each competition has its inclusion probability increased and the losing point has its decreased. Eventually n points have inclusion probabilities of one and N-n points have inclusion probabilities of zero. The resultant sample will have points separated spatially because it is unlikely two adjacent points will be included in the final sample of size n. The design can be computationally heavy with large N, and an alternative method, suboptimal LPM, was suggested for these situations (Grafström et al. 2014), where only a subset of neighbours are used for the comparison, rather than all possible points.

Balanced acceptance sampling, BAS, (Robertson et al. 2013) is a more recent design for spatially balanced sampling. The method uses the Halton sequence, a quasi-random number sequence (Halton 1960). In two-dimensional geographic space, Halton points are used to generate the geo-referenced locations of the sample units (starting from a randomly chosen position in the sequence). The design uses acceptance/rejection sampling to select sample units. If a generated sample point is beyond the edge of the sample space the sample unit is rejected, otherwise it is accepted. The design is straightforward computationally, and has better spatial balance than the comparable GRTS design (Robertson et al. 2013). The algorithm can be extended into more than two dimensions (e.g., up to five), and this is an appealing feature for some survey situations. One application is for surveys that involve repeat samples of the same population. Environmental monitoring is an example, where interest is in how population parameters change over time. This involves repeat visits to the survey area, often on an annual basis. Other examples can be in social economic surveys where there is interest in how indicators are changing over time. In these situations, the design can be viewed as having three dimension, the two dimensional georeferenced space and the third dimension that is time. Three dimensional spatial balance can be thought of as a way of ensuring that for any one survey there is spatial balance, and no one area is excessively over- or under-sampled. In addition, over the course of the repeat surveys (e.g., annual surveys for 10 years), there is no one area that is excessively repeatedly over - or under-sampled. Until now many survey designs rely on fixed intervals between site revisits, and here BAS offers a method for randomising the repeat interval.

Acknowledgments: Special Thanks to Peter Jaksons and Naeimeh Abi for their involvemnt in ongoing research, and helpful suggestions.

References

Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities. Lecture Notes in Statistics.* New York: Springer-Verlag.

Cochran, W.G. (1977). *Sampling techniques.* 3rd edition. John Wiley & Sons.

Deville, J.C. and Tille, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, **85**, 89 – 101.

Grafström , A, Saarela, S. and Ene, L. T. (2014).Efficient strategies for forest inventories by spreading the sample in auxiliary space. *Canadian Journal of Forest Research*, **44**, 1156 – 11164.

Grafström , A, Lundstom, N.L.P. and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, **68**, 514 – 520.

Halton, J.H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematic*, **2**, 84 – 90.

Robertson, B.L., Brown, J.A., McDonald, T. and Jaksons, P. (2013). BAS: Balanced acceptance sampling of natural resources. *Biometrics*, **69**, 776 – 784.

Stevens, D.L. and Olsen, A.R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, **14**, 593 – 610.

Stevens, D.L. and Olsen, A.R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, **99**, 262 – 278.

Wang, J.F., Stein, A., Gao, B.B. and Ge Y. (2012). A review of spatial sampling. *Spatial Statistics*, **2,** 1 – 14.

A comparison of wind power forecast models

Fraser Tough[1]

[1] Renewable Energy Systems

E-mail for correspondence: fraser.tough@res-group.com

Abstract: The forecast of wind speed and subsequent generated power is required to efficiently balance the electricity grid. Within this paper, a number of forecast models are applied to generate wind power predictions at a location in Scotland, UK. Conventional statistical models are compared with more complex machine learning techniques using all historic data as well as subsets of historic data. Local on-site variables collected from the turbines as well as Numerical Weather Predictions (NWP) are utilised within the framework.

Keywords: wind power forecasting, auto-regressive integrated moving average; generalised additive model; linear model; regularisation; neural network; auto-regressive neural network

1 Introduction

Once energy is generated it must be used immediately as it cannot be stored in large quantities. Supply must constantly meet demand to prevent outages. Therefore, grid operators must plan and balance the network accordingly. Conventional energy sources such as coal and gas can be controlled to a large degree and grid operators can request that operators of these power plants ramp-up or ramp-down output in real time. However, for renewable energy sources, the energy produced cannot be controlled to the same degree, although turbines can be shut down via curtailment to reduce power output. To help manage the grid, wind farm operators must supply energy production forecasts so that the grid can be efficiently balanced.

This paper compares a number of forecasting techniques applied to a 30 megawatt (MW) wind farm in Scotland, UK. The models are assessed using Mean Absolute Error (MAE) over a forecast horizon of 24 half hour periods.

2 Data and methodology

The data available for forecasts can be grouped into two categories: on-site historical data via Supervisory Control And Data Acquisition (SCADA), and Numerical Weather Prediction (NWP) via an external provider.

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.

SCADA data is collected via an automated system at the wind farm location. Wind speeds, power output and turbine availability are recorded on a 10 minute resolution then aggregated to half hours. From SCADA, lagged variables are of use for predictions as are time based variables such as month of the year and hour of the day which can be used to represent seasonal variation via a Fourier series.

NWP utilises mathematical models to make atmospheric forecasts via computer simulation. Forecasts are generally made at nodes which may be close but not at the wind farm location, at a certain height above the ground. Therefore, NWP wind speed forecasts may not reflect the actual weather conditions at the site and so are corrected utilising historical SCADA data via the statistical/machine learning models outlined in Section 3. Once relationships between the outcome variable and the NWP explanatory variables are established, the NWP forecasts and lagged SCADA data can be fed into the model to make predictions in real time.

NWP forecasts	Historical SCADA data	Time based
Wind speed	Wind speed	Time of day
Wind direction	Turbine availability	Month
Temperature	Power	
Pressure		

TABLE 1. Variables

Note that as future turbine availability is required, wind farm operators must submit this in advance of any forecasts.

3 Models

A number of methods were selected for model comparisons. Models were trained on data from the start of 2013 to the end of 2015 using k-fold cross validation for variable selection. Data from 2016 was used to compute MAEs, as shown in Figure 2.

- Persistence
- Linear models (LM)
- Linear models with elastic net regularisation (RLM)
- Auto-Regressive Integrated Moving Average (ARIMA)
- Generalised Additive Models (GAM)
- Neural Networks (NN)
- Auto-Regressive Neural Networks (ARNN)

The persistence/naive model is a commonly used benchmark, replicating the last known power such that $\hat{P}_{t+k|t} = P_t$. It is generally hard to improve upon the persistence model when forecast horizons are less than a few steps ahead. Powers of variables as well as interactions were used within the both the LMs and RLMs to quantify non-linear relationships. ARIMA models with exogenous variables (ARIMAX) were also applied but showed no value over linear models including AR terms and NWP inputs. GAM smooth terms were represented with thin plate regression splines. Finally, there was one hidden layer within the NN models and the number of nodes were chosen to minimise the MAE.

4 Analysis

Power can be predicted in two ways, a one stage process of forecasting power directly, or a two stage process, forecasting wind speed then applying a transformation to convert to power. Converting wind speed to power can be achieved via the theoretical power curve, or by estimating the on-site power curve while taking into account local conditions (Stathpoulos et al., 2013). The approach within this paper utilises only the latter, modelling the relationship between on-site wind speed, power and turbine availability via a Gaussian Process (GP) (see Figure 1) as it was found that forecasting power directly resulted in higher MAEs. The approach leaves open the opportunity to include other variables. This may be useful for forecasts where the transformation from wind speed to power is dependent on wind direction, due to the surrounding terrain, for example.

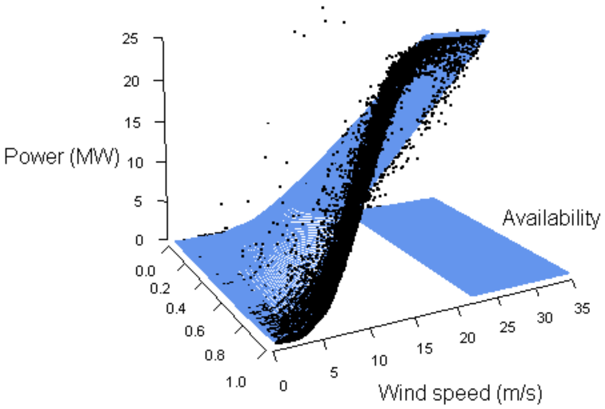


FIGURE 1. On-site power curve by availability with superimposed GP model

As wind speeds are Weibull distributed, the response variable was transformed to approximate normality by firstly fitting a Weibull distribution to the historical wind speed then dividing 3.6 by the shape parameter, α, rendering the data approximately normal (Dubey, 1967). Forecasts were back transformed as necessary.

Figure 2 shows MAEs by forecast horizon. It is clear that models based only on auto regressive terms (ARIMA and ARNN) do not perform as well as those which include NWP forecasts, out performing the naive approach only towards latter forecast horizons. The GAM does not perform as well as the others which may infer that random variation in the training dataset may have been modelled. RLM models outperform standard LMs and using more recent data (RLM1 and LM1) resulted in higher MAEs over all horizons. Note that the effect of regularisation is more prominent while computing parameters with the smaller dataset. Finally, the NNET model performed the best, outperforming all others over all forecast horizons.

5 Summary

The results show that all models tested performed well, with MAEs within roughly 100kW of each other by horizon 24. While the NNET model did outperform the others, it was only marginally better, is much more complex to interpret and takes much longer to train. This shows that the relationships between wind speed/power and the explanatory variables can be modelled adequately without

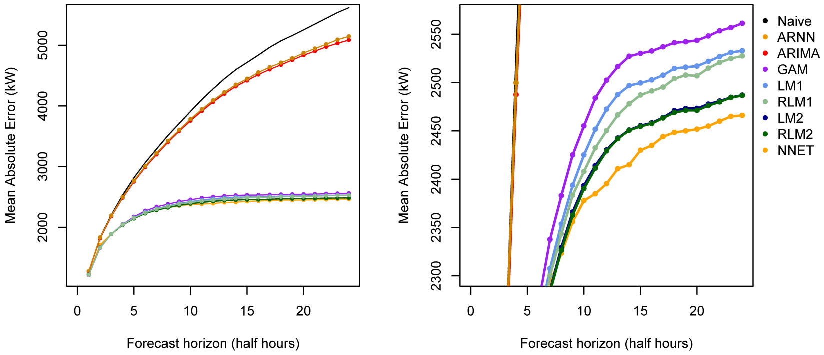


FIGURE 2. MAE by forecast horizon

the need for complex black box models such as neural networks. This can be carried out using linear models with or without regularisation, utilising powers of the input variables and interactions between them.

References

Stathopoulos, C., Kaperoni, A., Galanis, G., & Kallos, G. (2013). Wind power prediction based on numerical and statistical models. *Journal of Wind Engineering and Industrial Aerodynamics,* **112**, 2538.

Dubey, S. Y. D. (1967). Normal and Weibull distributions. *Naval Research Logistics (NRL),* **14(1)**, 69-79.

Beyond regression: what does really affect a football team on championships

Luiz R. Nakamura[1], Pedro H.R. Cerqueira[2], Thiago G. Ramires[23], Rodrigo R. Pescim[4], Robert A. Rigby[5], Dimitrios M. Stasinopoulos[5]

[1] Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Brazil
[2] Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Brazil
[3] Institute for Biostatistics and Statistical Bioinformatics, University of Hasselt, Belgium
[4] Departamento de Estatística, Universidade Estadual de Londrina, Brazil
[5] Centre of Communications Technology and Mathematics, London Metropolitan University, United Kingdom

E-mail for correspondence: `luiz.rn@gmail.com`

Abstract: In this paper we introduce a new regression model, based on the generalized additive models for location, scale and shape, in order to explain the points rate of football teams in the end of a championship from the four most important leagues in the world: Barclays Premier League (England), Bundesliga (Germany), Serie A (Italy) and BBVA league (Spain) during three different seasons (2011-2012, 2012-2013 and 2013-2014).

Keywords: Collective sports; GAMLSS; Match analysis; Soccer.

1 Introduction

Football can be considered the most popular and important collective sport in the world. This importance reflects on several aspects, among them the economic, since a lot of money is invested on football in different areas, such as betting, sponsorship, television rights and in particular the organization of major events. In this manner an issue that becomes paramount for managers is the assessment of which factors affect positive outcomes. To evaluate which variables might have such relevance, we model the final points rate results (i.e. the proportion of points

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.

scored to the maximum points possible) of each team in different national championships using the measured explanatory features. In order to do this task, we present a new distribution on support $(0, 1)$, called the *logit*SHASHo distribution and develop its regression model based on the generalized additive models for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005). We found that the new model outperformed the beta regression model and some other models on support (0,1).

2 GAMLSS framework

GAMLSS is a very flexible class of semi-parametric regression models that involves a distribution for the response variable and may involve parametric and/or non-parametric smoothing terms when modelling any or all of the parameters of the distribution as functions of a set of explanatory variables. This methodology is already implemented in the `gamlss` package (Stasinopoulos and Rigby, 2007) in `R`, which includes several distributions with up to four parameters (conveniently denoted by μ, σ, ν and τ).

In this work we will use only the parametric version of GAMLSS. Generically, let $Y \sim \mathcal{D}(\boldsymbol{\theta})$, where $\mathcal{D}$ represents the response variable distribution and $\boldsymbol{\theta}$ is its vector of parameters of length 4. The parametric GAMLSS model is given by $g_k(\theta_k) = \eta_k = \boldsymbol{X}_k\boldsymbol{\beta}_k$, for $k = 1, 2, 3, 4$, where $g_k(\cdot)$ is a known monotonic link function relating the distribution parameter θ_k to the predictor η_k, $\boldsymbol{X}_k$ is a known design matrix and $\boldsymbol{\beta}_k$ is a parameter vector of length J'_k.

2.1 The *logit*SHASHo distribution

As we can see in Hossain et al. (2016), any distribution on the range $-\infty < Z < \infty$ can be transformed to a restrictive range $0 < Y < 1$ using an inverse logit transformation, i.e. $Y = (1 + \exp\{-Z\})^{-1}$. In this paper we propose a new very flexible distribution to model response variables on the interval from zero to one, with four parameters using the logit transformation.

If $-\infty < Z < \infty$ follows a sinh-arcsinh distribution (Jones and Pewsey, 2009), then $Y = (1 + \exp\{-Z\})^{-1}$ will follow a logit sinh-arcsinh distribution, denoted by $Y \sim$ *logit*SHASHo(μ, σ, ν, τ) for $0 < Y < 1$, $-\infty < \mu < \infty$ is the location parameter, $\sigma > 0$ is the scale parameter, $-\infty < \nu < \infty$ determinates the skewness of the distribution (positive skewness corresponds to $\nu > 0$) and $\tau < 1$ and $\tau > 1$ yield heavier and lighter tails than the normal distribution, respectively.

3 Application

3.1 Data set

The following information about four European leagues (Barclays Premier League, from England, Bundesliga, from Germany, Serie A, from Italy and BBVA league, from Spain) were considered: points rate (response variable) and the covariates season, league, shots (Sh.Pg), shots on goal (ShG.Pg), clean sheet (ClSh), offsides (Off.Pg), dribbles (Drbl.pg), fouled, shots conceded (Sh.con.Pg), tackles (Tack.Pg), interceptions (Int.Pg), fouls (Fou.pg), yellow cards (YC), red cards (RC), possession (Poss) and passes (Pass).

3.2 Results and discussion

A stepwise selection of explanatory terms was performed for all parameters and values of global deviance (GD), Akaike information criterion (AIC) and Bayesian information criterion (BIC) were computed in order to compare all fitted models using the following distributions: *logit*SHASHo, *logit*SST, *logit*NO, beta and generalised beta. The results are displayed in Table 1 where we can see that the *logit*SHASHo model is the best fitted model when we use the GD and AIC values (-656.28 and -636.28, respectively), while the beta model is prefered when we use the BIC criterion.

TABLE 1. Statistics from the fitted models.

Model	Parameters	df	GD	AIC	BIC
*logit*SHASHo	4	10	**-656.28**	**-636.28**	-601.72
Beta	2	7	-645.19	-631.19	**-607.00**
*logit*SST	4	8	-646.19	-628.19	-597.09
*logit*NO	2	9	-644.72	-626.72	-595.62
Generalised beta	4	9	-642.48	-624.48	-593.38

As we will see in the residual plots, the *logit*SHASHo model is prefered. Hence, the final model from the *logit*SHASHo distribution under the GAMLSS parametric framework is given by

$$\begin{aligned}
\hat{\mu} &= -0.19 + 0.49 \text{ ShG.Pg} + 0.29 \text{ ClSh} - 0.13 \text{ Sh.Pg} - 0.05 \text{ Sh.con.Pg} \\
\hat{\sigma} &= \exp\left[-1.96 + 0.13 \text{ ShG.Pg}\right] \\
\hat{\nu} &= 0.07 - 0.12 \text{ YC} \\
\hat{\tau} &= 0.68.
\end{aligned}$$

ShG.Pg and ClSh are positively associated to μ. This relationship can be explained using simple game facts: a team that avoids being scored (ClSh) and creates real chances of scoring (ShG.Pg) has more chances to be successfully. Sh.Pg, surprisingly, has a negative effect on the location parameter μ. Statistically, this can be explained by using the value of the partial correlation between Sh.Pg and points rate given ShG.Pg, which is negative (-0.11). In practice, we may say that just arbitrarily shooting has a negative effect, since in many cases a football team trying to score from anywhere in the pitch, without presenting any real risk for the opposite team is less likely to score. The last covariate used to model μ is Sh.con.Pg which has a negative effect, since it increases the chances to concede goals and consequently to lose the match.

The value of the scale parameter σ increases according to the number of ShG.Pg by the football teams. Further, YC has a negative linear effect on the skewness parameter ν. Finally, the kurtosis parameter τ is a constant smaller than one, i.e. the final model from the *logit*SHASHo distribution presents heavy tails.

The worm plot for the residual analysis (Figure 1 (a)), for the model based on the *logit*SHASHo distribution does not present any trend (vertical shift, slope, quadratic or cubic shape), thus it fitted really well the skewness and kurtosis present in the response variable. As for the beta regression model, we can see from Figure 1(b) that the residuals present a cubic shape, indicating possible problems in the kurtosis, and three of the dots are not between the upper dotted curve, which act as 95% pointwise confidence interval. Hence, we conclude that the *logit*SHASHo model is the best model between the five used in this paper to explain the current data.

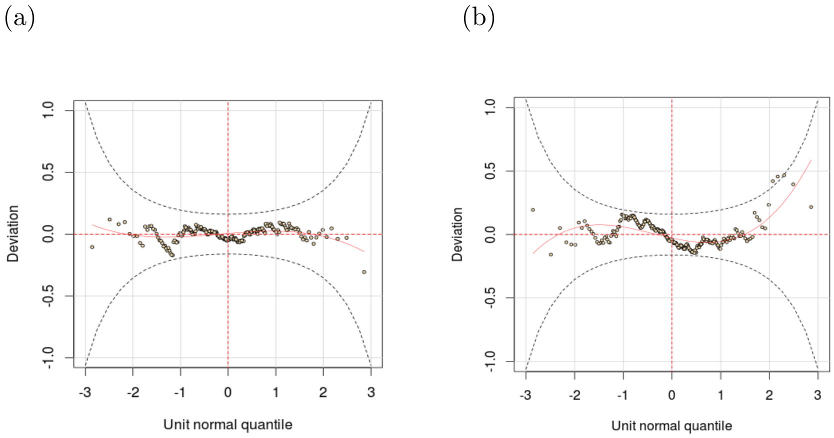


FIGURE 1. Worm plots of the (a) *logit*SHASHo and (b) beta GAMLSS models

References

Hossain, A., Rigby, R., Stasinopoulos, M. and Enea, M. (2016). Centile estimation for a proportion response variable. *Statistics in Medicine*, **35**, 895 – 904.

Jones, M.C. Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, **96**, 761 – 780.

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

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1 – 10.

A Continuous Time Multi-State model for Prostate Cancer Screening

Tiago M. de Carvalho [1], Ardo van den Hout [2], Jonathan Miles [3], Stephen Duffy [3], Nora Pashayan [1]

[1] Department of Applied Health Research, University College London, UK,
[2] Department of Statistical Science, University College London, UK
[3] Wolfson Institute of Preventive Medicine, Queen Mary's University London, UK

E-mail for correspondence: `t.marques@ucl.ac.uk`

Abstract: We use a continuous-time multi-state model to describe the natural history of prostate cancer. We apply the model to a two-arm randomized control trial of prostate cancer screening. Our main interest is to estimate the sojourn time (time between pre-clinical and clinical state). The challenge is that our observation scheme is unusual: for each individual we observe either the pre-clinical state, if the person is screen-detected or only the clinical state if the disease becomes symptomatic. We show that this can be done using standard software, by using *R* package *msm*.

Keywords: Continuous Time Multi-State Models; Prostate Cancer Screening; Sojourn Time; Survival

1 Introduction

We present a multi-state model, consisting of four states (healthy, pre-clinical, clinical and death) applied to interval-censored panel data, in a two-arm randomized controlled trial. We apply this model to prostate cancer screening, using data from the PLCO (Prostate, Lung, Colorectal and Ovarian) cancer screening trial (Andriole et al, 2012). We are interested in estimating the sojourn time, the transition time between pre-clinical and clinical states. This is the temporal window of opportunity for early detection of cancer. Sojourn time is useful to determine the frequency of screening and to estimate the potential benefits and harms of screening. However we only observe for each individual either the pre-clinical state, for men who are screen-detected (Screen Arm) or the clinical state for men who are clinically diagnosed (either cancer detected between screens or

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.

after the screening program ended or in the control arm). There is also censoring due to the limited time of follow-up of the trial. We show that this can be done using standard software, *R* package *msm*.

2 Continuous Time Multi-State Model

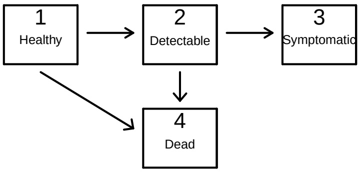


FIGURE 1. The 4-state model for the natural history of prostate cancer

A detailed description of this model can be found in, Jackson (2011) and van den Hout (2016). We describe here the model briefly. An individual is observed at times, $t_1, ..., t_J$, with corresponding observed states, $y_1, ..., y_J$, and a vector of covariates, $x_1, ..., x_J$. The set of health states is $S = \{1, 2, 3, 4\}$, where 1 denotes healthy, 2 pre-clinical disease, 3 clinical disease and 4 death. Additionally, $D = \{3, 4\}$ is a set of absorbing states. We define $p_{rs}(t_j, t_{j+1}) = P(Y_{j+1} = s|Y_j = r)$, where $r, s \in S$ and $j = 1, ..., J$. Then the likelihood contribution for each individual is $L = \prod_{j=2}^{J} L_j$, where L_j is given by,

$$L_j = p_{y_{j-1},y_j}(t_{j-1}, t_j), \tag{1}$$

for $y_j \in \{1, 2\}$. When $y_J = 4$ this becomes,

$$L_J = \sum_{m=1,2} p_{y_{J-1},m}(t_{J-1}, t_J)\ q_{m,y_J}, \tag{2}$$

since we observe the death time exactly, but we do not know the state immediately before death. When $y_J = 3$,

$$L_J = p_{y_{J-1},2}(t_{J-1}, t_J)\ q_{2,y_J}, \tag{3}$$

since by assumption a person must be in the pre-clinical state 2, before the clinical state 3 . Finally, at the end of follow-up we do not observe whether an individual is in state 1 or 2. The likelihood contribution then becomes,

$$L_J = p_{1,1}(t_{J-1}, t_J) + p_{1,2}(t_{J-1}, t_J). \tag{4}$$

Each L_j is composed of an entry of the transition probability matrix $P(t)$, $t = t_{j+1} - t_j$. For continuous time multi-state models this matrix equals, $P(t) = \exp(Qt)$, where Q is a matrix with non-diagonal entries, $q_{rs} = \exp(\sum \beta_{rs} x_{rs})$ (notice we could have a different set of covariates for each transition), and diagonal elements $q_{ss} = -\sum_{k \in S} q_{sk}$. If the transition is not possible, then we set *a priori*, $q_{rs} = 0$.

For this analysis, we consider two models. *Model I* denotes a model without covariates in the transition intensities. In *Model II*, we add age as a covariate, for

the transition between healthy and preclinical. This imposes a time-dependent transition intensity from healthy to preclinical.
Estimation is possible using R package msm (Jackson, 2011). The general idea is to maximize L with respect to the β parameters. For users of other statistical software, any general optimization algorithm could be used for instance, Nelder-Mead or Broyden-Fletcher-Goldfarb-Shanno.

3 Data

PLCO (Prostate, Lung, Colorectal and Ovarian) cancer screening trial is a multicenter randomized control trial. A total of $76,685$ men were randomized aged between 55 and 74. Randomization occured between 1993 and 2001. The screening protocol consisted of six annual screens, with a PSA test. The latest published report included results at 13 years of follow-up (Andriole et al, 2012). For the current analysis we use a random sample of 4000 men from the 13 years of follow-up data.

4 Simulation Experiment

The goal of this simulation experiment is to show that it is possible to obtain unbiased estimates of the q_{rs}'s and consequently, of the sojourn time, in the context of a two-arm trial. The basic idea of the simulation is for each individual, to simulate time spent in the k-th state, $k \in S$. If there are two states r and s, where one could move to, from state k, then the individual moves to the next state according to the rule, $min(T_r, T_s)$, where T_k is time until event k. The T_k's can be either Exponential or Gompertz distributed, with parameters given by $\beta_{rs} = (\beta_{rs,0}, \beta_{rs,1})$, where $\beta_{rs,1}$ is included if age is used as a covariate. We simulate the screen arm, by imposing a grid of observation times, with yearly time between observations (as in the PLCO data). We also add censoring, by ending follow-up at 13 years. We simulate two situations: *Without Mixed Design* denotes the natural history during follow-up, and *2-arm Mixed Design* denotes what we would observe in a screening trial. We do 100 simulations, with 1000 individuals each. In Table 1 we verify that we can obtain an unbiased estimate of the β parameter. The main difference lies in the standard errors, which, as expected, are larger for the mixed design. We see a similar story for the *Model II* (results not shown).

5 Preliminary Results

We used the PLCO data and the multi-state model to estimate prostate cancer sojourn time. We estimated a sojourn time of about 1.7 years for both models, with AIC value equal to 12669 (*Model I*) and 12653 (*Model II*). *Model II* seems to be better than *Model I*, i.e., the addition of the age covariate, did improve the model. The estimated sojourn time is substantially lower than what we would expect, given previous literature. This is the case since there is a substantial amount of missclassification in the control arm, namely, many men in state 3, are actually in the state 2, as they received the PSA screening test (Pinsky et

TABLE 1. Simulation Results for Model I

Parameter	Value	Mean	Bias	Relat. Bias	St. Dev.	St. Error
Without Mixed Design						
$\beta_{12,0}$	-2	-2.004	-0.004	0.002	0.034	0.037
$\beta_{14,0}$	-4	-3.998	0.002	0.000	0.112	0.109
$\beta_{23,0}$	-2	-1.993	0.007	0.004	0.049	0.051
$\beta_{24,0}$	-3	-3.010	-0.010	0.003	0.081	0.090
With 2-Arm Mixed Design						
$\beta_{12,0}$	-2	-2.003	-0.003	0.002	0.047	0.051
$\beta_{14,0}$	-4	-4.005	-0.005	0.001	0.159	0.154
$\beta_{23,0}$	-2	-1.993	0.007	0.004	0.099	0.098
$\beta_{24,0}$	-3	-3.028	-0.028	0.009	0.189	0.201

al, 2005). We aim to extend the model, to obtain an unbiased estimate of the sojourn time in the presence of missclassification.

Acknowledgments The research was funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care North Thames at Barts Health NHS Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

Andriole GL, Crawford ED, Grubb RL *et al* (2012). *Prostate cancer screening in the randomized PLCO Cancer Screening Trial: mortality results after 13 years of follow-up.* J Natl Cancer Inst. 104(2):125-32.

Jackson CH (2011). *Multi-State Models for Panel Data: The msm package for R.* J Stat Soft. 38(8).

Pinsky PF, Blacka A, Kramer BS, *et al* (2010). *Assessing contamination and compliance in the prostate component of the PLCO Cancer Screening Trial.* Clin Trials. 7(4):303-11.

van den Hout A (2016). *Multi-State Survival Models for Interval-Censored Data.* Chapman & Hall-CRC Monographs on Statistics & Applied Probability.

Multinomial and Discrete survival models to evaluate the performance of students

Hildete Pinheiro[1], Rafael Maia[1], Daniel Barboza[1]

[1] Department of Statistics, State University of Campinas, Brazil

E-mail for correspondence: hildete@g.unicamp.br

Abstract: A great interest in Public Policies in Brazil is the comparison of undergraduate students who come from Public High Schools (Pu HS) with those who come from Private High Schools (Pr HS). In this paper we analyze the performance of undergraduate students in Calculus I and Physics I. The database is from the University of Campinas (Unicamp), one of the top Public Research Universities in Brazil. First, we used a multinomial model to evaluate the proportion of students approved (in none, Calculus I only, Physics I only and in both subjects) according to socioeconomic status, demographic characteristics and pre-enrollment tests (e.g. SAT scores). Then, we studied the number of times the students took these courses until they pass. We used discrete survival models with the censored observations being those who did not pass the course until the end of the study. For the survival models, the analysis can be done separately with two univariate models, one for each course (Calculus I and Physics I), or as bivariate survival models.

Keywords: Multinomial model; Categorical data; Censored data; Discrete survival model; Survival analysis.

1 Introduction

A topic of great interest to educators and administrators of Universities is the performance of students in subjects like Calculus I and Physics I. They are usually required courses for all students in Engineering and Exact Sciences majors in the first year of enrollment in the university, but there are many questions regarding the factors which may contribute more to bad/good performance of the students. The database of this study is from the State University of Campinas (Unicamp), located in the state of São Paulo, Brazil, which is one of the top research universities in South America with a highly selective entrance exam. Unicamp has an entrance exam which consists of two phases. The first phase consists on a multiple choice test of general knowledge. The second phase consists on open questions 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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Mathematics, Portuguese, Geography, History, Biology, Chemistry, Physics and an Essay Writing. We will consider here only the scores on the second phase.

A great interest in Public Policies is the academic performance of those from Public High Schools (Pu HS) and those from Private High Schools (Pr HS). In Brazil, most of middle class students go to Pr HS, because the Public High School system is not very good. In view of this, Unicamp implemented an affirmative action program which gives a bonus in the second phase of the Entrance Exam Score (EES), when the candidate studied all High School years in Public Schools. Therefore, for the models considered here, the type of High School will be a very important covariate.

The primary goal of this study is to evaluate the performance of students in Calculus I and Physics I, especially when they are taking theses courses for the first time. Therefore, we would like to evaluate, when the students are taking these courses for the first time, the proportion who failed both, approved only in Calculus I or only in Physics and approved in both, according to a multinomial model with nominal response. We will be able to see what are the factors which can better explain the variation of performance of these students.

Our secondary goal is to model the average number of times a student need to take Calculus I/ Physics I until he/she passes. In this case, we can think of this as discrete survival model (Collet, 2003) and when the student did not pass the course, it is a censored observation. Note that there are university policies to determine when a student should be dropped out from the university, since public universities in Brazil are completely free for all enrolled students.

2 Data set

The original data set consists of 16,503 records of students who enrolled at Unicamp from 2009 to 2013. It was selected from the database only students with Calculus I and/or Physics I as required courses, having about 28% from Pu HS students and 72% from Pr HS. So, there are records for all the times the student took the subject with their respective grades. We also have entrance exam scores - EES (e.g., SAT scores) in each subject (Mathematics, Portuguese, Geography, History, Biology, Chemistry and Physics), some academic variables as well as socioeconomic status, which are considered as covariates in the models. Among the students who have Calculus/Physics as required courses, there are 1259 students who took only Calculus I or Physics I and 5384 who took both Calculus I and Physics I. Table 1 shows the number of times the students took the subjects according to type of High School.

Figure 1 shows the proportion of students approved (in none, only Calculus I, only Physics I and in both) according to type of High School when they are taking the courses for the first time. One can see that there are more students from Pr HS who were approved in both courses and more students from Pu HS who failed both courses. When comparing those who passed only in one of the courses, the difference is quite small.

3 Statistical models

The first model considered here is a generalized logit model for multinomial response (Agresti, 2013). Let Y_i be a r.v. indicating the number of subjects student

TABLE 1. Number of students taking Physics I and Calculus I by type of High School and number of times taking them.

		Number of times taking the course						
Subject	High School	1	2	3	4	5	6	≥ 7
Physics I	Public	1126	269	91	49	17	8	0
	Private	3275	519	143	60	25	5	0
Calculus I	Public	1273	328	113	48	23	8	5
	Private	3704	602	190	96	31	12	7

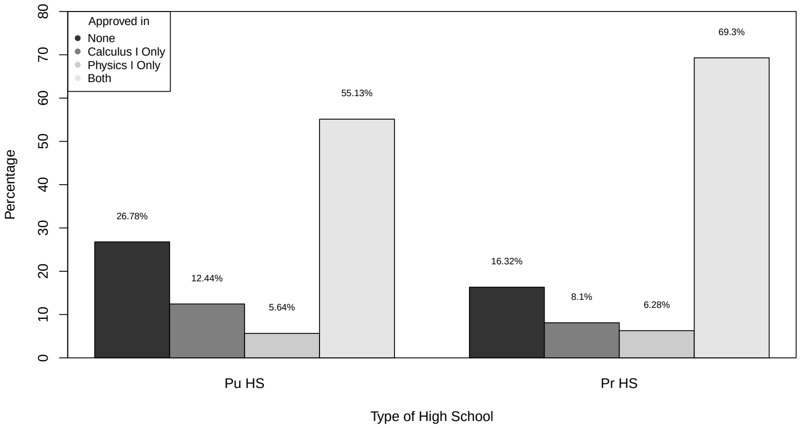


FIGURE 1. Bar plots of the proportion of students approved (in none, only Calculus I, only Physics I and in both) according to type of High School when they are taking the courses for the first time.

i approved, where$Y_i \in \{0, 1, 2, 3\}$, i.e., 0 if student i failed in Calculus I and Physics I; 1 if student i was approved in Calculus I only; 2 if student i was approved in Physics I only; and 3 if student i was approved in both subjects. Now, let π_{ij} denote the probability of response j for student i, $j = 1, 2, \ldots 4$. When the final category is the baseline, the $j - th$ logit is

$$\log\left(\frac{\pi_{ij}}{\pi_{i4}}\right) = \mathbf{x}_i \boldsymbol{\beta}_j, \text{ for } j = 1, 2, 3 \qquad (1)$$

with $\mathbf{x}_i$ being the i^{th} row of the design matrix and $\boldsymbol{\beta}_j$ the respective vector of fixed effect for logic j.

For the second analysis, we can use univariate discrete survival models (Collet, 2003; Hosmer and Lemeshow, 1989) The likelihood function here is written in terms of the probability of passing the course (Calculus/Physics) in the $j - th$ time, $j = 1, 2, \ldots, k$. Let T_i be the number of times student i took the course until being approved. Here $T_i \in \{1, 2, 3, \ldots\}$ is a discrete r.v. Let $p_j(\mathbf{x}_i) = P(T_i =$

$j \mid T_i \geq j-1, \mathbf{x}_i)$ be the probability of individual i pass the course in the $j - th$ time given that he did not pass the $(j-1) - th$ time. The likelihood is given by

$$\mathcal{L}(\boldsymbol{\beta}, \gamma) = \sum_{j=1}^{k} \sum_{i \in R_j} \left[\delta_{ij} \log \left(1 - \gamma_j^{\exp(\mathbf{x}_i' \boldsymbol{\beta})} \right) + (1 - \delta_{ij}) \log \left(\gamma_j^{\exp(\mathbf{x}_i' \boldsymbol{\beta})} \right) \right], \quad (2)$$

where $\delta_{ij} = 0$, if student i did not pass the course in the $j - th$ time and $\delta_{ij} = 1$, otherwise. $p_j(\mathbf{x}_i) = 1 - \gamma_j^{\exp(\mathbf{x}_i \boldsymbol{\beta})}$, with $\gamma_j = S_0(j)/S_0(j-1)$ and $S_0(.)$ is the baseline survival.

Acknowledgments: We would like to thank Renato Hirata and Flavio Ferraresi, database specialists at the Admissions Committee at Unicamp, and Silvio de Souza, database specialist at the Academic Records Division at Unicamp, for their support regarding the database.

References

Agresti, A. (2013). *Categorical Data Analysis.* New York: John Wiley & Sons.

Collet, D. (2003). *Modelling Survival Data in Medical Research.* London: Chapman & Hall.

Hosmer, P.J. and Lemeshow, B.W. (1989). *Applied Logistic Regression.* New York: John Wiley & Sons.

A Random Effect Mixture Model for Repeated Measurements of ordinal variables

Rosaria Simone[1], Gerhard Tutz[2], Maria Iannario[1]

[1] Department of Political Sciences, University of Naples Federico II, Italy
[2] Ludwig–Maximilians - Universität München, Germany

E-mail for correspondence: `maria.iannario@unina.it`

Abstract: The focus of the present contribution is on the modelling of subjective heterogeneity for repeated measurements of ordinal variables, which occur, for example, in questionnaires. The proposal is a multivariate random effects model based on CUB models, a class of mixture distributions for rating data that accounts for uncertainty of choices. Model performances are assessed on the basis of a survey on relational goods.

Keywords: Repeated Measurements; Subjective Heterogeneity; Mixture models

1 Introduction and Preliminaries

Within the framework of ordinal data analysis, a general mixture model with an uncertainty component as considered in Iannario and Piccolo (2016) and Tutz *et al.* (2016) is specified by:

$$P(R_i = r|\mathbf{x}_i) = \pi_i P_M(Y_i = r|\mathbf{x}_i) + (1-\pi_i)P_U(U_i = r), r = 1, \dots, m \qquad (1)$$

where R_i, $i = 1, \dots, n$, is the observed ordinal response and $\mathbf{x}_i$ is a vector of subjects' characteristics. The unobserved variables Y_i and U_i stand for the pure-preference choice and the uncertainty, respectively, both taking values in $\{1, \dots, m\}$. The distribution $P_M(Y_i = r|\mathbf{x}_i)$ of Y_i can be any ordinal model M, whereas $P_U(U_i = r) = \dfrac{1}{m}$ is assumed to follow a discrete Uniform distribution. The general model (1) is a CUP model: *C*ombination of *U*niform and *P*reference structures. In this contribution, we consider the traditional CUB model of Piccolo (2003), in which the distribution of Y_i is the (shifted) Binomial distribution $g_r(\xi_i)$ with *feeling* parameter ξ_i:

$$Pr(R_i = r|\mathbf{x}_i) = \pi_i g_r(\xi_i) + (1-\pi_i)\frac{1}{m}, \quad g_r(\xi_i) = \binom{m-1}{r-1}\xi_i^{m-r}(1-\xi_i)^{r-1}. \qquad (2)$$

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

In (2), parameters ξ_i and π_i are linked to the row vector of subjects' covariates $\mathbf{x}_i$ by:

$$logit(\xi_i) = \gamma_0 + \mathbf{x}_i\,\gamma, \qquad logit(\pi_i) = \beta_0 + \mathbf{x}_i\,\beta.$$

The mixing proportion π_i is referred to as the *uncertainty parameter*, since $1-\pi_i$ is the weight for the Uniform component. Thus, π_i measures the *individual propensity* to adhere to a meditated choice for the response, so that each choice takes place within the individual. As for any mixture model, π_i determines the distinction between the two separate sub-groups in the population: those responding driven by their *feeling/preferences*, and the others adhering to a *random* process for their choices. In the following a multivariate model is proposed that uses CUB models for the marginal responses and accounts for individual propensities across the repeated measurements. The model contains a subject-specific random effect in the mixing proportion π_i, which is considered fixed across the items. In this vein, CUB models gain a multidimensional perspective avoiding the shortcomings of an item-by-item analysis. Hierarchical CUB models with random effects for the feeling component have been introduced in Iannario (2012).

2 The Random CUB Models

We consider several measurements on the same subject on a given latent trait, as in questionnaire analysis. In order to simplify the presentation, covariates are not included. Let $R_{i1},\dots,R_{iK}$ denote the responses of the person i to a questionnaire with K items, for $i=1,\dots,n$, all measured on the same ordinal scale with m categories. The RCUB model (Random CUB Model) is defined as the following mixed model for $R_{i1},\dots,R_{iK}$:

$$Pr(R_{ij}=r|\eta^{'},b_i) = \pi_i g_r(\xi_j) + (1-\pi_i)\frac{1}{m}, \quad r=1,\dots m,\ j=1,\dots,K, \tag{3}$$

with $\eta^{'} = (\beta_0,\xi_1,\dots,\xi_K)^{'}$ and ξ_j being the CUB feeling parameter for item j. A subject-specific parameter b_i is included in the mixing proportion by:

$$\pi_i = \frac{\exp(\beta_0+b_i)}{1+\exp(\beta_0+b_i)}, \qquad b_i \sim \mathcal{N}(0,\sigma^2). \tag{4}$$

As usual in mixed models, the b_i's are assumed to be i.i.d. random variables and observations are conditionally independent given the random effect:

$$Pr(R_{i1}=r_{i1},\dots,R_{iK}=r_{iK}|\eta^{'},b_i) = \prod_{j=1}^{K} Pr(R_{ij}=r_{ij}|\eta_j^{'},b_i), \quad i=1,\dots,n,$$

where we set $\eta_j^{'} = (\beta_0,\xi_j)$. Note that $\sigma^2=0$ yields CUB (CUP) models with uncertainty parameter equal for all items to $1/(1+exp(-\beta_0))$. The rationale behind RCUB models is that the subject-specific tendency to responses determined by uncertainty is driven by a subject-specific parameter b_i, which is constant across items. Moreover, the parameter β_0 quantifies the common uncertainty among items and respondents, while σ accounts for the unspecified effects. This approach allows for a parsimonious parameterization of repeated ordinal measurements, since the covariance structure of the variables is explained by the variance component σ^2.

As customary for mixed-effect models, estimation of the fixed-effect parameters $\theta^{'} = (\eta^{'}, \sigma^2)^{'}$, with $\eta^{'} = (\beta_0, \xi^{'})$, $\xi^{'} = (\xi_1, \ldots, \xi_K)^{'}$ is pursued by integrating out the random parameters b_i of the full log-likelihood (Pawitan, 2001), thus obtaining the (marginal) log-likelihood to be optimized:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} log\, I_i, \quad I_i = \int_{-\infty}^{\infty} \prod_{j=1}^{K} Pr(R_{ij} = r_{ij} | \xi_j, \beta_0, b_i) \varphi(b_i; \sigma) d\, b_i, \tag{5}$$

where $\varphi(\cdot, \sigma)$ denotes the density function of the normal distribution $\mathcal{N}(0, \sigma^2)$. Then, after simple algebraic manipulations and the application of the Gauss-Hermite quadrature formula, the marginal log-likelihood (5) is maximized by means of the quasi-Newton method resulting from the Broyend-Fletcher-Goldfarb-Shanno (BFGS) algorithm.

3 The case study

In 2014, an observational study was carried out by the Department of Political Sciences, University of Naples Federico II to examine relational goods. Ratings were collected on a 10-point Likert scale from 1 = *"Never, Not at all"* to 10 = *"Always, Totally, Absolutely Yes"*). We fit an RCUB model to selected variables and report the results of an item-by item analysis with CUB models. It is seen that there is a significant variation of the tendency to the uncertainty ($\hat{\sigma} = 1.93$).

TABLE 1. Selected Questionnaire Items.

		CUB		RCUB		
		$\hat{\pi}_k$	$\hat{\xi}_k$	$\hat{\beta}_0$	$\hat{\xi}_k$	$\hat{\sigma}$
				-0.648		1.93
`Parents`	*How often do you speak with at least one of your parents?*	0.216	0.060		0.138	
`Friends`	*How good are your relationships with friends?*	0.742	0.169		0.140	
`Neighbours`	*How good are your relationships with neighbours?*	0.313	0.353		0.328	
`Safety`	*Do you feel safe in the place where you live?*	0.227	0.458		0.433	
`Familycond`	*Does your family easily make ends meet?*	0.427	0.341		0.324	

4 Final remarks

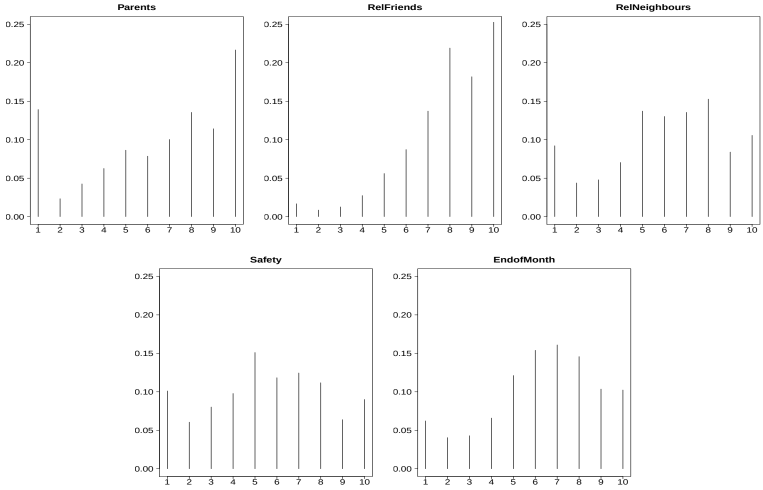
The RCUB proposal stems from the need of running multi-item analysis by focusing on subjective heterogeneity while maintaining the rationale implied by CUB models. For future work, this approach will be developed to specify the feeling component on item-basis to include possible overdispersion, shelter effect or, more generally, response-styles.

References

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85 – 104.

Iannario, M. and Piccolo, D. (2016). A generalized framework for modelling ordinal data. *Statistical Methods and Applications*, **25**, 163 – 189.

FIGURE 1. Selected Questionnaire Items



Iannario, M. (2012). Hierarchical CUB models for ordinal variables. *Communications in Statistics. Theory and Methods*, **41**(16-17), 3110 – 3125.

Pawitan, J. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford Science Publications, Clarendon Press.

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.

A Generic Method for Density Forecast Recalibration

Michael Richard[1][2], Christophe Hurlin[2], Jérôme Collet[1]

[1] EDF R&D division, OSIRIS department, France
[2] Laboratoire d'économie d'Orléans, France

E-mail for correspondence: michael-m.richard@edf.fr

Abstract: We address the calibration constraint of probability forecasting. We propose a generic method for recalibration, which allows to enforce this constraint. It remains to know the impact on forecast quality, measured by predictive distributions sharpness, or specific scores. We show that the impact on Pinball-Loss score expectation is weak under some hypotheses and that it is always positive under more restrictive conditions.

Keywords: Density forecasting; Rosenblatt transform; bias correction

1 Introduction

Due to increasing need for risk management, forecasting is shifting from point forecasts to probabilistic forecasts. According to T. Gneiting, probabilistic forecasting aims to maximize the sharpness of the predictive distributions, subject to calibration, on the basis of the available information set. It means we have a multi-objective problem, which is necessarily more difficult. That is why an important step in building a probabilistic forecast is calibration, either in a specific way, either in a more generic manner. We propose here a generic method, using Probability Integral Transform (PIT). PIT is usually a measure of forecast miscalibration, we show it can be used to recalibrate it.

2 Principle of the recalibration

One stands in the following case: let E be the set of all possible states of the world; for each forecasting time j the forecaster knows the current state of the world $e(j)$, and uses it to forecast. An estimated distribution G_e is associated to each state of the world.

Then, one can calculate the estimated PIT series, it is the series of the values $\left(G_{e(j)}(y_j)\right)_j$, where y_j is the realization for time j.

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

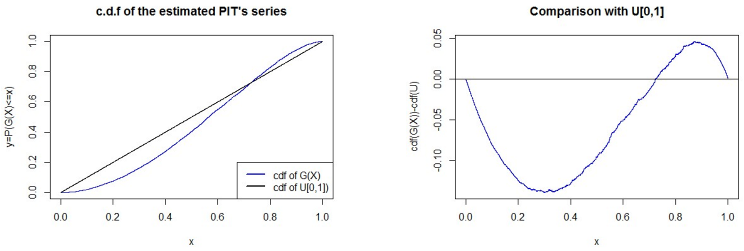


FIGURE 1. Empirical c.d.f of the estimated PIT series and comparison with the c.d.f of the uniform distribution on [0,1] respectively.

The theoretical c.d.f of the estimated PIT is the following:

$$\begin{aligned} C(y) &:= \Pr(\, G(\boldsymbol{Y}) \leq y) \\ &= \mathrm{E}\,[\,\mathrm{E}\,[\,\mathbf{1}_{\{G(\boldsymbol{Y})\leqslant y\}}/E\,]\,] \end{aligned}$$

$$= \begin{cases} \sum_e p_e\, F_e \circ G_e^{-1}(y) & \text{if } E \text{ is a discrete R.V.} \\ \int_{e\in\mathbb{R}} p(e)\, F_e \circ G_e^{-1}(y)\, de & \text{if } E \text{ is a continuous R.V.} \end{cases}$$

where p_e and $p(e)$ are respectively the frequency of appearance or the p.d.f of state e, and F_e and G_e the true and the estimated c.d.f of Y conditionally to the state e. Obviously, if $G_e = F_e$ for each e, then this c.d.f. is the c.d.f. of the uniform distribution on [0,1].

If there is a bias, as shown in figure 1, we propose to use the biased PIT series to correct the future forecasts.

That makes sense since :

$$\begin{aligned} \Pr(\, C \,\circ\, G(\,\boldsymbol{Y}) \leqslant y) &= \Pr(\, G(\,\boldsymbol{Y}) \leqslant C^{-1}(y)) \\ &= C \,\circ\, C^{-1}(y) \\ &= y \end{aligned}$$

Hence for instance, if we are interested in forecasting the τ quantile, we will use the model estimated for the $C^{-1}(\tau)$ quantile

It remains to study the impact of such a correction in finite sample, as Siegert does.

3 Impact on sharpness

We first present some theoretical results and then we show a case of positive bias correction.

3.1 Theoretical results

Under the two following hypotheses :

- F_e and G_e are close, *i.e* $\forall x, \forall e, \mid F_e(x) - G_e(x) \mid \leqslant \varepsilon$ with F and G the true and estimated distributions respectively
- the derivative of $G_e(x)$ are lower bounded on the following interval $I_g := [G_e^{-1}(\tau - \varepsilon)\,,\, G_e^{-1}(\tau + \varepsilon)]$, *i.e* $\forall x \in I_g, \forall e, \frac{1}{g_e(x)} \leqslant \xi$, with g_e the derivative of G_e

We prove that the impact of our correction on the absolute value of the Pinball-Loss score expectation is less than $2\,\varepsilon^2\,\xi$
Under more restrictive hypotheses :

- $\forall e, \epsilon_e := (G_e \circ F_e^{-1}(\tau) - \tau) \in [m_\epsilon \pm \frac{1}{\sqrt{2}} m_\epsilon]$ with m_ϵ the mean of the ϵ_e
- $\forall e, \frac{1}{f_e \circ F_e^{-1}(\tau)} \in [m_{f^{-1}} \pm \frac{1}{\sqrt{2}} m_{f^{-1}}]$ with $m_{f^{-1}}$ the mean of the $\frac{1}{f_e \circ F_e^{-1}(\tau)}$

We prove that the impact of our correction on the Pinball-Loss score expectation is always positive.

3.2 Illustration

- we simulate 10 000 independent realizations of a random variable $\boldsymbol{X}$, with standard normal distribution,
- For each realization x, one generates a realization of a random variable $\boldsymbol{Y}$ with Student distribution, with degree of freedom equal to 3, and lag parameter $2x$
- we model Y with a simple Gaußian linear model $\boldsymbol{Y} = 2\boldsymbol{X} + \varepsilon$, with $\varepsilon \sim \mathrm{N}(0, 1)$.

When we study the empirical c.d.f of the estimated PIT series of our simple Gaußian linear model, it fails to pass the validity test. Nevertheless, the correction allows to pass the validity test.
Note that it would be possible to pass the validity test with a ”climatological” forecast, which means to forecast, independently of x, the marginal distribution of Y. This forecast is reliable, but very inaccurate.

TABLE 1. Pinball-Loss score expectation for τ=0.5 and median quantile issued from different models.

	climatological	Gaußian	Gaußian, corrected	true
$\mathrm{E}[PL_{0.5}]$	1.24199	0.74155	0.74152	0.73720
ratio with true model	1.68474	1.00590	1.00586	1

As one can see in table 1, the climatological estimation has a greater Pinball-Loss score expectation than the other even if it passes the validity test, due to loss of information. Moreover, we remark that the score obtained with our correction is little better than the score obtained with the simple linear Gaußian model.

References

Bentzien, S., and Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, **140(683)**, 1924 – 1934.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, **135(643)**, 1512 – 1519.

Gneiting, T., and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, **1**, 125 – 151.

Gogonel, A., Collet, J., and Bar-Hen, A. (2013). Improving the calibration of the best member method using quantile regression to forecast extreme temperatures. *Natural Hazards and Earth System Sciences*, **13(5)**, 1161 – 1168.

Michelangeli, P-A., Vrac, M., and Loukos, H. (2009). Probabilistic downscaling approaches : Application to wind cumulative distribution functions. *Geophysical Research Letters*, **36**, L11708.

Siegert, S. (2016). Simplifying and generalising Murphy's Brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*

Siegert, S., Sansom, P. G., and Williams, R. (2015). Parameter uncertainty in forecast recalibration. *Quarterly Journal of the Royal Meteorological Society*

Analysis of social interaction data

Susan Fennell[1], Michael Quayle[2], James P. Gleeson[1,3], Kevin Burke[1]

[1] Department of Mathematics and Statistics, University of Limerick, Ireland
[2] Department of Psychology, University of Limerick, Ireland
[3] MACSI, University of Limerick, Ireland

E-mail for correspondence: Susan.Fennell@ul.ie

Abstract: VIAPPL is a software platform used to conduct experiments in social interaction. Participants interact by exchanging tokens. We would like to understand why participants choose certain players to interact with and how these reasons may change over the course of the experiment. We describe two preliminary models developed for the experimental data that indicate reciprocity, in-group favouritism and charity may occur in the experiments. We discuss why these models need to be improved upon.

Keywords: Linear Regression; Logistic Regression; Network Science; Social Psychology

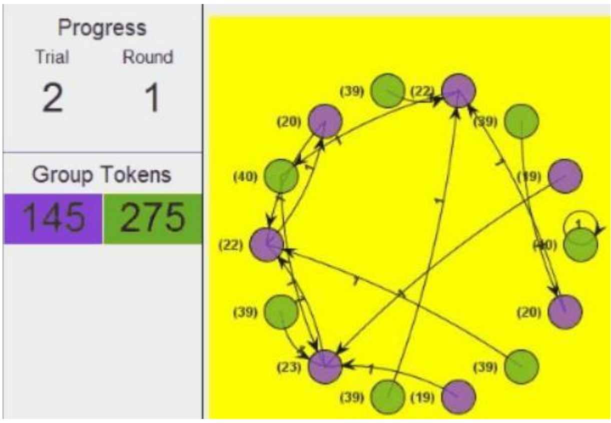
1 Introduction

Recent developments in social psychology suggest that social identities and norms develop and change through social interaction. A person does not only develop their identity through, and adopt the social norms of, the society that they are in, they can affect change in society and the social norms present. Durrheim et al. (2016a, 2016b) believed there was a need for a new type of experiment that incorporated this strategic nature of social interaction and so developed VIAPPL (Virtual Interaction APPLication), a software platform for carrying out these experiments.

Participants of the experiments are avatars in a game-like environment, and are referred to as players. They observe other players as nodes in a network, as shown in Figure 1, and they interact by exchanging tokens. At the start of a game players are randomly allocated to one of two groups, of which members are recognised by the colour of their node.

Each player begins with a number of tokens, which may depend on the group to which they have been assigned, and must choose a player to give a token to at each round of the game (there are 40 rounds in total). At the end of each round they observe the network shown in Figure 1. The number beside each node refers to the number of tokens that player has. Arrows between nodes indicate to whom

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.



Progress

Trial	Round
2	1

Group Tokens

145	275

FIGURE 1. VIAPPL environment.

each player has given their token in that round. The number of tokens belonging to each group is shown to the left of the screen. The VIAPPL technology allows for multiple trials per game but the games considered here have only one trial.
What we would like to discover from these games is whether relationships develop (i.e., do players reciprocate token giving), whether charity occurs (do wealthier players, in terms of token count, give more to poorer players than other wealthy players) and whether in-group favouritism is present in the games (do players give more often to players in their own group than the other group). We have data from 4 VIAPPL games, each of which has two groups of 7 players who start with 20 tokens.

2 Methods

In order to investigate whether there is reciprocity, etc. in the games, we consider what influences the number of tokens given from one player to another over the course of the game, over a number of rounds or even over one round (i.e., whether a player gives a token to another player in a given round). The variable of interest, $Y_{i,j,t}$, is the number of tokens player i receives from player j at, or up to, round t. The vector of covariates X_{t-1} used in the models for Y_t can contain any information from rounds 1 to $t-1$ of the game.
We consider first $Y_{i,j,t} = Y_{i,j,40}$ to be the total number of tokens player i receives from player j over the 40 rounds of the game. In this case the vector of covariates is $X_{t-1} = X_{39}$. A normal linear model for Y is assumed, $Y_{40} \sim N(X_{39}^{\mathrm{T}}\beta, \sigma^2)$, where β is the vector of unknown regression coefficients and σ^2 is the error variance. The covariates are the number of tokens player i had at the end of round 39, the number of tokens player j had at the end of round 39 and the number of tokens given from player i to player j in the first 39 rounds, as well as a categorical variable of which three were considered; the first describes whether players i and j are in the same group; the second whether j and i are the same person, in the same group but not the same person or in different groups; the third indicates player j's position in the network relative to player i.

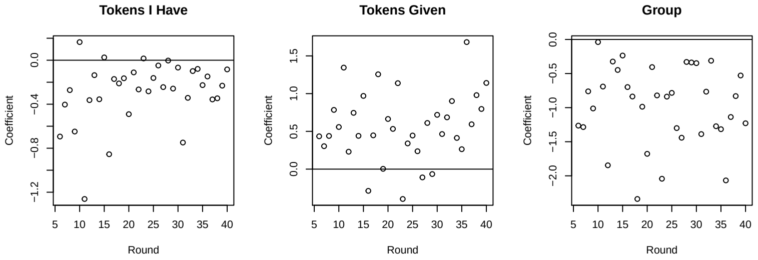


FIGURE 2. Coefficients for each variable in the logistic regression models for the first game.

The final model included the variable describing the number of tokens player i gave to player j and the categorical variable describing whether players i and j were in the same group. The model for the first game, which has $R^2 = 0.43$, is

$$\text{TokensReceived} = 1.82 + 0.59 \times \text{TokensGiven} - 1.21 \times \text{Group}.$$

The model predicts that the more tokens you give to another player, the more tokens you will receive in return and that on average you will receive at least 1 token more over the course of the game from a player in your group than from a player in the other group.

We also considered the case where $Y_{i,j,t}$ is the binary variable describing whether player i received a token from player j in round t. We assume $Y_t \sim \text{Bernoulli}(p_t)$, where the probability of receiving a token, p_t, is related to the covariates via $\text{logit}(p_t) = X_{t-1}^{\text{T}}\beta_t$.

A model was developed at each round t of the game, $6 \le t \le 40$, and the covariates used were similar to those in the linear model. An exception is the variable for the number of tokens given from player i to player j; in the model at round t this variable is the number of tokens given from player i to player j between rounds $t-5$ and $t-1$ (instead of between rounds 1 and $t-1$). This restriction to the previous 5 rounds assumes players have limited memory and will base their decisions on the most recent information.

The variables contained in the final logistic regression models are the number of tokens player i has, the number of tokens given from player i to player j and the categorical variable that tells us whether players i and j are in the same group. The coefficients of the variables in the final models for the first game are shown in Figure 2. In the majority of rounds the fewer tokens a player has, the more likely they are to receive a token, indicating that players are charitable. The more tokens player i has given to player j in the previous 5 rounds, the more likely player j is to reciprocate (again in the majority of rounds). The coefficients of the group variable are all negative (players are less likely to give to players in the other group) suggesting in-group favouritism is present in the game.

3 Discussion

The linear and logistic regression models indicate reciprocation and in-group favouritism are present in the games. The logistic models also show charity occurs.

However these simple models were used only as a first approach to modelling the data and can be improved upon.
Players only give one token per round which places a sum constraint on the response variable. This means observations are not independent. For example the linear model should have $\sum_{j=1}^{14} Y_{i,j} = 40$. We are therefore developing a model which respects the compositional constraint on the vector $Y_{j,t} = (y_{(1,j,t)}, \ldots, y_{(14,j,t)})$ whose elements describe the number of tokens player j has given to each other player up to, or at, round t.
The linear regression model ignored the temporal aspect of the game. While logistic regression models were constructed at each round and showed how the importance of each variable changed over time, having many models makes it difficult to interpret what is happening in the game as a whole. We are now starting to develop temporal network models to enable more detailed analysis of the dynamic social interaction structures in the data.

Acknowledgments: The first author is funded by the Irish Research Council.

References

Durrheim, K., Quayle, M., Titlestad, K. & Tooke, L. (2016a). The evolution of ingroup bias in a dynamic experimental environment: toward a social psychology of movement. viappl.org/wp/pubs.

Durrheim, K., Quayle, M., Tredoux, C. G., Titlestad, K. & Tooke, L. (2016b). Investigating the Evolution of Ingroup Favoritism Using a Minimal Group Interaction Paradigm: The Effects of Inter-and Intragroup Interdependence. *PLoS ONE*, **11**, 1–26. e0165974.

Nonlinear quantile regression with clustered data

Marco Geraci[1], Bo Cai[1]

[1] University of South Carolina, Columbia SC, USA

E-mail for correspondence: geraci@mailbox.sc.edu

Abstract: In this paper, we develop a novel approach to the estimation of nonlinear quantile functions when the data are clustered. The proposed methods are illustrated using a toy example taken from the nonlinear mixed-effects literature.

Keywords: Growth curves; Pharmacokinetics; Random effects.

1 Introduction

Quantile regression analysis of clustered data is a very active area of research. Since the seminal work of Koenker and Bassett (1978) on methods for cross-sectional observations, there have been a number of proposals on how to accommodate for the dependency induced by clustered (e.g., longitudinal) designs. As briefly outlined by Geraci and Bottai (2014) and then extensively reviewed by Marino and Farcomeni (2015), approaches to linear quantile regression with clustered data include distribution-free approaches (e.g., Koenker, 2004) and (pseudo) likelihood-based approaches. The latter mainly adopt the asymmetric Laplace (AL) density (Geraci and Bottai, 2007, 2014; Farcomeni, 2012).

We examined the statistical literature on nonlinear quantile regression with clustered data (in our review, we did not consider nonparametric smoothing). To the best of our knowledge, there seem to be only a handful of published articles. Karlsson (2008) considered nonlinear longitudinal data and proposed weighting the standard quantile regression estimator with pre-specified weights. Wang (2012), taking their cue from Geraci and Bottai (2007), used the AL distribution to define the likelihood of a Bayesian nonlinear quantile regression model. Finally, Oberhofer and Haupt (2016) established the consistency of the L_1-norm nonlinear quantile estimator under weak dependency.

In this paper, we propose an extension of Geraci and Bottai's (2014) linear quantile mixed models (LQMMs) to the nonlinear case.

TABLE 1. Estimated fixed parameters of logistic curves at quantile levels $\{0.05, 0.1, 0.5, 0.9, 0.95\}$ by genotype.

τ	0.05	0.1	0.5	0.9	0.95
			Genotype P		
β_1	16.24	16.97	19.51	24.02	30.66
β_2	55.18	55.35	52.98	52.97	56.37
β_3	7.96	8.40	8.08	8.76	9.73
			Genotype F		
β_1	8.97	9.73	17.77	20.72	21.48
β_2	53.31	51.85	55.06	54.14	53.56
β_3	7.38	6.93	8.10	8.34	8.31

2 Methods

Consider data from a two-level nested design in the form $(\mathbf{x}_{ij}^{\mathrm{T}}, y_{ij})$, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, M$, where $\mathbf{x}_{ij}$ is a given vector of covariates and y_{ij} is the jth observation of the response vector $\mathbf{y}_i = (y_{11}, \ldots, y_{1n_i})^{\mathrm{T}}$ for the ith cluster. We define the nonlinear conditional quantile regression function

$$Q_\tau(y_{ij}|\mathbf{u}_i) = f_{ij}\left(\boldsymbol{\beta}_\tau, \mathbf{u}_i, \mathbf{x}_{ij}\right), \tag{1}$$

where f is a nonlinear, smooth function of the $p \times 1$ fixed parameter $\boldsymbol{\beta}_\tau$, the $q \times 1$ random parameter $\mathbf{u}_i$, and the covariates $\mathbf{x}_{ij}$. For *given* $\mathbf{u}_i$, model (1) can be equivalently written as $y_{ij} = f_{ij}\left(\boldsymbol{\beta}_\tau, \mathbf{u}_i, \mathbf{x}_{ij}\right) + \epsilon_{\tau,ij}$, where $\epsilon_{\tau,ij} \sim \mathcal{AL}\left(0, \sigma_\tau\right)$, which denotes the AL distribution with location 0 and scale σ_τ. This convenient assumption leads to the quantile restriction $Q_\tau(\epsilon_{\tau,ij}) = 0$. We assume $\mathbf{u}_i \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\tau\right)$, independently from ϵ_{ij}, though, in principle, one can consider different distributions for the random effects. Our goal is to maximise the (pseudo) marginal log-likelihood

$$\ell\left(\boldsymbol{\beta}_\tau, \boldsymbol{\Sigma}_\tau, \sigma_\tau; \mathbf{y}\right) = \sum_i \log \int_{\mathbb{R}^q} p\left(\mathbf{y}_i|\mathbf{u}_i, \boldsymbol{\beta}_\tau, \sigma_\tau\right) p\left(\mathbf{u}_i|\boldsymbol{\Sigma}_\tau\right) \, \mathrm{d}\mathbf{u}_i \tag{2}$$

We applied the Laplacian and importance sampling (IS) approximations to compute the integral in (2). Since these approximations make use of second-order Taylor expansions, we first smoothed the absolute residuals in the AL's kernel exponent using the approximation of Chen (2007). As a result, we obtained approximated log-likelihoods akin to those used in nonlinear mixed models (Pinheiro and Bates, 1995). The R language (R Core Team, 2016) was used to implement the proposed methods and to carry out the data analysis in the next section.

3 Growth of Soybean Plants

We presents some results using the `Soybean` dataset, available in the R package `nlme`, which consists of measurements from an experiment to compare growth

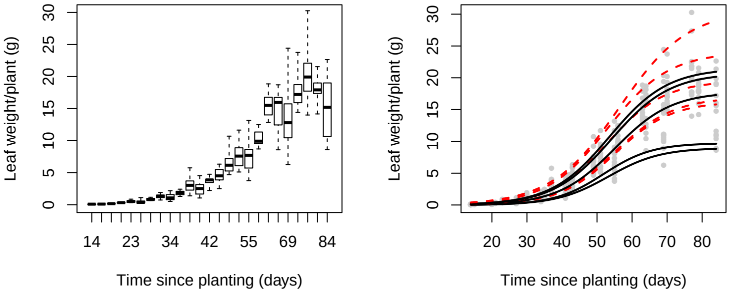


FIGURE 1. Left plot: Boxplots of average leaf weights per plant by day since planting. Right plot: Fitted logistic quantile curves at level $\tau \in \{0.05, 0.1, 0.5, 0.9, 0.95\}$ for genotypes P (dashed red line) and F (solid black line) superimposed to observed measurements (grey dots).

patterns of two genotypes of soybeans: an experimental strain (P) and a commercial variety (F). We considered the logistic function applied by Pinheiro and Bates (2000, p.289) to these data and specified the following quantile model:

$$y_{ij} = \frac{\beta_{\tau,1} + u_{1,i}}{1 + \exp\left\{(\beta_{\tau,2} + u_{2,i} - x_{ij})/(\beta_{\tau,3} + u_{3,i})\right\}} + \epsilon_{\tau,ij},$$

separately for the two strains, where y_{ij} is the average leaf weight (g) per plant in plot $i = 1, \ldots, 48$ at x_{ij} days after planting, measured on occasion $j = 1, \ldots, n_i$. We assumed a diagonal matrix $\boldsymbol{\Sigma}_\tau$ and then estimated five quantiles at levels $\{0.05, 0.1, 0.5, 0.9, 0.95\}$ using the Laplacian approximation (similar results were obtained with the IS approximation).
The estimates of the fixed effects are given in Table 1, while a graphical presentation of the results is shown in Figure 1. Genotype P showed larger asymptotes (β_1) than genotype F at all quantiles. However, asymptotes were more similar between genotypes at $\tau = 0.5$. Curves were steeper (as suggested by the reciprocal of β_3) at lower quantiles in both genotypes, and genotype F had faster growth than genotype P at all quantiles, except for the median. There was considerable variation in the growth curves among plots and heterogeneity differed by quantiles in both strains.

Acknowledgments: This research has been supported by the National Institutes of Health – National Institute of Child Health and Human Development (Grant Number: 1 R03 HD084807-01A1).

References

Chen, C. (2007). A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics*, 16, 136–164.

Farcomeni, A. (2012). Quantile regression for longitudinal data based on latent Markov subject-specific parameters. *Statistics and Computing*, 22, 141–152.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8, 140–154.

Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, 24, 461–479.

Karlsson, A. (2008). Nonlinear quantile regression estimation of longitudinal data. *Communications in Statistics-Simulation and Computation*, 37, 114–131.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91, 74–89.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Marino, M. F. and Farcomeni, A. (2015). Linear quantile regression models for longitudinal experiments: an overview. *METRON*, 73, 229–247.

Oberhofer, W. and Haupt, H. (2016). Asymptotic theory for nonlinear quantile regression under weak dependence. *Econometric Theory*, 32, 686–713.

Pinheiro, J. and Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4, 12–35.

Pinheiro, J. C. and Bates, D. M. (2000), *Mixed-effects models in S and S-PLUS*. New York: Springer Verlag.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Wang, J. (2012). Bayesian quantile regression for parametric nonlinear mixed effects models. *Statistical Methods & Applications*, 21, 279–295.

Statistical Inference in the Duffing System with the Unscented Kalman Filter

Michela Eugenia Pasetto[1], Dirk Husmeier[2], Umberto Noè[2], Alessandra Luati[1]

[1] Department of Statistics, University of Bologna, Italy
[2] School of Mathematics and Statistics, University of Glasgow, Scotland

E-mail for correspondence: michela.pasetto2@unibo.it

Abstract: We investigate the accuracy of inference in a chaotic dynamical system (Duffing oscillator) with the Unscented Kalman Filter, and quantify the dependence on the sample size, the signal to noise ratio and the initialization.

Keywords: Bayesian filtering, Unscented Kalman Filter, Chaotic dynamical system, Parameter estimation

1 Introduction

We focus on the analysis of the deterministic Duffing process, defined as

$$dx_{1t}/dt = x_{2t}, \qquad dx_{2t}/dt = -(cx_{2t} + \alpha x_{1t} + \beta x_{1t}^3), \tag{1}$$

where x_{1t} and x_{2t} are the position and the velocity, respectively, of the oscillation at time t, $g(x) = \alpha x_{1t} + \beta x_{1t}^3$ is a restoring force, α is the natural frequency of the vibration, β the mode of the restoring force (hard or soft spring), and c is the damping term. The Duffing system (1) describes a periodically forced oscillator with a nonlinear elasticity, and has been widely used in physics, economics and engineering (Kovacic and Brennan, 2011). A characteristic feature is its chaotic behaviour, which makes statistical inference challenging. In the present paper we present an approach based on the Unscented Kalman Filter (UKF).

2 Methodology

The UKF algorithm is a non-linear generalization of Kalman filter which relies on the unscented transform (Julier and Uhlmann (2004)) in order to construct a Gaussian approximation to the filtering distribution. The UKF performs a Bayesian estimation of a state-space model:

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}) + \boldsymbol{\varepsilon}, \qquad \boldsymbol{y}_t = h(\boldsymbol{x}_t) + \boldsymbol{\eta} \tag{2}$$

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.

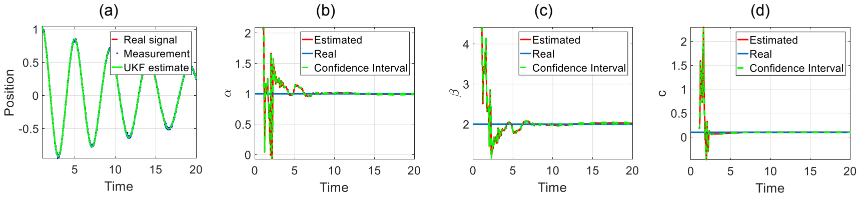


FIGURE 1. UKF estimates for the deterministic Duffing system with SNR=31 and $n = 1000$. (a) Signal estimate. (b) Estimate of parameter α. (c) Estimate of parameter β. (d) Estimate of parameter c.

where $\boldsymbol{x}_t \in \mathbb{R}^M$ is the (hidden) state at time t, $\boldsymbol{y}_t \in \mathbb{R}^D$ is the measurement, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$ is the Gaussian system noise and $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ is the Gaussian observation noise. The non-linear differentiable functions f and h are, respectively, the transition and observation models. UKF passes a deterministically chosen set of points (sigma points) through f to obtain the predictive distribution $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$. Then, the sigma points are transformed using model h to compute the filtering distribution $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$. As suggested in Sitz et al. (2002), we merge the signal with the parameter vector $\boldsymbol{\lambda} = [\alpha\ \beta\ c]^{\mathrm{T}}$ in a joint state vector $\boldsymbol{j}_t = [\boldsymbol{x}_t,\ \boldsymbol{\lambda}_t]^{\mathrm{T}} = [(f(\boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_{t-1}) + \boldsymbol{\varepsilon}),\ \boldsymbol{\lambda}_{t-1}]^{\mathrm{T}}$, and $\boldsymbol{y}_t = h(\boldsymbol{j}_t) + \boldsymbol{\eta}$. In our case, the function f of model (2) is given by the numerical solution of system (1), h is the identity function, and $\boldsymbol{\varepsilon} = \mathbf{0}$.

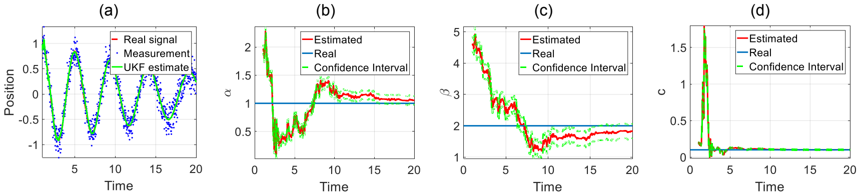


FIGURE 2. UKF estimates for the deterministic Duffing system with SNR=10 and $n = 1000$. (a) Signal estimate. (b) Estimate of parameter α. (c) Estimate of parameter β. (d) Estimate of parameter c.

3 Simulations

We simulate system (1) through the `ode23` MATLAB function with a stepsize of integration $\delta t = 0.01$ and starting values for the numerical integration $[1, 0]$. Measurements are obtained from the first component, x_{1t}, by adding observational noise $\eta_t \sim N(0, \sigma_\eta^2)$ with known variance. The time interval is $t = 1, \ldots, 20$, and the presented results are averaged over 10 simulations. The UKF algorithm is performed with the `EKF/UKF` toolbox of Hartikainen et al. (2011). To investigate the behaviour of the Duffing process and the UKF performance, we have simulated several scenarios, varying the Signal to Noise Ratio, SNR $\in \{30, 10, 1\}$,

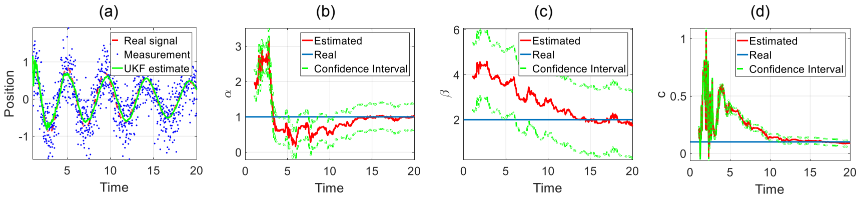


FIGURE 3. UKF estimates for the deterministic Duffing system with SNR=1 and $n = 1000$. (a) Signal estimate. (b) Estimate of parameter α. (c) Estimate of parameter β. (d) Estimate of parameter c.

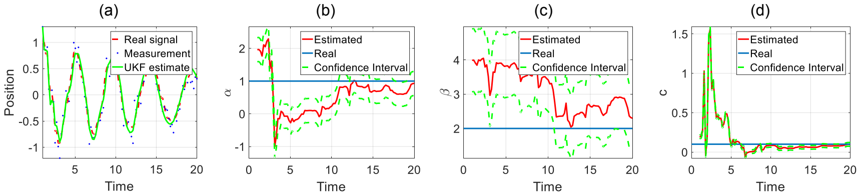


FIGURE 4. UKF estimates for the deterministic Duffing system with SNR=10 and $n = 100$. (a) Signal estimate. (b) Estimate of parameter α. (c) Estimate of parameter β. (d) Estimate of parameter c.

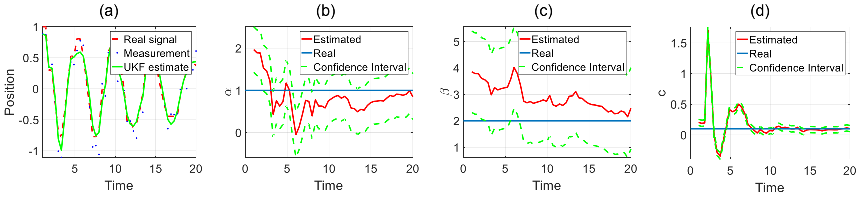


FIGURE 5. UKF estimates for the deterministic Duffing system with SNR=10 and $n = 50$. (a) Signal estimate. (b) Estimate of parameter α. (c) Estimate of parameter β. (d) Estimate of parameter c.

and the sample size, $n \in \{1000, 100, 50\}$ (Figures 1–5). To evaluate the impact of initialization, we considered different offsets as starting values for the parameters. The offsets are sampled randomly from a Gaussian distribution in which the mean is defined by a percentage deviation from the true parameter values and the variance is 10% of the mean (Table 1).

4 Results and Discussion

Figures 1–5 show that the UKF successfully learns the parameters from the noisy data, and that at the end of the filtering phase the true parameters always

TABLE 1. Impact of the initialization for the deterministic Duffing system for different offsets (as percentage of the true parameter values) in term of Euclidean norm prior inference and post inference.

	α		β		c	
	Prior	Post	Prior	Post	Prior	Post
100%	1.00	0.05	2.04	0.24	0.10	0.01
150%	1.52	0.12	3.02	0.50	0.15	0.01
200%	2.03	0.23	3.90	0.94	0.21	0.02
250%	2.48	0.65	4.61	2.12	0.25	0.04

lie within the predicted standard error around the estimate. This suggests that Bayesian filtering offers a successful paradigm for inference in chaotic dynamical systems. The prediction uncertainty depends on the sample size n, and the level of noise, quantified by the SNR. As one would expect, the uncertainty increases with decreasing n and decreasing SNR, i.e. as information in the data is lost, and our study allows a quantification of this trend. The increase in uncertainty particularly affects the parameter β, which is associated with the nonlinear term and the source of the chaotic behaviour. Table 1 shows the effect of the initialization, measured in terms of the Eucliedean distance in parameter space. This distance is consistently reduced in the filtering process, and the posterior distance (after filtering) is always smaller than the prior distance (before filtering). However, the posterior distance increases with the prior distance, suggesting that a good initialisation will improve the inference results.

References

Hartikainen, J., Solin, A., Särkkä, S. (2011). EKF/UKF Toolbox for MATLAB. `http://becs.aalto.fi/en/research/bayes/ekfukf/`.

Julier, S. J., and Uhlmann, J, K. (2004). Unscented Filtering and Nonlinear Estimation. *Proceedings of the IEEE*, 92(**3**), 401-422.

Kovacic I., Brennan, M. J., eds. (2011). *The Duffing Equation: Nonlinear Oscillators and their Behaviour*. New York: John Wiley & Sons.

Sitz, A., Schwarz, U., Kurths, J. and Voss, H. U. (2002). Estimation of parameters and unobserved components for nonlinear systems from noisy time series. *Phys. Rev.*, E 66(**1**), 016-210.

Effect of Health Risk Factors on All-Causes Mortality in Elderly People: A Joined Survival Analysis using Flexible Parametric Models of Cohorts coming from Brazil, Argentina and Italy.

Liciana Vaz de Arruda Silveira[1], Alberto Osella[2], Maria del Pilar Diaz[3]

[1] Biostatistics Department. Institute of Biosciences. Sao Paulo State University, Botucatu, Brazil.

[2] Laboratory of Epidemiology and Biostatistics, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Saverio de Bellis, Castellana Grotte, Italy.

[3] Biostatistics Unit, School of Nutrution and INICSA-CONICET, University of Córdoba, Córdoba, Argentina.

E-mail for correspondence: `pdiaz@fcm.unc.edu.ar`

Abstract: Aim: to estimate the effect of High Blood Pressure on mortality among people aged 65 years old in three scenarios with different socio-economic background and urbanization process by using flexible parametric survival models. Methods: Three cohorts coming from Brazil (n=365), Argentina (n=1800) and Italy (n=2472) were considered and only people with 65 years old included. Time to death (months) from enrolment and all-causes mortality were considered. Statistical analysis included Frailty Coxs Model and Flexible Parametric Survival Analysis. Due to the two-level structure of variability multilevel mixed-effect survival regression models were fitted. Main Results: Frailty Cox model showed significant positive effects of age and an effect modification of High Blood Pressure and Non-Communicable Diseases (NCD) but not effects proportionality. Multilevel modelling evidenced a positive statistically significant effect of Age, NCD and Smoking. There was also an effect modification of NCD on High Blood Pressure. Overall Multilevel Flexible Models estimates were more conservative with an increase in their precision. Conclusions: High Blood Pressure and NCD are important causal components and strong risk factors of cardiovascular mortality in these countries. The model obtained indicates a common causal model in three cohorts coming from very different environments.

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.

Keywords: Flexible survival model; Mortality; Elderly; Hypertension; Mixed model.

1 Introduction

The present work considers three different scenarios and proposes a joined survival analysis of elderly populations coming from different cohort studies conducted in Brazil, Argentina and Italy. The increase in the proportion of elderly population worldwide is one of the most important demographic facts both in developed countries and developing ones. Health problems affecting the elderly population are generally chronic, then factors associated with lower survival have been described and related to the risk of death. The aim of this work was to estimate the effect of High Blood Pressure (HBP) on all-causes mortality among people 65 years old in three scenarios with different socio-economic background and urbanization process by using flexible parametric survival models while considering countries variability.

1.1 Cohorts

Three cohorts coming from Brazil, Argentina and Italy were included. Brazil: individuals aged 60 years old were enrolled in 2003 and followed-up until 2010 at Botucatu (São Paulo State). A two-stage random sample (365 subjects) was used to collect information about socio-demographic characteristics, health status and personal history.

Argentina: Patients from a medical care center were enrolled in two cities of Córdoba province (January 2004 and still ongoing). Complete history about chronic diseases was recorded for 1,142 subjects.

Italy: The study was conducted at Castellana Grotte, (Bari province) enrolling 2472 subjects. A structured standard interview was performed to collect information about socio-demographic data, health status and personal history.

Time from date of enrolment to death, migration or end of study (in months), whichever occurred first. Follow-up ended at December 31st, 2010 in Brazil, June 30th, 2014 in Argentina and June 30th, 2015 in Italy. In this work only individuals of 65 years old were considered.

1.2 Modelling

To explore survival probabilities Kaplan-Meier method was applied. A Frailty Cox model assuming gamma-distributed latent random effects was fitted to handle unobserved heterogeneity due to pooling the studies. As proportional hazard assumption of Cox model was not satisfied, flexible parametric Royston-Parmar survival models (RPM, Royston & Parmar 2002) were chosen. This models use natural cubic splines to model baseline $g[S_0(t)]$ within the Aranda-Ordaz family of link function. Due to the two-level structure of variability (subjects nested in countries) of the overall dataset, multilevel mixed-effect survival regression models were fitted (Crowther et al. 2014), as it follows:

$$\begin{aligned} log[H_{ij}(t)] &= s(log(t)|\gamma, k_0) + X_{ij}^{'}\beta + Z_i^{'}b_i + \sum_{p=1}^{P} s(log(t)|\delta_p, k_p)X_{ijp}, \\ &= s(log(t)|\gamma, k_0) + X_{ij}^{'}\beta + b_i, \end{aligned}$$

with X_{ijp} covariate, $s(log(t)|\delta_p, k_p)$ spline and coefficients vector denoted as δ_p, random b_i intercept ($b_i \sim N(0, V)$). Gender, age, HBP, Diabetes Mellitus (DM), smoking habit and other NCDs were selected as covariates.
Post-estimation predicted hazards by categories of HBP at 65 and 80 years old were obtained. All statistical analyses were performed using Stata statistical software, version 14.1 (StataCorp LP, College Station, Texas).

2 Results and Conclusions

We refer to Table 1 for a summary of our main results of modelling. Overall Mixed-RPM estimates were more conservative with a gain in precision. Figure 1 shows predicted hazard for HBP for five years survival at 65 an 80 years old. There were increased hazards at 80 years old as evidenced on the vertical axis. Whereas there was an initial decreased hazard (which spans for different time periods) for men in Brazil and Italy, hazard for Argentina behaved in an irregular way.
The major findings of this study were the positive statistically significant main effect of Age, Smoking and NCDs on all-causes. There was also a strong effect modification of NCDs on HBP. These results were more plausible and attenuated estimates while gain in precision. The modelling process uncover a common causal model among different environments whereas maintain countries peculiarities, such as survival probabilities which remain substantially different.

TABLE 1. Frailty Cox Model and Two-level Mixed Survival Parametric Model. Brazil, Argentina and Italy.

	Frailty Cox Model		Multilevel Flexible Model	
Variable	HR	CI0.95	HR	CI0.95
Age	1.04**	1.01-1.08	1.06*	1.03-1.08
Sex(Female)	0.68	0.42-1.11	0.66	0.43-1.13
HBP(Yes)	0.99	0.59-1.91	1.07	0.61-1.87
DM(Yes)	0.51	0.27-2.13	0.95	0.54-1.66
NCD(Yes)	2.15**	1.31-3.56	2.12**	1.28-3.45
Smoking(Yes)	1.77*	1.05-2.98	1.61*	1.03-2.62
HBP*DM(Yes/Yes)	1.95	0.42-2.12	1.31	0.74-2.29
HBP*NCDs(Yes/Yes)	4.58**	1.14-18.37	2.30**	1.28-4.11
HBP*Smoking(Yes/Yes)	1.05	0.51-1.81	1.76	0.89-2.46

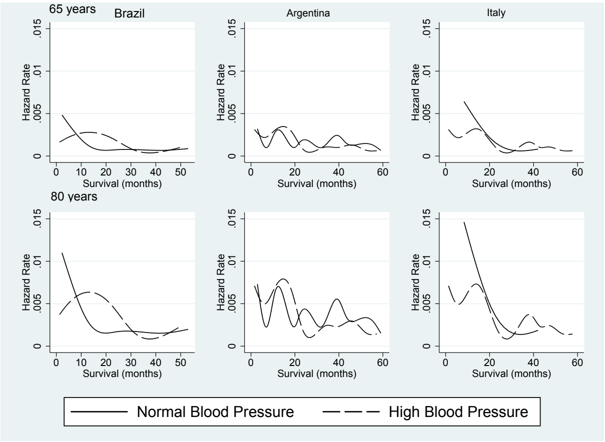


FIGURE 1. Hazard Rate by Country and Ages 65 and 80 years old for High Blood Pressure Status estimated using Multilevel Mixed Survival Parametric Model.

Acknowledgments: Special Thanks to all participants in the studies for their valuable time and information. This work was supported by São Paulo Research Foundation (FAPESP, Brazil) - Proc. No.2014/22856-7.

References

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata*. Texas: StataCorp.

Royston P., Parmar M.K.B. (2002). Flexible parametric proportional hazards and proportional odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, **21**, 2175 – 2197.

Crowther M.J., Look M.P., Riley R.D. (2014). Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Statistics in Medicine*, **33**, 3844 – 3858.

A new approach for clustering of effects in quantile regression

Gialuca Sottile[1] and Giada Adelfio[1]

[1] University of Palermo, Dip. di Scienze Economiche Aziendali e Statistiche, Italy

E-mail for correspondence: giada.adelfio@unipa.it

Abstract: In this paper we aim at finding similarities among the coefficients from a multivariate regression. Using a quantile regression coefficients modeling, the effect of each covariate, given a response (also multivariate) is a curve in the multidimensional space of the percentiles. Collecting all the curves, describing the effects of each covariate on each response variable, we could be able to assess if only one or more covariates have same effects on different responses.

Keywords: curves clustering; quantile regression coefficients modeling; multivariate analysis; functional data.

1 Introduction

Looking for curve similarity could be a complex issue characterized by subjective choices related to the continuous transformation of observed discrete data. Here, the alignment problem is handled introducing a new, simple and efficient procedure, based on a similarity measure between curves. A more general approach is based on the alignment of curves using a target function (Silverman, 1995). Adelfio et al. (2012) introduced a simple procedure to identify clusters of multivariate waveforms based on a simultaneous assignation and alignment procedure. This approach has been extended in Adelfio et al. (2016), where the authors focussed on finding clusters of multidimensional curves with spatio-temporal structure, applying a variant of a k-means algorithm based on the principal component rotation of data. Quantile regression can be used to fully describe the conditional distribution of an outcome and the effect of the covariates on it (Koenker and Bassett, 1978). In Frumento and Bottai (2015), the authors suggest adopting a parametric model for the coefficient function. They refer to this estimation approach as quantile regression coefficients modeling (QRCM).

In this paper we provide a new perspective of the curve similarity approach considering as curves the coefficient functions after applying a QRCM. A new

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.

algorithm is proposed to assess if two coefficients functions have the same behaviour. The rest of the article is organized as follows. Section 2 briefly presents the algorithm, Section 3 shows a simulation study and Section 4 provides conclusions.

2 Methods

In Frumento and Bottai (2015), the authors suggest to adopt a parametric model for the coefficient function of a quantile regression. Conversely to standard quantile regression which works in a quantile-by-quantile fashion, in the QRCM framework different quantiles are estimated one at the time. This modelling approach facilitates estimation, inference, and interpretation of the results, and generally yields a gain in terms of efficiency. Let us consider a response variable y and a set of q-covariates $\boldsymbol{x}$, the coefficients $\boldsymbol{\beta}(p)$ are defined as functions of $p \in (0,1)$ that depend on a finite-dimensional parameter $\boldsymbol{\theta}$,

$$\boldsymbol{\beta}(p \mid \boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(p),$$

where $\boldsymbol{b}(p) = [b_1(p), \dots, b_k(p)]^{\mathrm{T}}$ is a set of k known functions of p. In a multivariate framework, let $\boldsymbol{y} = [y_1, \dots, y_j, \dots, y_m]$ be a set of m response variables, correlated or not, and $\boldsymbol{x}$ be a set of q covariates. Applying the QRCM on each response variable, we estimate the coefficient functions $\beta_{1j}(p,\boldsymbol{\theta}), \dots, \beta_{qj}(p,\boldsymbol{\theta})$ over the percentiles. In this paper, starting from the QRCM estimation of curve effects, we propose a new algorithm to identify those covariates with the same effect on a single response, or, similarly, to identify the responses that are related by similar effect of a given covariate. In a generic framework, we investigate the similarities among n general curves, parametrized by $\beta_i(p)$, $i = 1, ..., n$. The clustering approach proposed in this paper is based on a new dissimilarity measures based both on shape and distance. More in the detail, we define a new dissimilarity measure, based on two measures accounting both for the shape and for the distance. Let $\beta_i(p)$ be the coefficient function approximated by a spline function $s_i(p)$, for $p = 1, ..., N_p$, $i = 1, ..., n$. Considering two different curves $\beta_i(p)$ and $\beta_{i'}(p)$ with $i \neq i'$, we define

$$\begin{aligned} d^{ii'}_{\text{shape}}(p) &= I(\text{sign}(s''_i(p)) \times \text{sign}(s''_{i'}(p)) = 1) \\ d^{ii'}_{\text{distance}}(p) &= I(|\beta_i(p) - \beta_{i'}(p)| \leq f(\alpha, \text{dist}(p))), \end{aligned}$$

where $s''_i(\cdot)$ is the second derivative of $\beta_i(\cdot)$ and $f(\cdot,\cdot)$ is a cut-off function, that depends on α, a probability value, and $\text{dist}(p)$, that is the vector of the distances between all the pairs of curves for each value of p. Computed the distribution of $\text{dist}(p)$ for each value of p, the cut-off function selects the corresponding $\alpha-$th percentile vector. Therefore, the proposed dissimilarity measure between two curves is defined as:

$$d^{ii'} = 1 - \frac{1}{N_p}\sum_{1=1}^{N_p}\left[d^{ii'}_{\text{shape}}(p) \cdot d^{ii'}_{\text{distance}}(p)\right] \tag{1}$$

In the proposed approach, the new dissimilarity measure is used to define a dissimilarity matrix, useful for the application of a hierarchical clustering method.

3 Simulation Study

Let us consider a multivariate scenario in which the quantile function is simulated as

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \beta_0(p \mid \boldsymbol{\theta}) + \beta_1(p \mid \boldsymbol{\theta})x_1 + \cdots + \beta_q(p \mid \boldsymbol{\theta})x_q,$$

where $x_1, x_2, \ldots, x_q$ are independent $\mathbb{U}(0,5)$ variables and $p \in \mathbb{U}(0,1)$. In the first simulation scenario, the intercept is modelled as a quantile normal distribution function (ϕ) for its flexibility. Other choices, as suggested in the original paper of Frumento and Bottai (2015), could be also considered. We use $q = 2$ covariates and define three groups of quantile functions

$$Q_1(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \phi(p)) + (.5 + .5p + p^2 + 2p^3)x_1 + (.5 + 2p + p^2 + .5p^3)x_2,$$
$$Q_2(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \phi(p)) + (-3 + .5p + p^2 + .5p^3)x_1 + (-1.5 - p - .5p^2 + p^3)x_2,$$
$$Q_3(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \phi(p)) + (.3 - .5p - p^2 + 2p^3)x_1 + (-.5 + p - .5p^2 - p^3)x_2,$$

Ten responses are generated for each quantile function (Q_1, Q_2, Q_3). Applying the QRCM method to these responses we obtained the 30 coefficients curves, namely curves effect, and their lower and upper bounds, useful to select the optimal number of clusters, for both covariates. The proposed algorithm is able to select the correct number of clusters and to discriminate all the curves effect. In Fig. 1, the curves for both covariates are represented in the three clusters. Results are summarized in terms of average cluster distances within clusters, which highlight closeness of curves, and silhouette widths to assess the cohesion of each curve compared to the other clusters. In particular, results show a valid clustering of curves, since silhouette widths are all greater than 0, and for clusters 2 and 3 these values are greater than 0.5. Moreover, all the average cluster distances are lower than or equal to 0.5.

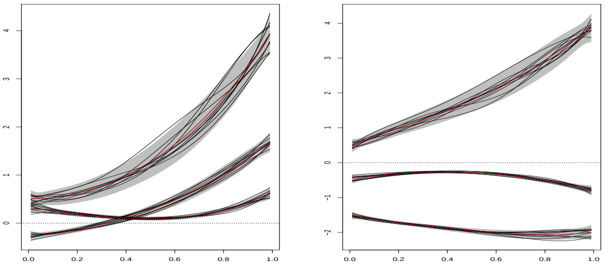


FIGURE 1. The simulated 30 curves (solid black lines) clustered in 3 clusters, conditioning to the first (left) and the second covariate (right), respectively. Red lines are the mean curves; the shaded areas are identified by the mean lower and upper bands within each cluster. The dotted line corresponds to zero.

4 Conclusions

The proposed method for curve-clustering provides a new perspective for the study of similarity, since it is used in a context of the coefficient functions of a quantile regression model. The proposed algorithm, can be actually used in any general context, since the main purpose is finding both close and similar shape curves. Although in this paper we briefly summarize some results, showing a good performance of the method, the curve similarity should be performed just on the subsets of quantiles where the effects are significant. This allows to account for a more comprehensive analysis of the relationship between variables along the distribution of the outcome variables.

Acknowledgments: This paper has been supported by the national grant of the MIUR for the PRIN-2015 program, "Prot. 20157PRZC4 - Research Project Title Complex space-time modeling and functional analysis for probabilistic forecast of seismic events. PI: Giada Adelfio"

References

Adelfio, G., Chiodi, M., DAlessandro, A., Luzio, D., DAnna, G. and Mangano, G. (2012). Simultaneous seismic wave clustering and registration. *Comput Geosci*, **44**, 60 – 69.

Adelfio, G., Di Salvo, F. Chiodi, M. (2016). Space-time FPCA clustering of multidimensional curves. In: *Proc. of the 48th Scientific Meeting of Ital. Stat. Soc.*, 1 – 12.

Frumento, P. and Bottai, M. (2015). Parametric modeling of quantile regression coefficient functions. *Biometrics*, **72** 74 – 84.

Koenker, R. and Bassett, G., Jr. (1978). Regression quantiles. *Econometrica*, **46**, 33 – 50.

Silverman, B.W. (1995). Incorporating parametric effects into functional principal components analysis. *J R Stat Soc Series B*, **57**, 673 – 689.

Social and material deprivation in the population aged over 50 in the Czech Republic in 2013

Ivana Malá

[1] University of Economics in Prague, Czech Republic

E-mail for correspondence: malai@vse.cz

Abstract: Finite normal mixture model is used to model the distribution of indices of deprivation in the Czech Republic. Two composite indicators are analysed (for material and social deprivation) based on the Survey of Health, Ageing and Retirement in Europe (survey SHARE) for the population aged more than 50 in the Czech Republic in 2013. The maximum likelihood estimates are compared to the moment method based on L-moments with the use of AIC information criterion. The numeric approach is used for the evaluation of quantiles of a mixture.

Keywords: L-moments; finite mixture of distributions; SHARE; deprivation

1 Methods

The finite normal mixture density is defined as (for K components)

$$f(x;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j f(x;\mu_j,\sigma_j),$$

where $f(x;\mu_j,\sigma_j), j = 1,...,K$ are component normal densities and π is a vector of weights of the components in the mixture. The vector of unknown parameters in the model $\boldsymbol{\psi}$ consists of the parameters (2K parameters) of component distributions $\mu_j, \sigma_j, j = 1,...,K$ and $K-1$ free parameters π_j fulfilling obvious constraints $0 \leq \pi_j \leq 1$, $\sum \pi_j \leq 1$. Usually the maximum likelihood estimates of parameters are used (in this text the package *mixtools*, Benaglia et all. (2009) was applied). In addition to this process, the moment method of estimation based on L-moments is proposed. For $3K-1$ parameters in the model the same number of "theoretical" L-moments are evaluated by the formula $(r = 1,...,3K-1)$

$$L_r = \int_0^1 Q(P) P^*_{r-1}(P) dP,$$

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

where Q_P is a quantile function (of the mixture) and P_r^* is the rth shifted Legendre polynomial. Unfortunately, there is not a closed formula for the quantile function of the mixture, however the weighted average of component quantiles is a useful approximation and initial value for a numeric method. Quantiles Q_P were evaluated numerically on the grid (as a root of an equation $F(Q_P) = P$, where F is a cumulative distribution function of the analysed mixture) and substituted into the formula for L-moments. Sample L-moments l_1 to l_{3K-1} were evaluated using *Lmoments* package (Karvanen, 2006) and equations $L_j = l_j, j = 1, ..., 3K - 1$ were solved with respect to unknown parameters in the model. Standard errors of estimates were evaluated using bootstrap. Both fits are compared with AIC criterion.

2 Data and Results

The Survey of Health, Ageing and Retirement in Europe (SHARE, Börsch-Supan (2016)) is a multidisciplinary and cross-national panel database of micro data of European population aged over 50. Data from 6 waves (from 2004 to 2015) are available at present. The module of deprivation is based on the work of (Adena at all., 2015). Values of two composite indicators, *depsoc* for social deprivation and *depmat* for material deprivation from this module are used to model the distribution of these indices in the Czech Republic in 2013. *Depmat* is an aggregate measure of material conditions of individuals aged over 50 in Europe using a set of 11 items that refer to two broad domains: the failure in the affordability of basic needs and financial difficulties. *Depsoc* is an index for measuring social deprivation, for this purpose 15 items from the survey were used. Answers with alternative values yes (in case of problems)/no are weighted into composite indices. Both indicators are transformed into $0 - 1$ scale, 0 means no deprivation and 1 is the highest deprivation.

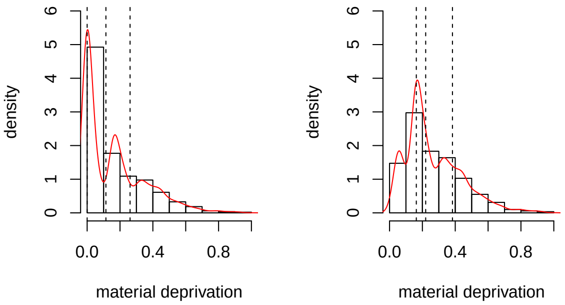


FIGURE 1. **Histogram and kernel density estimate for the index of material deprivation; the whole sample (left), positive values (right); quartiles are given with dashed lines.**

In the sample, n =4 091 respondents from the Czech republic aged above 50 are included (2 366 female and 1 725 male) with observed values of indices of

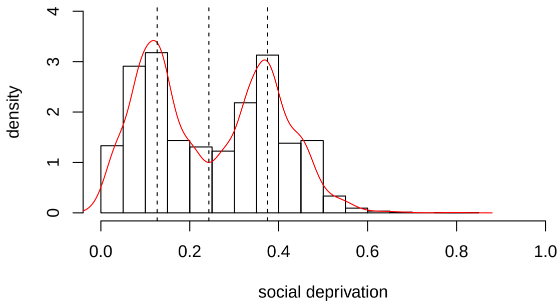


FIGURE 2. **Histogram and kernel density estimate for the index of social deprivation; quartiles are given with dashed lines.**

interest. A normal mixture with two components was selected for the index of social deprivation *depsoc* (see Figure 1). For the index *depmat* there is 41% of zeroes (no deprivation) in the data (Figure 2). For this reason, the model with a discrete part (value 0 with probability π_3) and a continuous part of $K = 2$ normal distributions for positive values. The value of π_3 was estimated a relative frequency of respondents with $depmat = 0$. Then for both models there are 5 parameters in the model and first five theoretical and sample L-moments must be evaluated to obtain five equations $L_j = l_j, j = 1, ..., 5$ to be solved with respect to unknown parameters. Both (MLE and moment method) approaches can be compared using AIC criterion. Fitted distributions are shown in the Figure 3. According to the AIC criterion both methods were comparable, however the chi-square test rejects the hypotheses of proper models ($P < 0.0001$). We consider both models to be acceptable for a description of probability distribution of analysed indices. The models enable identification of subgroups of individuals (Figure 3); two components with estimated expected values 0.113 (L-moments 0.106) and 0.367 (L-moments 0.356) for social deprivation and three with 0 (no material deprivation with estimated probability 0.406) and two components with 0.150 (L-moments 0.169) and 0.381 (L-moments 0.411).

The analysed indices, based on the SHARE data, correspond (at least for the Czech Republic) to the data regularly published by the Czech Statistical Office. Percentages of deprived (in all questions included in both surveys) are similar for age groups 50-65, 65+, that are used in official statistics. The mixture model is well applicable for the modelling of both distributions.

Acknowledgments: The institutional support of a long-term conceptual advancement of science and research number IP 400 040 at the Faculty of Informatics and Statistics, University of Economics is gladly acknowledged.

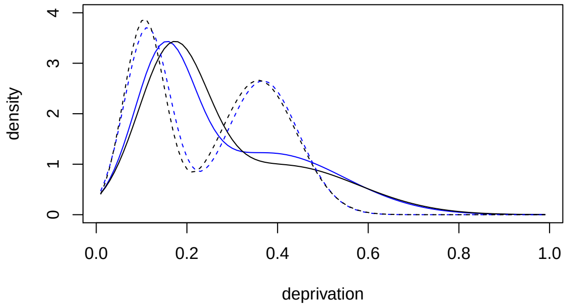


FIGURE 3. **Fitted probability distributions of the indices of deprivation; material deprivation solid lines, social deprivation dashed lines. Blue lines MLE, black lines moment method estimator.**

References

Adena, M., Myck, M., and Oczkowska, M. (2015). *Innovation for better understanding deprivation index.* In: *Ageing in Europe - Suporting Policies for an Inclusive Society*, De Gruyter.

Benaglia, T., Chauveau, D. and Hunter, D. R. (2009). mixtools: An R Package for Analyzing Finite Mixture Models *Journal of Statistical Software*, **32**, 1 – 29.

Börsch-Supan, A. (2016). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release version: 5.0.0. SHARE-ERIC. Data set.

Karvanen, J. (2006). Estimation of quantile mixtures via L-moments and trimmed L-moments. *Computational Statistics & Data Analysis*, **51**, 947 – 959.

On generalized additive partial linear models for correlated data

Roberto F. Manghi[1], Francisco J. A. Cysneiros[1], Gilberto A. Paula[2]

[1] Departamento de Estatística, Universidade Federal de Pernambuco, Brazil
[2] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

E-mail for correspondence: `giaapaula@ime.usp.br`

Abstract: In this paper we discuss estimation and diagnostic in generalized additive partial linear models (GAPLM) for analyzing correlated data. A reweighed iterative process based on the back-fitting algorithm is derived for the parameter estimation and discussions on the extension of some diagnostic procedures for GAPLM are given. A motivating data set is reanalyzed by the methodology developed in the paper.

Keywords: Diagnostic methods; B-splines; Generalized estimating equations; Semiparametric models.

1 Introduction

Generalized additive partial linear models (GAPLM) (see, for instance, Wang et al.,2014) comprise an important approach for modelling correlated data. Such models have the feature of jointly modelling the mean structure by parametric and nonparametric components, with the information only on the marginal distributions as well as on the within-subject correlation structure. The GAPLM class combines two well known approaches, generalized estimating equations (GEE) (Liang and Zeger, 1986) and generalized additive models (Hastie and Tibshirani, 1990).

The aim of this paper is to discuss estimation and diagnostic in GAPLM. The model is presented in section 2, whereas an iterative process for the joint estimation of the parametric and nonparametric parameters is derived in section 3. Discussions on the extension of some diagnostic procedures are also given. A motivating data set is reanalyzed in the last section.

2 The model

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^{\mathrm{T}}$ be the $m_i \times 1$ response vector for the i-th subject, for $i = 1, \ldots, n$. We will assume that the marginal distribution of y_{ij} belongs to the one-parametric exponential family of distributions, namely $y_{ij} \in \mathrm{EF}(\mu_{ij}, \phi)$, where $\mathrm{E}(y_{ij}) = \mu_{ij}$, $\mathrm{Var}(y_{ij}) = \phi^{-1}V_{ij}$, $\phi^{-1} > 0$ denotes the dispersion parameter and V_{ij} is the variance function. In addition, one has a link function $g(\cdot)$ such that

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=1}^{q} f_k(t_{ijk}),$$

where $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$ contains values of explanatory variables, $\boldsymbol{\beta}$ is a p-dimensional parameter vector and $f_k(\cdot)$ is an unknown smooth function of some continuous explanatory variable t_k. In matrix notation one has $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \sum_{k=1}^{q}\mathbf{N}_{ik}\mathbf{f}_k$, where $\mathbf{X}_i = (\mathbf{x}_{i1}^{\mathrm{T}}, \ldots, \mathbf{x}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{N}_{ik}$ is the k-th incidence matrix for the i-th subject and $\mathbf{f}_k = (f_k(t_{1k}^0), \ldots, f_k(t_{r_k k}^0))^{\mathrm{T}}$ with $(t_{1k}^0, \ldots, t_{r_k k}^0)$ denoting the distinct and ordered values of t_k. The working correlation matrix (within-subject) will be denoted by $\mathbf{R}_i(\boldsymbol{\rho})$, where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_s)^{\mathrm{T}}$, for $i = 1, \ldots, n$, $j = 1, \ldots, m_i$ and $k = 1, \ldots, q$.

3 Parameter estimation and diagnostics

Applying the Gauss-Seidel method to solve the penalized GEE the $(b+1)$-th step of the iterative process for obtaining the penalized estimates of $\boldsymbol{\beta}$ and $\mathbf{f}_1, \ldots, \mathbf{f}_q$, by fixing $\boldsymbol{\rho}$ and the smoothing parameters $\alpha_1, \ldots, \alpha_q$, may be expressed as

$$\begin{aligned}
\boldsymbol{\beta}^{(b+1)} &= \{\sum_{i=1}^{n}\mathbf{X}_i^{\mathrm{T}}\mathbf{W}_i^{(b)}\mathbf{X}_i\}^{-1}\sum_{i=1}^{n}[\mathbf{X}_i^{\mathrm{T}}\mathbf{W}_i^{(b)}\{\mathbf{z}_i^{(b)} - \sum_{\ell\neq 0}\mathbf{N}_{i\ell}\mathbf{f}_\ell^{(b+1)}\}] \\
\mathbf{f}_k^{(b+1)} &= \{\sum_{i=1}^{n}\mathbf{N}_{ik}^{\mathrm{T}}\mathbf{W}_i^{(b)}\mathbf{N}_{ik} + \alpha_k\mathbf{M}_k\}^{-1}\sum_{i=1}^{n}[\mathbf{N}_{ik}^{\mathrm{T}}\mathbf{W}_i^{(b)}\{\mathbf{z}_i^{(b)} - \sum_{\ell\neq k}\mathbf{N}_{i\ell}\mathbf{f}_\ell^{(b+1)}\}],
\end{aligned}$$

for $b = 0, 1, \ldots$ and $\ell = 0, 1, \ldots, q$, where $\mathbf{W}_i = \mathbf{D}_i\{\mathbf{V}_i^{\frac{1}{2}}\mathbf{R}_i(\boldsymbol{\rho})\mathbf{V}_i^{\frac{1}{2}}\}^{-1}\mathbf{D}_i$ denotes the weight matrix, $\mathbf{z}_i = \boldsymbol{\eta}_i + \mathbf{D}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$ is the modified dependent response with $\mathbf{D}_i$ denoting a diagonal matrix of derivatives $\mathrm{d}\mu_{ij}/\mathrm{d}\eta_{ij}$, $\mathbf{M}_k$ are related with the spline basis functions, $\mathbf{N}_{i0} = \mathbf{X}_i$ and $\mathbf{f}_0 = \boldsymbol{\beta}$, for $i = 1, \ldots, n$, $j = 1, \ldots, m_i$ and $k = 1, \ldots, q$. The iterative process above should be alternated with consistent estimates for $\boldsymbol{\rho}$ and the solution for $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_q)^{\mathrm{T}}$ leads to natural cubic splines. The smoothing parameters are estimated by the AIC criterion. Approximate standard errors are derived from the penalized estimated robust asymptotic variance-covariance matrix for $(\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\mathbf{f}}^{\mathrm{T}})^{\mathrm{T}}$.

Approximate leverage measures, normal conformal curvatures under some usual perturbation schemes and approximate standardized Pearson residuals are derived. Some simulation studies, under marginal distributions in the EF as well as within-subject correlation structures, indicate for a good agreement between the empirical distribution of the residuals and N(0,1).

4 Application

For illustration, we will consider a balanced study described in Myers et al. (2002, section 6.5) in which $n = 30$ rats had a leukemic condition induced. Three chemotherapy treatments were applied in groups of 10 rats, and each rat was observed in four time periods. The number of cancer cell colonies, the number of white blood cells (`wbc`) and the number of read blood cells (`rbc`) were observed at each period. Fig. 1(left) describes the dispersion graph between the log(number of cancer cells) and the number of `wbc`, and a nonlinear tendency may be observed. The dispersion graph between the log(number of cancer cells) and the number of `rbc` (omitted here) presents approximately a linear positive tendency.

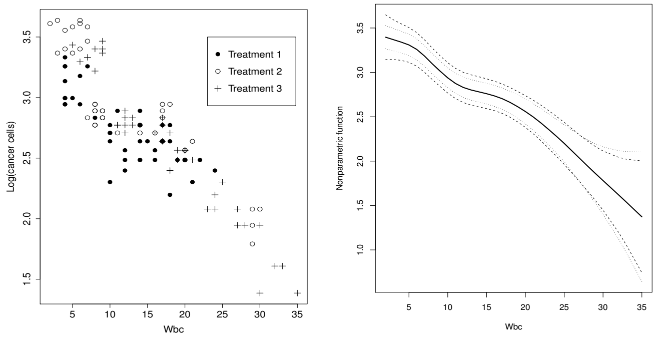


FIGURE 1. Dispersion graph between the log(number of cancer cells) and wbc (left) and the simultaneous 95% confidence intervals for $f(\texttt{wbc})$ (right).

Denoting by y_{ij} the number of cancer cells for the i-th rat at the j-th instant, we propose to explain $\mu_{ij} = \mathrm{E}(y_{ij})$ the following GAPLM: (i) $y_{ij} \sim \mathrm{P}(\mu_{ij})$, (ii) $\log(\mu_{ij}) = \beta_1 \texttt{Treat1}_i + \beta_2 \texttt{Treat2}_i + \gamma \texttt{rbc}_{ij} + f(\texttt{wbc}_{ij})$ and (iii) $\mathrm{corr}(y_{ij}, y_{ij'}) = \rho^{|j-j'|}$, for $i = 1, \ldots, 30$ and $j, j' = 1, 2, 3, 4$, where `Treat1` and `Treat2` denotes dummy variables.

The model was fitted in `R` (codes developed by the authors) and we found the parameter estimates (approximate standard errors) $\hat{\beta}_1 = -0.249(0.067)$, $\hat{\beta}_2 = -0.031(0.065)$, $\hat{\gamma} = 0.015(0.007)$ and $\hat{\rho} = 0.508$, with $f(\texttt{wbc})$ being significant. The smoothing parameter and the effective degrees of freedom were estimated as $\alpha = 636.84$ and $\mathrm{df}(\alpha) = 6.80(3 + 3.80)$, respectively. Fig. 1(right) presents the simultaneous 95% confidence intervals for $f(\texttt{wbc})$. The normal probability plot with the standardized Pearson residual (Fig. 2(left)) does not present unusual features and a measure related with rat #28 is pointed out in Fig. 2(right). This rat received treatment 3, and has the lowest number of cancer cells and the largest number of `wbc` counts.

Therefore, based on the above results, one may conclude that treatment 1 seems to reduce the average count of cancer cell colonies, whereas the other treatments do not seem to have significant effects. In addition, the average count of cancer cells colonies seems to decrease as the number of `wbc` increases and seems to increase as the number of `rbc` increases.

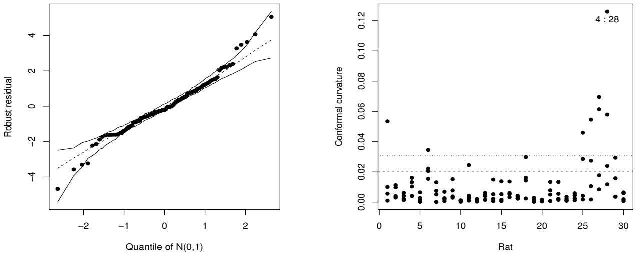


FIGURE 2. Normal probability plot for the standardized Pearson residual (left) and the index plot of the normal conformal curvature under the case-weight perturbation scheme (right).

Acknowledgments: Special Thanks to CNPq and FAPESP, Brazil.

References

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

Liang, K-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13 – 22.

Myers, R.H., Montgomery, D.C. and Vining, G.G. (2002). *Generalized Linear Models: with Applications in Engineering and the Sciences.* New York: Wiley.

Wang, L., Xue, L., Qu, A. and Liang, H. (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics*, **42**, 592 – 624.

What is the best approach to analyze longitudinal bounded scores? Application to Quality of Life data

Anne-Françoise Donneau[1], Cibele Russo[2], Nadia Dardenne[1], Murielle Mauer[3], Corneel Coens[3], Andrew Bottomley[3], Emmanuel Lesaffre[4]

[1] Biostatistics, University of Liège, CHU Sart Tilman, Liège, Belgium
[2] Department of Applied Mathematics and Statistics, University of Sao Paulo, Brazil
[3] EORTC Headquarters, Departments of statistics and quality of life, Brussels, Belgium
[4] L-Biostat, KU Leuven, Leuven, Belgium

E-mail for correspondence: afdonneau@ulg.ac.be

Abstract: The present study compares the linear mixed-effects model and the beta regression model to analyze longitudinal health-related quality of life bounded outcome scores.

Keywords: Bounded outcome scores; Longitudinal; Health-related quality of life.

1 Aims

In medical studies, it is a common practice to assess at prespecified time intervals the patient's quality of life (QoL) by means of Likert-type items questionnaire that covers different domains of the QoL (e.g., physical, emotional, social health issues). Usually, the items are summated and linearly transformed to construct a bounded score ranging from 0 to 100. The aim of the present study is to contrast different approaches, namely the linear mixed-effects model and the beta regression model, to analyze longitudinal bounded outcome scores.

2 Methods

The EORTC QLQ-C30 Fatigue scale and the Social Functioning scale from the 537 patients in the EORTC 26981 trial, a randomized multicenter phase III in

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.

glioblastoma evaluating the addition of temozolomide (TMZ) to radiotherapy (RT), were compared between the two arms using different statistical approaches for bounded outcome scores (see, for instance, Aaronson et al., 1993 and Taphoorn et al., 2005). Quality of life was assessed using the EORTC QLQC30 version 2 questionnaire: at baseline; during radiotherapy at week 4; 4 weeks after completion of radiotherapy; at the end of the third and sixth cycle of adjuvant temozolomide; and every 3 months thereafter until disease progression for patients allocated RT+TMZ, and at equivalent time points for those allocated radiotherapy alone. The statistical approaches selected in the present study included the beta regression model (for bounded values) from Ferrari and Cribari-Neto (2004) and the commonly used linear mixed-effects model (for continuous values) from Verbeke and Molenberghs (2000). Scores were divided by 100 to fit to the condition of application of the beta distribution.

3 Results

The magnitude of the P-values for the treatment difference derived under each statistical approach slightly varied at some assessment time points. However, the estimations of the treatment mean scores at each time point were comparable (See Table 1).

TABLE 1. Estimated mean scores in both arms and P-value related to treatment effect at each assessment time for the Fatigue scale.

	Linear mixed model			Beta regression model		
	RT	RT+TMZ	P-value	RT	RT+TMZ	P-value
Baseline	0.36	0.35	0.76	0.34	0.34	0.99
During RT	0.36	0.40	0.07	0.35	0.39	0.11
After RT	0.37	0.41	0.08	0.36	0.40	0.12
1st follow-up	0.33	0.40	0.05	0.31	0.40	0.04
2nd follow-up	0.31	0.32	0.90	0.28	0.30	0.63
3rd follow-up	0.29	0.31	0.74	0.30	0.34	0.54
4th follow-up	0.30	0.27	0.72	0.31	0.30	0.90
-2 Log Lik			-305.9			-785.6
AIC			-273.9			-753.6
BIC			-205.4			-685.0

RT = Radiotherapy / RT+TMZ = Radiotherapy and temozolomide

4 Conclusion

The preliminary analysis of these two QoL scales showed that both statistical approaches led to the same conclusion when considering the treatment effect, P-values and the mean scores. However, the beta regression model presented a better model fit for the QoL scales. This indicates that incorporating the bounded outcome assumption into the analysis methods can improve QoL hypothesis testing. Other models for longitudinal bounded outcome scores, such as truncated regression model and coarsening approach, will be investigated and compared to these preliminary results.

Acknowledgments: Second author thanks to Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, Brazil.

References

Aaronson, N. et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, **85**, 365 – 376.

Ferrari, S. and Cribari-Neto, F. (2004). Beta Regression or Modeling Rates and Proportions. *Journal of Applied Statistics*, **31(7)**, 799 – 815.

Taphoorn, M. et al. (2005). Health-related quality of life in patients with glioblastoma: A randomized controlled trial. *Lancet Oncology*, **6(12)**, 937 – 944.

Verbeke, G., and Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer.

Nonlinear mixed-effects models with scale mixtures of skew-normal distributions

Marcos Antonio Alves Pereira[1], Cibele Maria Russo[2]

[1] Federal University of Piauí, Floriano-Brazil
[2] University of São Paulo, São Carlos-Brazil

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

Abstract: This work presents a mixed-effects nonlinear regression model with a class of skewed distributions, which provide alternatives to the normal model and other symmetric distributions and may allow better fit to correlated nonlinear data. Many studies have been proposing models with scale mixtures of skew-normal (SMSN) family of distributions, including skew distributions with heavy-tailed. A real data analysis with this new proposal is performed.

Keywords: nonlinear model, scale mixtures of skew-normal distributions, skewness, mixed-effects.

1 Introduction

Models with mixed effects can be applied in several areas of knowledge, in particular when data are correlated. Nonlinear models with mixed effects have been explored recently because of their flexibility to deal with measures in areas such as economics and pharmacokinetics. Russo et al. (2009) developed a nonlinear model with random effects assuming elliptic distributions for the random components of the model. In this work we present an asymmetric nonlinear model with mixed-effects in which the associated error follows a distribution belonging to the scale mixtures of skew-normal (SMSN) family of distributions, which comprises distributions such as skew-normal, skew-Student-t, skew-contaminated normal, skew-exponential power and skew-slash and the symmetric versions of these distributions.

A random vector $\mathbf{y} \in R^p$ follows a SMSN distribution, as discussed in Zeller et al. (2011), $\mathbf{Y} \sim SMSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, H)$, with location vector $\boldsymbol{\mu} \in R^p$, matrix scale $\boldsymbol{\Sigma}$ positive definite with dimension $p \times p$ and skewness vector $\boldsymbol{\lambda} \in R^p$ if its probability density function (pdf) is given by

$$f(\mathbf{y}) = 2\int_0^\infty \phi_p(\mathbf{y}|\boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma})\Phi_1(\kappa^{-1/2}(u)\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu}))dH(u;\boldsymbol{\nu}),$$

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

where $\kappa(.)$ is a strictly positive function, U is a positive random variable with cumulative distribution function (cdf) $H(u;\boldsymbol{\nu})$, being $\boldsymbol{\nu}$ supposedly known, $d = (\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ is the Mahalanobis distance, $\phi_p(.|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the pdf of the p-variate normal distribution and $\Phi_1(.)$ is the cdf of the univariate standard normal distribution. In the SMSN family, if the vector of parameters $\boldsymbol{\lambda}$ is null, we will have a SMN family of distributions.

2 The model

A nonlinear regression model with mixed effects and under the SMSN family of distributions, where the observed responses vector $\mathbf{y}_i$, $i = 1,\dots,n$, of dimension $m_i \times 1$, according to Vonesh and Carter (1992), can be written as

$$\mathbf{y}_i = \mathbf{f}(\boldsymbol{\alpha},\,\mathbf{x}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_p)^\top$ is the vector of unknown parameters, $\mathbf{Z}_i$ is a $m_i \times r$ matrix of known constants. The random effects $\mathbf{b}_i = (b_{i1},\dots,b_{ir})^\top$ and the errors $\boldsymbol{\epsilon}_i = (\epsilon_{i1},\dots,\epsilon_{im_i})^\top$ are uncorrelated and follow multivariate SMSN, with null location vector and covariances matrices $\mathbf{D}$ and $\sigma^2\mathbf{I}_{m_i}$, respectively, and $\boldsymbol{\lambda}_{\boldsymbol{\epsilon}_i} = \mathbf{0}$. So we have the model

$$\begin{pmatrix}\mathbf{y}_i\\ \mathbf{b}_i\end{pmatrix} \sim SMSN_q\left(\begin{pmatrix}\mathbf{f}(\boldsymbol{\alpha},\,\mathbf{x}_i)\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_i & \mathbf{Z}_i\mathbf{D}\\ \mathbf{D}\mathbf{Z}_i^\top & \mathbf{D}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\lambda}_{\mathbf{y}_i}\\ \boldsymbol{\lambda}_{\mathbf{b}_i}\end{pmatrix}, H\right). \tag{2}$$

where $q = m_i + r$ and $\boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \sigma^2\mathbf{I}_{m_i}$ and $\mathbf{y}_i$ has marginal distribution $SMSN_{m_i}(\mathbf{f}(\boldsymbol{\alpha},\mathbf{x}_i),\boldsymbol{\Sigma}_i,\boldsymbol{\lambda}_{\mathbf{y}_i};H)$.

3 Maximum likelihood estimation

The estimation of the parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top,\sigma^2)^\top$ is obtained numerically via the EM-type algorithm and Fisher scoring algorithm. The model (2), with $\kappa(u_i) = 1/u_i$, can be represented hierarchically as

$$\mathbf{y}_i|T_i = t_i, U_i = u_i \overset{\text{ind.}}{\sim} N_{m_i}(\mathbf{f}(\boldsymbol{\alpha},\mathbf{x}_i) + \boldsymbol{\Delta}_i t_i, u_i^{-1}\sigma^2\mathbf{I}_{m_i}), \tag{3}$$

$$T_i|U_i = u_i \overset{\text{ind.}}{\sim} HN_1(0,u_i^{-1}) \text{ and} \tag{4}$$

$$U_i \overset{\text{ind.}}{\sim} h(u_i;\boldsymbol{\nu}), \tag{5}$$

where HN_1 denotes the univariate half normal distribution, $\boldsymbol{\Delta}_i = \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{\delta}_i$, $\boldsymbol{\delta}_i = \boldsymbol{\lambda}_i/(1+\boldsymbol{\lambda}_i^\top\boldsymbol{\lambda}_i)^{1/2}$ and $\sigma^2\mathbf{I}_{m_i} = \boldsymbol{\Sigma}_i - \boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^\top$, thus $\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top = \boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^\top$.
Let $\mathbf{y} = (\mathbf{y}_1^\top,\dots,\mathbf{y}_n^\top)^\top$, $\mathbf{u} = (u_1^\top,\dots,u_n^\top)^\top$ and $\mathbf{t} = (t_1^\top,\dots,t_n^\top)^\top$, under the hierarchical representation (3)-(5), the complete log-likelihood function associated with $\mathbf{y}_c = (\mathbf{y}^\top,\mathbf{u}^\top,\mathbf{t}^\top)^\top$ is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}|\mathbf{y}_c) \propto\ & \frac{1}{2}\sum_{i=1}^{n}\Big\{-\log|\sigma^2\mathbf{I}_{m_i}| - \frac{u_i}{\sigma^2}(\mathbf{y}_i - \mathbf{f}(\boldsymbol{\alpha},\mathbf{x}_i))^\top(\mathbf{y}_i - \mathbf{f}(\boldsymbol{\alpha},\mathbf{x}_i))\\ & + \frac{2ut_i}{\sigma^2}\boldsymbol{\Delta}_i^\top(\mathbf{y}_i - \mathbf{f}(\boldsymbol{\alpha},\mathbf{x}_i)) - \frac{ut_i^2}{\sigma^2}\boldsymbol{\Delta}_i^\top\boldsymbol{\Delta}_i\Big\}\end{aligned} \tag{6}$$

where $\widehat{u_i} = E(U_i|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i)$, $\widehat{ut_i} = E(U_iT_i|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i)$ and $\widehat{ut_i^2} = E(U_iT_i^2|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}_i)$ (see Lachos et al. 2010). Maximizing $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = E[\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k)}]$ in the k-th iteration of the EM-type algorithm, about $\widehat{\boldsymbol{\theta}}^{(k)}$, we obtain

$$\begin{aligned}
\widehat{\sigma^2}^{(k+1)} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{u}_i^{(k)}\widehat{\mathbf{r}}_i^{(k)\top}\widehat{\mathbf{r}}_i^{(k)} - 2\widehat{ut}_i^{(k)}\widehat{\boldsymbol{\Delta}}_i^{(k)\top}\widehat{\mathbf{r}}_i^{(k)} + \widehat{ut_i^2}^{(k)}\widehat{\boldsymbol{\Delta}}_i^{(k)\top}\widehat{\boldsymbol{\Delta}}_i^{(k)}\right\} \text{ and} \\
\widehat{\boldsymbol{\Delta}}_i^{(k+1)} &= \frac{\widehat{ut}_i^{(k)}\widehat{\mathbf{r}}_i^{(k)}}{\widehat{ut_i^2}^{(k)}}, \text{ where } \widehat{\mathbf{r}}_i^{(k)} = (\mathbf{y}_i - \mathbf{f}(\widehat{\boldsymbol{\alpha}}^{(k)}, \mathbf{x}_i)).
\end{aligned}$$

At each iteration of the EM-type algorithm, the estimates of $\boldsymbol{\alpha}^{(k)}$ are obtained by Fisher scoring algorithm

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}^{(k+1)} = \widehat{\boldsymbol{\alpha}}^{(k)} \quad + \quad & \left[\sum_{i=1}^{n}\left\{\widehat{u}_i^{(k)}\widehat{\mathbf{J}}_i^{(k)\top}\widehat{\mathbf{J}}_i^{(k)} + \widehat{ut}_i^{(k)}\widehat{\mathbf{L}}_i^{(k)\top}\widehat{\boldsymbol{\Delta}}_i^{(k)}\right\}\right]^{-1} \\
& \left[\sum_{i=1}^{n}\left\{\widehat{\mathbf{J}}_i^{(k)\top}\left(\widehat{u}_i^{(k)}\widehat{\mathbf{r}}_i^{(k)} - \widehat{ut}_i^{(k)}\widehat{\boldsymbol{\Delta}}_i^{(k)}\right)\right\}\right],
\end{aligned}$$

where $\widehat{\mathbf{J}}_i^{(k)} = \partial\mathbf{f}(\widehat{\boldsymbol{\alpha}}^{(k)}, \mathbf{x}_i)/\partial\widehat{\boldsymbol{\alpha}}^{(k)\top}$ and $\widehat{\mathbf{L}}_i^{(k)} = \partial^2\mathbf{f}(\widehat{\boldsymbol{\alpha}}^{(k)}, \mathbf{x}_i)/\partial\widehat{\boldsymbol{\alpha}}^{(k)}\partial\widehat{\boldsymbol{\alpha}}^{(k)\top}$. The prediction of the random effects $\mathbf{b}_i$ was obtained by empirical Bayes method.

4 Application

The data from a study of the kinetics of the theophylline anti-asthmatic agent presented in Pinheiro and Bates (2000) were analyzed with this new proposal and compared with symmetric distributions, thus, nonlinear models were fitted under skew-normal distribution and skew-Student-t distribution (skew-t) and compared to normal distribution and Student-t distribution (t), whose fits are presented in Russo (2010). We consider the models (1) and (2) where $\boldsymbol{\alpha} = (lk_e, lk_a, lc_l)$, with $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^\top$ the concentration measurement vector $C_i(t)$ and $\mathbf{x}_i = \mathbf{t}_i = (t_{i1}, \ldots, t_{im_i})^\top$ the vector representing the time measurements. The nonlinear model for theophylline concentration using time as a covariate can be represented by

$$y_{ij} = \frac{D_i(\mathrm{e}^{lk_e+lk_a+lc_l})}{\mathrm{e}^{lk_a} - \mathrm{e}^{lk_e}}(\mathrm{e}^{-t_{ij}\,\mathrm{e}^{lk_e}} - \mathrm{e}^{-t_{ij}\,\mathrm{e}^{lk_a}}) + \epsilon_{ij}, \; i = 1, \ldots, n \text{ and } j = 1, \ldots, m_i.$$

where D_i is the dose aplicated of the drug (in mg/kg) to the ith individual. We observe that the model under the skew-Student-t distribution presented the lowest standard errors for the parameters lk_e, lk_a and lc_l, but for the parameter σ^2 the smallest standard error occurred with the skew-normal distribution and according to the AIC and BIC information criteria obtained, the more appropriate distribution for this application is the skew-normal distribution because it presents lower values for these statistics.

5 Discussion

The mixed nonlinear model presented is a general case of nonlinear mixed models that include symmetric and asymmetric distributions. The proposed model can be used when atypical observations or asymmetry are present, due to the flexibility of the SMSN family of distributions.

Acknowledgments: Special thanks to FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil and to UFPI - Federal University of Piauí, Brazil.

References

Lachos, V. H., Ghosh, P. and Arellano-Valle, R. B. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica.* **20**(1), 303 – 322.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effect models in S and S-PLUS.* Springer.

Russo, C. M., Paula, G. A. and Aoki, R. (2009). Influence diagnostics in nonlinear mixed-effects elliptical models. *Comput. Statistics & Data Analysis*, **53**(12), 4143 – 4156.

Russo, C. M. (2010). *Modelos não lineares elípticos para dados correlacionados.* PhD thesis (in Portuguese), Universidade de São Paulo, Brasil.

Vonesh, E. F. and Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, **48**(1), 1 – 17.

Zeller, C. B., Lachos, V. H. and Vilca-Labra, F. E. (2011). Local influence analysis for regression models with scale mixtures of skew-normal distributions. *Journal of Applied Statistics*, **38**(2), 343 – 368

Modeling spatio-temporal earthquake dynamics using generalized functional additive regression

Alexander Bauer[1], Fabian Scheipl[1], Helmut Küchenhoff[1], Alice-Agnes Gabriel[2]

[1] Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany
[2] Department of Geophysics, Ludwig-Maximilians-Universität, Munich, Germany

E-mail for correspondence: `Alexander.Bauer@stat.uni-muenchen.de`

Abstract: We present a generalized functional additive regression model for spatio-temporal functional data representing surficial ground velocities of simulated earthquakes. The data come from a large-scale *in silico* experiment to investigate the effects of physical parameters describing the conditions at the triggering fault on the large-scale dynamics of earthquakes. We describe the performance of a recently developed efficient inference algorithm on this huge data set with a complex effect structure. The estimated effects are geophysically plausible and the methodological approach seems promising.

Keywords: functional regression; generalized additive model; penalized splines; geophysics; big data

1 Introduction and data

The physics of seismic waves emanated by earthquakes is well-understood, but challenging with respect to the natural complexity of topography, subsurface velocity structure and earthquake source effects on large scales relevant for earthquake hazard assessment. Furthermore, current earthquake models rarely include the effects of earthquake faulting dynamics. To tackle both of these challenges, we use a generalized functional additive regression to quantify how physical conditions at an earthquake fault affect the surficial ground velocity measured over time.

The simulated earthquake data we analyse are derived from a large scale computer experiment with the open source software SeisSol (Breuer, 2014) and are based on a real earthquake that took place in Northridge (USA) in 1994. Absolute ground velocities were simulated solving elastic wave equations and are defined as 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

(isotropic) L^2 norm of the ground velocities in east-west, north-south and vertical direction. Each of the simulations used a different set of physical conditions at the fault. Surficial ground velocities were then recorded at 6146 virtual seismometers. For the analysis, the first 15s of the absolute ground velocity measurements from 135 simulations were used in a resolution of 2Hz. Leading zeros were discarded until the first relevant observation (≥ 0.01).

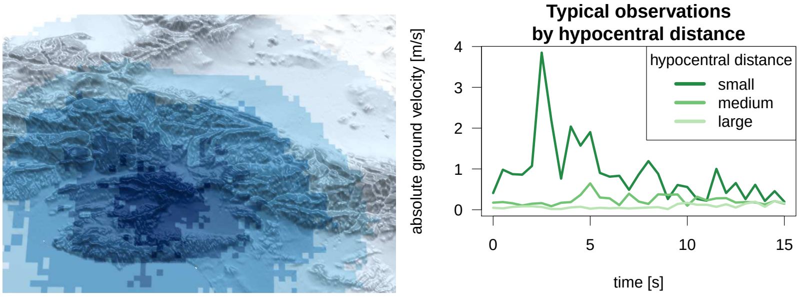


FIGURE 1. Left: Categorized mean absolute ground velocity in one simulation over the area under study, darker colours correspond to increased velocity. Right: Typical observations of absolute ground velocity over time. The initial peak of the ground velocity is delayed and smaller as hypocentral distance increases.

The evaluated predictors were all constant over time. Our main interest was in five physical parameters, which were pre-set in each simulation and which consisted of three frictional resistance variables, the direction of the regional tectonic background stress and the binary soil material of the whole area ($\{rock, sediment\}$). Additionally included were the moment magnitude as the classical measure of earthquake size, the local topography as well as the height and the hypocentral distance of each seismometer.

2 Methodology

Following Scheipl, Gertheiss and Greven (2016), the generalized functional additive model for function-on-scalar regression is written in the form $y_i(t_l) \sim F(\mu_{il}, \boldsymbol{\nu})$, where $g(\mu_{il}) = \sum_{r=1}^{R} f_r(\boldsymbol{\mathcal{X}}_{ri}, t_l)$, with seismometers $i = 1, \ldots, n$, absolute seismic ground velocity $y_i(t_l)$ recorded at time point t_l and the number of additive effects R with associated covariates $\boldsymbol{\mathcal{X}}_{ri}$. Conditional on the additive predictor and known response function $g(\cdot)$, $y_i(t_l)$ is assumed to come from some given distribution F with conditional expectation μ_{il} and dispersion and shape parameters $\boldsymbol{\nu}$. Each functional additive effect $f_r(\boldsymbol{\mathcal{X}}_{ri}, t)$ is represented in terms of a (tensor product) P-spline basis. A prediction error based approach was used for tuning basis sizes as the high-dimensional data dominate the penalization prior in the estimation. The model is fitted on 468,503 data points and comprises 18 smooth effects with a total of 633 spline coefficients and 25 smoothing parameters. We used a recently developed highly performant algorithm for penalized likelihood-based inference from Wood et al. (2016) that makes estimation of this complex model on such large data feasible. Its major advances are the use

of a highly efficient and parallelizable block-wise Cholesky decomposition and a compressed representation of the (marginal) spline basis.
Local small-scale (radius 300m) and large-scale (2000m) topography at each measuring station was included via the *Topographic Position Index* (TPI) introduced by Weiss (2001), defined as the height difference between the seismometer and the mean height in a circular neighborhood.

3 Results

A Gamma model with exponential response function $g(\cdot)$ explained 70.7% of the null deviance. Excluding moment magnitude from the model improves model performance measured by test set MSE. This is in line with a secondary finding that moment magnitude can be predicted quite precisely (98.2% explained deviance in a linear model with log-transformed response) based on the five parameters describing the simulation setup.

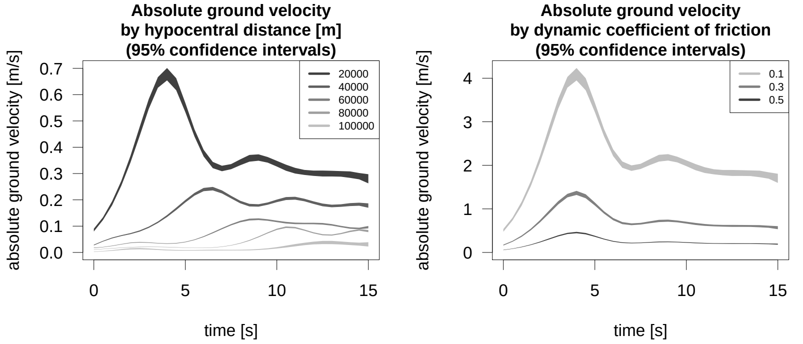


FIGURE 2. Pointwise 95% confidence intervals for the predicted mean of absolute ground velocity for various values of the two most important predictors, with remaining covariates being held constant at realistic values. Left: hypocentral distance. Right: dynamic coefficient of friction.

Among the remaining predictors the hypocentral distance and the dynamic coefficient of friction show by far the strongest effects, a visualization is given in Figure 2. Additional results can be found in Bauer (2016). All the estimated effects seem geophysically plausible.
In total the model has an acceptable fit, although absolute peak ground velocities are often underestimated. Remaining spatial structure of the mean residuals is shown in Figure 3.

4 Conclusion

Functional additive regression models are a promising approach in modeling surficial ground velocity. Major effects are the hypocentral distance and the dynamic coefficient of friction, with higher values leading to decreased ground velocities for both. The estimated effects all seem geophysically plausible. In future research,

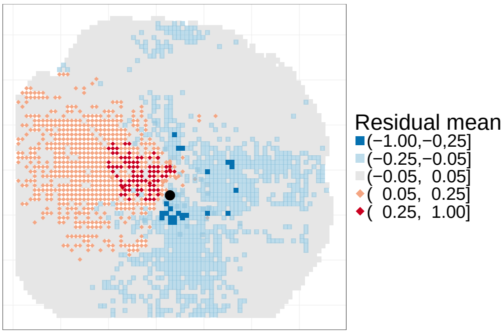


FIGURE 3. Mean residuals (averaged over time and over simulations) over the area under study. The black dot marks the epicenter.

the model will be refined further, e.g. by explicitly modeling spatial correlation and by relaxing the strict assumption of the hypocenter as fixed point source for all simulated earthquakes.

References

Bauer, A. (2016). *Auswirkungen der Erdbebenquelldynamik auf den zeitlichen Verlauf der Bodenbewegung*. MA thesis. Ludwig-Maximilians-Universität, Munich, Germany. Available: https://epub.ub.uni-muenchen .de/31976/

Breuer, A. et al. (2014). Sustained petascale performance of seismic simulations with seissol on supermuc. *Supercomputing - 29th International Conference, ISC 2014*, Leipzig, Germany, 1 – 18. Springer International Publishing.

Scheipl, F., Gertheiss, J., Greven, S. (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics*, **10.1**, 1455 – 1492.

Weiss, A. (2001). Topographic position and landforms analysis. *Poster presentation, ESRI user conference*, San Diego, CA, **200**.

Wood, S.N. et al. (2016) Generalized additive models for gigadata: modelling the UK black smoke network daily data. *Journal of the American Statistical Association*. DOI: 10.1080/01621459.2016.1195744.

A Destructive Series Power Cure Rate Model

Gladys D.C. Barriga[1], Vicente G. Cancho[2], Gauss M. Cordeiro[3], Edwin M. M. Ortega[1], Michael W. Kattan[4]

[1] Universidade de São Paulo, So Carlos, SP, Brazil.
[2] Universidade Estadual Paulista, Bauru, SP, Brazil.
[3] Universidade Federal de Pernambuco, Recife, PE, Brazil.
[4] Department of Quantitative Health Sciences-Cleveland Clinic, USA.

E-mail for correspondence: `glad@feb.unesp.br`

Abstract: A new flexible cure rate survival model is developed where the initial number of competing causes of the event of interest (say lesions or altered cells) follows a power series distribution. This model provides a realistic interpretation of the biological mechanism of the event of interest as it models a destructive process of the initial competing risk factors and records only the damaged portion of the original number of risk factors. Our proposed survival models are used for predicting breast carcinoma survival in women who underwent mastectomy. The postmastectomy survival rates are often based on previous outcomes of large numbers of women who had a disease, but they cannot predict what will happen in any particular patient's case. Pathologic explanatory variables such as disease multifocality, tumor size, tumor grade, lymphovascular invasion and enhanced lymph node staining are prognostically significant to predict these survival rates.

Keywords: Cure rate models; power series distribution; proportional hazard models; regression model; survival models

1 Introduction

The event of interest in many survival studies or cancer-relapse trials can be the death of a patient (due to various competing causes) or a tumor recurrence (attributed to metastasis-component tumor cells left active after an initial treatment). However with the recent advances in (cancer) treatment therapies, a high portion of the subjects are expected to be *cured*, i.e. remaining disease free after prolonged follow-ups. In this vein, there is now a vast literature on 'cure rate models' for survival data, though majority of these stems from either one of the *2-component Mixture Cure* model of Berkson & Gage(1952), or the *Bounded Cumulative Hazard* (or, BCH model) of Yakovlev & Tsodikov(1996).

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

Recently, Rodrigues *et al.*(2011) have extended the BCH model of Yakovlev & Tsodikov(1996) through a special case of the destructive (compound) weighted Poisson distribution. The main purpose of this study is to propose a general class of destructive survival cure rate model.

2 Model

For an individual in the population, let M denote the unobservable number of causes of the event of interest for this individual, for example, M can represent the number of altered cells before the treatment initial. Consider that M follows a power series (PS) distribution with probability mass function (pmf),

$$P(M = m; \theta) = \frac{a_m \theta^m}{A(\theta)}, \; m = 1, 2, \ldots, \tag{1}$$

where $a_m \geq 0$ and $A(\theta) = \sum_{m=1}^{\infty} a_m \theta^m$ and $\theta \in (0, s)$ (s can be ∞) is such that $A(\theta)$ is finite. The immediate consequence of a prolonged treatment is the formation (or not) of precancerous lesions. Given $M = m$, let W_j, $j = 1, \ldots, m$, be independent random variables (independent of M) following a Bernoulli distribution with success probability ϕ indicating presence of the j-th lesion (or competing cause), or the probability of an undestroyed clonogenic cell. The variable N, denoting the total number of altered cells among the M initial cells (competing causes), which are not destroyed or eliminated by the treatment, is defined by

$$N = W_1 + W_2 + \cdots + W_M. \tag{2}$$

By damaged or destruction, we mean $N \leq M$. The initial M competing causes is primary initiated malignant cells, where W_j in (2) denotes the number of living malignant cells that are descendants of the initiated malignant cell j during some time interval. In this case, N denotes the total number of living malignant cells at some specific time. The conditional distribution of N (given $M = m$) is referred to as the damaged distribution. The time to event for the jth competing cause is represented by Z_j, $j = 1, \ldots, N$. We assume that conditional on N, the Z_j are iid random variables with cumulative distribution function (cdf) $F(t)$ and survival function $S(t) = 1 - F(t)$ that does not depend on N. The total number of competing causes N and the time Z_j are not observable (latent). The observable time to the event of interest is defined by the random variable $T = \min(Z_1, \ldots, Z_N)$, and $T = \infty$ if $N = 0$ with $P(T = \infty | N = 0) = 1$. Under this setup, the survival function (Rodrigues *et al.*, 2009) for the entire population is given by

$$S_{\text{pop}}(t) = \frac{A(\theta[1 - \phi F(t)])}{A(\theta)}. \tag{3}$$

The cured fraction is given $p_0 = \lim_{t \to \infty} S_{\text{pop}}(y)$. From (2), we obtain $p_0 = \lim_{t \to \infty} S_{\text{pop}}(t) = \frac{A(\theta[1-\phi])}{A(\theta)} > 0$,
this indicating that (2) is not a proper survival function. The density and hazard function associated to (2) are given by $f_{\text{pop}}(t) = -S'_{\text{pop}}(t) = \frac{A'(\theta[1-\phi F(t)])}{A(\theta)} \, \phi \, \theta \, f(t)$, and $h_{\text{pop}}(t) = \frac{A'(\theta[1-\phi F(t)])}{A(\theta[1-\phi F(t)])} \, \phi \, \theta \, f(t)$ where $A'(\theta[1 - \phi F(t)]) = A'(\theta) \mid_{\theta(1-\phi F(t))}$ and $f(t) = -dS(t)/dt$.

3 Application

To illustrate our proposed modeling discussed so far, we consider the data set collected by Kattan *et al.*(2004), where a total of $n = 284$ women who had been treated with mastectomy and axillary lymph node dissection at Memorial Sloan-Kettering Cancer Center (New York, NY) between 1976 and 1979 and met the following requirements for study inclusion: confirmation of the presence of invasive mammary carcinoma, no receipt of neoadjuvant or adjuvant systemic therapy, no previous history of malignancy, and negative lymph node status as assessed on routine histopathologic examination. We report the survival times (T) until the patient's death or the censoring times at the end of the study. Some explanatory variables are associated with pathologic characteristics of the tumor. The tumor grading was performed using the standard modified Bloom-Richardson system. The lymphovascular invasion was obtained using morphologic criteria. The lymph node status was measured according to immunohistochemistry (IHC) and hematoxylin and eosin (H&E) stains. After deleting subjects with incomplete data and missing observation times, we have a subset of n = 365 patients with approximately 78% of censoring. We consider survival times until the patient's death (in years) as the response variable. The following variables were collected from each patient: t_i: observed time t_i (in years); x_{1i}: age (in years); x_{i2}: multifocality (0: no, 1:yes); x_{i3}: tumor size (in cm); x_{i4}: tumor grading (0: I, 1: II, II and lobular); x_{i5}: lymphovascular invasion (0: no, 1: yes) and x_{i6}: lymph node status (0: IHC+ IHC- and H&E-, 1: IHC+ and H&E+).
To these data, we fit some members of the model described in Section 2 such as the destructive Poisson, geometric and logarithmic models with all covariates on the proportions of undestroyed clonogenic cell (ϕ) and on the short-term survivors. i.e, $\phi_i = \frac{\exp(\beta_0+\beta_1 x_{1i}+\cdots+\beta_6 x_{i6})}{1+\exp(\beta_0+\beta_1 x_{1i}+\cdots+\beta_6 x_{i6})}$. Also, based on the proportional hazard function, we incorporate the covariates through $h(y|\boldsymbol{\lambda}) = h_0(y|\alpha)\exp\{\gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_6 x_{is}\}$, with $h_0(y|\alpha) = \alpha t^{\alpha-1}$. We apply the selection criteria $AIC = -2\ell(\widehat{\boldsymbol{\vartheta}})+2\#(\boldsymbol{\vartheta})$ for the candidate models, according to the AIC, we select the destructive geometric cure rate (DGCR) model as our working model. Considering the likelihood ratio statistic, we test the effect of these covariates in the probability of undestroyed causes and short-term survivors, i.e., $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = \beta_1 = \beta_2 = 0$ yielding $w_n = 0.179$ (p-value ≈ 1), indicating that the covariates age, multifocality and lymphovascular invasion are not significant for the probability of undestroyed causes and any of the covariates have a significant effect on short-term survivors. Hence, the MLEs and their standard errors for the parameters of the DGCR model with significant covariates are listed in Table 1. We conclude that the tumor size, tumor grading, lymphovascular invasion and lymph node status are significant prognostic variables for determining survival time and mortality risk in women with breast carcinoma.

References

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.

Johnson, L. N., Kemp, A. W. & Kotz, S. (2005). *Univariate Discrete Distribution*. Wiley, New York, NY, third edition.

TABLE 1. MLEs of the parameters for the reduced DGCR model.

Parameter	Estimate (est)	Standard error (se)	|est| / se
θ	0.005	0.076	—
α	1.933	0.205	—
γ_0	-4.075	0.463	8.805
β_0	-5.216	1.163	4.483
β_3	0.391	0.168	2.331
β_4	3.165	1.102	2.871
β_5	0.719	0.394	1.824
β_6	1.900	0.518	3.666

Rodrigues, J., Cancho, V. G., de Castro, M. & Louzada-Neto, F. (2009). On the unification of the long-term survival models. *Statistics & Probability Letters*, **79**, 753–759.

Rodrigues, J., de Castro, M., Balakrishnan, N. & Cancho, V. G. (2011). Destructive weighted poisson cure rate models. *Lifetime data analysis*, **17**(3), 333–346.

Yakovlev, A. Y. & Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore.

Latent vector autoregressive models: Application to evaluate evolution of oral health and general health

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

[1] KU Leuven, I-BioStat, Leuven, Belgium
[2] KU Leuven, Department of Oral Health Sciences, Belgium

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

Abstract: Vector autoregressive models (VAR) and random effects models are used to flexibly describe the evolution of a multivariate response. However, with a large number of responses at each occasion and many evaluations, a lot of parameters need to be estimated. We propose a latent vector autoregressive model that describes the evolution of the common factors used to summarize the responses via a combination of factor analysis (FA) models and random effects models. The latent VAR model makes use of q common factors for p ($p \gg q$) observed variables. The proposed model is applied to the BelRAI database to identify the effect of oral health (OH) and general health (GH) on the development of the other. Parameter estimation is done via Stan package.

Keywords: CFA; Longitudinal data; Oral and general health; VAR models.

1 Introduction

In most longitudinal studies, multiple outcomes are repeatedly measured over time. In the BelRAI database (De Almeida Mello et al. 2012) three OH and four GH items are recorded longitudinally on elderly people at baseline and regularly at every six months. It was of clinical interest to explore the joint evolution of oral and general health, and to know whether the information from OH provides additional information on GH and vice versa. To evaluate the evolution of all the items simultaneously, one could make use of the VAR models or random effects models. However, joint modeling of all responses leads to over-parameterization with VAR models (Koop and Korobilis, 2010) and computational problems with random effects models (Verbeke et al. 2014). To avoid this problem, we propose a latent VAR model which is a VAR model applied to the latent structure obtained

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.

from a model that combines confirmatory factor analysis and random effects models.

2 Motivating data set: BelRAI database

To obtain a picture of the joint dependence of OH and GH, we extracted 3450 individuals from the BelRAI database. The elderly subjects were examined on 2 to 6 occasions providing in total 8320 observations. Three binary OH indicators: non-intact teeth, chewing difficulty, and dry mouth were selected to represent OH and four ordinal GH measures: activity of Daily Living (ADL), cognitive performance scale (CPS), depression rating scale (DRS), and changes in health, end-stage disease, signs, and symptoms scale (CHESS) were recorded. In each of the responses there are missing values ranging from 5% to 20%, assumed to be at random.

3 VAR model for the latent structure

3.1 Basic VAR model

Let $\boldsymbol{Y}_t$ be a continuous $(p \times 1)$ multivariate response recorded at time t. The basic k-lag vector autoregressive (VAR(k)) model has the form:

$$\boldsymbol{Y}_t = \boldsymbol{c} + \Gamma_1 \boldsymbol{Y}_{t-1} + ... + \Gamma_k \boldsymbol{Y}_{t-k} + \boldsymbol{\delta}_t, \qquad t = 1, ..., T$$

where $\boldsymbol{c}$ is an unknown vector of intercepts, Γ_i $(i = 1, \dots, k)$ are coefficient matrices, and $\boldsymbol{\delta}_t$ is a zero mean, serially uncorrelated vector of errors with covariance matrix Δ. Thus a VAR model is a generalization of an autoregressive model by modeling the evolution of each variable based on its own lags and the lags of the others.

3.2 Latent VAR model with random effects for ordinal outcomes

For ordinal outcomes, we first assume that each of them is a manifestation of a latent standard normal distributed variable, Z, discretized by a set of cut-points. These latent variables are modeled using a FA model with q $(q \ll p)$ common factors combined with a random effects model:

$$\boldsymbol{Z}_t = \boldsymbol{\tau}_t + \Lambda_t \boldsymbol{\xi}_t + \boldsymbol{b} + \boldsymbol{\varepsilon}_t. \tag{1}$$

We then apply the VAR model to the latent structure of the common factors, i.e., at time point t:

$$\boldsymbol{\xi}_t = \Gamma_1 \boldsymbol{\xi}_{t-1} + ... + \Gamma_k \boldsymbol{\xi}_{t-k} + \boldsymbol{\delta}_t, \tag{2}$$

where $\boldsymbol{\xi}_t$ are the common factors, Λ_t is the factor loading matrix, $\boldsymbol{b}$, a vector of p random intercepts, $\boldsymbol{b} \sim N(0, D)$ with a diagonal matrix D, and $\boldsymbol{\varepsilon}_t \sim N(0, \Psi)$ with diagonal matrix Ψ. The restrictions $\tau_1 = ... = \tau_T$ and $\Lambda_1 = ... = \Lambda_T$ are applied to preserve the interpretation of the common factors over time. For identification, variances of the common factors at the first time point are kept at one.

4 Application of the latent VAR model to oral and general health

First, model (1) is used to summarize the OH and GH items in the BelRAI database using two common factors. VAR model (2) is then fitted to the common factor (Figure 1):

$$\begin{pmatrix} \xi_t \\ \eta_t \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} \xi_{t-1} \\ \eta_{t-1} \end{pmatrix} + \begin{pmatrix} \delta_\xi \\ \delta_\eta \end{pmatrix},$$

where $\begin{pmatrix} \xi_1 \\ \eta_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \delta_\xi \sim N(0, \sigma_\xi^2), \delta_\eta \sim N(0, \sigma_\eta^2)$. With this model, the cross-lagged parameter γ_{12} (resp. γ_{21}) indicates the amount of extra information that the current OH (resp. GH) indicator provides the future GH (resp. OH) indicator.

The model is implemented in R using the rstan package. Four chains are run and convergence is claimed when the estimated potential scale reduction factor is less than 1.10 and Monte Carlo standard error is less than 5% of the posterior standard deviation.

Estimates, provided in Table 1, indicate that given current GH, OH is still predictive of future GH and vice versa.

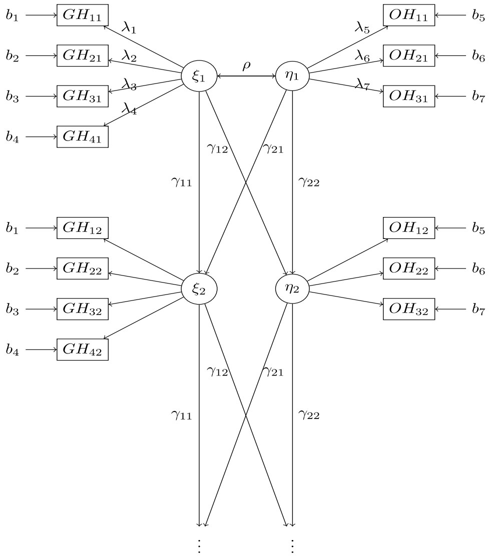


FIGURE 1. VAR(1) model for the latent factors derived from oral health (OH) and general health (GH) indicators.

5 Discussion

A similar idea in psychology was proposed by Oort (2001) for continuous variables with a balanced dataset. Our procedure allows for unbalanced data and works

TABLE 1. Posterior mean, SE and equal-tail 95% CI for the γ's.

Parameter	Est.	SE	95% CI	
γ_{11}	0.958	0.004	0.951	0.967
γ_{12}	0.032	0.009	0.015	0.048
γ_{21}	0.039	0.012	0.016	0.062
γ_{22}	0.994	0.001	0.991	0.996

also for ordinal data, making use of a latent structure as proposed by Liu and Hedeker (2006), Cagnone et al. (2009) but their approaches were restricted to just one latent factor.

References

Cagnone, S., Moustaki, I., Vasdekis, V. (2009). Latent variable models for multivariate longitudinal ordinal responses. *British Journal of Mathematical and Statistical Psychology*, **62**, 401 – 415.

De Almeida Mello, J., Van Durme, T., Macq, J., Declercq, A. (2012). Interventions to delay institutionalization of frail older persons: design of a longitudinal study in the home care setting. *BMC Public Health*, **12:615**.

Koop, G., Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, **3**, 267 – 358.

Liu, L., Hedeker, D. (2006). A mixed-Effects regression model for longitudinal multivariate ordinal data. *Biometrics*, **62**, 261 – 268.

Introduction to Multiple Time Series Analysis. Berlin: Springer-Verlag.

Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *British Journal of Mathematical and Statistical Psychology*, **54**, 49 – 78.

Verbeke, G., Fieuws, S., Molenberghs, G., Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, **23**, 42 – 59.

Multivariable fractional polynomials for correlated dependent variables using generalized estimating equations

Maren Vens[1][2], Jördis Stolpmann[2]

[1] Institut für Medizinische Biometrie und Epidemiologie, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany
[2] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

E-mail for correspondence: `m.vens@uke.de`

Abstract: Generalized estimating equations (GEE) are used for estimating the relationship between correlated outcome variables and several risk factors. For independent observations, non-linearity in continuous independent variables can be taken into account by multivariable fractional polynomials (MFP). We propose an MFP procedure for GEE models (GEE-MFP) which combines a function selection procedure with a stepwise elimination procedure. Quasi-likelihood information criteria are compared to a new Bayesian quasi-likelihood information criterion (BQIC) for model selection. GEE-MFP is illustrated by re-analyzing data on hypertension from the Framingham Heart Study. It is shown that the GEE-MFP algorithm using the BQIC provides the best fit to the real data, while the use of the established criteria overfit the data. The GEE-MFP algorithm provides a flexible approach for modeling the functional relationship between correlated dependent variables and several continuous independent variables.

Keywords: Akaike information criterion; Bayesian information criterion; Nonlinear relationship; Quasi likelihood information criterion; Stepwise selection.

1 Introduction

Generalized estimating equations (GEE) can adequately take into account the correlation between the dependent variables, e.g., repeated measurements or family members (Zeger and Liang, 1986). The main advantage is that only the mean structure needs to be correctly specified but the correlation between the dependent variables may be misspecified.Many extensions to GEE have been proposed (Ziegler, 2011). GEE is, however, still limited in its capability to model

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.

continuous independent variables. In all software packages, the functional form of the covariates needs to be specified in advance.
One approach to allow for nonlinearities in the independent variables are fractional polynomial transformations (FPs; Royston and Sauerbrei, 2008).
The advantages of GEE and FP can be combined to a GEE-MFP method. A first simple algorithm for two independent variables has been proposed by Cui et al. (2009), which we generalize to handle an arbitrary number of independent variables. Due to the large amount of possible different models, it is impossible to search the entire model space and there is a clear need for a simple search strategy to find the best-fitting model. To this end, we propose a simple feasible algorithm.

2 The GEE-MFP algorithm and the Bayesian QIC

The function selection procedure (FSP, Fig. 1, left) is combined with a P value-based stepwise procedure (SWP; Fig. 1, right).

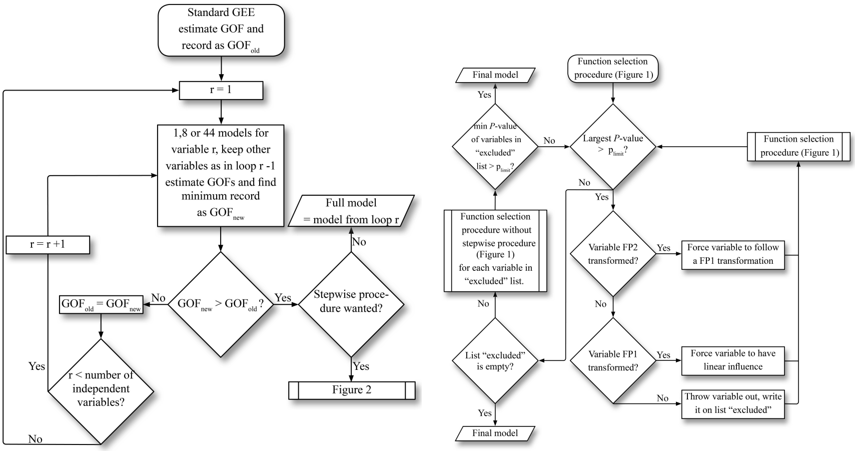

Standard GEE estimate GOF and record as GOF_{old}
r = 1
1,8 or 44 models for variable r, keep other variables as in loop r -1 estimate GOFs and find minimum record as GOF_{new}
r = r +1
GOF_{old} = GOF_{new}
No
GOF_{new} > GOF_{old} ?
Yes
Full model = model from loop r
No
Stepwise procedure wanted?
Yes
Figure 2
Yes
r < number of independent variables?
No
Final model
Yes
min *P*-value of variables in "excluded" list > p_{limit}?
No
Function selection procedure (Figure 1)
Largest *P*-value > p_{limit}?
Function selection procedure (Figure 1)
No
Yes
Function selection procedure without stepwise procedure (Figure 1) for each variable in "excluded" list.
Variable FP2 transformed?
Yes
Force variable to follow a FP1 transformation
No
No
List "excluded" is empty?
Variable FP1 transformed?
Yes
Force variable to have linear influence
No
Throw variable out, write it on list "excluded"
Yes
Final model

FIGURE 1. Left: Function selection procedure of the proposed GEE-MFP approach. Right: Stepwise procedure of the proposed GEE-MFP approach.

To measure the goodness-of-fit (GOF) of a GEE model, Pan (2001) proposed the QIC, an extension of the Akaike information criterion. GOF is investigated by measuring the difference between the estimated model-based covariance matrix $\hat{\boldsymbol{A}}$ and the estimated robust covariance matrix $\hat{\boldsymbol{C}}$. Therefore, the QIC is estimated by $\widehat{\mathrm{QIC}}(\boldsymbol{R}) = -2\hat{\varphi}Q(\hat{\mu}) + 2\mathrm{tr}\left(\hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{C}}\right)$ where $\hat{\varphi}Q(\hat{\mu})$ is the quasi likelihood function, φ the dispersion parameter, $\hat{\mu}$ and $Q(\hat{\mu})$ are estimated using the model with working correlation matrix $\boldsymbol{R}$ (Pan, 2001). For model selection of the mean structure, the QIC is simplified by assuming that $\boldsymbol{Q}$ equals the independency matrix (QIC_u).

Cui et al. (2009) used the QIC_u for selecting the best FP transformations. Disadvantages of the AIC are that models have a large number of independent variables, and tend to overfit the data. We propose to use the Bayesian quasi likelihood information criterion: $\widehat{\mathrm{BQIC}_\mathrm{u}}(\boldsymbol{R}) = -2\hat{\varphi}Q(\hat{\mu}) + p\ln(N)$, where N is the total number of observations in the study. BQIC_u penalizes additional parameters more strictly if the sample size is larger than 8. As a result, fewer FP2 transformations will be observed in the model building process and overfitting is prevented.

3 Application to the Framingham Heart Study

To illustrate the GEE-MFP approach, data from the Framingham Heart Study, cohort 2 as provided for the Genetic Analysis Workshop 13, was re-analyzed (Cupples et al., 2003). The dichotomous outcome high blood pressure at any investigation (HBP) was modeled using explanatory binary, categorical, and continuous variables observed at baseline.

Analysis showed that models derived using BQIC_u included fewer variable variable transformations compared with QIC and QIC_u models. FP2 transformation were never for one of the covariates using BQIC_u, and transformations were more stable in the sense that less different FP transformations were chosen for each covariate. The SWP using BQIC_u yielded in lower FP transformations for each covariate and less covariates were chosen.

The effect of different transformations on the model fit is illustrated for the variable age (Fig. 2). The age distribution of study participants was divided into deciles, and the proportion of subjects with HBP within each decile was plotted against the median age per age decile group. A good fit would be close to these points.

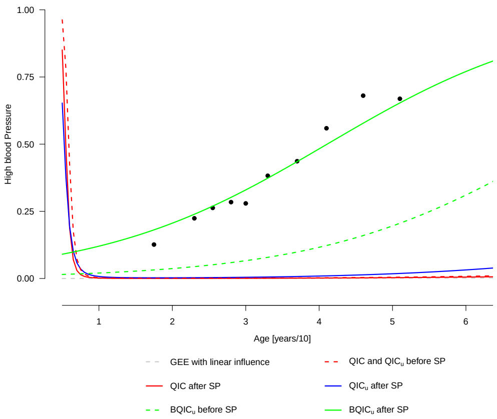


FIGURE 2. Comparison of fits using the different selection and goodness of fit criteria before and after stepwise procedure.

QIC and QIC_u models (FP2 for age) were not close to the estimated proportions. Models using BQIC_u (linear in age) provided a good fit to the data. Results were similar for other variables kept in the models after SWP.

4 Conclusion

We propose to use fractional polynomial transformations together with GEE for the joint analysis of an arbitrary number of independent binary, categorical, and continuous variables together on correlated possibly non-normal dependent outcome. The approach may be combined with a SWP using a quasi-likelihood criterion based on the BIC.

Acknowledgments: Authors acknowledge the Framingham Heart Study of the National Heart, Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine, supported by the National Heart, Lung and Blood Institutes Framingham Heart Study (Contract No. N01 HC 25195) for providing data for independent analysis by our group. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

References

Cui, J. et al. (2009). Fractional polynomials and model selection in generalized estimating equations analysis, with an application to a longitudinal study in Australia. *American Journal of Epidemiology*, **169**, 113 – 121.

Cupples, L.A. et al. (2003). Description of the Framingham Heart Study data for Genetic Analysis Workshop 13. *BMC Genetics*, **4 Suppl 1**, S2.

Pan, W. (2001). Model selection in estimating equations. *Biometrics*, **57**, 529 – 534.

Royston, P. and Sauerbrei W. (2008). *Multivariable model-building.* New York: John Wiley & Sons.

Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121 – 130.

Ziegler, A. (2011). *Generalized estimating equations.* New York: Springer.

An application of mixture models to estimate the size of Salmonella infected flock data through the EM algorithm using validation information

Carla Azevedo[1], Dankmar Böhning[1], Mark Arnold[2], Antonello Maruotti[1]

[1] University of Southampton, United Kingdom
[2] Animal and Plant Health Agency, United Kingdom

E-mail for correspondence: `cfa1e14@soton.ac.uk`

Abstract: Capture-recapture methods are used to estimate the total size N of a population when it is partially observed. Due to an incomplete identification/registration mechanism in real life applications, we observe only the positive counts representing the number of repeated identifications, and in order to estimate N, we wish to predict the number of units of the unobserved part. Sometimes, a validation sample is available in the study, providing complete information on the unobserved units. The estimate of the total population size can be obtained by fitting jointly a zero-truncated distribution to the truncated data and an untruncated distribution of the same class to the untruncated data by means of the EM algorithm. In this paper, we consider a flexible non-parametric mixture model approach allowing the heterogeneity of the data by means of a nested EM algorithm using validation information. A simulation study illustrates the major ideas of this application.

Keywords: capture-recapture; validation information; mixture models.

1 Introduction and background

The aim of this work is to determine the size N of an elusive target population. Let us assume that the members of the population are identified at m observational occasions where m is considered fixed. For each member i the count of identifications X_i returns a count in $0, 1, ..., m$ and i takes values from 1 to N. It is assumed that X_i is observed if unit i has been identified for at least one occasion. We have then that X_i is observed and let $X_1, ..., X_n$ denote the observed counts with n representing the total number of recorded individuals. We assume w.l.o.g. that $X_{n+1} = ... = X_N = 0$. Let f_x be the frequency of units with count $X = x$. The population can be described by a probability density function $p_x(\theta)$ which denotes the probability of exactly x identifications for a generic unit where $p_x \geq 0$ and $\sum_{x=0}^{\infty} p_x = 1$. It is possible to incorporate a validation sample into the modelling and decrease the bias in the estimation process. In this sample all the counts are observed. We will denote by $g_0, g_1, ..., g_m$ the frequency distribution associated with this sample. Situations of heterogeneity in the population can be modelled using non-parametric mixture models. We start with the simple homogeneous case of one component using the binomial distribution through the EM algorithm. Mixture models were utilized to allow for more components and include validation information. Simulation studies allowed to enhance the value of including validation information into the modelling to estimate the total size of the population.

2 Case study - Salmonella data

The following data was provided by the Animal and Plant Health Agency, UK. A European Union wide baseline survey of *Salmonella* infection was carried out between October 2004 and September 2005. The results of that survey were used as a basis for setting flock prevalence reduction targets for *Salmonella* national control programmes in each member state of the EU. In the UK, a randomized sample of 454 commercial layer flock holdings was tested for *Salmonella.* In order to be able to monitor the progress of control measures for *Salmonella*, it is important to be able to obtain an accurate estimate of the initial prevalence at the time of the EU baseline survey. The goal of this study is to determine the number of farms which had *Salmonella* infected poultry but for which result in the survey was negative. 53 holdings tested positive for *Salmonella* in one or more samples of the survey using a EU survey method which consists of a total of 7 tests, so each farm could have 0,1,...,7 positives as Table 1 (second row) shows.

TABLE 1. Positive and validation samples for salmonella data.

x	0	1	2	3	4	5	6	7
f_x	?	17	9	5	6	5	5	6
g_x	3	1	3	2	3	3	4	2

The same method was conducted in 21 of the infected farms which established the validation sample as shown in Table 1 - third row. This sample also allows to observe infected holdings with all repeated tests negative.

3 Non-parametric mixture models

Our interest lies in estimating the size N of a target population knowing that no zero counts have been captured. To model $p_x = p_x(\theta)$ we need to find an estimate $\hat{\theta}$ for θ so that $\hat{p}_x = p_x(\hat{\theta})$. Since we are dealing with a fixed number of sampling occasions $m = 7$, a binomial distribution to model the data seems appropriate and we can estimate θ using the EM algorithm. However, specific assumptions such as independence of observations and homogeneity are required by the simple binomial model. Also, the benefits of having a validation sample have not yet been considered. Mixture models allow relaxing these assumptions offering a flexible approach in modelling heterogeneity. Let $f(x, \theta)$ denote the simple, parametric binomial density function. The finite mixture distribution is given by $p_x = p(x, Q) = \sum_{j=1}^{k} f(x|\theta_j)q_j$ and appears as the marginal distribution with respect to some variable Z having distribution Q, where Q is the mixing distribution giving non-negative weights q_j to θ_j. Notice that $\sum_{j=1}^{k} q_j = 1$. The mixing distribution can be seen as the heterogeneity distribution of the listing parameter of the population. If Q is available, we can estimate N, for example, by means of the Horvitz-Thompson estimator as $\frac{n}{(1-p(0,Q))}$. Therefore, we need to estimate Q and this will be done by maximum likelihood. We have two likelihoods for capture-recapture modelling: the unconditional likelihood $L^U(\theta, N) = \prod_{j=0}^{m} \left(\sum_{l=1}^{k} \binom{m}{l} \theta_j^l (1-\theta_j)^{m-l} \right)^{f_j+g_j}$ and the conditional, observed joint likelihood $L^C(\theta) = \prod_{j=1}^{m} \left(\sum_{l=1}^{k} \frac{\binom{m}{l} \theta_j^l (1-\theta_j)^{m-l} q_l}{1-\sum_{j=1}^{k} q_j (1-\theta_j)^m} \right)^{f_j} \times$ $\prod_{j=0}^{m} \left(\sum_{l=1}^{k} \binom{m}{l} \theta_j^l (1-\theta_j)^{m-l} \right)^{g_j}$. The estimation of the unconditional likelihood would imply the maximization of the unknown f_0. However, $L^U(Q) = L^C(Q) \times B(N, Q)$ where $B(N, Q) = \binom{N}{n} p_0^{f_0} (1-p_0)^n$. Hence, the unconditional likelihood depends to a large extent from the conditional likelihood which the EM algorithm for mixtures maximizes very easily. In the following we will consider only the conditional likelihood.

4 Application of the EM algorithm with mixtures to the case study

This theory was applied to the Salmonella data using a mixture model with 2 and 3 components using just the positive sample (Pos) or both samples (Pos-Val). Note that only the positive models and the positive with validation information, respectively, are comparable. Simulation studies allow to conclude that we get an estimate for the population size more accurate and with less bias using a validation sample (details not reported here).

5 Conclusion

The mixture model approach using only the positive sample can be extended to include information from the validation sample, the untruncated sample including also zero counts which are not observed in conventional capture-recapture settings. It was done here using the binomial mixture model with 2 and 3 components. We have focused on the model selection criteria AIC and BIC to choose

TABLE 2. Details of the obtained mixture models. The first column indicates the samples used in the model, $\hat{f}_0$ is the number of unreported cases, $\hat{N}$ is total size of the population, k is the number of components of the mixture model, $\hat{\theta}_j$ and $\hat{q}_j$ are the parameters of the mixture model.

Model	$\hat{f}_0$	$\hat{N}$	k	$\hat{\theta}_j$	$\hat{q}_j$	log-likelihood	AIC	BIC
Pos	1	54	1	0.45	-	-146.05	294.1	296.40
Pos-Val	1	54	1	0.48	-	-187.79	377.58	379.88
Pos	9	62	2	5.47 1.25	0.34 0.66	-98.75	203.49	209.40
Pos-Val	7	60	2	5.47 1.46	0.41 0.59	-144.64	295.28	301.19
Pos	20	73	3	6.49 3.99 0.78	0.13 0.25 0.62	-96.91	203.82	213.67
Pos-Val	9	62	3	4.24 6.38 1.16	0.31 0.17 0.52	-143.13	296.26	306.11

the more appropriate model. According to that, the mixture model with 2 components might be used for both samples. More trust can be developed in the model for the unobserved part using validation information.

References

Arnold, M.E., Martelli, F., McLaren, I., Davies, R.H. (2014). *Estimation of the rate of eggs contamination from salmonella infected chickens.* Zoonoses and Public Health 61: 18-27.

Böhning, D., Dietz, E., Kuhnert, R., Schön, D. (2005). *Mixture models for capture-recapture count data.* Statistical Methods and Applications DOI: 10.1007/BF02511573, 14: 29–43.

Retrospective estimation of growth of Chilean common hake based on the otolith measurements

Cristian Villegas[1], Emmanuel Lesaffre[2], Víctor Espejo[3]

[1] Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Piracicaba, Brazil
[2] L-Biostat, KU Leuven, Leuven, Belgium
[3] Subsecretaría de Pesca y Acuicultura, Valparaíso, Chile

E-mail for correspondence: `clobos@usp.br`

Abstract: The age-at-capture and length-at-capture of common hake (*Merluccius gayi gayi*) was collected in San Antonio and Talcahuano areas in Chile between 1967 and 2010. In addition, the number of annual rings in sagittal otolith, a common estimate of the age of the fish, is determined. This estimate is, however, interval-censored. Research questions such as the comparison of the age-at-capture over the years and the relationship of age- with length-at-capture require appropriate modeling techniques. It was also of interest to know whether there are different growth patterns inferred from the otolith radii in the hake and whether there is a trend in growth over the many years data were collected. This involves mixture modeling of longitudinal data. Because of the complexity of the data, a Bayesian approach was used to address the research questions.

Keywords: common hake; radius-at-capture; length-at-capture; Bayesian inference; repeated measurements

1 Introduction

Measurements were taken on the common hake (*Merluccius gayi gayi*) in mainly two maritime zones in Chile, i.e. between the IV (central zone) and the X Region (south zone) in Chile. It is of scientific but also commercial interest to determine growth parameters of the hake and their evolution over time. At capture, the length of the fish is determined as well as the otolith is extracted (a destructive measurement) to determine the age of the fish. Based on the number of rings, but also the lengths of the different otolith radii, it is of interest to estimate the length of the hake at different ages. It is also of interest to examine the relationship of the otolith radii and the length of the fish, and to explore trends

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 this relationship over time. Knowledge of fish age characteristics is necessary for stock assessments, and to develop management or conservation plans. The latter allows for prediction of the volume of the future captures. Some of the research questions involve techniques that allow for interval-censored responses and/or interval-censored covariates. But also mixture modeling is involved, and hierarchical modelling of longitudinal measurements. Because of the complexity of the data, Bayesian modeling is envisaged.

2 Background and description of the hake data set

The unpublished data set consists of measurements from 8347 common hake otoliths from the San Antonio and Talcahuano areas in Chile from Fisheries Development Institute (IFOP). The data set is composed of hake lengths, and measurements taken on otoliths collected during the years 1967-2010. Ages-at-capture fish range from 1 to 20 years for females and from 1 to 15 years for males. An historical analysis of otolith growth could help to describe the ways in which a given population grows and shrinks over time, as controlled by birth, death, and migration. It is the basis for understanding changing fishery patterns and issues such as habitat destruction, predation and optimal harvesting rates. The population dynamics of fisheries is used by fisheries scientists to determine sustainable yields.

For illustrative purposes, we present in Table 1 the length-at-capture, radius-at-capture and radii of the otolith for each annulus for a female fish caught in 2002. In Figure 1 we show the length-at-capture as a function of age-at-capture for female and male hake and the fitted growth curve model.

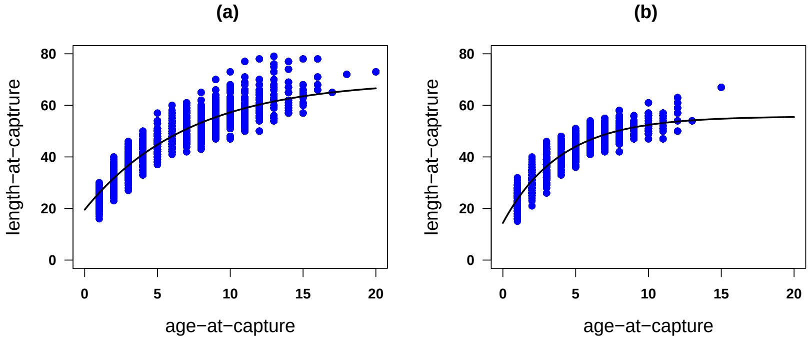


FIGURE 1. Hake study: Length-at-capture of fish as a function of age-at-capture ((a) = females, (b) = males) obtained from number of rings in otolith, together with estimated length obtained from von Bertalanffy growth model ignoring interval-censored character of age measurement.

TABLE 1. Hake study: Measurements taken at capture of a four years old female fish caught in 2002.

		Radius at year (μm)			
Length-at-capture (cm)	Radius-at-capture (μm)	1	2	3	4
44	6.9	3	5	6	6.7

3 Statistical modelling approach

A popular growth curve model in biology is the von Bertalanffy growth model given by

$$\mathrm{L}_i = L_\infty \left(1 - \exp\{-K\,(\mathrm{t}_i - t_0)\}\right) + \varepsilon_i \qquad \text{with} \quad \varepsilon_i \sim N(0, \sigma^2),$$

where L_i is the length-at-capture and t_i is the age-at-capture for the ith fish, respectively. The parameters of the von Bertalanffy growth model are interpreted as follows: L_∞ is the asymptotic (average) length, K is the growth rate coefficient (units are years^{-1}), and t_0 is the age when (average) length was zero. To estimate the length of the hake as a function of sex, year of capture and age determined by the number of rings in the otolith necessitates a regression model with an interval-censored covariate, see e.g. Bogaerts, Komarek and Lesaffre (2017). Model estimation, selection and checking will be done using a Bayesian approach. Another possibility is to regress on the otolith maximal radius-at-capture. This regression model will be compared to the previous model with respect to fit to the fish lengths.

There is also interest in exploring the annual radii and to see how that has evolved over time during the study period. This analysis involves a time series model on longitudinal profiles where year is taken as time measurement and the annual radii in the hake as longitudinal measurements. Notice that these longitudinal measurements are unbalanced across the fish.

All calculations were and will be done using R in combination with Bayesian software such as JAGS and Stan.

4 Conclusion

This rich data set allows for many interesting explorations and complex modelling using a Bayesian approach, which will be illustrated at the meeting.

Acknowledgments: FAPESP from Brazil and IFOP from Chile.

References

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

Forecasting Quality of Hard Disk Drive with Logistic Regression and Neural Networks

Somsri Banditvilai[1], Chammaree Chubuathong[1]

[1] King Mongkut's Institute of Technology Ladkrabang, Thailand

E-mail for correspondence: somsri_b2000@yahoo.com

Abstract: The purpose of this research is to develop a model to predict the quality of hard disk drives in an outgoing reliability test process. This would reduce the loss of resources, such as cost reduction and testing time. The amount of data in this research is huge, therefore 2 data mining techniques which are logistic regression analysis and neural networks are employed in this study. SAS Enterprise Miner Workstation 13.2 was used to analyze the data. The data is separated into 3 sets which are training set 60%, validation set 20% and test set 20%. There are 197 quantitative independent variables. In order to reduce the independent variables, the relationship between each independent variable and the quality of the hard disk drive was tested and it was found that only 25 variables were related to the quality of the hard disk drive. Therefore, 25 independent variables were used to build the neural network model. Since logistic regression needs to eliminate independent variables that creates multicollinearity problem. Following this 17 variables were left to build a logistic regression model. The results show that the neural network model has 95.98% accuracy while the logistic regression model has 72.37% accuracy. Therefore, it is seen that the neural network model performs better predictions for hard disk drive quality than the logistic regression model.

Keywords: Hard Disk Drive; Logistic Regression; Neural Networks.

1 Introduction

Currently, the hard disk drive which is an electronic storage device, is very important. It is used for storing information and its size is growing exponentially. The hard disk industry has expanded from the computers market to other markets, such as mobile phone, DVD player, TV and CCTV. In this research, the researchers focus on the prevention of waste in the process of outgoing reliability test (ORT) before deliver them to the customers. It was found that there are many types of abnormalities in read/write heads. Therefore, the researchers

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

would like to study and create a model to predict the hard disk drive whether the read/write heads were fault. Since there are 197 independent variables and data is huge. Therefore, 2 data mining techniques which are logistic regression and neural networks are employed. Logistic regression model is a regression model where the dependent variable is categorical. It used to predict whether an event will occur. By configuring one or more variables that are expected to affect the occurrence of the event. (Freedman. D. A. 2009) In the case of binary logistic regression, two values are defined as 0 and 1. The error term e is a random variable. It is assumed that $E(e) = 0$, the error are uncorrelated, and no multicollinearity among independent variables. Logistic regression requires sample size n larger than regular regression analysis. The sample size is $n>30p$, where p is the number of independent variables. Artificial neural networks has been inspired by attempt to simulate human biological neural systems. The human brain consists mainly of neurons, The linking of neurons through fibers called axons is used to deliver signal from one neuron to another. Whenever nerve cells are stimulated, nerve cells are connected to other axon via dendrites. The connectors between dendrites and axons are called synapses. Artificial neural network is widely used in classification, modeling, forecasting, controlling, and clustering. (Tan, P.N.2005)

2 Methodology

2.1 The study of outgoing reliability test.

Examine hard disk abnormalities occurring in the pre-delivery test process and found that the error most occur in the read/write heads set. Therefore, this research focuses on modeling prediction of hard disk test results of malfunction at read/write heads set. The data is divided into 60 % learning set, the 20 % verification set and the 20% testing set. Each set has the same ratio of normal hard disk and hard disk failure. The data in the learning set is used in modeling. The data in the validation set is used to evaluate the accuracy and the test set is used to measure the accuracy of the final model.

2.2 Data preparation.

Select relevant information, verification of data integrity, consolidate data from multiple sources and transform the data into the format that is appropriate for the modeling techniques.

2.3 Build the model.

The model is build with SAS Enterprise Miner workstation 13.2, and using logistics regression analysis and artificial neural networks. Using data from April to November 2015 (8 months). Study the factors affecting the hard disk drive quality. There are 197 independent variables that affect or correlate with malfunction of the read/write head set in post-assembly hard disk test and using the chi-square test of association to eliminate the independent variables. In this research, if the Pearson correlation value is greater than 0.8, then the independent variables are strongly correlated.(Campbell, I. 2007) The dependent variable refers to the hard disk drive quality which can be defined to be 1 as ”PASS” and 0 as ”FAIL”. After testing the relationship between independent variables and hard disk quality, it was found that there were only 25 independent variables were used to build the model. For logistic regression analysis, in order to save all 25 independent variables and eliminate multicollinearity, this research used principal component analysis to regroup independent variables, and left 17 independent variables to construct the logistic regression model. Based on the stepwise method, only 7 independent variables had a p-value < 0.05, indicating that these variables were effectively or significantly correlated with hard disk quality. For artificial neural networks, we employed multilayer neural networks and backpropagation training. There is no specific theory for the number of nodes in hidden layer. However, (Satish K, 2012) proved that it should be between 1 and n-1 where n is the number of independent variables. Therefore, this research set the nodes in hidden layer from 1-24. The final model had 1 input layer with 25 input nodes which are equal to the number of independent variables, and one hidden layer with 5 hidden nodes, and 1 output layer with 1 node and defined the learning rate of 0.1.

2.4 Select of the models

The results of learning set, validation set, and test set were compared. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized in the table. This research uses accurary which is the ratio of correct predictions to total predictions made. The correct predictions are composed of True Positive (TP) which is hit and True Negative(TN) which is correct rejection. Incorrect predictions are composed of False Positive (FP)which is Type I error, and False Negative(FN) which is Type II error.

3 Result and conclusion

The results from artificial neural networks had 95.98% accuracy as shown in Table 2 and the logistic regression analysis had 72.37% accuracy as shown in Table 1. From Table 1, 2, the accuracy of training set, validation set and testing set are similar for both logistic regression and neural networks. There is no over fitting problem occurred (Chen, et. al., 2004). Therefore, the neural networks analysis was used to test the hard disk drive before the outgoing reliability test. It was found that the model was able to prevent the hard disk with malfunction read/write heads. The results showed that 90 % of the predicted hard disk failure are correct and it could prevent scraps from hard disk drive test process by 90%. Thus, it could be concluded that the model derived from neural network analysis had the ability to predict the quality of the hard disk effectively.

References

Campbell, I. (2007). *Chi-squared and Fisher-Irsin tests of two-by-two tables with small sample recommendation: Statistics in Medicine.* 3661 − 3675.

Chen, C. Liaw, A. and Breiman, L. (2009). *Using Random Forest to Learn Imbalanced Data.* California: University of California. Report ID:666

TABLE 1. Logistic regression results

Data set	FN	TN	FP	TP	Accuracy
Training set	110	72337	27977	335	72.37 %
Validation set	40	24152	9286	108	73.34%
Testing set	41	24089	9351	108	73.30%
Total	191	120578	46614	551	72.37%

TABLE 2. Neural networks results

Data set	FN	TN	FP	TP	Accuracy
Training set	13	95823	4511	432	95.92 %
Validation set	12	32010	1428	136	96.10%
Testing set	14	31999	1441	135	96.05%
Total	39	159812	7380	703	95.98 %

Freedman, D. A. (2009). *Statistical Models:Theory and Practice*. Cambridge: Cambridge University Press. p.128.

Satish, K. (2012). *Neural Networks: A Classroom Approach (Second Edition)* Columbus: McGraw Hill Education.

Tan, P.N., Steinbac, M. and Kumar, V. (2005). *Introduction to Data Mining*. Boston: Pearson International, Inc.

Growing Bigger and More Accurate with GSBPM (part one)

Arbi Setiawan[1], Ni Lessari[2]

[1] National Statistical Training Center, Indonesia
[2] STIS National Statistical College, Indonesia

E-mail for correspondence: nilessari25@gmail.com, hanik.devia@gmail.com

Abstract: Indonesia has single government agency disseminating official data so that national statistics accuracy can be as simple as mostly a matter of single agency. Higher national statistics accuracy is viable through elements of generic statistical business process model [gsbpm]. GSBPM provides pathways to be more accurate with bigger data. Big data in layman's sense could be a trade-off between quantity and quality (including accuracy). With a wider availability of smart phones, guide book and manual soft copies can easily be loaded and retrieved. Question libraries can be made more handy. There are more than three hundred local languages in our country. Currently all variable descriptions available only in single national language. Interviewer may not be able to accurately translate variable descriptions from single national language to local language. This incapability can lead to different interpretation by respondent and in turn different raw data obtained. Accuracy can be improved with bilingual variable descriptions. Paper questionnaire space is limited for identified potential problems, errors, discrepancies to be hand written. A personal computer on every desk offers far more ease and flexibility for review, validate, edit sub-process. Distributive computing power will make possible a repository of data and relevant metadata in almost all subject matter specialist office. Currently archive is meaningful only for small number of permanent staff members. To our thought with minimal cost more staff members with limited skill can be exposed to data and relevant metadata. Every year there are around half a thousand new statistics under-graduate level employees having personally-owned notebook computer at home and available for after office hours limited back-up. There are staff members having more than twenty years tenure and ready for assessing how well the data reflect their initial expectations. It can be done by viewing the data from different perspectives using different tools and media. Our thought is that we can introduce a formal written documentation in addition to informal oral assessment.

Keywords: Big data; Smart phones; Distributive Computing.

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

Advertised landed house price data among other things can be used to evaluate interest rate.

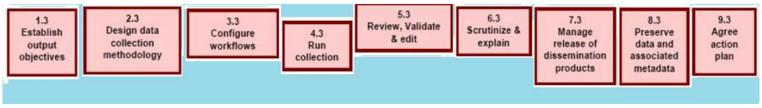

1.3 Establish output objectives
2.3 Design data collection methodology
3.3 Configure workflows
4.3 Run collection
5.3 Review, Validate & edit
6.3 Scrutinize & explain
7.3 Manage release of dissemination products
8.3 Preserve data and associated metadata
9.3 Agree action plan

FIGURE 1. sub process 1·3 2·3 3·3 4·3 5·3 6·3 7·3 8·3 9·3 gsbpm version four.

1.3 Any organization which is serving central bank may establish output objectives such how many regions will be covered. Will it be all five hundred Indonesian regions ? To be accurate will it be of equal importance for all regions ? Otherwise do some regions having central bank office are more important ? To be accurate will it be of equal dead line for all regions ? Otherwise do some regions having central bank office are less tolerant to delay ? That is some regions expectation is well ahead of deadline.

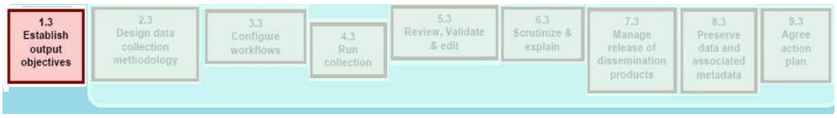

1.3 Establish output objectives
2.3 Design data collection methodology
3.3 Configure workflows
4.3 Run collection
5.3 Review, Validate & edit
6.3 Scrutinize & explain
7.3 Manage release of dissemination products
8.3 Preserve data and associated metadata
9.3 Agree action plan

FIGURE 2. Sub process 1·3 only gsbpm version four.

2.3 Will same primary raw data collection methodology be applied to all regions ? If same primary raw data collection methodology is applied will it be person to person interview ? Will it be paper-based interview ? Otherwise do some regions having central bank office may use cellular-phone-based personal interview ? For further accuracy do some regions having central bank office may use advertised house price eliminating direct personal contact with primary raw data provider ?

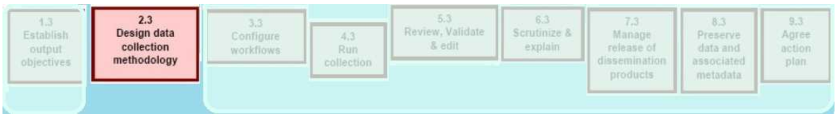

1.3 Establish output objectives
2.3 Design data collection methodology
3.3 Configure workflows
4.3 Run collection
5.3 Review, Validate & edit
6.3 Scrutinize & explain
7.3 Manage release of dissemination products
8.3 Preserve data and associated metadata
9.3 Agree action plan

FIGURE 3. Sub process 2·3 only gsbpm version four.

3.3 Configure workflows for more accurate primary raw data collection. In general the smaller a region the less accurate primary raw data in any given collection time. In general the more remote a region the less accurate primary raw data in any given collection time. Remote small region understandably has least accurate primary raw data in any given collection time. Therefore accuracy improvement is different for workflow in accessible regions with good transport. Workflow in limited-communication regions will be different.

4.3 Accuracy improvement during run collection of cellular-phone-based personal interview can be achieved in several subtle steps such as adherence to standard operating procedure improvement. Always recharge before power supply depletion. Prepare fully recharged portable power bank. Refresh cellular-phone memory to

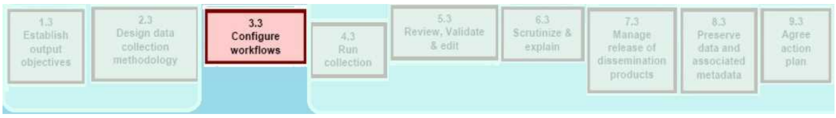

1.3 Establish output objectives
2.3 Design data collection methodology
3.3 Configure workflows
4.3 Run collection
5.3 Review, Validate & edit
6.3 Scrutinize & explain
7.3 Manage release of dissemination products
8.3 Preserve data and associated metadata
9.3 Agree action plan

FIGURE 4. Sub process 3·3 only gsbpm version four.

accommodate raw data. Prepare secondary portable magnetic storage. Remote small region needs back up of second generation (2G or 2·5G) cellular-phone.

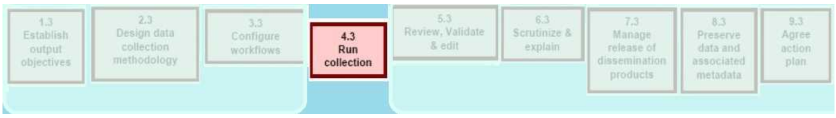

1.3 Establish output objectives
2.3 Design data collection methodology
3.3 Configure workflows
4.3 Run collection
5.3 Review, Validate & edit
6.3 Scrutinize & explain
7.3 Manage release of dissemination products
8.3 Preserve data and associated metadata
9.3 Agree action plan

FIGURE 5. Sub process 4·3 only gsbpm version four.

5.3 One acceptable fact on big data is that any particular raw data may have accumulated in some sort of series. Depending on intelligence of raw data analyst this sort of series can be used to review for accuracy improvement.

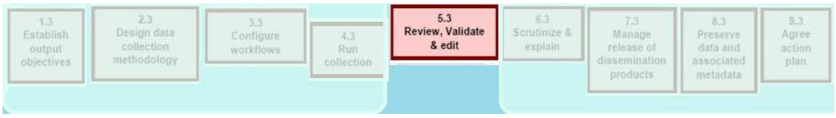

1.3 Establish output objectives
2.3 Design data collection methodology
3.3 Configure workflows
4.3 Run collection
5.3 Review, Validate & edit
6.3 Scrutinize & explain
7.3 Manage release of dissemination products
8.3 Preserve data and associated metadata
9.3 Agree action plan

FIGURE 6. Sub process 5·3 only gsbpm version four.

House price data in any region may be associated to opening of newly-built non-intersection road. House price data in other region may be associated to addition of toll-road access. House price may be related more to development of commuter railway and less to development of high speed train railway.

6.3 First sub process establish output objectives expect a responsibility to scrutinize raw data. Contrasting objectives against preliminary product is carried out. Raw data originating from unusual circumstances deserves different treatment to improve accuracy. Unusual circumstances applies both to more favorable condition as well as less favorable condition. Preliminary product scrutiny is done with ease in case of way ahead of schedule activities or any expense far less than activities budget.

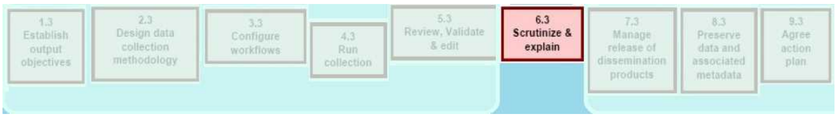

1.3 Establish output objectives
2.3 Design data collection methodology
3.3 Configure workflows
4.3 Run collection
5.3 Review, Validate & edit
6.3 Scrutinize & explain
7.3 Manage release of dissemination products
8.3 Preserve data and associated metadata
9.3 Agree action plan

FIGURE 7. Sub process 6·3 only gsbpm version four.

7.3 Central bank publishes result of residential property survey for primary house. First page of report has shown methodology of quarterly survey covering sixteen cities spread over 15 regional offices. The report does not mention stratification explicitly. Reader of report may to certain extent interpret that Balikpapan represents a stratum of Eastern Kalimantan and and newly renamed Northern Kalimantan. Interpret that Manado represents a stratum of Northern Sulawesi and Gorontalo and Western Sulawesi. To be more accurate central bank may conduct a press conference.

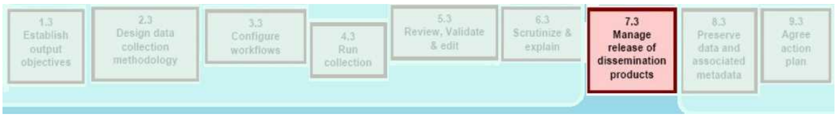

1.3
Establish output objectives
2.3
Design data collection methodology
3.3
Configure workflows
4.3
Run collection
5.3
Review, Validate & edit
6.3
Scrutinize & explain
7.3
Manage release of dissemination products
8.3
Preserve data and associated metadata
9.3
Agree action plan

FIGURE 8. Sub process 7·3 only gsbpm version four.

Acknowledgments: The author and co-authors declared there is not any potential conflict of interest with respect to the research, authorship, and/or publication of this article. The views expressed here are those of the individual author and co-authors and not necessarily those of STIS or its board, or officers, or staff.

References

McCullagh, Peter. (2002). What Is A Statistical Model?. In: *Proceedings of The Annals of Statistics, Volume 30*, University of Chicago, 5, 1225 – 1310 see page 1247.

UNECE Secretariat. (2009). *Generic Statistical Business Process Model Version 4.0*. UNECE Secretariat Steven Vale (steven.vale@unece.org), based on previous work by Statistics New Zealand (for the first seven phases) and Statistics Canada (for the Archive phase), with considerable input and feedback from the members of the METIS group.

Imputation in High-dimensional Mixed-Type data by Nearest Neighbors

Shahla Faisal[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: `shahla.ramzan@stat.uni-muenchen.de`

Abstract: In modern industrial and biomedical research, the data often contains a large number of variables measured at mixed data types (continuous, nominal or binary) but the information on some variables is missing. Imputation is a common solution where the downstream analyses require a complete data matrix. A number of imputation methods are available that work under some distributional assumptions. We propose an improvement over the popular nonparametric nearest neighbor imputation method that requires no particular assumptions. The proposed method makes practical and effective use of the information on association among the variables. In particular we propose a weighted version of L_q distance for mixed-type data that uses the information from a subset of important variables only. The performance of the proposed method is investigated under a variety of data settings. The results show a smaller imputation error and better performance when compared to other approaches. It is shown that the proposed imputation method works efficiently even when the number of samples is less than the number of variables.

Keywords: Weighted Nearest neighbors; Missing values; Mixed-type data; Kernel function; High-dimensional data.

1 Introduction

Missing values are a common phenomenon in practical research. A number of different methods are available that can be used to fill the missing values in metrically scaled data only (e.g. Troyanskaya, et al, 2001) or handle nominal data only (e.g. Schwender, 2012). Since one has to deal with combination of continuous and nominal variables in many real world applications, the methods to impute mixed data become more crucial. Since the multiple imputation techniques fail to impute high dimensional missing data, the nonparametric simgle imoutation methods are gaining more popularity. We propose an improved version of the popular nonparametric nearaest neighbors method which uses information only

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

on potentially relevant neighbors to impute missinng values. More specifically, We introduce a distance function that is more appropriate for mixed data. It is an extension of Tutz and Ramzan (2015) and uses information on association among variables.

2 Weighted Distance for Mixed-type Data

Let $\mathcal{R} = (R_{is})$ be the $n \times (p+m)$ data matrix with p continuous and m categorical covariates defined by $\mathcal{R} = (\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{X} = (x_{ig})$, $g = 1, \cdots, p$, with x_{ig} denoting i^{th} observation of the g^{th} continuous covariate, and $\boldsymbol{Z} = (z_{il})$, $l = 1, \cdots, m$, with z_{il} denoting i^{th} observation of the l^{th} categorical covariate. Let $\boldsymbol{O} = (o_{is})$ is the corresponding $n \times (p+m)$ matrix with $o_{is} = 1$ if R_{is} is observed, otherwise $o_{is} = 0$.

The categorical observations z_{il} in the data matrix, can assume values $c_l \in \{1, \ldots, k_l\}, l = 1, \ldots, m$, where k_l is the number of categories that the l^{th} attribute can take. Then the i^{th} row of the data matrix $\mathcal{R}$ can be written as

$$\boldsymbol{R}_i^T = (\boldsymbol{x_i}^T, \boldsymbol{z_i}^T)$$

with $\boldsymbol{x_i}^T = (x_{i1}, \cdots, x_{ip})$ and $\boldsymbol{z_i}^T = (z_{i1}, \cdots, z_{im})$

For the computation of distances, the categorical variables are transformed into binary variables. Thus the observation z_{il} becomes a vector, $\boldsymbol{z}_{il}^T = (z_{il1}, \ldots, z_{ilk_s})$ with $z_{ilr} = 1$ if $Z_{il} = r$. Thus, the ith row of the transformed matrix $\mathcal{R}^D$ has the form

$$[(x_{i1}, \ldots, x_{ip}), (\boldsymbol{z}_{i1}^T, \cdots, \boldsymbol{z}_{im}^T)]^T$$

with dummy vectors $\boldsymbol{z}_{il}$, $l = 1, \cdots, m$.

Let R_{is} be a missing entry in the data matrix $\mathcal{R}$, that is $O_{is} = 0$. Then the distance between the i-th and the j-th rows specific for a missing value in the sth covariate is defined by

$$\begin{aligned} d(\boldsymbol{R}_i, \boldsymbol{R}_j) \quad = \quad & \Bigg(\gamma_1 \sum_{g=1}^{p} |x_{ig} - x_{jg}|^q I_{(o_{ig}=1)} I_{(o_{jg}=1)} . C(\delta_{sg}) + \\ & \gamma_2 \sum_{l=1}^{m} \sum_{c=1}^{k_l} |z_{ilc} - z_{jlc}|^q I_{(o_{il}=1)} I_{(o_{jl}=1)} . C(\delta_{sl}) \Bigg)^{1/q}, \end{aligned} \qquad (1)$$

where $I_{(.)}.I_{(.)}$ is the number of *valid* components. $C(.)$ is a convex function defined on the interval $[-1, 1]$. We use $C(\delta_{sl}) = |\rho_{sl}|^c$, where c is a tuning parameter. γ_1 and γ_2 are tuning parameters with $\gamma_1 = 1 - \gamma_2$ and δ_{sl} is a measure of association between covariates s and l.

3 Weighted Imputation by Nearest Neighbors

Let a value is missing in the i_{th} row of the data matrix $\mathcal{R}$. One has two possible options: (i) metric covariate has the missing value (ii) a categorical covariate has the missing value. One finds k nearest neighbor observation vectors $\boldsymbol{R}_k$ based on the distances

$$\boldsymbol{R}_{(1)}^D, \ldots, \boldsymbol{R}_{(k)}^D \quad \text{with} \quad d(\boldsymbol{R}_i, \boldsymbol{R}_{(1)}) \leq \cdots \leq d(\boldsymbol{R}_i, \boldsymbol{R}_{(k)})$$

Imputing Categorical Missing Value For the imputation of the values z_{is}, we use the weighted estimator

$$\hat{\pi}_{isc} = \sum_{j=1}^{k} w(\boldsymbol{R}_i, \boldsymbol{R}_{(j)}) z_{(j)sc}, \tag{2}$$

where $c = 1, \ldots, k_s$. The weighted imputation estimate of a categorical missing value z_{is} is

$$\hat{z}_{is} = \text{arg max}_{c=1}^{k_s} \hat{\pi}_{isc}, \tag{3}$$

i.e. the value of $c \in \{1, \ldots, k_s\}$ with highest value of $\hat{\pi}$ is .

Imputing Continuous Missing Value The weighted imputation estimate of continuous missing value, x_{is}, is defined by

$$\hat{x}_{is} = \sum_{j=1}^{k} w(\boldsymbol{R}_i, \boldsymbol{R}_{(k)}) x_{(j)s} \tag{4}$$

The weights in equation (2) and (4) are defined by

$$w(\boldsymbol{R}_i, \boldsymbol{R}_j) = \frac{k(d(\boldsymbol{R}_i, \boldsymbol{R}_j)/\lambda)}{\sum_{l=1}^{k} k(d(\boldsymbol{R}_i, \boldsymbol{R}_j)/\lambda)}, \tag{5}$$

where $K(.)$ is a kernel function and λ is a tuning parameter. We use Gaussian kernel as in the initial simulations it yielded smaller imputation errors. The tuning parameter λ is chosen by cross validation.

4 Simulation Study

We generated $S = 200$ samples of size $n = 100$ for $p = 30$ predictors drawn from a multivariate normal distribution with $N(\mathbf{0}, \boldsymbol{\Sigma})$. The correlation matrix $\boldsymbol{\Sigma}$ has an autoregressive type of order 1 with $\rho = 0.9$. Some randomly selected variables are converted to nominal scale. We construct categories from the continuous data by setting cut points. In each sample, $miss$ = 10%, 20%, 30% of the total values were replaced by missing values completely at random (MCAR).
As initial step we investigate the performance of different kernel functions (Gaussian and Triangular), value of q (q =1,2) and the convex functions (linear and power). The results showed that the Gaussian kernel for $q = 2$ using power function provides smallest imputation error. We use these findings to compare the performance of *wNNSel$_{mix}$* with existing methods. For comparison purpose, we compute imputation errors for continuous and categorical variables separately and add them to get final imputation error.
The weighted nearest neighbors approach (*wNNSel$_{mix}$*) is compared with k-nearest nrighbors (kNN), multiple imputation by chained equations (MICE) and random forest (RF) methods. Figure 1 shows the results for 10% missing data. Clearly, the proposed weghted nearest neighbors method provides smallest imputation errors.

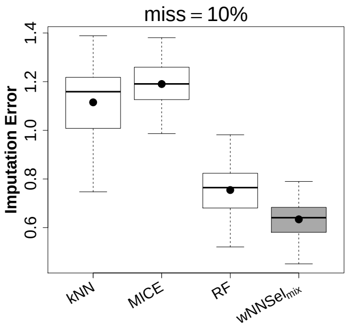


FIGURE 1. Boxplots of average imputation error obtained by different imputation methods. Solid circles within boxes show mean values.

References

Tutz, G. and Ramzan, S. (2015). Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics and Data Analysis*, **90**, 84 – 99.

Schwender, H. (2012). Imputing missing genotypes with weighted k nearest neighbors. *Journal of Toxicology and Environmental Health*, Part A, **75(8-10)**, 438 – 446.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17 (6)**, 520 – 525.

Using hierarchical hidden Markov models for joint inference at multiple temporal scales

Timo Adam[1], Vianey Leos-Barajas[2], Roland Langrock[1], Floris M. van Beest[3]

[1] Bielefeld University, Germany
[2] Iowa State University, USA
[3] Aarhus University, Denmark

E-mail for correspondence: timo.adam@uni-bielefeld.de

Abstract: Hidden Markov models are prevalent in the field of animal movement modeling, where they can be used to infer behavioral modes from various types of movement data. Due to substantial improvements in tagging technology, such data can nowadays be collected at much finer time scales than only a few years ago. Behavioral modes, however, do not necessarily manifest themselves at the fine scales at which the data are collected, but may effectively operate on much cruder time scales. To address the mismatch between data resolution on the one hand and biologically meaningful resolution on the other hand, we discuss a modeling framework that allows to jointly infer behavioral modes at multiple temporal scales. The approach is illustrated by modeling vertical movements of a harbor porpoise throughout the northeastern part of the North Sea.

Keywords: animal movement; statistical ecology; time series.

1 Introduction

In recent years, hidden Markov models (HMMs) and related state-switching models have emerged as increasingly popular tools for the analysis of animal behavior data (see, e.g., Leos-Barajas *et al.*, 2017, DeRuiter *et al.*, 2017), where they provide a versatile framework to infer behavioral modes (e.g. foraging or traveling) from various types of movement or general behavior data. Some types of animal behavior of potential interest, e.g. migratory behavior, may however not directly manifest themselves at the fine temporal grids at which data nowadays tend to be collected, but may rather be operating on much cruder scales.
As an example, the vertical movements of a harbor porpoise can be observed at a dive-by-dive resolution. However, behavioral modes such as foraging or traveling cannot easily be inferred directly from single dives (which may for example

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.

be shallow or deep), but rather become apparent from collections of dives (e.g. a sequence of many shallow dives, occasionally interspersed with deeper dives, typically indicates traveling behavior).

Here we discuss an extension to the basic HMM framework that regards the time series observations as stemming from two different underlying behavioral processes that operate on different time scales. The first process, which we call the *internal state* process, determines the behavioral mode at the crude scale, while the second process, which we call the *production state* process, determines the fine-scale mode within the crude-scale mode.

2 Methodology

A basic HMM comprises an observable *state-dependent* process $\{Y_t\}_{t=1}^T$ which is driven by an unobservable *state* process $\{S_t\}_{t=1}^T$. The state process is determined by an N-state Markov chain with transition probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{ij})$, where $\gamma_{ij} = \Pr(S_t = j \mid S_{t-1} = i)$, $i, j = 1, \ldots, N$. The initial distribution $\boldsymbol{\delta} = (\delta_i)$, where $\delta_i = \Pr(S_1 = i)$, gives the probability of the first observation y_1 belonging to state i. Given $S_t = i$, the state-dependent process is assumed to generate an observation from a state-dependent distribution with density (or probability) function $f_i(y_t)$.

To extend the basic framework in a way that allows for joint inference at multiple temporal scales, we initially segment the fine scale (e.g. dive-by-dive) observations into M distinct (e.g. hourly) chunks, denoted by $\boldsymbol{y}_m$, $m = 1, \ldots, M$. Each chunk is connected to one of K *production* HMMs, which generate the observations at the fine scale. The likelihood $L_{k,m}^{(P)}$ of the m^{th} chunk being generated by the k^{th} production HMM is given by

$$L_{k,m}^{(P)} = \boldsymbol{\delta}_k^{(P)} \mathbf{P}^{(P)}(y_1) \prod_{t=2}^{T} \mathbf{\Gamma}_k^{(P)} \mathbf{P}^{(P)}(y_t) \mathbf{1}, \tag{1}$$

where $\mathbf{P}^{(P)}(y_t) = \operatorname{diag}(f_1(y_t), \ldots, f_N(y_t))$. Considering the likelihoods $L_{k,m}^{(P)}$, an additional, *internal* K-state Markov chain $\{H_m\}_{m=1}^M$ is considered to model the switches among the K production HMMs, so that the likelihood $L^{(I)}$ of the hierarchical HMM is given by

$$L^{(I)} = \boldsymbol{\delta}^{(I)} \mathbf{P}^{(I)}(\boldsymbol{y}_1) \prod_{m=2}^{M} \mathbf{\Gamma}^{(I)} \mathbf{P}^{(I)}(\boldsymbol{y}_m) \mathbf{1}, \tag{2}$$

where $\mathbf{P}^{(I)}(\boldsymbol{y}_m) = \operatorname{diag}(L_{1,m}^{(P)}, \ldots, L_{K,m}^{(P)})$. The model structure is illustrated in Figure 1. Notably, due to the simplifying conditional independence assumptions, the likelihood can efficiently be evaluated and hence directly (numerically) maximized (Zucchini *et al.*, 2016).

3 Illustrating example

To illustrate the application of the suggested methodology, we model vertical movements of a harbor porpoise in the northeastern part of the North Sea. Raw

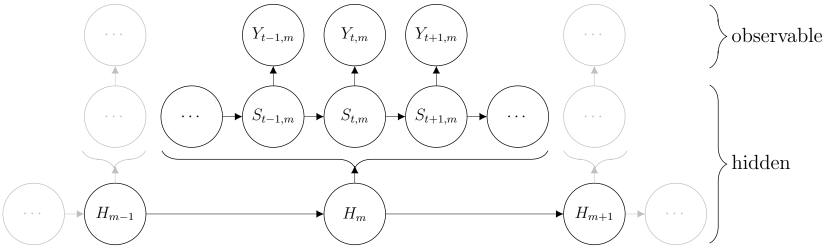


FIGURE 1. Dependence structure in hierarchical HMMs.

data on dive depth were processed into measures of the dive duration, the maximum depth and the dive wiggliness (as represented by the absolute vertical distance covered at the bottom of each dive) to characterize the porpoise's vertical movements at a dive-by-dive resolution.
The fitted state-dependent gamma distributions (extended by a zero point mass for the dive wiggliness) displayed in Figure 2 suggest three different dive types. State 1 captures the shortest, shallowest and smoothest dives with small variance. State 3 exhibits opposite features, namely very long, deep and wiggly dives, with high variation. State 2 lies in-between these two extremes, with all three variables taking on moderately large values.
The off-diagonal transition probabilities for the internal state process were estimated as $\hat{\gamma}_{12} = 0.211$ and $\hat{\gamma}_{21} = 0.219$. The corresponding stationary distribution is $(0.509, 0.491)$, indicating that about half of the observations are generated by each of the two production HMMs. The t.p.m.s for the two production HMMs, determining the switching behavior among the three production states, were estimated as

$$\hat{\mathbf{\Gamma}}_1^{(P)} = \begin{pmatrix} 0.406 & 0.443 & 0.150 \\ 0.240 & 0.600 & 0.159 \\ 0.196 & 0.366 & 0.437 \end{pmatrix} \text{ and } \hat{\mathbf{\Gamma}}_2^{(P)} = \begin{pmatrix} 0.277 & 0.153 & 0.570 \\ 0.124 & 0.248 & 0.628 \\ 0.057 & 0.087 & 0.856 \end{pmatrix},$$

with corresponding stationary distributions $(0.277, 0.506, 0.217)$ and $(0.083, 0.110, 0.807)$, respectively. These figures imply that when the first HMM is active, then in the long run about 28%, 51% and 22% of the observations are generated in production states 1, 2 and 3, respectively, whereas when the second HMM is active, then about 8%, 11% and 81% of the dives are generated in the respective production states. Biologically, the second production HMM can be seen as a proxy of foraging behavior, which particularly manifests itself in the extensive dive wiggliness suggesting prey-chasing. The first production HMM, which is characterized by a large proportion of short, shallow and smooth dives, could be associated with resting or traveling behavior. Further differentiation would require the inclusion of additional variables, e.g. the horizontal step length performed during dive.

4 Conclusion

The proposed hierarchically structured HMM constitutes a novel framework allowing for joint inference at multiple temporal scales. Environmental covariates

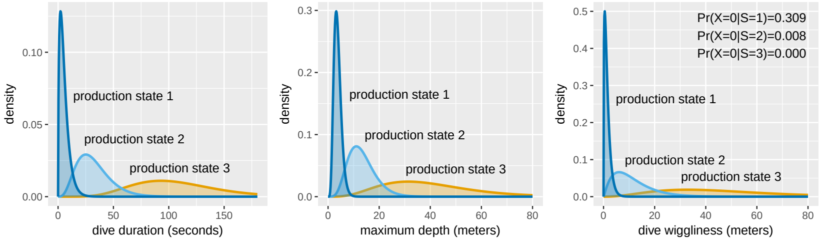


FIGURE 2. Fitted state-dependent distributions.

may be included in the model to investigate their effects on state occupancy and the dynamics of variation in behavioral modes at cruder scales than those at which the data are collected. The suggested methodology can be extended to allow for simultaneous modeling of multiple data streams collected at different scales (e.g. dive depths, which are observed at a dive-by-dive resolution, and hourly step lengths).

Acknowledgments: The harbor porpoise movement data were collected as part of the DEPONS project (www.depons.au.dk) funded by the offshore wind developers Vattenfall, Forewind, SMart Wind, ENECO Luchterduinen, East Anglia Offshore Wind and DONG Energy.

References

DeRuiter, S.L., Langrock, R., Skirbutas, T., Goldbogen, J.A., Calambokidis, J., Friedlaender, A.S. and Southall, B.L. (2017). A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *Annals of Applied Statistics*, in press.

Leos-Barajas, V., Photopoulou, T., Langrock, R., Patterson, T.A. Murgatroyd, M., Watanabe, Y.Y. and Papastamatiou, Y.P. (2017). Analysis of accelerometer data using hidden Markov models. *Methods in Ecology and Evolution*, in press.

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

Bayesian cure rate models induced by frailty in survival analysis

Vicente Garibay Cancho[1], Daiane de Souza[1], Josemar Rodrigues[1], N. Balakrishnan [2]

[1] Department of Applied Mathematics and Statistics, University of São Paulo, São Carlos, Brazil

[2] Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada

E-mail for correspondence: `garibay@icmc.usp.br`

Abstract: Frailty models provide a convenient way of modeling unobserved dependence and heterogeneity in survival data which, if not accounted for duly, would result incorrect inference. Gamma frailty models are commonly used for this purpose, but alternative continuous distributions are possible as well. However, with cure rate being present in survival data, these continuous distributions may not be appropriate since individuals with long-term survival times encompass zero frailty. So, we propose here a flexible probability distribution induced by a discrete frailty, and then present some special discrete probability distributions. We specifically focus on a special hyper-Poisson (hP) distribution and then develop the corresponding Bayesian simulation, influence diagnostics and an application to real dataset by means of intensive Markov chain Monte Carlo (MCMC) algorithm. These illustrate the usefulness of the proposed model as well as the inferential results developed here.

Keywords: Bayesian inference, frailty models, proportional hazard models, long-term survivors, hyper-Poison distribution.

1 Introduction

Studies on the frailty models generally assume a non-negative and continuous frailty random variable (Hougaard ,1984). Usually, the gamma distribution, inverse Gaussian and positive stable are used to model the frailty random variable. From the computational point of view, these distributions are convenient since it is easy to derive closed-form expressions for the survival, density and hazard functions using the Laplace transform. However, continuous frailty distribution

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

do not allow zero risk. Zero frailty indicates that there is a subgroup of non-susceptible individuals among whom the event of interest has not happened even after a long period of observation. The event of interest in many survival studies or cancer-relapse trials can be the death of a patient or a tumor recurrence. However, due to recent advances in (cancer) treatment therapies, a high portion of the subjects are expected to be cured, i.e., remaining disease-free after prolonged follow-ups. For this reason, there is a vast literature on cure rate models for survival data, also called survival models with a surviving fraction or long-term survival models. Most of these models were obtained in a competing risks scenario (Tsodikov *et al.*,2003), but they can also be obtained from the proportional hazard models with discrete frailty distributions (Caroni *et al.*,2010). In this paper, we focus on this direction from the Bayesian point of view.

2 The model

The frailty model assumes a proportional hazard structure conditional on the random effect Z. This random effect is a non-negative frailty variable which indicates the individual level of risk. So, the frailty model is basically specified by the following hazard rate function (hrf) $h(t|Z) = Z\ h_B(t)$, where $h_B(t)$ is the baseline hrf that can be equal to $h_B(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta})$ in the proportional hazard scenary (Cox & Oakes, 1984). The corresponding survival function, conditional on Z, is given by

$$S(t|Z) = P(T > t|Z) = \exp\{-ZH_B(t)\} = S_B(t)^Z, \tag{1}$$

where $H_B(t) = \int_0^t h_B(u)du$ is the baseline cumulative hazard function and $S_B(t)$ is the corresponding survival function. A new survival model is obtained when we assume that the frailty variable Z in (1) follows a hyper-Poisson distribution with probability mass function

$$P(Z = z; \eta, \theta) = p(z; \eta, \theta) = \frac{1}{{}_1F_1(1; \eta; \theta)} \frac{\theta^z}{(\eta)_z}; z = 0, 1, \ldots,$$

where $\eta, \theta > 0$, $(a)_r = a(a + 1)\ldots(a + r - 1) = \frac{\Gamma(a+r)}{\Gamma(a)}$ for $a > 0$, r is a positive integer, and ${}_1F_1(a; b; w) = \sum_{r=0}^{\infty} \frac{(a)_r}{(b)_r} \frac{w^r}{r!}$ is the confluent hypergeometric series. Thus, the unconditional survival function can be expressed as

$$S(t) = \frac{{}_1F_1(1; \eta; \theta S_B(t))}{{}_1F_1(1; \eta; \theta)} \tag{2}$$

and the fraction of "cured" individuals is given by $p_0 = \lim_{t\to\infty} S(t)$. From (2), $p_0 = \lim_{t\to\infty} S(t) = \frac{1}{{}_1F_1(1;\eta;\theta)} > 0$ implying that (2) is not a proper survival function. The density function associated with (2) is given by $f(t) = \frac{{}_1F_1(2;\eta+1;\theta S_B(t))}{{}_1F_1(1;\eta;\theta)} \frac{\theta}{\eta} f_B(t)$, where $f_B(t) = -dS_B(t)/dt$. The corresponding hazard function is given by $h(t) = \frac{{}_1F_1(2;\eta+1;\theta S_B(t))}{{}_1F_1(1;\eta;\theta S_B(t))} \frac{\theta}{\eta} f_B(t)$. Note that when $\eta = 1$, the model reduces to the cure rate survival model investigated by Yakovlev & Tsodikov(1996).

3 Application

We illustrate the application of the proposed model to a data set from a Phase III cutaneous melanoma clinical trial conducted by the Eastern Cooperative Oncology Group (Kirkwood *et al.*,2000). These data are part of an assay for the evaluation of postoperative treatment performance with a high dose of a certain drug (interferon alpha - 2b) in order prevent recurrence. Patients were included in the study from 1991 to 1995, and follow-up was conducted until 1998. After deleting subjects with incomplete data and missing observation times, we have a subset of $n = 417$ patients with approximately 56% of censoring. The observed time has mean equal to 3.18 and standard deviation equal to 1.69. The following variables were associated with each participant, $i = 1, \ldots, 417$: t_i: observed time (in years); x_{i1}: treatment (0: observation, $n = 204$; 1: interferon, $n = 213$); x_{i2}: age (0: ≥ 48 years, $n = 197$; 1: < 48 years, $n = 220$); x_{i3}: nodule (nodule category is coded from the number of lymph nodes involved in the disease: 1: $n = 111$; 2: $n = 137$; 3: $n = 87$; 4: $n = 82$); x_{i4}: sex (0: male, $n = 263$; 1: female, $n = 154$); x_{i5}: p.s (performance status-patient's functional capacity scale as regards the daily activities: 0: fully active, $n = 363$; 1: other, $n = 54$); x_{i6}: tumor (tumor thickness in tenth of a millimeter).

We initially consider the hP cure rate model with with all regression variables, $\theta_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_6 x_{i6}\}, \quad i = 1, \ldots, 417$.

For the Bayesian analysis, we assume a Weibull distribution for the baseline distribution function in (1), with $S_B(t; \boldsymbol{\gamma}) = \exp\{(-t/\gamma_2)^{\gamma_1}\}$. From posterior summaries for the parameters of the hP model with all covariates, we observe through the percentiles that the variable x_3 (nodule category) has a significant effect in the model since the interval does not contain zero.

TABLE 1. Posterior summaries for the hP model with only nodule category as covariate fitted to the melanoma dataset.

Parameter	Mean	Standard deviation	Percentile 2.5%	Percentile 97.5%
η	6.014	0.5414	5.000	7.000
γ_1	1.783	0.121	1.549	2.022
γ_2	2.867	0.274	2.477	3.549
β_0	0.485	0.169	0.130	0.808
β_3	0.293	0.053	0.189	0.400

Table 1 shows posterior summaries of the parameters of the simplified hP model retaining only the nodule category (x_3) as the covariate.

We finally conclude our melanoma analysis by considering the estimation of the cure rate (p_0) by employing the hP cure rate model. We stratified the patients according to the nodule category. From Section 2, we have $p_{0_j} = \frac{1}{{}_1F_1(1,\eta,\theta_j)}$, with $\theta_j = \exp(\beta_0 + j\beta_3)$ for $j = 1, \ldots, 4$. In Table 2, the posterior summaries reveal that cure rate (p_{0_j}) is decreasing with respect to the j-th nodule category, for $j = 1, 2, 3, 4$.

TABLE 2. Posterior summaries of the cure rate p_0 for the hP model according to nodule category (1-4)

Cure rate	Mean	Standard	Percentile	
		deviation	2.5%	97.5%
p_{0_1}	0.654	0.040	0.594	0.748
p_{0_2}	0.552	0.037	0.497	0.633
p_{0_3}	0.428	0.039	0.349	0.491
p_{0_4}	0.289	0.051	0.183	0.393

References

Caroni, C., Crowder, M. & Kimber, A. (2010). Proportional hazards models with discrete frailty. *Lifetime Data Analysis*, **16**, 374–384.

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

Hougaard, P. (1984) (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, **71**, 75–83.

Kirkwood *et al.* (2000).. High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of Intergroup Trial E1690/S9111/C9190. *Journal of Clinical Oncology*, **18**, 2444–2458.

Tsodikov, A. D., Ibrahim, J. G. & Yakovlev, A. Y. (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**, 1063–1078.

Yakovlev, A. Y. & Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*

Robust Generalised Autoregressive Score Models

Bernardi Mauro[1], Ruli Erlis[1]

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

E-mail for correspondence: mauro.bernardi@unipd.it

Abstract: Dynamic conditional score models are observation driven alternatives to classical state space models which are optimal in the sense of reducing the local Kullback–Liebler divergence between the true and the model implied conditional density. In this paper we apply the density power divergence approach of Basu et al. (1998) to robustify the dynamic conditional score equation with respect to outliers. The density power divergence approach minimises an appropriately modified divergence governed by an additional parameter that controls for the efficiency–robustness trade–off. Our approach combines the optimality of the score dynamics with the robustness properties of the class of M–estimators and it is a viable alternative to the Student–t approach of Harvey and Luati (2014). The proposed approach is illustrated through two examples of time series being strongly affected by outlying observations.

Keywords: Generalised autoregressive score models, robustness, minimum distance estimation, estimating equations.

1 Introduction

The class of Generalised Autoregressive Score (GAS) processes, recently introduced by Creal et al. (2013) and Harvey (2013), is becoming increasingly popular tool for modeling the time varying behaviour of unobservable parameters in an observation driven environment. The key feature of GAS models is that the predictive score of the conditional density is used as forcing variable into the updating equation of a time–varying parameter. Two main reasons for adopting the GAS updating mechanism has been advocated by the literature. Firstly, the GAS dynamics can be seen as an approximation to a filter for a model driven by a stochastic latent parameter that is by definition unobservable. Secondly, the conditional score can be considered as a steepest ascent direction for improving the model's local fit given the current parameter, position as usually happens into a Newton–Raphson type algorithm. Moreover, the class of GAS models nests a large

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.

number of observation driven models, such as the ARCH–type models of Bollerslev (1986). As a practical justification the GAS updating mechanism circumvents the problem of choosing a non–adequate forcing variable. Score driven processes have been proved to be effectively used in many applications. Harvey and Luati (2014) and Caivano and Harvey (2014) motivate the use of score dynamics by the need to reduce the potentially relevant effect of outliers on the extracted signal by assuming a fat–tailed distribution for the conditional innovation. In this paper, we adopt a different perspective and we robustify the dynamic conditional score equation to incidental influential observations and outliers by means of the Density Power Divergence (DPD) approach of Basu et al. (1998). Our approach combines the optimality of the score dynamics with robustness properties of the class of M–estimators and it is a viable alternative to the Student–t approach of Harvey and Luati (2014).

2 GAS models

Formally, assume that the observed random variable $y_t \sim p\left(y_t \mid \vartheta_{t|t-1}, \eta\right)$, where $p\left(\cdot\right)$ is a probability density and $\vartheta_{t|t-1}$ is a set of time–varying parameters and η is a vector of time–independent parameters. The GAS(1,1) updating equation for the time–varying parameters $\vartheta_{t|t-1}$ is

$$\begin{aligned} \vartheta_{t+1|t} &= \omega + \varphi\vartheta_{t|t-1} + \kappa s_t & (1)\\ s_t &= S_t\left(\vartheta_{t|t-1}, \eta\right)\nabla\left(y_t, \vartheta_{t|t-1}, \eta\right), & (2) \end{aligned}$$

where $\nabla\left(y_t, \vartheta_{t|t-1}, \eta\right)$ is the score of the conditional density function of the observed random variable evaluated at $\vartheta_{t|t-1}$, i.e.

$$\nabla\left(y_t, \vartheta_{t|t-1}, \eta\right) = \frac{\partial \ln p\left(y_t, \vartheta_{t|t-1}, \eta\right)}{\partial \vartheta_{t|t-1}},$$

and $S_t\left(\vartheta_{t|t-1}, \eta\right)$ is a positive definite, possible parameter–dependent scaling matrix. A convenient choice for $S_t\left(\vartheta_{t|t-1}, \eta\right)$ is usually given by $S_t\left(\vartheta_{t|t-1}, \eta\right) = \left[\mathcal{I}\left(\vartheta_{t|t-1}, \eta\right)\right]^{-\alpha}$, where $\mathcal{I}\left(\vartheta_{t|t-1}\right)$ is the Fisher Information matrix and $\alpha = \left(0, \frac{1}{2}, 1\right)$.

3 Robust extension

The density power procedure has been introduced by Basu et al. (1998) as alternatives to the classical maximum likelihood approach for robust parameter estimation. Specifically, let $p_\vartheta\left(Y\right)$ be a parametric family of distributions indexed by the parameter θ and let $q_{\vartheta_0}\left(Y\right)$ be the true unknown density of the random variable Y. To estimate the unknown parameter ϑ_0, Basu et al. (1998) proposed the family of density power divergences which depends on the tuning parameter $\beta > 0$, defined as

$$\mathcal{D}_\beta\left(p_\vartheta, q_{\vartheta_0}\right) = \begin{cases} \int p_\vartheta^{\beta+1}\left(Y\right) - \left(\frac{1+\beta}{\beta}\right) q_{\vartheta_0}\left(Y\right) p_\vartheta^\beta\left(Y\right) + \frac{q_{\vartheta_0}^{1+\beta}\left(Y\right)}{\beta} dY, & \beta > 0 \\ \int q_{\vartheta_0}\left(Y\right)\left(\log q_{\vartheta_0}\left(Y\right) - \log p_\vartheta\left(Y\right)\right) dY, & \beta = 0, \end{cases} \quad (3)$$

and discussed the corresponding estimator obtained by minimising the empirical counterpart of equation (3). The tuning parameter β controls for the trade–off between robustness and efficiency. Here, we apply the methodology of Basu et al. (1998) to robustify dynamic conditional score models. The advantages of robustifying the DCS models using the approach of Basu et al. (1998) are twofold. First, it provides a general estimation methodology that bridges the gap between the maximum likelihood and the robust L_2 estimation where deviations from efficiency are controlled for by the tuning parameter β. Second, it allows for the definition of a robustified score entering the transition equation as forcing variable. Furthermore, coherently with the theory of the DCS models, the robustified version of the score is defined as the first derivative of the probability density function of the observed variable and the robustified score is redescending. The robustified version of the GAS updating equation for the time–varying parameters defined in equations (1)–(2) becomes

$$\vartheta_{t+1|t} = \omega + \varphi\vartheta_{t|t-1} + \kappa s_t^* \tag{4}$$

$$s_t^* = S_t^*\left(\vartheta_{t|t-1}, \eta\right)\left(\nabla^*\left(\vartheta_{t|t-1}, \eta\right) - \xi\right), \tag{5}$$

where $\nabla^*\left(y_t, \vartheta_{t|t-1}, \eta\right) = \nabla\left(y_t, \vartheta_{t|t-1}, \eta\right) p\left(y_t, \vartheta_{t|t-1}, \eta\right)^\beta$ is the roubstified score of the conditional density function, evaluated at $\vartheta_{t|t-1}$, $\xi = \int \nabla^*\left(y_t, \vartheta_{t|t-1}, \eta\right) p\left(y_t, \vartheta_{t|t-1}, \eta\right) dy_t$ is the mean of the robustified score and $S_t^*\left(\vartheta_{t|t-1}, \eta\right) = J_\beta\left(\vartheta_{t|t-1}, \eta\right)$ is a positive defined, possible parameter–dependent scaling matrix with

$$J_\beta\left(\vartheta_{t|t-1}, \eta\right) = \int \nabla^*\left(z, \vartheta_{t|t-1}, \vartheta\right) \nabla^*\left(z, \vartheta_{t|t-1}, \eta\right)' p\left(z, \vartheta_{t|t-1}, \eta\right) dz - \xi\xi'.$$

4 Application

We apply the robust GAS to the problem of signal extraction from two series which are characterised by the presence of extreme outliers: the daily series of electricity prices from the Nord Pool Energy Market from January 1, 2013 to November 30, 2016 and the series of US average weekly hours in manufacturing from 1940 to 2016. The two series are plotted in the top panels of Figure 1. The superimposed blue and red lines denote the signal extracted using a GAS(1,1) model with Gaussian innovations and its roubstified version with $\beta = 0.05$, respectively. The optimal tuning parameter has been selected as suggested by Ghosh and Basu (2013). Bottom panels plot the dynamic evolution of the scaled score that enters the GAS updating equation for the two models. It is evident that the signal extracted by the robust GAS filter is less influenced by the presence of outliers than the filter based on the Gaussian GAS model.

References

Basu, A., Harris, I.R., Hjort, N.L., and Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549 – 559.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307 – 327.

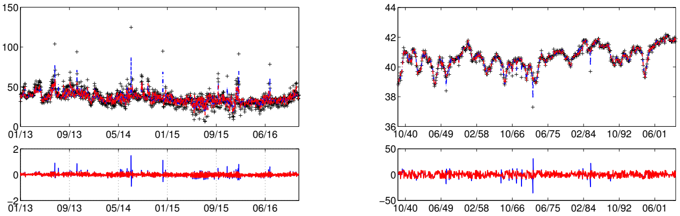


FIGURE 1. Signal extraction form the series of electricity price for the Nord Pool market and the series of US average weekly hours in manufacturing.

Caivano, M. and Harvey, A. (2014). *Time series models with an EGB2 conditional distribution.* Banca d'Italia.

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

Ghosh A. and Basu, A. (2013) Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics*, **7**, 2420 – 2456.

Harvey, A. and Luati, A. (2014). Filtering with heavy tails. *Journal of the American Statistical Association*, **109**, 1112 – 1122.

Harvey, A.C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*, volume 52. Cambridge University Press.

Monte Carlo modified profile likelihood in survival models for clustered censored data

Claudia Di Caterina[1], Giuliana Cortese[1], Nicola Sartori[1]

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

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

Abstract: When analyzing failure time datasets within stratified contexts, the main focus is usually not on the clustering variables and hence the group-specific parameters are treated as nuisance. If a fixed effects formulation is preferred and the total number of clusters is large relative to the single-stratum sizes, standard frequentist techniques are often misleading and the use of adjustments to make reliable inference on the parameter of interest is complicated by the presence of censored data. Here we show how Monte Carlo simulation may be exploited to compute a modification of the profile likelihood in general regression settings for survival models with unspecified censoring mechanism.

Keywords: Fixed effects; Incidental parameters; Monte Carlo simulation.

1 Introduction

Stratification of time-to-event data subject to censoring is frequent in many applied areas, although the primary concern of the study is typically not the inter-cluster variability. To avoid the assumptions imposed on the survival analysis by frailty models, a fixed effects approach considering the group-related traits as unknown nuisance parameters can be adopted.

Under such a formulation, if the amount of groups is much larger than the within-cluster sizes, usual asymptotic results are no longer valid and maximum likelihood inference on the component of interest may be inaccurate (Sartori, 2003). One solution to tackle this issue is offered by the modified profile likelihood (Severini, 1998). Its good performance has been proved in the literature for various models with incidental parameters, but in regression contexts with failure data it is unclear how to properly deal with random censoring schemes (Pierce and Bellio, 2006). We present here a convenient strategy to calculate the modified profile likelihood in stratified regressions for survival times with unspecified censoring mechanism.

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

Cortese and Sartori (2016) opted for eliminating the group-specific parameters from the likelihood of Weibull-distributed survival data by integration. The same model, within an extended regression framework, is studied below to illustrate our less computationally intensive procedure, whose basic principle is yet immediately applicable to different supposed distributions.

2 Modified profile likelihood in survival models under unknown random censoring scheme

2.1 Weibull model

Let independent grouped failure times $\tilde{y}_{it} \geq 0$, $i = 1, \ldots, N, t = 1, \ldots, T_i$, be realizations of a Weibull random variable with probability density function

$$p_{it}(\tilde{y}_{it}; \psi; \lambda_i) = \eta_{it} \xi \big(\eta_{it} \tilde{y}_{it}\big)^{\xi - 1} \exp\{-\big(\eta_{it} \tilde{y}_{it}\big)^{\xi}\}, \qquad (1)$$

where $\eta_{it} = e^{-(\lambda_i + x_{it}^{\mathrm{T}} \beta)}$. The parameter of interest is $\psi = (\xi, \beta)$, while the nuisance component consists of the single intercepts $\lambda = (\lambda_1, \ldots, \lambda_N)$. One can see that $\xi > 0$ is the common shape parameter and $\beta = (\beta_1, \ldots, \beta_k)$ contains the regression coefficients associated with the k-dimensional vector of fixed covariates x_{it}.
Failures are right-censored, so data are observations from the random pair $\big(Y_{it}, \Delta_{it}\big)$, where $Y_{it} = \min\big(\widetilde{Y}_{it}, C_{it}\big)$ with C_{it} censoring time and Δ_{it} is the censoring indicator such that $\Delta_{it} = 1(0)$ if $\widetilde{Y}_{it} \leq (>) C_{it}$. The censoring mechanism is only hypothesized to be independent and non-informative, thus each C_{it} is unrelated to the other survival or censoring times and its continuous distribution does not depend on $\theta = (\psi, \lambda)$. In particular, we avoid to parametrically specify the distribution of C_{it}.
For ease of exposition, suppose $T_1 = \ldots = T_N = T$. Writing $\delta_{i\cdot} = \sum_{t=1}^{T} \delta_{it}$ and $\delta_{\cdot\cdot} = \sum_{i=1}^{N} \delta_{i\cdot}$ allows to express the profile log-likelihood for ψ as

$$\begin{aligned} l_P(\psi) = & \sum_{i=1}^{N} \delta_{i\cdot} \left\{ \log \delta_{i\cdot} - \log \sum_{t=1}^{T} y_{it}^{\xi} e^{-\xi (x_{it}^{\mathrm{T}} \beta)} \right\} - \xi \sum_{i=1}^{N} \sum_{t=1}^{T} \delta_{it} (x_{it}^{\mathrm{T}} \beta) \\ & + \delta_{\cdot\cdot} (\log \xi - 1) + (\xi - 1) \sum_{i=1}^{N} \sum_{t=1}^{T} \delta_{it} \log y_{it}, \end{aligned}$$

with maximizer $\hat{\psi} = (\hat{\xi}, \hat{\beta})$. Note that the ith element of the constrained estimate $\hat{\lambda}_{\psi} = (\hat{\lambda}_{1\psi}, \ldots, \hat{\lambda}_{N\psi})$ is obtained as a function of ψ by equating to 0 the scalar partial score

$$l_{\lambda_i}(\theta) = -\xi \delta_{i\cdot} + \xi \sum_{t=1}^{T} (\eta_{it} y_{it})^{\xi}, \qquad i = 1, \ldots, N, \qquad (2)$$

and solving the equation for λ_i. We define then $\hat{\theta}_{\psi} = (\psi, \hat{\lambda}_{\psi})$ and $\hat{\theta} = (\hat{\psi}, \hat{\lambda}_{\hat{\psi}})$.

2.2 Monte Carlo modified profile likelihood

The modified profile log-likelihood (MPL) of Severini (1998) takes the form $l_M(\psi) = l_P(\psi) + M(\psi)$, where the modification term is expressed by

$$M(\psi) = \sum_{i=1}^{N} \left\{ \frac{1}{2} \log j_{\lambda_i\lambda_i}(\hat{\theta}_\psi) - \log I_{\lambda_i\lambda_i}(\hat{\theta}_\psi; \hat{\theta}) \right\}. \tag{3}$$

While j_{λ_i,λ_i} is readily available from the double differentiation of the log-likelihood with respect to λ_i, explicitly calculating the expected value $I_{\lambda_i\lambda_i}(\hat{\theta}_\psi;\hat{\theta})=E_{\hat{\theta}}\{l_{\lambda_i}(\hat{\theta}_\psi)l_{\lambda_i}(\hat{\theta})\}$ requires a parametric assumption on the distribution of C_{it}. Such restriction is not needed to compute the MPL via Monte Carlo simulation instead, adapting the general idea already used by Bartolucci et al. (2016) in econometric fixed effects models for panel data.

The expectation in (3) can be approximated through the empirical mean

$$I^*_{\lambda_i\lambda_i}(\hat{\theta}_\psi; \hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} l^r_{\lambda_i}(\hat{\theta}_\psi) l^r_{\lambda_i}(\hat{\theta}), \qquad i = 1, \dots, N, \tag{4}$$

where, in our case, $l^r_{\lambda_i}$ is the partial score (2) for the rth Monte Carlo sample $(y^r_{it}, \delta^r_{it})$, derived nonparametrically via the Kaplan-Meier estimator. Eventually, we have $y^r_{it} = \min(\tilde{y}^r_{it}, c^r_{it})$ with corresponding indicators δ^r_{it}.

3 Simulation studies

Two experiments with 2000 iterations are run to discuss inference on ψ under diverse overall proportion of censored data ($P_c = 0.2, 0.4$) in the survival model described above, with $k=2$. Subjects of comparison are $l_P(\psi)$ and its Monte Carlo adjustment $l_{M^*}(\psi) = l_P(\psi) + M^*(\psi)$, where $M^*(\psi)$ equals (3) with $I_{\lambda_i\lambda_i}(\hat{\theta}_\psi;\hat{\theta})$ replaced by (4). Dimensions of the artificial datasets are $T = 4, 6, 10$ and $N = 50, 100, 250$. In every group, the first covariate is set to 0 if $t = 1, \dots, T/2$ and to 1 otherwise, the second is sampled independently from a $N(0, 1)$ distribution. We suppose $\xi = 1.5$ and $\beta = (-1, 1)$, while the independent incidental parameters are drawn as $\lambda_i \sim N(0.5, 0.5^2)$. Censoring times are realizations of an $Exp(\varsigma)$ random variable, with ς chosen according to the selected P_c like in Cortese and Sartori (2016, Sec. 5).

Table 1 shows some measures of inferential accuracy for ξ from the study referred to $P_c = 0.2$. Empirical bias (B), root mean squared error (RMSE) and ratio SE/SD, where SE is the average over simulations of likelihood-based estimated standard errors and SD is the standard deviation of the estimates, are reported along with actual coverages of 0.95 Wald confidence intervals (CI). These results, similar to those recorded when $P_c = 0.4$ and for β, attest how the Monte Carlo modification remarkably improves point and interval estimations supplied by the profile likelihood.

References

Bartolucci, F., Bellio, R., Salvan, A. and Sartori, N. (2016). Modified profile likelihood for fixed-effects panel data models. *Econometric Reviews*, **35**, 1271 – 1289.

Cortese, G. and Sartori, N. (2016). Integrated likelihoods in parametric survival models for highly clustered censored data. *Lifetime Data Analysis*, **22**, 382 – 404.

Davison, A.C. and Hinkley, D.V. (1987). *Bootstrap Methods and their Application.* Cambridge University Press.

TABLE 1. Inference on ξ in the stratified Weibull regression model with unknown censoring distribution and $P_c = 0.2$. Figures based on 2000 trials and $R = 500$.

N	T	Method	B	RMSE	SE/SD	0.95 CI
50	4	$l_P(\psi)$	0.392	0.418	0.858	0.111
		$l_{M^*}(\psi)$	0.010	0.112	0.979	0.956
	6	$l_P(\psi)$	0.231	0.252	0.884	0.291
		$l_{M^*}(\psi)$	0.008	0.088	0.964	0.943
	10	$l_P(\psi)$	0.124	0.141	0.976	0.517
		$l_{M^*}(\psi)$	0.005	0.061	1.029	0.961
100	4	$l_P(\psi)$	0.371	0.385	0.840	0.015
		$l_{M^*}(\psi)$	-0.006	0.079	0.966	0.936
	6	$l_P(\psi)$	0.219	0.230	0.903	0.063
		$l_{M^*}(\psi)$	-0.000	0.060	0.987	0.947
	10	$l_P(\psi)$	0.119	0.128	0.938	0.259
		$l_{M^*}(\psi)$	0.001	0.044	0.989	0.943
250	4	$l_P(\psi)$	0.366	0.372	0.847	0.000
		$l_{M^*}(\psi)$	-0.009	0.050	0.972	0.939
	6	$l_P(\psi)$	0.214	0.218	0.890	0.000
		$l_{M^*}(\psi)$	-0.005	0.039	0.972	0.934
	10	$l_P(\psi)$	0.116	0.120	0.943	0.018
		$l_{M^*}(\psi)$	-0.002	0.028	0.993	0.949

Pierce, D.A. and Bellio, R. (2006). Effects of the reference set on frequentist inferences. *Biometrika*, **93**, 425 – 438.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika*, **90**, 533 – 549.

Severini, T.A. (1998). An approximation to the modified profile likelihood function. *Biometrika*, **85**, 403 – 411.

Using State Space models to track Measles Infections

Kirsten Eilertson[1], Matthew Ferrari[2], John Fricks[3]

[1] Penn State Statistics Department, USA
[2] Penn State Center for Infectious Disease Dynamics, USA
[3] Arizona State Statistics Department, USA

E-mail for correspondence: eilertson@psu.edu

Abstract: We present a state space model for estimating the measles disease burden at the country level. Our approach estimates model parameters of a dynamic epidemic model using surveillance data from each country, and builds off of previous work by incorporating age distribution information and explicitly estimating the uncertainty in the estimates for use in forward projections. The primary aim of this work is to improve our understanding of the impact of public health measures.

Keywords: state space model; disease burden; particle filters.

1 Introduction

We present a state space or hidden markov model of measles disease burden, and estimate parameters of the model from surveillance data at the country level collected by the World Health Organization (WHO). The aim is to predict the unobserved values of the number of individuals in a population who are infected and susceptible each year using the observed information of number of cases reported. From this we hope to provide valueable information on the impact of regular and supplemental vaccine campaigns both historically and in the near future. Fitting the model to each of 193 countries will enable the setting of goals for vaccination programs, allocation of resources, and evaluation of program success on a country by country basis.

Our method builds off of earlier work by Chen, Fricks and Ferrari (2012). Like this previous work it combines expert knowledge in the dynamics of measles disease burden with surveillance data, but additionally it incorporates age distribution of the popluation and modifies the model so that parameter estimation and forward projections are tenable. First we present the basic model for progression 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.

measles through time and then discuss our approach to fitting the model using a particle filter.

2 Model

The model form is a dynamic non-linear epidemic model known as susceptible-infected-recovered or an SIR model. In this framework individuals are assumed to belong to one of three classes: susceptible (S, after birth), infected (I), and recovered (R, who are immune to subsequent infection) (Anderson and May (1991), Bjornstad et al. (2002)). The number of susceptibles in year t can be modeled as a function of the number of susceptibles from year $t-1$, the number of births (B) that year, the impact of vaccination campaigns, and infections from the previous year.

$$S_t = S_{t-1} + B_t - V_t(B_t, S_{t-1}) - I_{t-1}$$

Our model specifies that the impact of vaccines on the number of susceptibles $V_t(B_t, S_{t-1})$ incorporates the age distribution of the population, the efficacy of vaccinations based on age of administration in that country, and "pulsed" or suplemental vaccine campaigns which target specific age groups.

The resulting number of infections in the year's susceptible population is a function of the susceptible population size relative to the total population S_t/N_t and other factors captured with the term e_t where $e_t \sim N(0, \sigma_e)$.

$$I_t \sim Bin(S_t, \pi_t(S_t, N_t, e_t))$$
$$\pi_t = \text{logit}^{-1}(\beta_0 + \beta_1 \frac{S_t}{N_t} + e_t)$$

Using this logistic form to model the probability of infection as a function of the proportion of susceptibles in the population is phenomonologically consistent with herd immunity.

The surveillance data reflects the number of reported cases, C_t, a subset of the number of actual cases.

$$C_t \sim Bin(I_t, p_t)$$

Our model states that cases are reported with probability p, where p is specific to that country and genearlly assumed to be constant. For some countries, some years have been identified (based on outside factors) as "high reporting years". These are years where due to an epidemic or other reasons a country is thought to have a higher than usual probability of reporting cases. Thus, the probability of reporting may vary across these "high reporting" years. These years are usually less than 10% of the years observed.

3 Fitting

We are primarily interested in the number of individuals infected each year. We would like to "predict" these values both for each year in the past I_t (where t is a year within the span of surveillance data) conditioning on all years of observed of data. Additionally, if our parameters (β_0, β_1, σ_e) are well-estimated than we can use the model to run forward projections under different scenarios of vaccination campaigns.

To this end we will use a particle filter. A filter in this context allows us to use the conditional density of the state of a system (the pair $[S, I]_t$ in our case) at time t given the observed part of the system up to time t $(C_1, ..., C_t])$. Briefly, the filtering step is as follows:

1. draw proposals $[S, I]_t^1, ..., [S, I]_t^R$ from $f([S, I])$ for $j = 1, ..., R$
2. evaluate conditional likelihood $f(C_t|[S, I]_t^j)$ to calulate weight w_j
3. resample M observations from $[S, I]_t^1, ..., [S, I]_t^R$ according to weights w_j

These steps are then repeated to progress through the data. The evaluation of the filter provides the value of the likelihood function of the system for a given set parameter values. This will allow us to use the particle filter to numerically optimize this likelihood function to give maximum likelihood estimators for the parameters. In addition to the filter, a smooth may also be calculated where all of the observed values of C both past and future states are used to predict I_t. In this way, the particle filter approach allows us a “best guess” for the state at time t given the observations through the conditional expectation; as well as a means to obtain forward projections under various vaccine campaign scenarios.

4 Simulation Example

Using simulated data we examine the reliability and sensitivity of our estimation procedure. Simulated data is generated by sampling values for β_0, β_1, and p, as well as sampling a vaccine campaign history. Using preliminary fits to real country data from 193 countries, we obtained a range of parameter values to explore. Refer to Figure 1 for an example of a simulated country history.

From simulations such as these we will establish expected coverage of particle clouds for past years as well as future projections.

References

Anderson, R. M. and R. M. May (1991). *Infectious diseases of humans: dynamics and control.* Oxford: Oxford University Press.

Bjornstad, O. N., B. F. Finkenstadt, and B. T. Grenfell (2002). *Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series sir model.* Ecological Monographs **72(2)**, 169 – 184.

Chen, S., Fricks, J., Ferrari, M (2012). *Tracking measles infection through non-linear state space models.* Journal of the Royal Statistical Society: Series C (Applied Statistics) **61.1** 117-134.

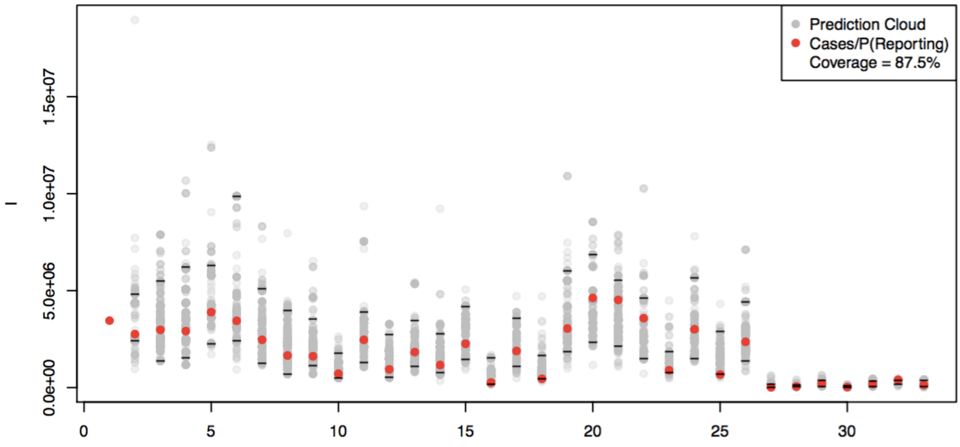


FIGURE 1. Particle cloud coverage of number of infections. On the x-axis is year, and on the y-axis is number of infections. The red dots indicate the “true” simulated number of infections. The dramatic reduction in measles cases in year 27 is a result of a supplemental vaccine campaign included in the simulation.

DiNAR: web application for Differential Network Analysis and visualisation in R

M. ZAGORŠČAK[1], A. BLEJEC[1], Ž. RAMŠAK[1], M. PETEK[1], K. GRUDEN[1]

[1] National Institute of Biology, Slovenia

E-mail for correspondence: maja.zagorscak@nib.si

Abstract: To these days most widely used approach for comparing biological samples under multiple condition is differential expression (DE) analysis. As the generated expression data sets are becoming larger and more complex, rather than to investigate change in the gene behaviour as an isolated occurrence, more popular question to answer is which parts of system are perturbed and rewired during an altered response. To access and understand difference in signalling relationships at the system level differential network analysis (DiNA) is used. DiNA requires as an input reference and condition specific networks. Accurate reconstruction of true biological network topology due to a smaller number of samples than the number of genes is quite a complicated problem. Majority of existing methods are relying on correlation based analysis for construction of gene regulatory networks which entails noise like spurious inference, reaction oversimplification and missed latent and mediated effects. Other more accurate methods are computationally exhaustive even when focused only on two conditions and in the case of long time series experiments completely impractical. As a simple and straightforward approach we propose to integrate abundant biological knowledge and condition specific experimental data by mapping obtained DE onto verified signalling pathways.
Here we present DiNAR, a Shiny web application that allows researcher to analyse multiple condition data sets in the biological network context using an interactive graphical interface with user defined thresholds and colour palettes. As a reference network the large knowledge network of *Arabidopsis thaliana*, created by merging expert knowledge based interactions with large scale experimentally validated interactions was introduced. Translation tables based on PLAZA orthologues information allow the automated transfer of knowledge from model organism to some non-model plant species. Condition specific networks are automatically generated when experimental data files are uploaded. Changes in processes of interest can be monitored through a simple animation.
DiNAR is written in R and customised using JavaScript, HTML and CSS. App 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

accessible over the web at *https://nib-si.shinyapps.io/DiNAR* or it can be run directly from the GitHub repository using *shiny:::runGitHub("NIB-SI/DiNARscripts", "NIB-SI")* command.
Although DiNAR is demonstrated on high-throughput biological data it can handle any properly formatted custom network and associated quantitative data set.

Keywords: Condition specific network; Differential network analysis; Shiny

References

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, **8**, 565

Henry,V.J., Bandrowski, A.E., Pepin, A.-S. et al. (2014). OMICtools: an informative directory for multi-omic data analysis. *Database (Oxford)*, **2014**, bau069

Banf M. and Rhee S.Y. (2017). Computational inference of gene regulatory networks: Approaches, limitations and opportunities. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms.*, **1860**, 41 – 52.

Median unbiased estimator for the two-parameter logistic model

Euloge Clovis Kenne Pagui[1], Edoardo Michielon[2], Alessandra Salvan[1], Nicola Sartori[1]

[1] Department of Statistical Sciences, University of Padua, Italy
[2] SDG Consulting, Milan, Italy

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

Abstract: We propose a median modification of the score equation for large-dimensional parameter estimation in fixed effects one-parameter and two-parameter logistic models, well known as Rasch models. These models are of great importance in item response theory for the analysis of data generated by aptitude tests or questionnaires. The number of parameters grows together with the number of subjects or items and complicates the estimation procedure. Similarly to logistic regression, the model often presents cases of complete data separation. As a consequence, all or some components of the maximum likelihood estimate may be infinite. Moreover, the maximum likelihood estimates are often not available for the two-parameter logistic model due to numerical irregularities of the likelihood. The proposed method estimates the whole vector of parameters and also solves the infinite estimate problem. The performance of the approach is evaluated by simulation studies and compared with mean bias reduction.

Keywords: Infinite estimates; Likelihood; Logistic model; Median unbiased; Modified score

1 Introduction

The two-parameter logistic (2PL) model is a generalization of Rasch (1960) one-parameter logistic (1PL) model, originally proposed in the context of reading ability tests. These models are of great importance in item response theory for the analysis of data generated by aptitude tests or questionnaires. Without latent variables assumptions, the number of parameters grows together with the number of subjects or items and complicates the estimation procedure. Additionally, the data could present cases of complete or quasi-complete separation implying nonexistence of maximum likelihood estimates. Many authors have studied 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

properties of the standard estimators of these models, such as marginal, conditional and joint maximum likelihood estimators (see for instance Andersen, 1980, Ch. 6). Although joint maximum likelihood was the preferred method of estimation due to its intrinsic robustness, it soon became clear that numerical issues were involved in the estimation of the 2PL model (Baker, 1987, 1988). For this reason, marginal maximum likelihood with a random effects assumption on the subjects abilities is nowadays the standard choice.

Firth (1993) proposed an implicit method for bias reduction of the maximum likelihood estimator by adding a modification term to the score function. The corresponding modified estimating equation does not depend explicitly on the maximum likelihood estimate and has been found to overcome infinite estimate problems. An alternative modification of the score equation is proposed in Kenne Pagui et al. (2016), aiming at median centering of the estimator. This method respects equivariance under componentwise reparameterizations and is also effective in preventing infinite estimates. The modification is obtained by equating each score component to the approximate median of its profiled version. Both bias and median bias reduction allow fixed effects estimation of the 2PL model, thus avoiding the latent variable assumption.

2 The 1PL and 2PL models

Let y_{si} be the answer for subject s to item $i, s = 1, \ldots, S,\ i = 1, \ldots, I$. Assume that the observations y_{si} are realisations of independent random variables $Y_{si} \sim Bin(1, \pi_{si})$. Here, π_{si} is viewed as the probability that subject s succeeds in item i of an achievement test. Under the 1PL model, $\log\{\pi_{si}/(1-\pi_{si})\} = \gamma_s + \alpha_i$, where γ_s is interpreted as a measure of the ability of subject s, while α_i is a measure of easiness of item i. Thus, a large value of α_i corresponds to high probability of a correct response. When π_{si} is viewed as a function of γ_s, it is called the item characteristic curve (ICC). The constraint $\alpha_1 = 0$ allows identifiability of the model so that there are $S + I - 1$ unknown parameters.

The 2PL model generalises the 1PL model by assuming $\log\{\pi_{si}/(1 - \pi_{si})\} = \beta_i\gamma_s + \alpha_i$, where β_i is interpreted as the discrimination parameter of item i. The larger the discrimination parameter, the steeper the ICC curve is. Here, with the constraints $\alpha_1 = 0, \beta_1 = 1$, the model is identifiable and has $S+2(I-1)$ unknown parameters (San Martin et al., 2015).

The number of parameters drastically increases with the number of subjects and hence a suitable estimation method is needed. For the 1PL model, the conditional likelihood works well in estimating the item parameters since the maximum likelihood estimator is asymptotically biased. However, in the 2PL model, maximum likelihood estimates are often affected by numerical instabilities and no conditional likelihood is available. Therefore, the standard solution assumes γ_s as a random effect, usually $N(0, 1)$. In Section 3, we propose a method that makes no assumption on γ_s.

3 Median bias reduction method

Consider a regular statistical model with parameter $\theta = (\theta_1, \ldots, \theta_p)$. We denote by $\ell(\theta)$ the log likelihood, $i(\theta)$ Fisher information, and $\hat{\theta}$ the maximum likelihood estimator. Let i_{rs} be a generic entry of $i(\theta)$ and i^{rs} an entry of its inverse, $r, s, \ldots = 1, \ldots, p$. Let U_r and U_{rs} be partial derivatives of $\ell(\theta)$ with respect to elements of θ with indices r and s. In particular, U_r is the rth component of the score function. Moreover, expected values of log likelihood derivatives are denoted as $\nu_{r,st} = E_\theta(U_r U_{st})$ and $\nu_{r,s,t} = E_\theta(U_r U_s U_t)$.

In the following, summation is understood over repeated indices. The median bias reduced estimator, $\tilde\theta$, is obtained as a simultaneous solution of the modified score equations

$$\tilde U_r = A_{rs}U_s - \kappa_{1r} + \tfrac{1}{6}\tfrac{\kappa_{3r}}{\kappa_{2r}},\ r = 1, \ldots, p,\ \text{with}\ A_{rs} = i^{rs}/i^{rr},$$

where κ_{jr}, $j = 1, 2, 3$, are approximate cumulants of the profiled version of U_r, with the following expressions $\kappa_{1r} = -\frac{1}{2} i^{rs}\nu^{tu}(\nu_{s,tu} + \nu_{s,t,u})/i^{rr}$, $\kappa_{2r} = 1/i^{rr}$ and $\kappa_{3r} = i^{rs}i^{rt}i^{ru}\nu_{s,t,u}/(i^{rr})^3$, where $\nu^{tu} = i^{tu} - i^{tr}i^{ur}/i^{rr}$. See Kenne Pagui et al. (2016) for details and properties. In particular, the asymptotic distribution of $\tilde\theta$ is the same as that of $\hat\theta$. For implementation in 1PL and 2PL models, we vectorise the $S \times I$ data matrix obtaining a $n \times 1$ vector of the form $y = (y_{11}, \ldots, y_{S1}, y_{12}, \ldots, y_{S2}, \ldots, y_{1I}, \ldots, y_{SI})^T$, with a $n \times p$ design matrix X and where $n = S \times I$. Due to the special structure of Rasch models, summation over three indices in $\tilde U_r$ can be avoided by exploiting sparsity of arrays involved in the modification. A further significant computational gain is achieved by implementing the modification term using `Rcpp` rather than plain `R`.

4 Numerical studies

We conducted a simulation study to assess the performance of the median bias reduced (MBR) estimator in 2PL model. We compare the new estimator with the mean bias reduced (BR) estimator of Firth (1993) in terms of error distribution. The BR estimates are obtained through the `R` package `brRasch` available on `GitHub`. Here, the simulation is under the usual assumption of a random effects model. In particular, the results are obtained with 1,000 replications, $I = 10$, $S = 100$ and assuming $\gamma_s \sim N(0, 1)$. From Figure 1, it appears that the BR and MBR estimators of α are almost comparable in terms of estimation error, while the MBR estimator of β is preferable. Unreported simulation results in the more extreme settings show that the method still gives reasonable results.

References

Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications.* North-Holland, Amsterdam.

Baker, F. B. (1987). *Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models.* Applied Psychological Measurement, 11, 111 – 141.

Baker, F. B. (1988). *The item log-likelihood surface for two-and three-parameter item characteristic curve models.* Applied Psychological Measurement, 12, 387 – 395.

Firth, D. (1993). *Bias reduction of maximum likelihood estimates.* Biometrika, 80, 27 – 38.

Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2016). *Median bias reduction of maximum likelihood estimates. http://arxiv.org/abs/1604.04768.*

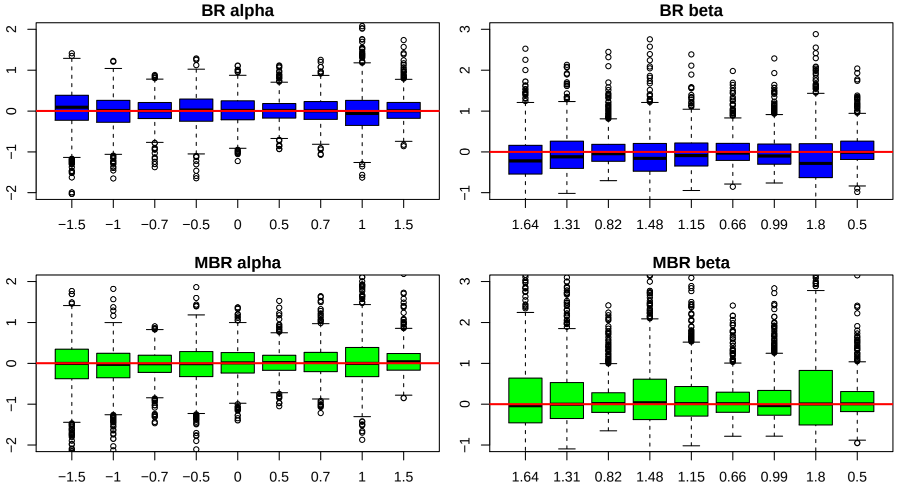


FIGURE 1. Estimation error simulated distribution for the BR and MBR estimators of (α, β) with $I = 10$, $S = 100$ and $\gamma_s \sim N(0, 1)$.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Studies in Mathematical Psychology I. Danish Inst. Educational Research, Copenhagen.

San Martín, E., González, J., and Tuerlinckx, F. (2015). *On the unidentifiability of the fixed-effects 3PL model.* Psychometrika, 80, 450 – 467.

Modelling and categorisation of seismic waves

Duncan McGregor[1], Andrew Bell[2], Bruce J. Worton[1]

[1] School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, Edinburgh, UK
[2] School of GeoSciences, The University of Edinburgh, Edinburgh, UK

E-mail for correspondence: Bruce.Worton@ed.ac.uk

Abstract: A key problem in seismology is assessing the similarity or difference between events, and constructing categorisations based on these measures. Clustering algorithms remain an active area of research, but many approaches are well documented. This paper assesses the suitability of a selection of these approaches to the problem of seismic waves, with reference to a data set taken from Tungurahua, Ecuador between 6–13 April 2015. In addition, approaches to modelling the waves, both parametric and nonparametric are fitted and assessed, and the suitability of certain data transformations are considered.

Keywords: Seismic waves, Clustering, Nonparametric.

1 Introduction

To better understand the processes occurring within volcanoes, seismologists study the seismic waves generated within — these are vibrations that propagate through the earth, and are measured by seismometers located near the volcano. The raw seismic data is processed to identify distinct ‘seismic events’ where the seismic activity rises above some level. There are many aspects to the study of these events, however one key problem concerns their categorisation with the intention of identifying events that are similar and likely to have arisen from a common source. Seismologists can then use this categorisation to trace the evolution of the number of events occurring before, during, and after periods of activity, and make inferences about the nature of the events giving rise to activity. This process of categorisation is currently not well-defined, with a variety of similarity measures and ad-hoc approaches to clustering in use. Plausible approaches exist and are in wide use, but these are computationally expensive and have practical shortcomings we will discuss. The analysis is further complicated by the presence of ambient vibrations arising from natural and man-made sources, and any categorisation must allow for the presence of significant noise in the event signals.

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

2 Data set and modelling

We study modelling seismic data for 4805 events between 06/04/2015 and 13/04/2015 recorded at the station next to the Tungurahua volcano. Each event data item records the velocity of the vibration at 3001 distinct equally spaced points in time over a period of 30 seconds.

Our attempts to fit flexible parametric models to the data proved very challenging due to computational issues. An exhaustive search of suitable models would be impossible, but whilst the failure of our efforts cannot confirm the task is impossible. However, we have considered various possible nonparametric models (Hastie et al, 2009). All appear to give reasonable fits to the data, but wavelets appear to best capture the behaviour of the model (Donoho and Johnstone, 1994), and lend themselves readily to the smoothing of the wave (to remove ambient noise) and downsampling to reduce the dimension of the problem.

3 Results

Figure 1 gives an example of an event. Note the noisy nature of the data which leads to problems with parametric fits.

The gap statistic analysis provides useful summary information concerning the clusters. For example, on Day One there is no strong evidence for more than 6 groups. A typical example of a dendrogram (see Figure 2) for a day of events is rather crowded (making it difficult to identify individual events). However the high level structure is clearly visible and it is instructive to compare the distance of the different numbers of clusters with the Gap Statistic.

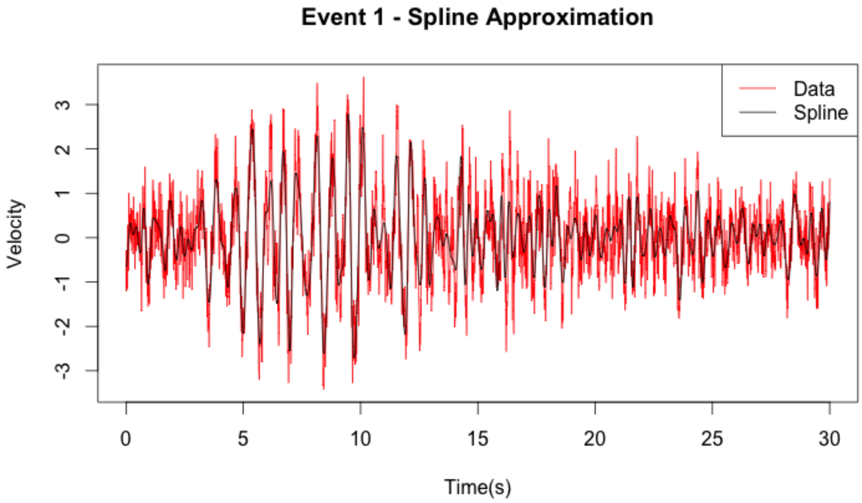


FIGURE 1. An event: example of the type of data being modelled.

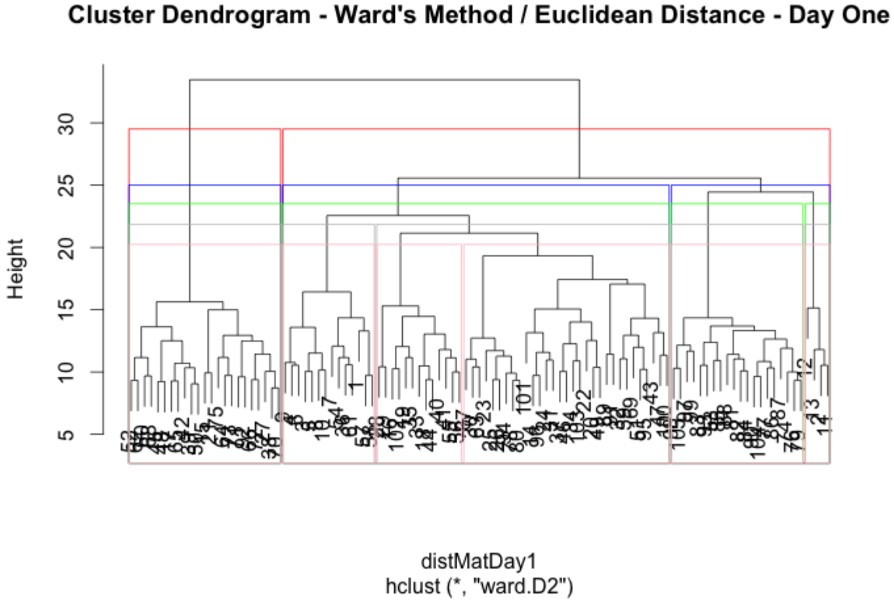


FIGURE 2. Spectral data, Euclidean distance, Ward's method on Day 1. This suggests 2 to 6 groups.

4 Conclusions

We started this work with the intention of investigating methods of categorising seismic waves. At the end, we have arrived at the following conclusions.

1. Seismic events are complex, and not readily modelled using a simple parametric approach.
2. Smoothing of events can be performed through a variety of non-parametric techniques, or by low-band pass.
3. Seismic events data require transformation before they can be compared directly. Transformations include: (a) Scaling to match amplitude of events; (b) Translation to align events; (c) Windowing to eliminate regions where the signal to noise ratio is too low to be useful; (d) Smoothing to eliminate high frequency noise; (e) Fourier Transform to shift the event from the time domain to the frequency domain; (f) Principal Component Analysis to reduce the dimension of the problem.
4. Scaling and translation are essential for clustering techniques using conventional distance measures.
5. Windowing appear to stabilise the clustering under different clustering methods and distance measures.

6. Clustering methods give quite different groupings for smoothed signals. Relatedly, groupings are not invariant under decimation of the wavelet smoothed signal (Donoho and Johnstone, 1994).
7. Principal Component Analysis can reduce the dimension of our data from 3001 dimensions to 200 and retain over 80% of the observed variance.
8. The existing technique of cross-correlation looks sensible and fit for purpose, however we propose an alternative technique of carrying our Gap Analysis on the Spectral Intensity Data (Tibshirani et al, 2001). The brief simulation study we carried out suggests the technique requires refinement, and that spectral intensity data may not be the optimal choice, and applying Gap analysis on the raw data using cross correlation as a similarity measure may be a superior technique.

Acknowledgments: We are particularly grateful to The Instituto Geofisico of the Escuela Politecnica Nacional (IGEPN) of Ecuador for all their hard monitoring work and providing the data.

References

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425 – 455.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Second Edition. New York: Springer.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B*, **63**, 411 – 423.

Bayesian Modelling of the Exponential Random Graph Models with Non-observed Networks

Abdolreza Mohammadi[1], Ernst C. Wit[2]

[1] Tilburg University, Netherlands
[2] University of Groningen, Netherlands

E-mail for correspondence: a.mohammadi@uvt.nl

Abstract: Across the sciences, one of the main objectives is network modelling to discover complex relationships among many variables. The most promising statistical models that can be used for network modelling is the Exponential Random Graph Models (ERGMs). These models provide an insightful probabilistic model to represent a variety of structural tendencies that define complicated dependence patterns hardly modelled by other probabilistic models. However, they are restricted to the models that regarded the network as given, observed networks data. In the present paper, we develop a novel Bayesian statistical framework which combines the class of ERGMs with graphical models capable of modelling non-observed networks. Our proposed method greatly extends the scope of the ERGMs to more applied research areas. We discuss possible extensions of the method.

Keywords: Bayesian inference; Exponential random graph models; Graphical models; Birth-death process; Markov chain Monte Carlo.

1 Introduction

Network modeling pervades all of science since one of the main objectives of science is to discover complex relationships among large numbers of variables. For the prevention of epidemics, social science relies on a keen understanding of interactions among individuals. One way to describe these kinds of complex relationships is by means of an abstract network.

Exponential random graph models are promising and flexible family of statistical models for modelling network topology. These models have been used mainly in the social science literature since they allow to statistically account for 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.

complexity inherent in many network data (Snijders 2002). In ERGMs, the basic assumption is that the topological structure of an observed network can be explained by the a vector of network statistics that capture the generative structures in the network (Snijders et al. 2006). However, up till now, these models are restricted to the observed network data. In most applications, we have multivariate data which are not the observed networks, like in biology in neuroscience. The main question is that is it possible to extend the ERGMs to the multivariate data that are not observed network? A possible solution is combining the class of ERGMs with graphical models.

Graphical models Lauritzen(1996) provide an appealing and insightful way to obtain uncertainty estimates for inferring network structure. The close relationship between the topology of the underlying graphs and their probabilistic properties is a main aspect in graphical models, and it provides the potential tools to interpret the underling graph structure. In this regard, Bayesian approaches provide a mainly straightforward tools, and much recent progress has been made in graphical models (Mohammadi and Wit 2015). However, graphical models are powerful approaches only for estimating the underlying graph structure, they are not designed for network modeling.

In this paper, we develop a new Bayesian statistical framework for ERGMs, which is capable not only network modeling but also estimating underlining graph for multivariate data which are not observed networks. The proposed method greatly extends the scope of the ERGMs to more applied research areas, which not limited only in social science. In our method, to apply the ERGMs to non-observed networks data, we combine the class of ERGMs with graphical models capable of modelling non-observed networks. In particular, in our Bayesian framework, we design a computationally efficient search algorithm to explore all the graph space to distinguish not only important edges but also key features and detect the underlying graph structure. This search algorithm is based on birth-death Markov chain Monte Carlo algorithm proposed by Mohammadi and Wit (2015) for Gaussian graphical models.

2 Exponential families and graphical models

Exponential random graph models models are the families of statistical models that provide a flexible way to model the complex dependence structure of networks. The aim to model data as observed networks consisting of nodes and edges, which in the social network context represent actors and relationships between these actors, respectively.

In an ERGMs, the random matrix $G = \{g_{ij}\}$ is defined over the graph space on a set of p nodes, with each variable in G representing the presence or absence of a particular edge ($g_{ij} = 1$ if there is a link from i to j, and $g_{ij} = 0$ otherwise). Edges connecting a node to itself are not allowed so $g_{ii} = 0$. For a graph, the ERGM is then given by

$$P(G|\theta) = \frac{1}{Z(\theta)} \exp\{\theta^t S(G)\}, \tag{1}$$

where $\theta \in \Theta$ represents a vector of unknown parameters, and $Z(\theta)$ is a normalizing constant and $S(G)$ term is a network statistic of interest that gives the

ERGMs much of its explanatory power. The vector $S(G)$ can contain statistics to capture the generative structures of connectivity in the network.

Note, in ERGMs data are observed networks, which is a strong limitation. In most of the applications, data are measured on variables, such as gene expression data and cell signalling data. The question we intend to answer is whether it is possible to extend the idea of ERGMs to those types of data? We extend the ERGMs to multivariate data (that are not observed networks) by combining it with graphical models.

Graphical models Lauritzen(1996) use a graph concept to represent conditional dependence relationships among random variables, as non-observed networks. When observed data come from noisy measurements of the variables, then graphical models present an appealing and insightful way to describe graph-based dependencies between the random variables. A graph $G = (V, E)$ denotes a set of vertices $V = \{1, 2, ..., p\}$ – where each node corresponds with a random variable – and a set of existing edges E. In this class of models, nodes in the graph G correspond to the random variables. The absence of an edge between two nodes determines the two corresponding variables are conditionally independent given the remaining variables. Graphical models that follow the multivariate Gaussian distribution are called Gaussian graphical models (GGMs), also known as covariance selection models.

3 Bayesian hierarchical model for ERGMs with non-observed networks

We can display the hierarchical model schematically as below

$$\theta \longrightarrow G \longrightarrow K \longrightarrow \mathbf{X} = (X_1, ..., X_n).$$

Thus, we consider the joint posterior distribution of the parameters as bellow

$$P(\theta, G, K \mid \mathbf{X}) \propto P(\mathbf{X} \mid \theta, G, K)\ P(K \mid G)\ P(G \mid \theta)\ P(\theta). \tag{2}$$

In our methodology, we assume the observed data follows a multivariate Gaussian distribution.

For prior specification on graph, by consider the idea of exponential random graph, we use a prior on the graph as follows

$$P(G \mid \theta) = \frac{1}{Z(\theta)} \exp\{\theta^t S(G)\} \tag{3}$$

where $S(G)$ is a vector of statistics of the graph (e.g., the number of edges, triangles, etc.) and $\theta \in \Theta$ denotes the parameter vector of the model. For the prior distribution of the precision matrix, we use the G-Wishart distribution.

3.1 MCMC sampling scheme

The MCMC algorithm is in three steps as follows

Step 1: Sample from θ, based on exchange algorithm (Murray et al. 2012).

Step 2: Sample from graph space, based on birth-death MCMC sampling algorithm proposed by Mohammadi and Wit (2015).

Step 3: Sample from precision matrix, based on exact sampling algorithm form G-Wishart distribution.

For step 1, We sample from conditional distribution of θ based on exchange algorithm (Murray et al. 2012). For step 2, we using computationally efficient birth-death MCMC sampler proposed by Mohammadi and Wit (2015) for Gaussian graphical models. Their algorithm explores the graph space by adding or deleting an edge in a birth or death event, in which the events are based on a continuous time birth-death Markov process.

References

Lauritzen, S.L. (1996). *Graphical models.* Oxford University Press.

Mohammadi, A. and Wit, E. (2015). Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian Analysis*, **10(1)**, 109 – 138.

Mohammadi, A. and Wit, E. (2016). BDgraph: An R Package for Bayesian Structure Learning in Graphical Models. *Journal of Statistical Software.*

Murray, I. and Ghahramani, Z., David (2012). MCMC for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848.*

Snijders, T. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, **3(2)**, 1 – 40.

Snijders, T. and Pattison, P. and Robins, G. and Handcock, M. (2006). New specifications for exponential random graph models. *Sociological methodology*, **36(1)**, 99 – 153.

Analysis of differential effects in multi-group regression methods

Adrian Quintero[1], Emmanuel Lesaffre[1], Geert Verbeke[1],
Benedict Orindi[1], Luk Bruyneel[1]

[1] KU Leuven, I-BioStat, Leuven, Belgium

E-mail for correspondence: luisadrian.quinterosarmiento@kuleuven.be

Abstract: In regression analysis, the data sample is often composed by sub-populations and the impact of some of the covariates may differ across groups. A model with group interactions can be fit to avoid biased estimates, but the power to test for significant effects is importantly reduced, especially when the number of sub-populations is large. We propose a prior distribution which combines the information of the groups with a similar covariate effect. This increases importantly the power whilst allowing to study differential effects across sub-populations. The method is applied to analyse patients' satisfaction in seven European countries.

Keywords: Differential effects; Multiple groups; MCMC methods.

1 Introduction

In regression analysis, the response is usually modelled as a function of the explanatory variables assuming that the effect of each covariate is the same for all observations. Such a model can be expressed as

$$y_i = \beta_0 x_{0,i} + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i, \tag{1}$$

where the covariate x_0 corresponds to the intercept and the error is assumed to follow $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, n$. However, data often come from populations with well-defined groups such as geographical region, level of education, ethnicity or gender. Ignoring group effects may have serious consequences on the estimation of regression coefficients. As a simple example, let us consider the case in which a specific covariate has a strong negative effect on the response for women and a strong positive effect for men. Fitting a model with a common regression coefficient for both groups would lead to a biased estimate and perhaps inferring non-significance of the covariate effect. Therefore, the following model with interaction terms may be more appropriate in multi-group studies

$$y_{i(g)} = \beta_{0(g)}x_{0,i(g)} + \beta_{1(g)}x_{1,i(g)} + \cdots + \beta_{p(g)}x_{p,i(g)} + \varepsilon_{i(g)}, \tag{2}$$

where $y_{i(g)}$ is the response for individual $i = 1, \ldots, n_g$ in group $g = 1, \ldots, G$. The regressor $x_{j,i(g)}$ corresponds to the interaction between covariate $x_{j,i}$ and group membership and $\varepsilon_{i(g)} \sim \mathcal{N}(0, \sigma^2)$. Hence, the effect of covariate x_j in group g is $\beta_{j(g)}$ for $j = 0, \ldots, p$.

However, model (2) embraces low power to detect significant effects when the number of groups is large. To illustrate this, let us assume that the variance σ^2 is known and that all covariates in model (1) are mutually independent, i.e., $\text{Cov}(x_j, x_{j'}) = 0$ for $j \neq j'$. If the covariates are mean centered and standardized, the variance of the regression coefficient estimators under model (1) is $\text{Var}(\hat{\beta}_j) = \sigma^2$ for $j = 1, \ldots, p$. On the other hand, with the interaction model (2) the variance of the estimators is substantially larger, namely $\text{Var}(\hat{\beta}_{j(g)}) = G\sigma^2$ for $j = 1, \ldots, p$. This equality holds assuming that $\sum_i x^2_{j,i(g)} = \sum_i x^2_{j,i(g')}$, $\sum_i x_{j,i(g)} = 0$ for all groups and that the interaction cross-products are null, i.e., $\sum_i x_{j,i(g)}x_{j',i(g)} = 0$ for $j \neq j'$. Therefore, including the interaction terms is unfavourable when a covariate has the same effect across groups, because the model without interactions describes adequately the data and presents higher power for significance testing.

We propose an approach that is a compromise between the two alternatives, allowing for a differential effect across groups but taking into account that the covariate effect may be similar for some or all of the sub-populations. Using a Bayesian hierarchical model, we assume that for a given covariate x_j, the interaction coefficients across groups $\beta_{j(1)}, \ldots, \beta_{j(g)}$ can be classified into 3 components: sub-populations where the effect of the covariate is negative, groups where there is no effect on the response and sub-populations where the impact is positive. If no differential effects are present for the covariate, all the interaction terms can be classified in the same component. The proposed method combines the information of the groups with a similar covariate effect, which improves the power in significance testing, whilst allowing to study differential effects across sub-populations.

2 Proposed Bayesian hierarchical model

A latent variable $\gamma_{j(g)} = \{-1, 0, +1\}$ is introduced to classify the interaction coefficients of the model in the three defined components: the covariate x_j has no impact in group g when $\gamma_{j(g)} = 0$, the effect is negative if $\gamma_{j(g)} = -1$ and the effect is positive when $\gamma_{j(g)} = +1$. The distribution of the interaction coefficients conditionally on the latent variable is defined as

$$\begin{aligned} \left[\beta_{j(g)}|\gamma_{j(g)} = -1\right] &\sim \mathcal{N}\left(\mu_{j(-)}, \sigma^2_{j(-)}\right), \\ \left[\beta_{j(g)}|\gamma_{j(g)} = +1\right] &\sim \mathcal{N}\left(\mu_{j(+)}, \sigma^2_{j(+)}\right), \end{aligned} \tag{3}$$

where $\mu_{j(-)} < 0$ and $\mu_{j(+)} > 0$ for $j = 0, \ldots, p$. If $\gamma_{j(g)} = 0$ we have a degenerate distribution and $Pr(\beta_{j(g)} = 0) = 1$. For each covariate, the proportion of subgroups where x_j has no impact is $\pi_{j(0)} = Pr(\gamma_{j(g)} = 0)$, whereas the overall proportions with a negative and positive effect are respectively $\pi_{j(-)} = Pr(\gamma_{j(g)} =$

$-1)$ and $\pi_{j(+)} = Pr(\gamma_{j(g)} = +1)$. The model specification is completed by assuming independence for $\gamma_{j(g)}$'s across covariates and sub-populations conditionally on $\boldsymbol{\pi}_j = (\pi_{j(-)}, \pi_{j(0)}, \pi_{j(+)})$.

This model for the analysis of differential effects can be estimated via MCMC methods. We select as prior distribution for the component probabilities $\boldsymbol{\pi}_j \sim \text{Dir}(0.25, 0.5, 0.25)$ to have $Pr(\beta_{j(g)} = 0) = Pr(\beta_{j(g)} \neq 0)$. A vague inverse gamma density $IG(0.001, 0.001)$ is assigned to the variance of the error. In the method, the response variable y is mean centered and standardized to facilitate the choice of priors for the parameters in the mixture model (3). We select a priori $\mu_{j(-)} \sim U(-2, 0)$ and $\mu_{j(+)} \sim U(0, 2)$ for $j = 0, \ldots, p$, since the case $\beta_j = \pm 1$ would imply a perfect correlation between the covariate x_j and the response. Finally, for the variances of the mixture components we specify $\sigma_{j(-)}, \sigma_{j(+)} \sim U(0, 1)$.

After running the MCMC algorithm, a sample of the indicator component $\gamma_{j(g)}$ is obtained for each interaction in the model. If the proportion of MCMC iterations in which $\gamma_{j(g)} = -1$ is larger than 0.5, this indicates that the effect of x_j in group g is negative. Similarly, if the proportion in which $\gamma_{j(g)} = +1$ is larger than 0.5, the effect is positive. In any other case, we conclude that there is no impact of the covariate in that group.

3 Simulation study

To assess the proposed method, a simulation study was carried out based on the model $y_{i(g)} = \beta_{0(g)} x_{0,i(g)} + \beta_{1(g)} x_{1,i(g)} + \beta_{2(g)} x_{2,i(g)} + \beta_{3(g)} x_{3,i(g)} + \varepsilon_{i(g)}$ with $n_g = 20$ observations for each sub-population $g = 1, \ldots, 30$. The intercept takes the values $\beta_{0(g)} = \{-3, -2.8, -2.6, \ldots, 2.8\}$ across groups, whereas $\beta_{1(g)} = 0$ and $\beta_{3(g)} = 3$ for all sub-populations. The covariate effect of x_2 is $\beta_{2(g)} = 0$ for $g = 1, \ldots, 15$ and $\beta_{2(g)} = 3$ for $g = 16, \ldots, 30$.

The proposed model (3) is compared with the method advocated by Benjamini and Hochberg (1995) for multiple testing. The so called BH approach controls the false discovery rate (FDR) which is defined as the proportion of hypothesis $H_0 : \beta_{j(g)} = 0$ mistakenly rejected among all (mistakenly and accurately) rejected hypothesis. This allows to control the familywise error rate in a weak sense and admits more powerful procedures compared to Bonferroni-type methods, which are well-known to be conservative.

In the simulation study, 100 replicated data sets were generated from the model. Besides the FDR, we compared the methods based on: false non-discovery rate (FNDR), type I error rate, the power and total number of interaction effects that are misclassified (TotalMiss). The results in Table 1 show that BH presents a lower average FDR but the power is 0.44, very low compared to 0.92 in the proposed model (3). When looking at TotalMiss, we see that around 35% of the 120 interactions are being misclassified by BH whereas in our proposed method it is only 9%.

4 Analysis of patient overall satisfaction

In the registered nurse forecasting (RN4CAST) study, patients were asked to rate their hospitals on a scale from 0 (worst) to 10 (best). Additionally, the subjects responded 16 items related to doctor communication, nurse communication,

TABLE 1. Results from the simulation study.

Method	FDR	FNDR	Type I	Power	TotalMiss
BH	0.017	0.476	0.013	0.446	42.370
Prop.	0.055	0.120	0.093	0.917	10.440

physical environment, pain control and discharge information. Hence, it is of interest to understand how the patient experiences (reflected in the 16 items) relate to the overall rating given by patients to the hospitals.

Furthermore, the RN4CAST is a multi-country study, so the 2942 patients in the sample (after discarding observations with missing information) belong to seven different countries: Belgium, Switzerland, Spain, Finland, Greece, Ireland and Poland. Therefore, it is important to determine if the impact of the 16 items on overall satisfaction is the same across countries or, on the contrary, whether any patient experiences are more relevant in some specific nations.

Fitting model (1) indicates that 13 of the items are related to overall patient satisfaction. To analyse differential effects, the BH method is applied to model (2) finding that in 5 of those items, none of the interactions is significant. On the other hand, the proposed method (3) indicates that all of the 13 items revealed by model (1) have an important effect in at least four of the seven countries.

References

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

Gene regulatory network reconstruction with prior knowledge over mRNA data for COPD patients and controls

Victor Bernal [12], Corry A. Brandsma[3], Alen Faiz[5], Victor Guryev[4], Wim Timens[3], Maarten van den Berge[5], Rainer Bischoff[2], Marco Grzegorczyk [1], Peter Horvatovich[2]

[1] Johann Bernoulli Institute (JBI), Groningen Rijksuniversiteit, Groningen, NL.
[2] Department of Pharmacy, Analytical Biochemistry, Rijksuniversiteit Groningen, NL.
[3] Universitair Medisch Centrum Groningen (UMCG), Department of Pathology and Medical Biology, Rijksuniversiteit Groningen, NL.
[4] Universitair Medisch Centrum Groningen (UMCG), ERIBA, Rijksuniversiteit Groningen, NL.
[5] Universitair Medisch Centrum Groningen (UMCG), Department of Pulmonary Diseases, Rijksuniversiteit Groningen, NL.

E-mail for correspondence: `v.a.bernal.arzola@rug.nl`

Abstract: Chronic obstructive pulmonary disease (COPD) is a type of lung disease characterized by persistent bronchitis and emphysema. Current therapy is restricted to alleviate lung tissue inflammation, but is not able to stabilize or improve lung function of patients making necessary to understand the underlying molecular mechanisms of COPD. Genome-wide gene expression of lung tissue provides a powerful tool to elucidate molecular mechanism of COPD patients. In particular, Bayesian Networks (BNs) have been applied to infer genetic regulatory interactions from microarray gene expression data. In this study we aim obtain a clearer understanding of the genes interaction in COPD patients by learning a BN over microarray expression data. A subset of genes was selected for the study fulfilling that i) the genes were significantly expressed in COPD stage 4 and ii) there is reported gene-gene experimental association. The reported associations are introduced as prior biological knowledge in the reconstruction.

Keywords: COPD; Bayesian Networks; Microarray Gene Expression; Introduction of prior knowledge.

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 Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by distinct phenotypes as, emphysema, chronic bronchitis and fibrosis. The exact nature of the abnormal lung tissue repair processes in COPD lungs is unknown, hence a better understanding the underlying molecular mechanisms is thus crucial. Bayesian Networks (BNs) have been applied to infer genetic regulatory interactions from microarray gene expression data. This inference problem is particularly hard because of the relatively small size of the data sets. Moreover the asessment of the inference for COPD results is unfeasible as there is a lack of known gold standards from the biological literature. The aim of the present study is the reconstruction of a regulatory network from a microarray gene expression data set for COPD (stage 4) patients using BNs with prior biological knowledge coming from STRING (database of known and predicted protein-protein interactions). Inference is done using with Markov Chain Monte Carlo (MCMC) sampling following the approach of Werhli et al. (2007). Two networks are obtained and compared, one for the controls and one for the COPD data.

2 Bayesian Networks with prior knowledge

Bayesian Networks (BNs) represent probabilistic relationships between interacting agents (e.g. genes) in the form of conditional independence. For BNs the model is a Directed Acyclic Graph (DAG), a set of nodes that are connected by (directed) edges without cycles. We denote by G the network structure, which is the set of nodes and edges, and by q the interaction strength. Our objective is to learn the network structure G directly from the scratch microarray data D. Nodes are associated with genes and the edges indicate interactions between the genes. A Bayesian learning approach uses the posterior $P(G|D)$

$$P(G|D) = P(G|D)P(G)/P(D)$$

where $P(G)$ is called the prior over graphs, $P(D|G)$ the marginal likelihood, $P(D)$ is a scaling constant and $P(G|D)$ is called the posterior probability. The term $P(D|G)$ is the result of integrating the likelihood $P(D, q|G)$ over all the possible values of the parameters q. Calculating $P(G|D)$ for all possible network structure G is computationally impossible due to $P(D)$ is intractable. Gene expression data implies a large number of genes with only a relative small number of measurements. As a consequence the posterior $P(G|D)$ will not have a clear maximum G^* . A solution is to use a MCMC sampling based on Metropolis Hasting algorithm (MH). This produces a Markov Chain that reaches $P(G|D)$ as its stationary distribution (under some conditions). Generally different DAGs can represent the same probabilistic relationships (i.e they can be equivalent). The set of equivalent DAGs is called the equivalence class and is represented by a CPDAG.

Introducing prior knowledge for learning BNs is done following Werhli et al. (2007). The prior knowledge is encoded in a matrix B whose entries $B_{ij} \in [0, 1]$ represent the belief about the presence/absence of an edge. The elements B_{ij} greater than/less than 0.5 reflects evidence in favor/against the presence of the edge G_{ij}, and $B_{ij} = 0.5$ as uninformative. The prior over graphs takes the form of a Gibbs distribution,

$$P(G|\beta) = e^{-\beta E(G)}/Z(\beta)$$

The function $E(G)$ measures the agreement between a sampled network G and B (e.g. using the Hamming distance). The parameter β indicates the weight of the respective source of knowledge to the data, and $Z(\beta)$ is a normalization constant.

3 Evaluation of the COPD dataset and Implementation

The data consist of genome-wide gene expression profiling in lung tissue samples of 72 subjects, where 48 exhibited COPD (stage 4). The others 24 are non-COPD tissues (Controls). The 60 most significant genes based on p-value and $log_2(FC) > 2$ were checked in STRING data base, obtaining that 7 genes (FGG, RORC, FGA, OSMR, NR1D1, CSF3, THBS1) had reported experimental and/or data base associations. This set of genes is augmented as follows: i) the Pearson's correlation matrix ρ is built for Controls and COPD, and ii) a pair of genes is selected if $|\rho_{controls}| < 0.25$ and $|\rho_{COPD}| > 0.75$. In this way 8 additional genes are obtained, namely MT1M, IL1RL1, MT1P3, MT1A, SLCO4A, GPA33, TTN, ZBED2. In total 2 networks were reconstructed for the forementioned genes i) Controls with prior knowledge, and ii) COPD with prior knowledge. The coupling parameter β is defined in $[0, 30]$. The MCMC was implemented with the REV move from Grzegorczyk (2008) for number of 9x10^5 iterations, with thining every 1000 DAGs. The first half of the sample was discarded as part of the burn in phase and Model Averaging was performed over the final set of 501 CPDAGs.

4 Results and Conclusions

After the burn in phase the 75th percentile of β (in Controls and COPD cases) is less than 1. This suggest that COPD's gene regulation has no apparent commonalities with previously reported interaction. Table 1 shows a summary of the interactions that considerably changed between Controls and COPD. In particular interactions like FGG-FGA, and MT1P3-MT1A are reasonable as they belong to the same family. On the other hand NR1D1 has no a well known association with fibrogenes, this interaction (FGG-NR1D1) is present in Controls (mediated by FGA) and COPD (directly). Our next research plan is to extend the gene set. We will provide more precise biological interpretations in the presentation.

References

Grzegorczyk, M., Husmeier, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, **71**, 265 – 305.

Werhli A., Husmeier, D. (2007). Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge.*Stat Appl Genet Mol Biol*, **6**(1), 1 – 45.

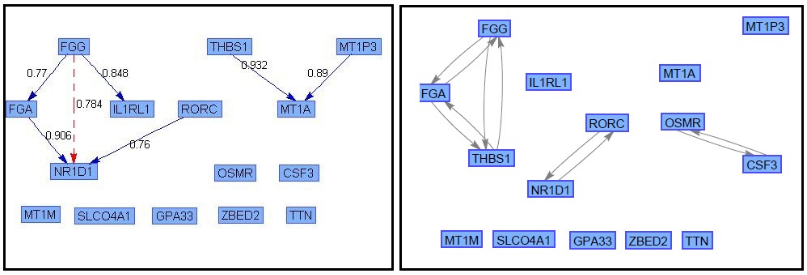


FIGURE 1. **Reconstructed regulatory network for Control and COPD.** On the right: Network from STRING with experimental and database associations. Doble arrows stand for reported associations. On the left: Reconstructed network. The blue edges stand for gene interactions in the Controls. The segmented red edges stand for gene interactions in the COPD. There are no common edges. Only edges with a frequency above 0.75 are shown.

TABLE 1. **Marginal edge frequency changes** This table presents a set of relevant gene-gene interaction whose edge frequency had a notable change between the network for the Controls and the one for COPD.

	Controls	COPD
FGG-FGA	0.770	0.314
FGA-NR1D1	0.906	0.078
MT1P3-MT1A	0.890	0.614
THBS1-MT1A	0.932	0.502
RORC-NR1D1	0.760	0.114

Resolving the Lord's Paradox

Priyantha Wijayatunga[1]

[1] Department of Statistics, Umeå University, Umeå SE-90187, Sweden

E-mail for correspondence: priyantha.wijayatunga@umu.se

Abstract: An explanation to Lord's paradox using ordinary least square regression models is given. It is not a paradox at all, if the regression parameters are interpreted as predictive or as causal with stricter conditions and be aware of laws of averages. We use derivation of a super-model from a given sub-model, when its residuals can be modelled with other potential predictors as a solution.

Keywords: Effect; Predictive; Causal; Confounding.

1 Introduction

In 1967 Frederic Lord posed following question (see Lord 1967 and Pearl 2016) that became a paradox among applied statistical community. To see effects and if there is any sex difference of diet provided in a university weights of students at time of their arrival and those a year later are recorded. The data are independently examined by two statisticians. The first examines the mean weight of the girls at the beginning and at the end of the year, and finds that they are to be identical, i.e., frequency distribution of the weight for the girls is not changed, so is for the boys. The second statistician finds that the slope of the regression line of the final weight on the initial weight is essentially the same for both sexes but the regression coefficient of the variable sex to be statistically significant and concludes that the boys showed significantly more gain in weight than the girls when proper allowance is made for differences for initial weight.

Conclusions of the two statisticians seem to contradict with each other; the first is predictive and the second is both predictive, and causal if the initial weight is the only confounder of causal relation between the sex and the final weight. The second has given causal effect of the sex on the final weight (weight gain) by a regression coefficient. In fact, to give it by comparing, two supports of the confounder of both sexes should coincide. But one can assume that the population initial weight ranges of boys and girls coincide even though sample counterparts differ (so, extrapolation is meaningful).

Let the initial weight, final weight and sex are denoted by W_I, W_F, S respectively ($S = 0$, a girl and $S = 1$, a boy) and weight gain be $D = W_F - W_I$. If the effect 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 copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

S on D is found by difference of conditional means, $E\{D|S=1\}-E\{D|S=0\}$ then it is no effect. This can be found by running regression of D on S. Note that $E\{W_F|S=1\}=E\{W_I|S=1\}$, (say, μ_B) and $E\{W_F|S=0\}=E\{W_I|S=0\}$, (say, μ_G). If $E\{D|W_I=i,S=1\}$ and $E\{D|W_I=i,S=0\}$ are calculated simply by partitioning the data by taking W_I to be discrete or as a functions of i, then the difference $E\{D|W_I=i,S=1\}-E\{D|W_I=i,S=0\}$ may not be zero for each i, so may be difference of their weighted means, $\sum_i E\{D|W_I=i,S=1\}p(W_I=i)-\sum_i E\{D|W_I=i,S=0\}p(W_I=i)$. If the effect of S on W_F is calculated by it then it is different from former value (paradoxical!).
Now let us see why two types of differences of averages differ by simple algebra, that will say that they should have two different interpretations. First assume that we have a number of subgroups of boys and, for simplicity, the same is true for girls. Let D^1_{ij} be the weight gain of the j-th boy in the i-th subgroup of boys where sub-group size is n_i and D^0_{ij} be that of the girls where sub-group size is m_i and furthermore, let $f^1_i=n_i/\sum_k n_k$, $f^0_i=m_i/\sum_k m_k$ and $f_i=(n_i+m_i)/\sum_k(n_k+m_k)$ for $j=1,...,n_i$ and $i=1,...,a$. And let A_1 be difference of the average weight gain of the boys and the girls, $\bar{D}^1_i=\sum_j D^1_{ij}/n_i$ and $\bar{D}^0_i=\sum_j D^0_{ij}/m_i$ for $i=1,...,a$. So,

$$\begin{aligned}A_1 &= \frac{\sum_{i=1}^a\sum_{j=1}^{n_i}D^1_{ij}}{\sum_{i=1}^a n_i}-\frac{\sum_{i=1}^a\sum_{j=1}^{m_i}D^0_{ij}}{\sum_{i=1}^a m_i}=\sum_{i=1}^a\bar{D}^1_i f^1_i-\sum_{i=1}^a\bar{D}^0_i f^0_i\\ &\neq \frac{1}{2}\Big\{\sum_{i=1}^a\bar{D}^1_i f^1_i+\sum_{i=1}^a\bar{D}^1_i f^0_i-\sum_{i=1}^a\bar{D}^0_i f^0_i-\sum_{i=1}^a\bar{D}^0_i f^1_i\Big\};\text{ generally}\\ &= \frac{1}{2}\Big\{\sum_{i=1}^a(\bar{D}^1_i-\bar{D}^0_i)(f^1_i+f^0_i)\Big\}\neq\sum_{i=1}^a(\bar{D}^1_i-\bar{D}^0_i)f_i=A_2\end{aligned}$$

where $f_i=\alpha f^1_i+(1-\alpha)f^0_i$ for $i=1,...,a$ such that $\alpha=\sum_{i=1}^a n_i/\sum_{i=1}^a(n_i+m_i)$ and A_2 is the difference of weighted averages of the sub-group weight gain averages. So, the difference of group averages A_1 (which is zero in our case) is different from the difference of pooled-weighted average of the sub-group averages A_2. The second statistician compares the boys and the girls subgroup-wise and finds that it is a constant gain for the boys over the girls across the subgroups, i.e., $\bar{D}^1_i-\bar{D}^0_i$ is constant for all initial weight i. Therefore he finds that the boys gain more weight than the girls in corresponding sub-groups. Note that for simplicity we have taken initial weights as discrete values. In fact, $A_2=\sum_i E\{D|S=1,W_I=i\}p(W_I=i)-\sum_i E\{D|S=0,W_I=i\}p(W_I=i)$ is the causal effect of S on D if W_I is the only confounder, under the linear assumption. It is different from A_1 unless $E\{D|S,W_I\}=E\{D|S\}$. The confounding effect (A_2-A_1) depends on how different f^1 and f^0 are (can have a measure from them).

2 Regression Solution

Now we define interpretation of ordinary least square (OLS) estimates of the regression coefficients (parameters). The OLS estimation is based on the variation of the response variable Y for a given functional form of the values of explanatory factors. Regression coefficients are estimated so that sum of squared prediction errors for the data in the sample is the minimum. So, reverse regression is not generally obtainable from forward regression and may not be consistent with the latter. For simple linear regression one can easily establish that the reverse regression and the forward regression are consistent with each other if and only if one of the regressions have symmetric residuals about and uni-modal at conditional expectation of response, that implies other regression too.

Now consider the OLS linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, then linear effect of X_1 on Y when X_2 is held unchanged is given by β_1 if Y values are symmetric about and uni-modal at $\beta_0 + \beta_1 X_1 + \beta_2 X_2$. It is clear that the supports of X_2 for each value of X_1 are the same (or extrapolation is meaningful if empirical supports differ). Symmetry and uni-modality of Y values for given values of X_1 and X_2 are observed if all other factors that affect or are associated, but are not taken into consideration are allowed to vary pure randomly. This is a fundamental assumption used in statistical modelling often implicitly.

Let us do a regression of W_F on the binary variable S. Then we get the model $W_F = \mu_G + (\mu_B - \mu_G)S + \epsilon_1$. where the regression co-efficient of S is the predictive effect of S on W_F provided that above requirement is fulfilled. The residuals of the model are just individual values of D, i.e., $\epsilon_1 = D$ for each subject and it is easy to see in Fig. 1 of Lord 1967, that the residuals are predictive by W_I for each sex category separately, $\epsilon_1 \not\perp W_I|S$. However, it may be that $\epsilon_1 \perp S$. So, if the two clusters of values of W_F for two sexes are symmetric about and uni-modal at the respective means then the effect of S on W_F is the regression coefficient of S in the model. But it is uncontrolled confounders that are associated with W_F, then it should be interpreted accordingly. That is, it is the predictive effect of sex differences and causal if there are no confounders such as W_I. And we see that we get zero predictive effect from the meal change since the regression coefficient is the same as that when the girls and boys had previous meal type.

Let we can write the distribution of residuals for each value s of S, say, $f(\epsilon_1|s)$ as a mixture, $f(\epsilon_1|s) = \int g(\epsilon_1|x,s)\pi(x,s)dx$ for some random variable X, and for each value x of X the component distribution $g(\epsilon_1|x,s)$ may have non-zero mean such that $E\{\epsilon_1|s\} = \int E\{\epsilon_1|x,s\}\pi(x,s)dx = 0$ and then we have that $Var\{\epsilon_1|x,s\} \leq Var\{\epsilon_1|s\}$ where $\pi(x,s) = h(x|s)p(s)$; here $h(x|s)$ is the conditional probability density of X given $S = s$ and $p(s)$ is the marginal probability distribution of S. If X could be identified meaningfully, then model should include such feature variables too. In this case, X could be identified as the initial weight W_I. Then one should accept the upgraded model that includes W_I too. It has residuals that have a smaller conditional standard deviation given W_I and S. Furthermore, if W_I is the only confounding factor and when it is also included in the model the the coefficient of S is the causal effect of S on W_F.

Let the residual ϵ_1' corresponds to the context that $W_I = w_I$ and $S = s$ and then it can be written as $\epsilon_1' = \mu_{\epsilon_1}^{w_I,s} + \epsilon_2$ where $\mu_{\epsilon_1}^{w_I,s}$ is the expectation of it. So, we have $E\{\epsilon_2|W_I = w_I, S = s\} = 0$ and also that $Var\{\epsilon_2|W_I = w_I, S = s\} \leq Var\{\epsilon_1|S = s\}$. And furthermore, we can have that $\mu_{\epsilon_1}^{w_I,0} = a_0 + b_0 w_I$ for $s = 0$ and $\mu_{\epsilon_1}^{w_I,1} = a_1 + b_0 w_I$ for $s = 1$ where a_0, b_0 and a_1 are constants. Now, given that $W_I = w_I$ and $S = s$, for $s = 0, 1$, and $I(A) = 1$ when A is a true statement and $I(A) = 0$ otherwise, we have

$$\begin{aligned}
W_F &= W_F = \mu_G + (\mu_B - \mu_G)s + \mu_{\epsilon_1}^{w_I, s} + \epsilon_2 \\
&= \mu_G + (\mu_B - \mu_G)s + (a_0 + b_0 w_I)I(S = 0) + (a_1 + b_0 w_I)I(S = 1) + \epsilon_2 \\
&= \mu_G + (\mu_B - \mu_G)s + a_0 I(S = 0) + a_1 I(S = 1) + b_0 w_I + \epsilon_2 \\
&= \mu_G + (\mu_B - \mu_G)s + a_0(1 - s) + a_1 s + b_0 w_I + \epsilon_2 \\
&= \mu_G + a_0 + (\mu_B - \mu_G - a_0 + a_1)s + b_0 w_I + \epsilon_2
\end{aligned}$$

So we can obtain a super-model (regression) from a given regression model (it is a sub-model of the former) as long as its residuals are predictive (linearly in this case) with another explanatory variable. The predictive effect of S on W_F when controlled for W_I is $\mu_B - \mu_G - a_0 + a_1$ that is generally different from earlier value of $\mu_B - \mu_G$ and for each individual model prediction is more accurate than that of the previous model, therefore new model is preferred to the previous one. If W_I is only a confounder but not an intermediate variable between the causal pathway between S and W_F, and has a common support for all values of S, then β_1 is the average causal effect of S on W_F in the linear case. In our example, sample supports of W_I for $S = 1$ and $S = 0$ differ but we can assume that they are the same in the population (so, extrapolation is meaningful). Note that the above arguments can be generalised. For restrictions of space, we avoid presenting solution to the paradox, that is based on causal diagrams. We object recent solution by Pearl. Our explanations comply with Lord's initial comments.

References

Lord, F. M. (1967). A Paradox in the Interpretation of Group Comparisons. *Psychological Bulletin*, **68**(5), 304–305.

Pearl, J. (2016). Lord's Paradox Revisted - (Oh Lord Kumbaya!). *Journal of Causal Inference*, **4**(2). DOI: 10.1515/jci-2016-0021.

On a mixture regression model for left-censored data

Mário F. Desousa[1], Helton Saulo[2], Manoel Santos-Neto[3], Víctor Leiva[4]

[1] Department of Statistics, Universidade Estadual de Campinas, Brazil
[2] Department of Statistics, Universidade de Brasília, Brazil
[3] Department of Statistics, Universidade Federal de Campina Grande, Brazil
[4] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Chile

E-mail for correspondence: heltonsaulo@gmail.com

Abstract: Most of the models designed for modeling censored data rely on the assumption of normality for the error distribution. It is well known that not all applications are well modeled by this distribution. Some efforts have relaxed the normality assumption by considering more flexible distributions such as t and log-alpha-power. Nevertheless, these models do not consider partial observations from the assumed distribution which potentially leads to biased inference. We have explored a real data example of measles vaccine in Haiti and confirmed both the possibility of partial observation and asymmetry problems. Then, to solve such problems, we propose a mixture model consisting of the Birnbaum-Saunders and Bernoulli distributions. We discuss estimation of the model parameters based on the maximum likelihood method. We then carry out a Monte Carlo simulation study to evaluate the performance of the maximum likelihood estimators. We use the R software in all computations and the results favor the proposed methodology.

Keywords: Mixture model; Censoring; Birnbaum-Saunders Distribution.

1 Introduction

The determination of antibody concentration by quantitative assays is a very important topic of research, because there is always a concentration value ($\mathcal{T}$) below which an exact measurement cannot be obtained regardless of the employed technique. Nevertheless, this antibody concentration value ($\mathcal{T}$) is a function of the associated assay. When left-censoring is present in data from an assay, 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

lower detection limit (LDL) can be used to substitute a value for the censored observation, namely, the value $\mathcal{T}$. In special, this substitution is applied in a safety and immunogenecity study of measles vaccine in Haiti presented by Moulton and Halsey (1995), an example explored in this paper.

A two-part model for the situation of zero excess was considered by Cragg (1971). The Cragg model considers the possibility of having observations from the assumed distribution f and from the point mass distribution. In this model, the log-normal distribution was considered for the independent variable. The Cragg model, however, does not consider the existence of both a lower limit and some observations below this limit. Moulton and Halsey (1995) proposed a straightforward generalization of the two-part model, called Bernoulli/lognormal model, by considering the possibility of some limiting responses resulting from interval censoring associated with f. The generalized two-part model allows the possibility of a observation i, if located below $\mathcal{T}$ to be either a partial observation from f or a realization of the point mass distribution.

The main objective of this paper is to propose a regression model for censored data based on the mixture between the Birnbaum-Saunders (BS) and Bernoulli distributions, that is, a censored continuous distribution and a point mass distribution located below the detection limit. The Birnbaum-Saunders (BS) distribution is positively skewed and has a failure rate with upside-down bathtub shape and a close relation with the normal distribution; see Birnbaum and Saunders (1969) and Johnson et al. (1995). The proposed model extends to the BS case the Moulton and Halsey (1995)'s Bernoulli/lognormal model.

2 The Bernoulli/BS mixture model

2.1 Formulation

We propose a mixture model between the Bernoulli and BS distributions (Bernoulli/BS), that is,

$$g(y_i) = \left[p + (1-p)\,\Phi\left(\zeta^{\mathrm{c}}_{i2}\right)\right] \mathrm{I_i} + (1-p)\left[\frac{\mathrm{c}_1}{\alpha}\cosh\left(\frac{y-\mu_1}{2}\right)\exp\left(-\frac{2}{\alpha^2}\sinh^2\left(\frac{y-\mu_1}{2}\right)\right)\right](1-\mathrm{I_i}), \tag{1}$$

where $\mathrm{c}_1 = 1/\sqrt{2\pi}$, $\mu_1 = \boldsymbol{x}_{(1)}^{\top}\boldsymbol{\beta}_{(1)}$, $\zeta^{\mathrm{c}}_{i2} = (2/\alpha)\sinh\left((y_0 - \boldsymbol{x}_{(1)}^{\top}\boldsymbol{\beta}_{(1)})/2\right)$,

$$\mathrm{I}_i = \begin{cases} 1, & \text{if } y \le y_0, \\ 0, & \text{if } y > y_0, \end{cases} \tag{2}$$

and $\Phi(\cdot)$ is the CDF of the standard normal distribution, with $\boldsymbol{x}_{(1)}$ being the covariates associated with $\boldsymbol{\beta}_{(1)}$. We assume a logit link for the random variable D, thus it is possible to include covariates as follows

$$\operatorname{logit}\left[\mathbb{P}\left(D = 1 | \boldsymbol{x}_{(2)}\right)\right] = {\boldsymbol{x}_{(2)}}^{\top}\boldsymbol{\beta}_{(2)}, \tag{3}$$

where $\boldsymbol{x}_{(2)}$ are the covariates associated with $\boldsymbol{\beta}_{(2)}$. The formulation of the logit link defined in Equation (3) becomes,

$$\tau_i = 1 - p_i = \frac{\exp\left({\boldsymbol{x}_{(2)}}^{\top}\boldsymbol{\beta}_{(2)}\right)}{1 + \exp\left({\boldsymbol{x}_{(2)}}^{\top}\boldsymbol{\beta}_{(2)}\right)}, \tag{4}$$

where the vector $\boldsymbol{x}_{(2)}$ has a dimension q, which can be, and usually is, different from the dimension of the vector $\boldsymbol{x}_{(1)}$.

The log-likelihood function for the mixture Bernoulli/BS is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}) = -(n-m)\log(2) - (n-m)\tfrac{\log(2\pi)}{2} + \qquad (5)\\ \textstyle\sum_{i=1}^{m} \mathrm{I}_\mathrm{i} \left[\log\left(1+\exp(\boldsymbol{x}_{(2)}^{\top}\boldsymbol{\beta}_{(2)})\left[\Phi\left(\zeta_{i2}^{\mathrm{c}}\right)-1\right]\right) - \log\left(1+\exp\left(\boldsymbol{x}_{(2)}^{\top}\boldsymbol{\beta}_{(2)}\right)\right)\right]\\ +\textstyle\sum_{m+1}^{n}\left(1-\mathrm{I}_\mathrm{i}\right)\left[\boldsymbol{x}_{(2)}^{\top}\boldsymbol{\beta}_{(2)} + \log\left(\zeta_{i1}\right) - \tfrac{1}{2}\zeta_{i2}^{2} - \log\left(1+\exp(\boldsymbol{x}_{(2)}^{\top}\boldsymbol{\beta}_{(2)})\right)\right],\end{aligned}$$

where ζ_{i2}^{c} is defined in Equation (1) and ζ_{i1} and ζ_{i2} are defined in Equation (6),

$$\zeta_{i1} = \frac{2}{\alpha}\cosh\left(\frac{y_i - \boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)}}{2}\right), \quad \zeta_{i2} = \frac{2}{\alpha}\sinh\left(\frac{y_i - \boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)}}{2}\right). \qquad (6)$$

3 Application

We analyse a data set provided by Moulton and Halsey (1995) from a study of measles vaccines. Neutralization antibody levels were collected from 330 children at 12 months of age. The LDL was 0.1 international units (IU) or −2.306 in logarithm scale. Around 86 (26.1%) of the observations fell below the LDL and then recorded as 0.1. The following covariates were considered: X_1 indicates the type of vaccine used (0 if Schwartz and 1 if Edmonston-Zagreb); X_2 is the level of the dosage (0 if medium and 1 if high); and X_3 is the gender where 0 is male and 1 is female.

We here present the estimation results for the Bernoulli/BS model along with those of the standard tobit, tobit-BS (Chapter 2) and Bernoulli/LPN Martínez-Flórez et al. (2013) models. The Bernoulli/BS and Bernoulli/LPN have both a logit link. The covariates EZ and HI were used only in the logit component, and covariate FEM entered only in the continuous component of the models. Table 1 shows the ML estimates, Akaike information criterion (AIC) values and standard errors for the considered models. A glance at the results indicates that, in the Bernoulli/BS model, the receiver of Edmonston-Zagreb strain does not contribute to the odds ratio of being above the detection limit, however, the receiver of a high dose impacts $\exp(1.499) = 4.472$ in the odds of being above the detection limit. Moreover, the Bernoulli/BS model suggests that girls have $\exp(-0.078) - 1 = -0.075$ less concentration of measles antibody concentration than boys. We observe that the Bernoulli/LPN and Bernoulli/BS models do not agree on the sign of the coefficient corresponding to the FEM variable, while the first model indicates that girls have a higher measles antibody concentration than boys, the other indicates the opposite. Also from Table 1, we note that the Bernoulli/BS model provides a better fit compared to the other models based on the AIC values.

Acknowledgments: This study was partially supported by CAPES from the Brazilian government.

References

Birnbaum, Z.W., and Saunders, S.C. (1969). A new family of life distributions. *Journal of Applied Probability*, **6**, 319 − 327.

TABLE 1. ML estimates (with SE in parentheses) and AIC values for the indicated models with the measles vaccine data

			Continuous component			
Model	AIC	α	INT	EZ	HI	FEM
tobit	1299.27	0.945***	0.597**	0.225	−0.228	0.271
		(0.047)	(0.288)	(0.297)	(0.295)	(0.296)
tobit-BS	1168.60	1.545***	−0.910***	0.188*	0.074	0.121
		(0.048)	(0.105)	(0.111)	(0.109)	(0.110)
Bernoulli/LPN	976.48	8.918**	−2.869***			0.222*
		(3.922)	(0.582)			(0.134)
Bernoulli/BS	760.64	1.560***	0.123***			−0.078***
		(0.108)	(<0.001)			(<0.001)

Obs: Rejects H_0 at *10% of significance,** 5% of significance and ***1% of significance.

Cragg, J.G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, **39**, 829 – 844.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions.* Volume 2. New York: Wiley.

Martínez-Flórez, G., Bolfarine, H., and Gómez, H.W. (2013). Asymmetric regression models with limited responses with an application to antibody response to vaccine. *Biometrical Journal*, **55**, 156 – 172.

Moulton, L.H., and Halsey, N.A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, **51**, 1570 – 1578.

On asymmetry models and decompositions of symmetry for square contingency tables with ordered categories

Kouji Tahata[1]

[1] Tokyo University of Science, Japan

E-mail for correspondence: kouji_tahata@is.noda.tus.ac.jp

Abstract: To analyze square contingency tables with ordered categories, herein a model that indicates the structure of asymmetry for cell probabilities is proposed. This model is the closest to the symmetry model in terms of the f-divergence under certain conditions and incorporates various types of asymmetry. It is shown that the symmetry model can be separated into the proposed model and other models. This may reveal the origin for the poor fit of the symmetry model when it occurs for a real dataset.

Keywords: f-divergence, moment equality, quasi-symmetry.

1 Introduction

To analyze square contingency tables with ordered categories, an issue of symmetry (rather than independence) arises naturally. So, we are interested in considering a structure of symmetry. Many other statisticians, including Kateri and Papaioannou (1997), Kateri and Agresti (2007), and Tahata and Tomizawa (2011), have proposed various models of symmetry and asymmetry. In the present paper, we propose a model that is the closest to the symmetry model in terms of the f-divergence (Csiszár and Shields, 2004) under certain conditions. Our model is a generalization of the symmetry model in the sense that it includes the various types of asymmetry models.

Caussinus (1965) proved the theorem that the symmetry model holds if and only if both the quasi symmetry model and the marginal homogeneity model hold. When the symmetry model does not fit for a real dataset, the separation of symmetry may be useful to identify the origin for the poor fit of symmetry. For example, Kateri and Papaioannou (1997), Tahata and Tomizawa (2011), and Saigusa, Tahata and Tomizawa (2015) have given the separation of the symmetry model. In the present paper, we show that the symmetry model can be separated into some models. One is the proposed model and others are concerned with the equality of marginal moments. The results include the separation of the symmetry model given by Caussinus (1965) as a special case.

The present paper is organized as follows. Section 2 proposes the new model and gives the separation of the symmetry model. Section 3 provides a numerical example.

2 Asymmetry models

Consider an $r \times r$ square contingency table with ordered categories. Let π_{ij} denote the probability that an observation will fall in the (i, j)th cell of the contingency table $(i = 1, \ldots, r; j = 1, \ldots, r)$. Let $\pi = (\pi_{ij})$ and $q = (q_{ij})$ be two bivariate probability distributions. The f-divergence between π and q is defined as

$$I^C(\pi : q) = \sum_i \sum_j q_{ij} f\left(\frac{\pi_{ij}}{q_{ij}}\right),$$

where f is a convex function on $(0, +\infty)$ with $f(1) = 0$. Also, we take $f(0) = \lim_{t \to 0} f(t)$, $0 \cdot f(0/0) = 0$, and $0 \cdot f(a/0) = a \lim_{t \to \infty}[f(t)/t]$ (Csiszár and Shields, 2004).

Let $\{u_i\}$ be a set of known scores $u_1 \leq u_2 \leq \cdots \leq u_r$ (with $u_1 < u_r$). In an analogous manner to Kateri and Papaioannou (1997), for a given k $(k = 1, \ldots, r-1)$ we propose that the asymmetry based on the f-divergence (AS$_k[f]$) model is defined as

$$\pi_{ij} = \pi_{ij}^S F^{-1}\left(\sum_{h=1}^{k} u_i^h \alpha_h + \gamma_{ij}\right) \quad (i = 1, \ldots, r; j = 1, \ldots, r),$$

where $\gamma_{ij} = \gamma_{ji}$, $\pi_{ij}^S = (\pi_{ij} + \pi_{ji})/2$, f is a twice-differential and strictly convex function, and $F(t) = f'(t)$. From the relation $\pi_{ij}^S = (\pi_{ij} + \pi_{ji})/2$, the parameters of the AS$_k[f]$ model must satisfy

$$F^{-1}\left(\sum_{h=1}^{k} u_i^h \alpha_h + \gamma_{ij}\right) + F^{-1}\left(\sum_{h=1}^{k} u_j^h \alpha_h + \gamma_{ji}\right) = 2.$$

Note that when $\alpha_1 = \cdots = \alpha_k = 0$, (i) the AS$_k[f]$ model is reduced to the symmetry (S) model (Bowker, 1948), (ii) the AS$_1[f]$ model is reduced to the OQS$[f]$ model (Kateri and Agresti, 2007) and (iii) the AS$_{r-1}[f]$ model with $u_1 < u_2 < \cdots < u_r$ is reduced to the QS$[f]$ model (Kateri and Papaioannou, 1997). The AS$_k[f]$ model is the closest model to the S model in terms of the f-divergence under the condition that $\sum_i \sum_j u_i^h \pi_{ij}$ (or $\sum_i \sum_j u_j^h \pi_{ij}$) for $h = 1, \ldots, k$ as well as the sums $\pi_{ij} + \pi_{ji}$ for $i = 1, \ldots, r; j = 1, \ldots, r$ are given.

When we set $f(t) = t \log(t)$, $t > 0$, the AS$_k[f]$ model can be expressed as

$$\frac{\pi_{ij}}{\pi_{ji}} = \prod_{h=1}^{k} \beta_h^{u_i^h - u_j^h},$$

where $\beta_h = \exp[\alpha_h]$. Above equation with $\{u_i = i\}$ is the k-th linear asymmetry model proposed by Tahata and Tomizawa (2011).

Let X and Y denote the row and column variables, respectively. Assume that a set of known scores $\{u_s\}$ can be assigned to both the rows and the columns. Let $X_1 = u_i$ when $X = i$ $(i = 1, \dots, r)$, and $X_2 = u_j$ when $Y = j$ $(j = 1, \dots, r)$. For a given positive integer k $(k = 1, \dots, r-1)$, consider a model defined by

$$E(X_1^h) = E(X_2^h) \quad (h = 1, \dots, k),$$

where $E(X_1^h) = \sum_{s=1}^{r} \sum_{t=1}^{r} u_s^h \pi_{st}$, $E(X_2^h) = \sum_{s=1}^{r} \sum_{t=1}^{r} u_t^h \pi_{st}$. For a given k, we shall refer to this model as the marginal kth moment equality model for the scores $\{u_s\}$ (denoted by ME$_k$).
This leads to the following theorem.

Theorem 1. For a fixed k $(k = 1, \dots, r-1)$, the S model holds if and only if both the AS$_k[f]$ and ME$_k$ models hold.

Assume that a set of known scores $\{u_i = i\}$ is assigned. We now obtain that the test statistic for the S model is asymptotically equivalent to the sum of those for the AS$_k[f]$ model and the ME$_k$ model under the S model.

3 Example

Table 1, which is taken directly from Stuart (1955), is the cross-classification of the unaided distance vision of 7477 women aged 30–39 employed in Royal Ordnance factories in Britain from 1943 to 1946. The S model fits the data poorly, yielding the likelihood ratio statistic $G^2 = 19.249$ with 6 degrees of freedom (df).

TABLE 1. Unaided distance vision of 7477 women aged 30–39 employed in Royal Ordnance factories in Britain from 1943 to 1946; from Stuart (1955).

Right eye	Left eye grade				
grade	Highest (1)	Second (2)	Third (3)	Lowest (4)	Total
Highest (1)	1520	266	124	66	1976
Second (2)	234	1512	432	78	2256
Third (3)	117	362	1772	205	2456
Lowest (4)	36	82	179	492	789
Total	1907	2222	2507	841	7477

We set $f(t) = (1-t)^2$ and used the integer scores $\{u_i = i\}$. Consider the hypothesis that the AS$_2[f]$ model holds under the assumption that the AS$_3[f]$ model holds; namely, the hypothesis that $\alpha_3 = 0$. According to the test based on the difference between the G^2 values for the AS$_2[f]$ and AS$_3[f]$ models, the hypothesis is accepted at the 0.05 level because $7.267 - 7.262 = 0.005$ with 1 df. In a similar manner, the hypothesis that the AS$_1[f]$ model holds under the assumption that the AS$_2[f]$ model holds is accepted at the 0.05 level because $7.271 - 7.267 = 0.004$ with 1 df. Therefore, the AS$_1[f]$ model may be preferable to the AS$_2[f]$ and AS$_3[f]$ models. The AS$_1[f]$ model fits the data well. On the other hand, the ME$_1$ model fits the data poorly, yielding $G^2 = 11.978$ with 1 df. From Theorem 1 with $k = 1$, the poor fit of the S model is due to the lack of structure of the ME$_1$ model (rather than the AS$_1[f]$ model).

References

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, **43**, 572 – 574.

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des Sciences de l'Université de Toulouse*, Série 4, **29**, 77 – 182.

Csiszár, I. and Shields, P. C. (2004). *Information Theory and Statistics: A Tutorial.* Now Publishers, Hanover, Massachusetts.

Kateri, M. and Agresti, A. (2007). A class of ordinal quasi-symmetry models for square contingency tables. *Statistics and Probability Letters*, **77**, 598 – 603.

Kateri, M. and Papaioannou, T. (1997). Asymmetry models for contingency tables. *Journal of the American Statistical Association*, **92**, 1124 – 1131.

Saigusa, Y., Tahata, K. and Tomizawa, S. (2015). Orthogonal decomposition of symmetry model using the ordinal quasi-symmetry model based on f-divergence for square contingency tables. *Statistics and Probability Letters*, **101**, 33 – 37.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412 – 416.

Tahata, K. and Tomizawa, S. (2011). Generalized linear asymmetry model and decomposition of symmetry for multiway contingency tables. *Biometrics and Biostatistics*, **2**, 1 – 6.

Analysis of the t linear mixed model and evaluation of a direct estimation method

M. Regis[1], N. Nooraee[1], A. Brini[1], E. R. van den Heuvel[1]

[1] Eindhoven University of Technology, Department of Mathematics and Computer Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

E-mail for correspondence: m.regis@tue.nl

Abstract: The t linear mixed model naturally occurs from the linear mixed model with inverse gamma distributed heterogeneous variances. Various publications emerged with the aim of proving superiority with respect to traditional linear mixed models, extending to more general settings and proposing more efficient estimation methods. However, little attention has been paid to the mathematical properties of the model itself and to the evaluation of the proposed estimation methods in finite samples with many repeated outcomes. In this paper we propose an in depth analysis of the t linear mixed model, with particular focus on its identifiability and on the evaluation of a maximum likelihood estimation method via an intensive simulation study.

Keywords: Heterogeneous variances; outliers; latent variable; model identifiability; variance components.

1 Introduction

The linear mixed model (LMM) introduced by Laird and Ware (1982) has been largely investigated and applied to diverse longitudinal studies for their appealing feature of capturing both the intra- and inter-subject variability, together with a broad set of possible correlation structures. However, they rely on strong normality assumptions for both the random effects and the residuals, properties often not satisfied in real datasets which can be heavy-tailed, with outlying observations and heteroscedastic residual variances. The t linear mixed model (tLMM) is a particular extension of the classical LMM in which the response variable is jointly t distributed, which makes it robust to outliers. The literature on tLMM is vast, but little has been explored on the properties of the model itself and the accuracy of the associated estimation methods in finite samples with many repeated outcomes. In this paper we thoroughly investigate tLMMs, study its properties and derive identifiability conditions. According to these restrictions, we conduct an intensive simulation study to assess the accuracy of the estimation method under different scenarios. We implement the maximum likelihood (ML) estimation method introduced by Lin and Lee (2006), extending it by allowing the inclusion of both random intercept, slope and AR(1) correlation structure at the same time.

2 The t linear mixed model

We consider the t linear mixed model defined as

$$y_i = X_i\zeta + Z_iu_i + e_i,$$

where y_i is the vector of n_i observations for subject $i = 1, \dots, m$, X_i and Z_i are known $n_i \times p$ and $n_i \times q$ matrices corresponding to the vectors ζ and u_i of the fixed and random effects respectively, and e_i the vector of residuals. The model extends the traditional LMM by introducing a latent variable γ_i for the variance of the residuals and of the random effects,

$$\gamma_i = 1/v_i \sim Gamma(\alpha, \beta).$$

Given γ_i, the random effects are assumed to be independent of the residuals with $u_i|\gamma_i \sim N(0, v_iG)$ and $e_i|\gamma_i \sim N(0, v_iR)$. Here G is the covariance matrix for the random effects which may take a general unstructured form

$$G = \begin{pmatrix} \tau_1^2 & \eta_{1,2}\tau_1\tau_2 & \dots & \eta_{1,q}\tau_1\tau_q \\ \eta_{1,2}\tau_1\tau_2 & \tau_2^2 & \dots & \eta_{2,q}\tau_2\tau_q \\ \dots & \dots & \dots & \dots \\ \eta_{1,q}\tau_1\tau_q & \eta_{2,q}\tau_2\tau_q & \dots & \tau_q^2 \end{pmatrix}$$

and R is an AR(1) correlation matrix. Under the stated distributional assumptions, y_i is t distributed with 2α degrees of freedom and variance-covariance matrix $\mathrm{Var}(y_i) = v_i\Omega_i = v_i(Z_iGZ_i^T + R)$.

3 Identifiability

We assume here the definition of identifiability conditions as restrictions on the parameter space $\Theta = \{\theta = (\zeta_1, \dots, \zeta_p, \tau_1, \dots, \tau_q, \eta, \sigma^2, \rho_1, \dots, \rho_{n_i}, \alpha, \beta)\}$ such that the mapping $\theta \longrightarrow L_i(\theta)$ is one to one, namely

$$\log(L_i(\theta_1)) = \log(L_i(\theta_2)) \iff \theta_1 = \theta_2,$$

where L_i is the contribution to the total likelihood from subject $i = 1, \dots, m$. In general, in case of n observations, identifiability conditions require the number N of parameters of the variance-covariance structure to be less or equal to $n(n+1)/2$ since this is the maximum number of distinct elements in Ω_i. Assuming this condition to be satisfied, X_i and Z_i to be full rank, and the variance-covariance matrices G and R themselves identifiable with respect to the corresponding parameters, the problem reduces to the identifiability of

$$f_{y_i} = \frac{\Gamma(\alpha + \frac{ni}{2})}{\Gamma(\alpha)} \frac{1}{(2\pi\beta)^{n_i/2}|\Omega_i|^{1/2}} \left(1 + \frac{1}{2\beta}\Delta^2\right)^{-(\alpha+n_i/2)}.$$

Here $\Delta^2 = (y_i - X_i\zeta)^T \Omega_i^{-1}(y_i - X_i\zeta)$, and so formulated the problem is not identifiable. In the literature it is commonly assumed $\alpha = \beta$ to overcome this issue and reduce identifiability conditions to those for classical LMM. Although a requirement on the ratio between α and β is necessary, that can be somewhat looser, as the parameter β can be easily be compensated by Ω_i. For example by reparametrizing $\tilde{\Omega}_i = \omega\ \Omega_i$, and imposing a constraint on the ratio $\omega = \beta/\alpha$, it is possible to show that for specific combinations of G and R, the model is identifiable.

4 Simulation study

We introduce a parametrization in compliance with identifiability conditions and previous literature, with the latent variable γ_i gamma distributed with both parameters equal to α and an extra positive parameter σ^2 to rescale the variance of the residuals $e_i|\gamma_i \sim N(0, v_i\sigma^2 R)$. We consider a tLMM including three fixed effects (e.g. gender, age and time) and two random effects (intercept and slope),

$$y_{ij} = \zeta_0 + \zeta_g x_{ig} + \zeta_a x_{ia} + \zeta_t t_j + \delta_i + \omega_i t_j + e_{ij}.$$

The coefficients for the fixed effects are set to determined values, while we choose a set of possible values for each of the variance-covariance parameters, the combination of which results in 64 simulation settings. We initially investigate the method accuracy with 50 subjects and then increase it to 100 for the critical cases. Each setting is simulated 1000 times to ensure stability. In Figure 1 we show the relative bias in estimating the variance-covariance parameters, as the fixed effects can be estimated with good accuracy. The results from the simulation study show that overall the method estimates the parameters with high accuracy, and the estimation in the critical cases can be consistently improved by increasing the number of subjects involved in the study.

5 Conclusion

In this work we derive identifiability conditions for the tLMM and test the accuracy of a ML-based estimation method for tLMM via an intensive simulation study. Our results show that the model should be applied with caution, as one may occur in the wrong estimates by simply using routines implemented in the available softwares. When identifiability conditions are accounted for, the proposed estimation method is shown to be accurate.

Acknowledgments: Research performed within the framework of the Data Science Flagship.

References

Laird, N.M., and Ware, J.H. (1982) Random-effects models for longitudinal data. *Biometrics*, 963 – 974.

Lin, T.I. and Lee, J.C. (2006) A robust approach to t linear mixed models applied to multiple sclerosis data. *Statistics in medicine*, **25**(8), 1397 – 1412.

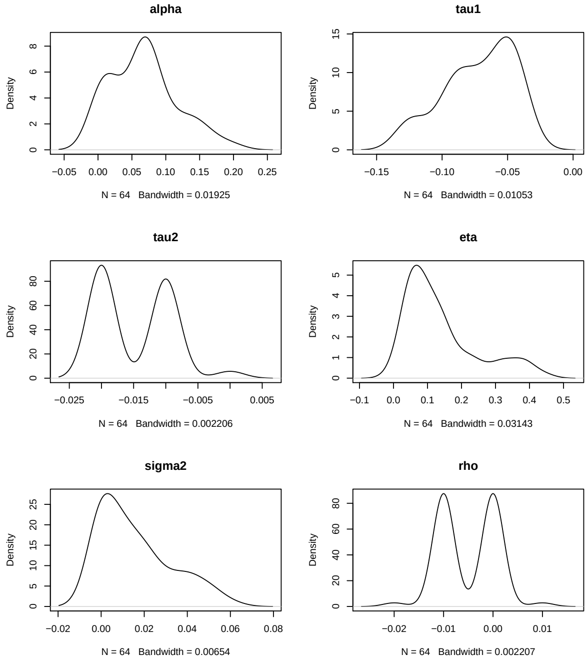


FIGURE 1. Relative bias in estimating the variance-covariance parameters in all the settings.

Combining information at different scales in spatial epidemiology: a Composite Link Generalized Additive Model approach

Diego Ayma[1], Dae-Jin Lee[2], María Durbán[1]

[1] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain
[2] Basque Center for Applied Mathematics, Bilbao, Spain

E-mail for correspondence: `dlee@bcamath.org`

Abstract: Mortality data provide valuable information for the study of the spatial distribution of mortality risk, in disciplines such as spatial epidemiology, medical demography, and public health. However, they are often available in an aggregated form over irregular geographical units, hindering the visualization of the underlying mortality risk and the detection of meaningful patterns. It is also common that factors that may affect mortality are usually measured in a different spatial resolution than mortality data. In this paper, we propose the use of the composite link models to combine information such as covariates measured at different spatial scales (i.e. municipalities, districts or census tracts levels) within a generalized additive framework. We called this model CL-GAM (*"Composite Link - Generalized Additive Model"*). We illustrate our proposal with the analysis of deaths by cardiovascular diseases in Madrid, Spain (period 1996–2003).

Keywords: Disease mapping; penalized Composite Link Models; Mixed models.

1 Introduction

Disease maps deal with public health data that are usually available in an aggregated form over geographical units, like counties, districts, and municipalities. Epidemiologists, health care practitioners, and other related researchers use these data to study the spatial distribution of mortality caused by an specific disease, and thus identify areas of excess and their potential risk factors. In this paper, we propose the extension of the spatial penalized composite link model (P-CLM) in Ayma et al. (2016) for the case of spatial area-to-area (ATA) disaggregation (i.e. from coarse geographical units to smaller units), where mortality maps from aggregated counts are combined with information at different scales.

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.

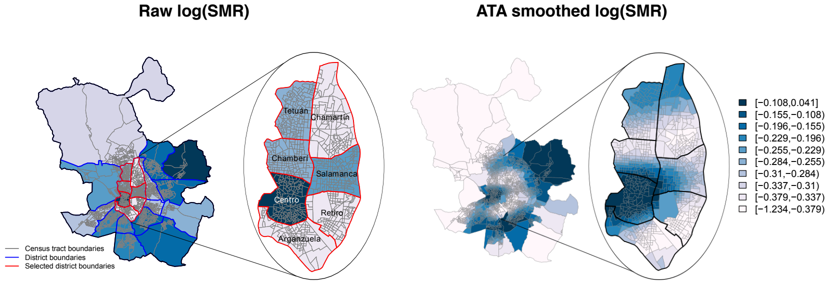


FIGURE 1. Left: raw log(SMR) in Madrid. Right: ATA (from municipalities to census tracts disaggregation) smooth log(SMR).

2 Deaths by cardiovascular disease in the Community of Madrid (1996–2003)

The data correspond to the number of observed and expected female deaths by cardiovascular diseases in the community of Madrid, Spain, over the period 1996–2003, which are available at different (aggregated) spatial levels. We consider the area-to-area case (i.e. from municipalities to census tracts). Let $\boldsymbol{y}$ be the vector of observed deaths and assume that $\boldsymbol{y}$ is Poisson distributed with mean $\boldsymbol{\mu}$. Also, consider the longitude and latitude centroid coordinates of units v_j^*, (x_{1j}, x_{2j}), given by the longitude and latitude vectors $\boldsymbol{x}_1 = (x_{11}, ..., x_{1m})^{\mathrm{T}}$ and $\boldsymbol{x}_2 = (x_{21}, ..., x_{2m})^{\mathrm{T}}$, respectively. Then, the P-CLM for the area-to-area case is given by:

$$\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma} = \mathbf{C}\left(\boldsymbol{e}_{\mathrm{f}} * \exp(f(\boldsymbol{x}_1, \boldsymbol{x}_2))\right), \tag{1}$$

where $\boldsymbol{\gamma}$ denotes the vector of latent expectations at the fine unit level and $\mathbf{C}$ is an area-to-area composition matrix, whose entries are obtained as $c_{ij} = 1$ if v_j^* is contained in unit v_i and 0 otherwise. The vector $\boldsymbol{\gamma}$ in model (1) is expressed in terms of the vector of exposures at the fine unit level ($\boldsymbol{e}_{\mathrm{f}}$). Hence, we can use the *standardised mortality ratio* (SMR), i.e. $SMR_i = y_i/e_i$, instead of counts. The function $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$, represents the latent spatial trend that is modelled via a Tensor product bivariate P-spline.

Ayma et al. (2016), derived the mixed model reparameterization of the P-CLM, obtaining the mixed model equations with the so-called working matrices: $\breve{\mathbf{X}} = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$ and $\breve{\mathbf{Z}} = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$, with $\mathbf{W} = \mathrm{diag}(\boldsymbol{\mu})$ and $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. The covariance matrix of the random effects $\mathbf{G}$ depends on two variance components τ_1 and τ_2 that controls the spatial smoothness.

Finally, defining the working vector: $\boldsymbol{z} = \breve{\mathbf{X}}\boldsymbol{\beta} + \breve{\mathbf{Z}}\boldsymbol{\alpha} + \mathbf{W}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. Hence, given $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$, we use penalized quasi-likehood (PQL), the solution is achieved by iteration between $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\alpha}}$ and the restricted or residual maximum likelihood (REML) criteria until convergence. Figure 1, illustrates the ATA P-CLM allowing for a better visualisation of the central districts of Madrid.

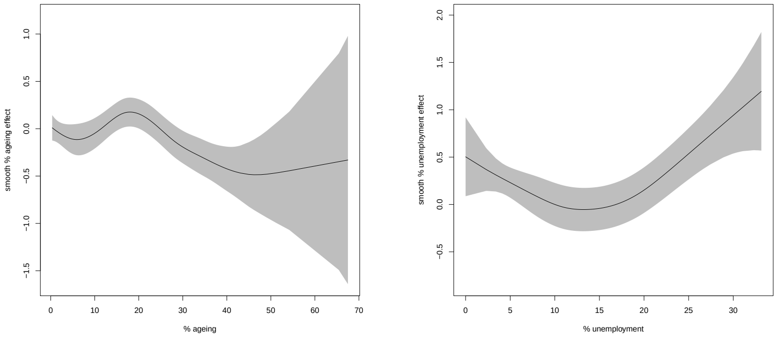


FIGURE 2. Smooth effects for covariates measured at municipality level for ATA disaggregation of mortality rates by cardiovascular disease. Left: % of population > 65. Right: % of unemployment.

3 Extending the CLMM to combine information at different spatial scales

In many occasions it might be of interest to include covariates in the model which are collected at a different scales of the response variable, or covariates measured at different scales. The CLM provides a framework in which both situations can be easily resolved, yielding what we call composite link generalized additive mixed model (CL-GAMM). Given that in all cases we are interested in spatial disaggregation of counts, we can consider different situations: (a) the covariates are measured at the fine level; (b) the covariates are measured at the aggregated level; and (c) when we have covariates measured both at aggregated and fine scale levels, which generalizes the previous cases a) and b).

Including covariates measured at census tract level. In order to study relationships between different socio-economic indicators and mortality rates, we considered two covariates: i) an indicator of ageing (people whose ages are greater or equal than 65 years old); ii) an indicator of unemployment for people whose ages are greater or equal than 16 years old. Figure 2, shows the smoothed additive effects.

Comparing mortality of two populations. Let $\boldsymbol{y} = (\boldsymbol{y}_{\text{female}}, \boldsymbol{y}_{\text{male}})^{\mathrm{T}}$ be the vector of observed number of counts for females and males, and $\boldsymbol{e}_{\text{f}} = (\boldsymbol{e}_{\text{f-female}}, \boldsymbol{e}_{\text{f-male}})^{\mathrm{T}}$ be the vector of expected number of counts for females an males at the fine unit level. The model is extended to include a factor variable for sex and hence able to compare both populations. Figure 3, shows the resulting *'contrast map'* where the difference of the mortality surfaces at census tract levels between females and males are classified into three categories based on the smoothed spatial trends and standard errors.

Acknowledgments: This research was supported by the Basque Government through the BERC 360 2014-2017 and by the Spanish Ministry of Economy and Competitiveness MINECO and FEDER: BCAM Severo Ochoa excellence accred-

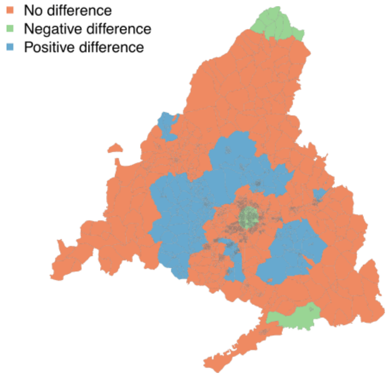


FIGURE 3. Classification of mortality differences between females and males into three categories.

itation SEV-2013-0323 and MTM2016-74931-P.

References

Ayma, D., Durbán, M., Lee, D.-J., and Eilers, P.H.C. (2016). Penalized composite link models for aggregated spatial count data: a mixed model approach. *Spatial Statistics*, **17**, 179 – 198.

Eilers, P.H.C. (2007). Ill–posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, **7**(3): 239 – 254.

Detecting Match-Fixing in the Italian Serie B Using Flexible Regression

Marius Ötting[1], Christian Deutscher[1], Roland Langrock[1]

[1] Bielefeld University, Germany

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

Abstract: In recent years, several match-fixing scandals in soccer took place in Europe and all over the world. In order to avoid match-fixing, the literature and fraud detection systems analyse odds movements from bookmakers. In our work, we analyse not only odds movements but also total volume placed on bets. For the volume of money placed on bets, we use data from the betting exchange platform Betfair. We focus on the second division in Italian Soccer (Serie B) from season 2009/10 until 2015/16, since for this league it has effectively been proven that some matches were fixed. To model both the mean and the variance of the betting volume, a Generalized Additive Model for Location, Shape and Scale (GAMLSS) is employed. Both the mean and the variance of the betting volume are related to several explanatory variables such as the type of bet, the matchday or the popularity of the teams. Possible smooth functions are used to estimate effects of the non-categorical variables. Suspicious matches are detected via normalised quantile residuals. Using this approach, our model obtains a true positive rate of 21%, i.e. 21% of the fixed matches are detected.

Keywords: detecting corruption; soccer; flexible regression

1 Introduction

Between 2009 and 2015, several match-fixing occurrences were detected in Italian soccer, especially in the second division (Serie B). This resulted in forced relegation and point deduction for various teams, potentially endangering the integrity of this league. Detecting match-fixing before the start of the game is challenging. Fraud detection systems typically analyse odds movements from bookmakers. Reade and Akie (2013) and Feustel and Rodenberg (2015) argue that in the absence of any attempt to fix a match, the probability of a particular outcome stated by bookmakers should be very close to the probability stated by a statistical model, because most of the relevant information for estimating the probability is available, e.g. position in the league table or home field advantage. Feustel

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

and Rodenberg (2015) estimate odds for certain outcomes in football matches in England, USA, France and Italy and compare these odds with the odds from bookmakers.

In this work, we propose to model both betting odds and the total volume of bets placed, instead of considering only the former. We demonstrate our approach using betting volume data on the Serie B obtained from the betting exchange platform Betfair. The aim of the empirical analysis is to demonstrate that modelling betting volumes, in addition to the usual consideration of odds movements, can help to more accurately identify suspicious matches. In this regard, the Serie B is a useful case study, since here for several matches it has effectively been proven that they were fixed. Furthermore, we try to minimise false positives, i.e. flagging of matches which were not proven to be fixed.

2 Data

In order to find matches with a strong indication of match-fixing activities, pregame data from the betting exchange platform Betfair are analysed. The data include the seasons 2009/10 to 2015/16. In this period, 3234 matches were played, from which 3224 matches offer full information and are included in the data set. In addition to betting volumes, several explanatory variables are included in the data, such as home and away team, the date of the match, the type of bet and the matchday. For the type of the bet, we focus on bets on the winning team and on the number of goals scored per match. These type of bets are very popular compared to others, e.g. bets on a particular player scoring a goal, and therefore the total volume of these bets is typically higher on average. Due to the high volumes, placing heavy bets on these markets by match fixers is not as conspicuous as placing high volume bets on less popular markets. For the number of goals scored in a match, the betting types over/under 1.5/2.5/3.5 goals are considered.

To account for popularity effects and the importance of the match, additional variables are collected, e.g. Facebook likes of the teams and a dummy variable indicating whether the match is important for the team in terms of the outcome of the season. There is a large variation between the betting volumes placed on individual games and types of bets, ranging from a few Euro to ~ 3 million Euro.

3 Methodology

From visual inspection of the data, we found that the distribution of the betting volumes varies substantially across several non-categorical covariates, and that it would be difficult to accommodate these effects within a linear model. Thus, we use the very flexible class of Generalized Additive Models for Location, Scale and Shape (GAMLSS), which allows 1) to simultaneously model several parameters of the distribution of the response variable (e.g. mean and variance) and 2) to estimate smooth functional effects of non-categorical covariates (Stasinopoulos and Rigby, 2007). We apply the semi-parametric additive formulation of GAMLSS formulated by Stasinopoulos and Rigby (2007), given by

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} h_{jk}(x_{jk}), \tag{1}$$

where θ_k is a parameter of the distribution assumed for the response variable Y, $g_k(\cdot)$ is a known link function, X_k is an $n \times J_k^{'}$ design matrix, β_k is a vector of regression coefficients of length $J_k^{'}$, and the h_{jk} are unknown smooth functions. In this work, we assume a normal distribution for the response variable Y, here the logarithm of the betting volumes. Thus, the explanatory variables are linked to two parameters of the distribution of Y, namely the mean $\mu\,(=\theta_1)$ and the standard deviation $\sigma\,(=\theta_2)$, leading to the following special case of (1):

$$g_1(\mu) = \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}(x_{j1}) \tag{2}$$

$$g_2(\sigma) = \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} h_{j2}(x_{j2}) \tag{3}$$

The explanatory variables described in Section 2 enter the model in (2) and (3) in different parts. Dummy variables for each season, for the type of bet, for the day of the week and for indicating whether the match is important are included in the linear parts both in (2) and (3). The effects of the non-categorical variables, i.e. the matchday and the Facebook likes of the teams, are estimated using P-spline smoothers. A major advantage of the GAMLSS framework here is the option to model both the mean and the variance. The betting volume data is heteroscedastic, as can be seen for example by comparing the betting volumes across the 42 matchdays. The variance in the betting volume increases for the last matchdays, which can be accommodated within the GAMLSS framework by modelling the variance via the covariate matchday. By being able to use smooth functional effects for the non-categorical covariates — the second major advantage of the GAMLSS framework — the model fit is substantially improved.

We apply this model is to detect outliers, i.e. suspicious matches. Outlier detection is done by using the normalised quantile residuals, which are standard normally distributed provided the model is correct (Dunn and Smyth, 1996). Therefore, observations with normalised quantile residuals larger than three are flagged as outliers and hence as suspicious matches.

4 Preliminary Results

Out of 19 matches that have effectively been proven to have been fixed, our model flags four, equalling a true positive rate of 21%. There is little information available on how many matches exactly were fixed in the Serie B. Thus, there may very well have been more than 19 fixed matches from 2009 to 2015. In total, 59 matches out of 3224 matches are flagged by the GAMLSS applied. From these, 55 are not currently known to have been fixed. Assuming that these matches were indeed free from irregularities, the GAMLSS model hence returns $55/3224 \approx 1.7\%$ false positives, i.e. matches to be flagged as suspicious but unconfirmed to be fixed.

5 Discussion

Future research should attempt to further increase the number of true positives, and to reduce the number of false positives. For example, it may be beneficial to additionally model betting odds using game and team characteristics, in order to identify deviation between actual and fair betting odds. Betting odds were modelled for example in Reade and Akie (2013) and in Feustel and Rodenberg (2015), and are also considered within the procedures commonly applied in fraud detection systems. Odds can be modelled for example by fitting Poisson regression models to the number of scored goals by a team in a match, then calculating the odds of winning by simulation. Extensions of the Poisson distribution, such as the zero-inflated Poisson or the bivariate Poisson, may be useful here. These distributions involve several parameters, such that the GAMLSS framework is again a good candidate.

References

Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics.*

Feustel, E.D. and Rodenberg, R.M. (2015). Sports (betting) integrity: detecting match-fixing in soccer. *Gaming Law Review and Economics.*

Reade, J.J. and Akie, S. (2013). Using forecasting to detect corruption in international football. *Proceedings of the 4th International Conference on Mathematics in Sport.*

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

Copula VAR models

Lucia Benetazzo[1], Ernst Wit[2]

[1] Università degli studi di Padova, Italy
[2] JBI, University of Groningen, Netherlands

E-mail for correspondence: lucia.benetazzo@gmail.com

Abstract: We propose a VAR graphical model for the analysis of time series genomic data; our goal is to relax the normality assumption of Gaussian graphical models employing a semiparametric Gaussian copula. We adopt the lasso penalized likelihood inference approach in order to obtain sparse estimates.

Keywords: Chain graphical model; Time series; Gaussian copula; Lasso penalty; Gene expression.

1 Introduction

In the last decades, significant developments in DNA microarray technology have allowed us to collect observations of gene expression levels over time. Time-series data provide a deeper insight on the biological process under study. This work can be considered as an extension of the Time Series Chain Graphical Model (TSCGM) by Abegaz and Wit (2013), since we aim to tackle the issue of non-normality with the use of the Gaussian copula.

2 Methodology

2.1 Time series chain graphical models

TSCGMs assume longitudinal microarray data in which n replications indexed by $i = 1, \dots, n$ of continuous measurements across p genes indexed by $j = 1, \dots, p$ are repeated T times. Our aim is to model both contemporaneous and dynamic (delayed) relationships among the genes. This will be accomplished with a time series chain graph $G = (V, E)$, where V is a finite set of *vertices* and the set of *edges* E is a subset of the set $V \times V$ of ordered pairs of distinct vertices. The time series chain graph is based on the partitioning of V into a number of blocks $\bigcup_{t=1}^{T} V_t$. Links within a time step are undirected and represent contemporaneous

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

interactions; links between time steps are directed and they explain dynamic or delayed interactions between genes in time.
Let $X_t = (X_{1t}, \dots X_{pt})$ be a p-variate random variable associated with the nodes in V_t. X_t represents the gene expression levels at time t. We assume that X_t follows a Markovian dynamics, which can be translated into a vector autoregression process of order 1 (VAR(1)):

$$X_t = \Delta X_{t-1} + \epsilon_t. \tag{1}$$

In the TSCGM setting, the following normality assumption is made: $\epsilon_t \sim \mathcal{N}_p(0, \Omega^{-1})$. In this project, our aim is to make no assumption about the distribution of gene expressions. Particularly, in order to weaken the gaussian assumption on ϵ_t, we will exploit the Gaussian copula.

2.2 Copula VAR models

Let $Z_t = (Z_{1t}, \dots Z_{pt})$ be a multivariate latent variable. The Gaussian copula transformation is defined as:

$$Z_{jt} = \Phi^{-1}(F_{jt}(X_{jt})), \quad \text{for } j = 1, \dots, p, \quad t = 1, \dots, T, \tag{2}$$

where X_{jt} is the gene expression level for gene j at time t, F_{jt} is the marginal CDF of gene j at time t and Φ^{-1} is the quantile function of the univariate normal distribution $N(0, 1)$. The latent variable Z_t follows a dynamics equivalent to that in (1):

$$Z_t = \Gamma Z_{t-1} + \eta_t,$$

where $Z_t | Z_{t-1} \sim \mathcal{N}_p(\Gamma Z_{t-1}, \Theta^{-1})$. The matrices Γ and Θ are in general different from Δ and Ω respectively, but they have the same nonzero entries, resulting in the same graphs. Therefore, Γ and Θ contain all the information about the dependencies among the variables. In particular, nonzero entries in Γ are equivalent to directed edges; likewise, contemporaneous interactions (undirected edges) are related to nonzero entries of the concentration matrix, also known as precision matrix, Θ:

$$(\alpha, \beta) \in V_{t-1} \times V_t \Leftrightarrow \Gamma_{\alpha\beta} \neq 0, \quad (\alpha, \beta) \in V_t \times V_t \Leftrightarrow \Theta_{\alpha\beta} \neq 0.$$

We now define the log-likelihood for n replicates each at T time steps as

$$\ell(\Gamma, \Theta) = -\frac{npT}{2} \log(2\pi) + \frac{nT}{2} \log|\Theta| - \frac{nT}{2} Tr(S_\Gamma \Theta) + D \tag{3}$$

where D is the sum of the logs of the Jacobians of the transformation in (2) and $S_\Gamma = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (Z_{it} - \Gamma Z_{i,t-1})(Z_{it} - \Gamma Z_{i,t-1})^\top$. We exploit the marginal rank likelihood by Hoff (2007), which can be seen as a type of marginal likelihood function for estimation in the presence of nuisance parameters (the marginal distributions). This allows us to neglect the term D in (3).

2.3 Kernel density estimate

The marginal CDFs involved in the transformation given in (2) are unknown: we estimate them in a nonparametric fashion via weighted kernel CDF estimation. Weights exponentially decay to zero as the time interval becomes larger: i.e., in

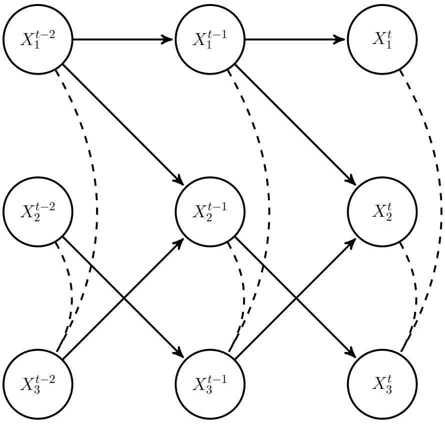


FIGURE 1. Example of graphical representation of a Copula VAR model of order 1 for $p = 3$ time series variables. Directed edges represent nonzero entries of the transition matrix Γ and undirected edges (dashed links) characterize the nonzero entries of Θ.

order to estimate the CDF in a time point, we use observations from the future and from the past giving more importance to those observations closer in time. Hence, the CDF estimate for the j-th gene at time t is

$$\hat{F}_{jt}(x_{ijt}) = \sum_{l=1}^{T} \sum_{k=1}^{n} \Phi\bigg(\frac{x_{ijt} - x_{kjl}}{h_j}\bigg) w_{tl},$$

where $w_{tl} = \frac{\omega^{|t-l|}}{\sum_{j=1}^{T} \omega^{|t-j|}}$, Φ is the CDF of the standard normal (which is the chosen kernel), h_j is the bandwidth parameter (different for each gene) and ω is a constant which determines the decay velocity. We choose $\omega = 0.925$.

2.4 Penalized log-likelihood

Gene networks are known to be sparse, i.e. a gene will not interact with all the genes in the network but only with a relatively small subset of them. This results in the matrices Γ and Θ to have many zero entries. For the purpose of obtaining sparse estimates, we add two lasso penalties to the likelihood, for both Γ and Θ. Therefore, the objective function for optimization is defined as

$$\ell_{pen}(\Gamma, \Theta) = \log\big|\Theta\big| - Tr(S_\Gamma \Theta) - \lambda_1 \sum_{i \neq j}^{p} |\theta_{i,j}| - \lambda_2 \sum_{i \neq j}^{p} |\gamma_{i,j}|$$

where $\theta_{i,j}$ and $\gamma_{i,j}$ are the entries of Θ and Γ. The estimation is accomplished via a two-stage iterated procedure as in Rothman (2010). The first stage consists of the following optimization problem

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \bigg\{ \log\big|\Theta\big| - Tr(S_\Gamma \Theta) - \lambda_1 \sum_{i \neq j}^{p} |\theta_{i,j}| \bigg\} \quad \text{with } \Gamma \text{ fixed.}$$

We use the graphical lasso algorithm to solve it. The second stage is given by

$$\hat{\Gamma} = \underset{\Gamma}{\operatorname{argmax}} \left\{ -Tr(S_{\Gamma}\Theta) - \lambda_2 \sum_{i=j}^{p} |\gamma_{i,j}| \right\} \quad \text{with } \Theta \text{ fixed.}$$

The solution is computed using a cyclical-coordinate descent algorithm.

2.5 Model selection

The regularization parameters of the lasso penalty must be carefully tuned. One can adopt the Bayesian information criterion (BIC) to select the parameters, which is defined as

$$BIC(\lambda_1) = -nT\{\log|\hat{\Theta}_{\lambda_1}| - Tr(S_{\hat{\Gamma}_{\lambda_2}} \hat{\Theta}_{\lambda_1})\} + k \log(nT),$$

where k is the number of nonzero estimated parameters. The values of λ_1 and λ_2 that jointly minimize the BIC criterion have to be chosen.

References

Abegaz, F. and Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, **14**, 3, 586 – 599.

Hoff, P.D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 265 – 283.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19.4, 947 – 962.

An exact test for comparing two predictive values in small-size clinical trials

Kanae Takahashi[1], Kouji Yamamoto[1]

[1] Department of Medical Statistics, Osaka City University Graduate School of Medicine, Japan

E-mail for correspondence: takahashi.kanae@med.osaka-cu.ac.jp

Abstract: The positive predictive value and the negative predictive value describe the performance of a diagnostic test. There are several methods to test the equality of predictive values of two binary diagnostic tests. However, these methods were premised on large sample size theory and they may not be suitable for small-size clinical trials.

In this study, we propose an exact test for conducting a small-size clinical trial that investigates the equality of predictive values of two binary diagnostic tests. In addition, we execute simulation studies to evaluate the performance of the proposed exact test and existing methods in small-size clinical trials.

The proposed test can calculate exact p-value and, as a result of simulations, the performance of it is not much different from the other methods. Therefore, it is considered that the proposed exact test may be useful for small-size clinical trials.

Keywords: exact method; negative predictive value; positive predictive value; small-size clinical trials.

1 Introduction

In medicine, diagnostic tests are important for early detection and treatment of disease. The positive predictive value (PPV) and the negative predictive value (NPV) describe the performance of a diagnostic test. The PPV is the probability of having the disease when the diagnostic test result is positive, and the NPV is the probability of not having the disease when the diagnostic test result is negative. The PPV and NPV are useful clinically, and may influence the treatment decision.

There are several methods to test the equality of predictive values of two binary diagnostic tests (Leisenring et al, 2000; Moskowitz and Pepe, 2006; Wang et.al, 2006; Kosinski, 2013). Some researchers use regression frameworks, and others

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.

utilize the multivariate central limit theorem and the delta-method. These methods were premised on large sample size theory. For this reason, these methods may not be suitable for small-size clinical trials.
In this study, we propose an exact test for conducting a small-size clinical trial that investigates the equality of predictive values of two binary diagnostic tests. Furthermore, we execute simulation studies to evaluate the performance of the proposed exact test and existing methods in small-size clinical trials.

2 Methods

We performed a simulation study to assess the performance of two existing methods; one method is based on the delta-method with identical link (Tm) and the other is weighted generalized score statistic (Twgs) in a regression framework. The performance measure were actual type 1 error rate and empirical power. We executed 1,000,000 repeated simulations for each method. The simulation was conducted in the same way as Kosinskis article. The odds ratio of diseased group was set to OR_{D+}=5, and OR_{D-}=2 for the odds ratio of non-diseased group. The total sample size N was assumed to be 30, 40 and50 because this study assumes a small-size clinical trial. PPV for new diagnostic test (PPV1) was set to 0.75 or 0.85 and PPV for the existing diagnostic test (PPV2) was 0.75. NPV for new diagnostic test (NPV1) was set to 0.85 or 0.95, NPV for the existing diagnostic test (NPV2) was 0.95. The disease prevalence was assumed to be 0.4. The nominal type 1 error rate was set to 0.05 (two-sided test).
In addition, we performed a simulation study to assess the performance of the proposed exact test. The exact test is based on the permutation test. The simulation was conducted in the same way as above.

TABLE 1. Actual type 1 error rate (PPV1=PPV2=0.75)

N	Tm	Tgs	Exact
30	0.078	0.043	0.025
40	0.068	0.048	0.034
50	0.064	0.049	0.037

TABLE 2. Actual type 1 error rate (NPV1=NPV2=0.80)

N	Tm	Tgs	Exact
30	0.044	0.035	0.024
40	0.051	0.042	0.032
50	0.053	0.045	0.037

TABLE 3. Empirical power (PPV1=0.85 PPV2=0.75)

N	Tm	Tgs	Exact
30	0.133	0.101	0.089
40	0.143	0.130	0.124
50	0.162	0.154	0.148

TABLE 4. Empirical power (NPV1=0.90 NPV2=0.80)

N	Tm	Tgs	Exact
30	0.190	0.162	0.130
40	0.274	0.248	0.217
50	0.346	0.321	0.294

3 Results

Table 1 shows the actual Type 1 error rate for comparing two PPVs. It shows that the actual type 1 error rate of Twgs and Exact test does not exceed the nominal type 1 error rate 0.05. Table 2 shows the actual Type 1 error rate for comparison of NPVs. It shows almost same findings as the case of PPV.

Table 3 shows the empirical power of PPVs. It shows that the empirical power of Tm is the highest among three methods, and the empirical power of proposed exact test is not much different from that of Twgs. Table 4 shows the empirical power of NPVs. It also shows similar findings to the case of PPVs.

4 Discussion

By the result of the simulation studies, we consider that the proposed exact test may be useful for small-size clinical trials because it can calculate exact p-value, and the exact test does not have so less power than other methods.

References

Kosinski, A.S. (2013). A weighted generalized score statistic for comparison of predictive values of diagnostic tests. *Statistics in medicine*, **32**, 964 – 977.

Leisenring, W. et al. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics*, **25**, 2215 – 2229.

Moskowitz, C.S. and Pepe, M.S. (2006). Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. *Clinical Trials*, **3**, 272 – 279.

Wang, W. et al. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. *Statistics in Medicine*, **25**, 2215 – 2229.

Comparison of Two Testing Procedures in Clinical Trials with Multiple Binary Endpoints

Takuma Ishihara[1], Kouji Yamamoto[1]

[1] Department of Medical Statistics Osaka City University Graduate School of Medicine, Japan

E-mail for correspondence: ishihara.takuma@med.osaka-cu.ac.jp

Abstract: The test treatment in confirmatory clinical trials is often assessed by using multiple primary endpoints. We considered the efficacy of the trials that is confirmed only when non-inferiority for all endpoints and superiority for at least one endpoint. Perlman and Wui2004jand Nakazuru et al.i2014jproposed testing procedures when the multiple primary endpoints are continuous variables. However, it is not yet discussed the case that a trial has multiple binary endpoints. In this presentation, we consider a testing procedure when the multiple primary endpoints are binary, and show the performance through Monte Carlo simulations.

Keywords: Clinical trial; Multivariate Bernoulli distribution; Non-inferiority; Superiority.

1 Introduction

In confirmatory clinical trials, the efficacy of a test treatment are sometimes assessed by using multiple primary endpoints. However, it is difficult to demonstrate that each endpoint is significantly when the number of endpoints is not small. For the reason, this manuscript deals with a case where it is superior for at least one of the endpoint and not clinically inferior for the remaining endpoints. Perlman and Wu (2004) proposed a testing procedure that is applicable to the above case. This procedure has a disadvantage that cause the inflation of type I error when correlation coefficients among the endpoints are highly positive. Moreover, Nakazuru et al. (2014) proposed a modified testing procedure for Perlman and Wu method. These procedures are able to be applied only when endpoints are continuous variables. Therefore, we propose a new testing procedure when the case that multiple primary endpoints are binary.

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 Notations and Hypotheses

To simplify this case, we consider comparing p binary endpoints with 2 treatment groups comprising n_1 and n_2 subjects. Let $Y_{ijk}(i = 1, 2; j = 1, ..., p; k = 1, ..., n_i)$ denote the response variable of the kth subject to the ith treatment at the jth endpoint, $\bar{Y}_{ij}(i = 1, 2; j = 1, ..., p)$ is the sample mean for the jth endpoint to the ith treatment, and set $\mathbf{X} = (X_1, ..., X_p)^t$ with $X_j = (\bar{Y}_{1j} - \bar{Y}_{2j})(j = 1, ..., p)$, where the superscript t denotes transpose.
A null hypothesis H_0 is expressed by

$$H_0 : \left\{ \max_{1 \leq j \leq p} \mu_j \leq 0 \right\} \cup \left\{ \min_{1 \leq j \leq p} (\mu_j + \epsilon_j) \leq 0 \right\},$$

where $\mu_j = \mu_{1j} - \mu_{2j}$ is mean of the treatment effect, and $\epsilon_j(j = 1, ..., p)$ is non-inferiority margin of jth endpoint that denotes prespecified positive constants.

3 Proposal of a New Procedure

3.1 Perlman and Wu Procedure

Perlman and Wui2004jproposed a testing procedure that rejects above null hypothesis H_0 if and only if

$$\begin{aligned} \|\mathbf{X} - \pi_s(\mathbf{X}; \mathcal{N}^p)\|_s^2 > c_\alpha^* \quad & \text{and} \\ c_{n_1,n_2}(X_j + \epsilon_j)/\sqrt{\hat{\sigma}_{jj}} > t_\alpha \quad & \text{for} \quad j = 1, ..., p, \end{aligned}$$

where $c_{n_1,n_2}^2 = n_1 n_2/(n_1 + n_2)$, and let $\hat{\sigma}_{jj}$ denote jth diagonal element of the pooled sample covariance matrix $\hat{\Sigma}$, then $\boldsymbol{S} \equiv (n_1 + n_2 - 2)\hat{\Sigma}$ is distributed as a Wishart distribution $W_p(\Sigma, n_1 + n_2 - 2)$ when $n_1 + n_2 > p + 2$. Here $\mathcal{N}^p$ is the non-positive orthant in $\mathcal{R}^p$(p-dimensional Euclidean space), $\|\boldsymbol{x}\|_s^2 \equiv \boldsymbol{x}^t \boldsymbol{S}^{-1} \boldsymbol{x}$ is the Euclidean norm determined by $\boldsymbol{S}$, $\pi_s(\mathbf{X}; \mathcal{N}^p)$ is the projection of $\mathbf{X}$ onto $\mathcal{N}^p$ with respect to this norm. Then c_α^* is the critical value of size-α one-sided LRT (likelihood ratio test) for superiority part of H_0 determined by 1

$$\alpha = \frac{1}{2}\Pr\left[\frac{\chi_{p-1}^2}{\chi_{n_1+n_2-p}^2} > c_\alpha^*\right] + \frac{1}{2}\Pr\left[\frac{\chi_p^2}{\chi_{n_1+n_2-p-1}^2} > c_\alpha^*\right],$$

where χ_n^2 is a chi-square variable with n degrees of freedom.

3.2 Nakazuru et al.'s Procedure

Nakazuru et al.i2014jproposed a more powerful procedure rather than Perlman and Wu procedure by modifying the Glimm et al. (2002)'s method, and it rejects the null hypothesis if and only if

$$\begin{aligned} min(\bar{u}_A^2, \bar{u}_B^2) > c \quad & \text{and} \\ c_{n_1,n_2}(X_j + \epsilon_j)/\sqrt{\hat{\sigma}_{jj}} > t_\alpha \quad & \text{for} \quad j = 1, ..., p. \end{aligned}$$

Here $\mathbf{u}_A \equiv (u_{A1}, ..., u_{Ap})^t = c_{n_1,n_2}\mathbf{AX}$ and $\mathbf{u}_B \equiv (u_{B1}, ..., u_{Bp})^t$ $= (\frac{det\mathbf{A}}{det\mathbf{B}})^{2/p} c_{n_1,n_2}\mathbf{BX}$ are the test statistics that distributed as a p variable standard normally distribution. Then $\bar{u}_A^2$ and $\bar{u}_B^2$ are as follows:

$$\bar{u}_A^2 = \sum_{j=1}^{p} max(u_{Aj}, 0)^2$$
$$\bar{u}_B^2 = \sum_{j=1}^{p} max(u_{Bj}, 0)^2.$$

The matrix **A** and **B** are omitted here. For further details, see Nakazuru et al. (2014). Then c is determined by

$$\frac{1}{2^p}\sum_{j=0}^{p-1}\frac{p!}{j!(p-j)!}\frac{1}{B(\frac{p-j}{2},\frac{n_1+n_2-1-p+j}{2})}\cdot$$
$$\int_0^{c/(n_1+n_2-1)} r^{\frac{p-j}{2}-1}(1-r)^{\frac{n_1+n_2-1-p+j}{2}-1}dr + \frac{1}{2^p} = 1-\alpha$$

where $B(a,b)$ is beta function defined with gamma function.

3.3 A new testing procedure

However, these methods are discussed when the efficacy of a test treatment are continuous variables. Suppose that the vector of responses $Y_{ik} = (Y_{i1k}, ..., Y_{ipk})$ are distributed as a P-variate Bernoulli distribution with $E(Y_{ijk}) = \pi_{ij}$, $V(Y_{ijk}) = \pi_{ij}(1-\pi_{ij})$, and $corr(Y_{ijk}, Y_{ij'k}) = \rho_i^{jj'}$ for all $j \neq j'$ $(1 \leq j < j' \leq p)$. Then, we can apply the same way as Nakazuru et al.'s procedure to the case that multiple primary endpoints are binary by using an approximation based on a multivariate central limit theorem. Let $\hat{\pi}_{ij}(i = 1,2; j = 1,...,p)$ be a sample proportion of the jth endpoint to the ith treatment, and $\Delta_j = \hat{\pi}_{1j} - \hat{\pi}_{2j}$ denote the difference of jth endpoint. It rejects the null hypothesis if and only if

$$min(\bar{u}_A^2, \bar{u}_B^2) > c \quad \text{and}$$
$$\frac{\Delta_j + \epsilon_j}{\sqrt{\frac{\hat{\pi}_{1j}(1-\hat{\pi}_{1j})}{n_1} + \frac{\hat{\pi}_{2j}(1-\hat{\pi}_{2j})}{n_2}}} > Z_\alpha \quad \text{for} \quad j = 1,...,p,$$

where Z_α is the upper α-percentile of standard normally distribution. On the day of the presentation, we introduce the other procedure that is applied Westfall PH and Troendle JF (2008) method to the part of superiority in test statistics, and compare the performance of two procedures.

3.4 Numerical comparison

The performance of the proposed procedure was examined through Monte Carlo simulations. In this simulation, we confirmed difference of 2 proportions in the level of significance was set at 0.05. For evaluating powers, we assumed that $(\pi_{11}, \pi_{12}) = (0.6, 0.4)$, $(\pi_{21}, \pi_{22}) = (0.4, 0.4)$, the simulated data was repeated 10,000 times. In addition, we assumed that $(\pi_{11}, \pi_{12}) = (0.5, 0.5), (\pi_{21}, \pi_{22}) = (0.5, 0.5)$, the simulated data was repeated 100,000 times for evaluating type I error rates. The sample size is set as $n = 30$ and $n = 100$. The correlation coefficient is set as $\rho = 0$, $\rho = 0.4$ and $\rho = 0.8$ in each case.

TABLE 1. Estimated powers and type I error rates.

	Powers		Type I error rates	
ρ	$n = 30$	$n = 100$	$n = 30$	$n = 100$
0.8	0.5717	0.9132	0.04540	0.00025
0.4	0.4775	0.8986	0.03696	0.00026
0	0.4728	0.9009	0.03036	0.00026

References

Nakazuru, Y., Sozu, T., Hamada, C. and Yoshimura, I. (2014). A new procedure of one-sided test in clinical trials with multiple endpoints. *Japanese Journal of Biometrics,* **35**, 17-35.

Glimm, E., Srivastava, M. and Lauter, J. (2002). Multivariate tests of normal mean vectors with restricted alternatives. *Communications in Statistics,* **B 31**, 589-604.

Perlman, M.D. (1969). One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics,* **40**, 549-567.

Perlman, M.D. and Wu, L. (2004). A note on one-sided tests with multiple endpoints. *Biometrics,* **60**, 276-280. **26**, 1193-1207.

Westfall PH and Troendle JF. (2008). Multiple testing with minimal assumptions. *Biometrical Journal,* **50(5)**, 745-755.

Parameter Inference in the Pulmonary Blood 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 focus on parameter inference in a pulmonary blood circulation model for mice. It utilizes a fluid dynamics network model that takes some 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 the progression of pulmonary hypertension. Constraint nonlinear optimization is successfully used to learn the parameters.

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 indirectly estimate parameters from the measured hemodynamic 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). 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 diagnostic and treatment purposes. The ultimate goal behind this model, 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. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

hence the motivation behind inferring the parameters, is to minimize the number of invasive procedures for PH patients, as well as to assist the clinicians in devising better treatment strategies. Thus, this study focuses on inference of key parameters pertinent to disease detection and treatment. We have shown that by using our statistical method we can improve BP prediction in both healthy and hypoxic mice. This leads to enhanced reliability of key parameter estimates obtained using the model.

2 Mathematical Model

The 1D fluids model studied here is derived from the incompressible axisymmetric Navier-Stokes equations for a Newtonian fluid, and coupled with a constitutive wall model predicting stiffness of the blood vessels. 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)). In order 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 MPA whereas conservation of blood flow and continuity of pressure are ensured across the bifurcations. At the 7 terminal outlet vessels, 3-element Windkessel models (represented by two resistors R_1, R_2 and a capacitor C) are attached. The outflow boundary conditions account for the lumped effects of pulmonary hemodynamics beyond the truncated network of large arteries. The model takes several 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, k, which significantly changes its behaviour during initial stages of PH.

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, $x_i \in \mathbf{x}$ denote other input variables and $\boldsymbol{\epsilon}$ are the errors, which we assume are i.i.d following a Gaussian distribution. The objective function to be minimised using Constraint Nonlinear Optimization is the Residual Sum of Squares:

$$RSS = (\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta}))^2 = \sum_i (y_i - f(x_i; \boldsymbol{\theta}))^2 \tag{1}$$

A Sequential Quadratic Programming (SQP) gradient-based method is used to minimise the RSS (Wilson, 1967).

4 Simulations

Simulations are set up to mimic experimental waveforms, which are recorded in the main pulmonary artery in healthy and hypoxic mice (Tabima et al., 2012). The parameter set to be inferred includes: $\boldsymbol{\theta} = (k, r_1, r_2, c)$, where r_1, r_2, c are resistances and capacitor factors used to predict parameters assigned at the outlet and k is the elastance factor used to predict stiffness in all vessels. Since the parameters are on different scales, to avoid having an ill-conditioned problem, we rescale the parameters to have the same magnitude. There are certain parameter configurations that violate the model assumptions, these are marked by setting RSS to a high value (10^{10}). In this study, the RSS is calculated for pressure and we aim to find parameters that minimise the RSS. The initial parameter values examined by the SQP algorithm 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. $|\boldsymbol{\theta_i} - \boldsymbol{\theta_{i+1}}| < 1e - 11$.

5 Results and Discussion

Regardless of the initial value, the algorithm converges towards the same parameter values for both the healthy and hypoxic mouse. Figure 1 shows our optimised pressure waveform, plotted along the measured and the reference pressure. 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)). As 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 hypothetise 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. Therefore, future work could include improvements in the fluids model and inferring geometry measurements. Finally, we also aim to apply our statistical methods presented here 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 (SoftMech), 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.

References

Bratley, Bennett (1988), Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator. *ACM Trans Math Softw*, **14(1)**, 88 – 100.

Qureshi, Haider, Chesler, Olufsen (2017), Simulating effects of hypoxia on pulmonary haemodynamics in mice. In: *Proc CMBE*. 271 – 274.

Tabima, D.M. et al. (2012), Persistent vascular collagen accumulation alters hemodynamic recovery from chronic hypoxia. *J. Biomech*, **45(5)**, 799 – 804.

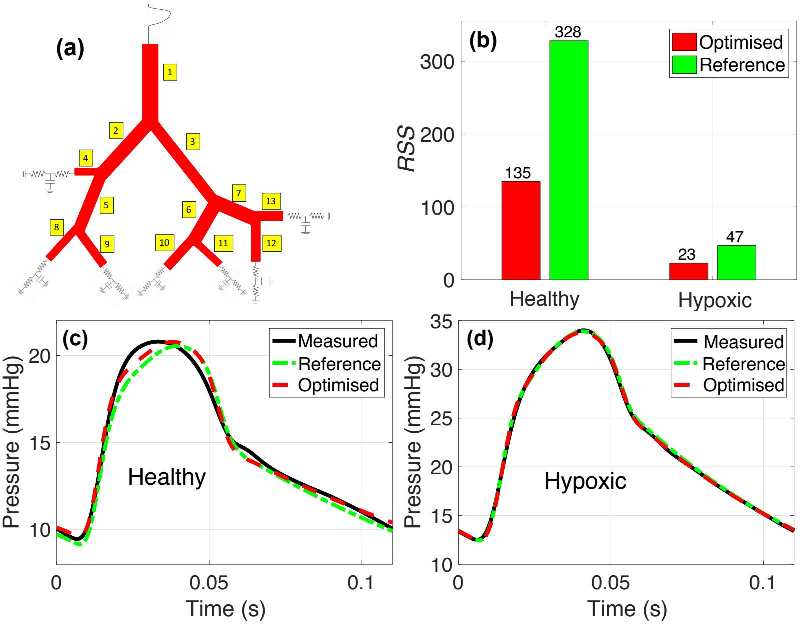


FIGURE 1. (a): The arterial network for the fluid dynamical model, (b): comparison of RSS between reference and optimised pressure simulations, (c) & (d): comparison of simulated pressure using reference and optimised parameters values for the healthy and hypoxic mice.

Wilson, R.B. (1967). *Simplicial Method for Convex Programming.* PhD thesis, Harvard University.

Exploring brain-behaviour interactions during auditory oddball detection with generalized additive modeling

Toivo Glatz[1][2], Wim Tops[1][2], Natasha Maurits[2][3], Ben Maassen[1][2]

[1] Center for Language and Cognition (CLCG), University of Groningen, The Netherlands
[2] Research School of Behavioural and Cognitive Neurosciences (BCN), University of Groningen, The Netherlands
[3] Department of Neurology, University Medical Center Groningen (UMCG), The Netherlands

E-mail for correspondence: `t.k.glatz@rug.nl`

Abstract: We conducted an auditory oddball experiment with Dutch speaking first grade children using electroencephalography. While conventional analyses usually involve averaging over items, subjects, and time windows, we used generalized additive modeling with single trial data and explored possible interactions with continuous behavioral predictors.

Keywords: ERP; MMN; LDN; GAM.

1 Background and Objective

The mismatch negativity (MMN) is an event related potential (ERP) component, which is widely considered to be an indicator of auditory discriminatory capabilities (Näätänen et al., 2007). The MMN has been shown to correlate with behavioral judgements, as well as to reflect behavioural training effects over time (Lovio et al., 2012). In addition, the MMN acquired from newborns was predictive for their reading fluency many years later (Leppänen et al., 2010). The MMN usually peaks around 100-230ms after stimulus deviancy in an oddball paradigm, and it is sometimes followed by a long-lasting component named the late discriminative negativity (LDN) from 250ms onwards (see Figure 1). As the nature of these two components is not fully understood, the latter is also sometimes called MMN; alternatively, both together may be referred to as mismatch response (MMR).

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.

In this work, which is part of a bigger study, we are investigating possible interactions of the MMR components with other behavioral (e.g. IQ, phonological awareness (PA)) and biographical measures (e.g. age, gender, handedness, familial risk for reading difficulties) of Dutch speaking first grade children using a novel approach with generalized additive modeling.

2 Design and Methods

We recruited 40 first grade children who played a computer game, which either trained reading related skills such as PA (N=20), or simple arithmetic (N=20), on a daily basis for six weeks. Behavioral tests measured reading-related abilities before, during, and after the gaming. Electroencephalography (EEG) was recorded before and after the training to study the MMR to syllabic speech sounds as a potential correlate of game-based phoneme discrimination improvement.
The EEG data were preprocessed using the EEGLAB toolbox for Matlab, applying bandpass filtering (0.3 - 30Hz), re-referencing to the average of the mastoids, ICA-based ocular artifact correction, as well as pre-stimulus baseline correction. For the analysis of the single trial ERP data, we used generalized additive modeling (GAM; mgcv package in R; Wood, 2006). Subsequently, we tested which behavioural measures interact with the MMR to consonant, vowel and duration deviants of speech sounds in a stepwise procedure. Starting with a basic model including smooths over time per condition, we added a random effects structure for participants, and subsequently tested all behavioural measures as main effects and interactions.

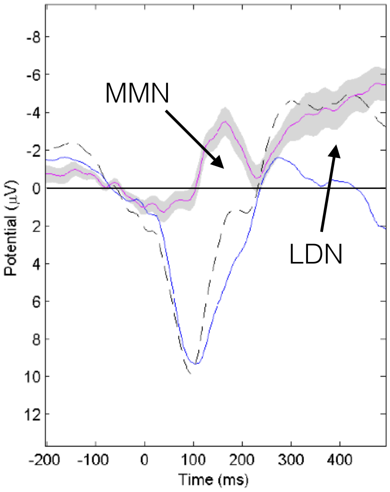


FIGURE 1. Conventional grand-average ERP analysis averaged over items and participants. Solid line: standard stimulus; dashed line: deviant stimulus; red line: mismatch response; shaded area: confidence interval for mismatch response.

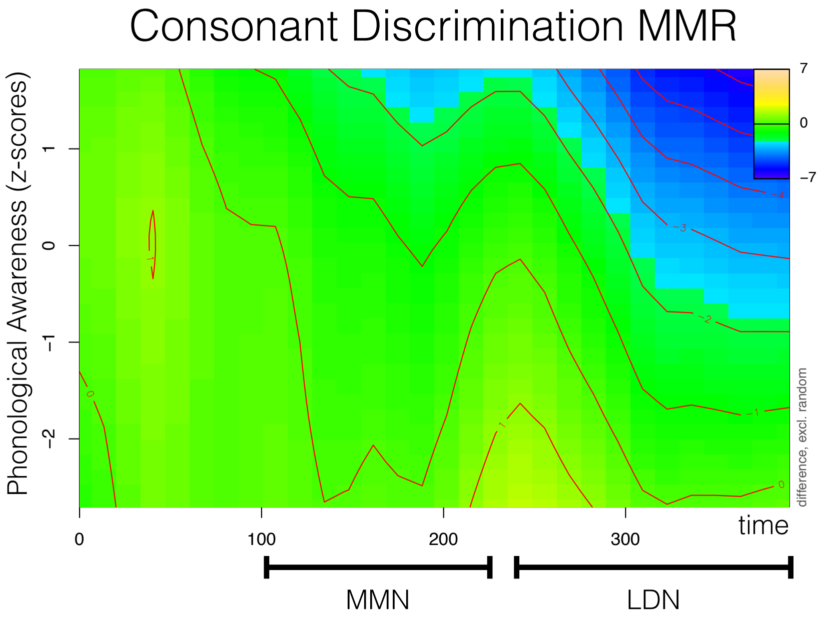


FIGURE 2. GAM predicted tensor smooth for the nonlinear interaction of the consonant mismatch response (deviant subtracted from standard, color-coded) and phonological awareness.

3 Results and Discussion

We found a significant nonlinear 3-way interaction of PA skills and condition (standard vs. deviant syllable) over time, when predicting single trial ERP amplitudes (see Figure 2). Interestingly, this interaction was only found for the consonant deviancy condition (i.e. /ti/ and /pe/ changing to /pi/ and /te/, respectively), but not for the vowel or duration change. While we expected to find an interaction of MMN with PA skills, we saw that only the LDN is modulated by PA skills. More specifically, only the LDN to consonant change was modulated by PA skills, which is interestingly also the condition with the smallest MMN response.

It is not yet fully understood what cognitive processes the early and late components of the MMR reflect. To some extent, this may be due to the fact that the vast majority of ERP studies rely on signal averaging over items, subjects, as well as time points (i.e., when averaging over time windows instead of considering single sampling points) when analyzing the MMR together with behavioural data. Generalized additive modeling has recently started to be used with EEG data (e.g. Meulman et al., 2015) as it allows to consider single trial data, and to integrate continuous behavioural predictors into the analysis in a way that had not been possible before. This should allow researchers to further narrow down the functions which certain ERP components reflect.

Acknowledgments: We are very thankful to Martijn Wieling, Iris te Stroete, Maaike Rietdijk, Jenny Aussum, and all the children who participated in our study!

References

Leppänen, P. H., Hämälainen, J. A., Salminen, H. K., Eklund, K. M., Guttorm, T. K., Lohvansuu, K., ... & Lyytinen, H. (2010). Newborn brain event-related potentials revealing atypical processing of sound frequency and the subsequent association with later literacy skills in children with familial dyslexia. *Cortex, 46*(10), 1362-1376.

Lovio, R., Halttunen, A., Lyytinen, H., Näätänen, R., & Kujala, T. (2012). Reading skill and neural processing accuracy improvement after a 3-hour intervention in preschoolers with difficulties in reading-related skills. *Brain research, 1448*, 42-55.

Meulman, N., Wieling, M., Sprenger, S. A., Stowe, L. A., & Schmid, M. S. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. *PloS one, 10*(12), e0143328.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology, 118*(12), 2544-2590.

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

Interval Shrinkage Estimation Parameter of Burr type III Distribution

Parviz Nasiri

[1] Department of Statistics, University of Payam Noor, 19395-4697 Tehran, I.R. Iran

Abstract: In this article, the Burr type III distribution is defined and for unknown parameter, different point estimates are derived. In shrinkage estimation, we believe that the parameter belong to a finite interval. So we propose for such situations an interval shrinkage approach which combines in a coherent way an unbiased conventional estimator and non-sample information about the range of plausible parameter values. At the end, we conclude those interval shrinkage estimators are better than the maximum likelihood estimation.

Keywords: Maximum likelihood estimation, Interval information, Shrinkage estimation, Means square error

1 Introduction

Burr in 1942 introduced a family of the twelve cumulative distribution functions for modeling lifetime data. In 2011, Asgharzadeh and Fallah done estimation and prediction for exponential family of distributions based on records. The two important members of the family are Burr type III and XII. Burr type III distribution allows for a wider region for skewness and kurtosis plane, which covers several distributions including the log-logistics, and the Weibull and Burr type XII distributions. Let us consider the cumulative distribution function, probability density function of the Burr type III distribution are given respectively by,

$$F(x;\theta,c) = (1-x^{-c})^{-\theta} \quad , \quad x>0 \quad , \quad \theta>0 \quad , \quad c>0 \quad (1)$$

$$f(x;\theta,c) = \theta c x^{-(c+1)}(1-x^{-c})^{-(\theta+1)} \quad , \quad x>0 \qquad (2)$$

where the parameters $c > 0$ and $\theta > 0$ are the shape parameters of the distribution. Our approach is closely related to the interval shrinkage estimators. The idea of shrinkage is providing a balanced trade-off between a conventional estimator and a shrinkage target (see Pandey, B. N. (1983)). Recently Nasiri (2016)

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

introduces Interval shrinkage estimators for location parameter of the exponential distribution.
The rest of the paper is organized as follows. In section 2 and 3, we deal with the maximum likelihood and feasible interval shrinkage estimation of parameter. In section 4, we compare the Bias and MSE of the estimates empirically.

2 Maximum Likelihood Estimation

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n of the Burr type III distribution. The log-likelihood function for based on a random sample is

$$l(\theta) = n\log(\theta) - (\theta+1)\sum_{i=1}^{n}\log(1+x_i^{-c}) + n\log(c) + (c+1)\sum_{i=1}^{n}\log(x_i) \quad (3)$$

The maximum likelihood estimate of the unknown parameter is obtained by maximizing the log-likelihood function $l(\theta)$ with respect to θ. The likelihood equation which is obtained from the derivatives of $l(\theta)$ with respect to the parameter, become

$$\frac{\partial l(\theta)}{\partial\theta} = \frac{n}{\theta} - \sum_{i=1}^{n}\log(1+x_i^{-c}) \quad (4)$$

The maximum likelihood estimator $\hat{\theta}$ of θ can be obtained by solving the likelihood equation $\frac{\partial l(\theta)}{\partial\theta} = 0$, so

$$\hat{\theta} = \frac{n}{\sum\limits_{i=1}^{n}\log(1+x_i^{-c})} \quad (5)$$

3 Feasible interval shrinkage estimator

Let $(X_1, X_2, \ldots, X_n)$ be the random sample of size n taken form Burr type III distribution. Proposed shrinkage estimator and its properties following Thompson (1968), a shrinkage estimator for the parameter θ and θ_g, a guess values of θ is available, is defined as

$$\hat{\theta}_{sh} = \theta_g + \omega(\hat{\theta} - \theta_g) \quad (6)$$

A shrinkage factor is defined based on guessed value. Let $\hat{\theta}_{sh}$ be the shrinkage estimation of scale parameter. Then

$$\hat{\theta}_{sh} = \theta_g + \omega(\hat{\theta} - \theta_g)$$

To find we have to consider MSE of estimator as:

$$\begin{aligned} MSE(\hat{\theta}_{sh}) &= \{(\hat{\theta}_{sh} - \theta)^2\} \\ &= \omega^2 MSE(\hat{\theta}, \theta) + (\omega-1)^2(\theta-\theta_g)^2 + 2\omega(\omega-1)(\theta-\theta_g)E(\hat{\theta}-\theta) \end{aligned}$$

Now, we have to minimize the $(\hat{\theta}_{sh})$,

$$\frac{dMSE(\hat{\theta}_{sh})}{d\omega}=0$$

$$\omega=\frac{(\theta_g-\theta)^2+(\theta_g-\theta)E(\dot{\theta}-\theta)}{MSE(\hat{\theta}-\theta)+(\theta_g-\theta)^2+2(\theta_g-\theta)E(\hat{\theta}-\theta)} \quad (7)$$

In 2011, Golosony and Liesenfeld show the shrinkage estimator towards the interval $\theta \in [\theta_0,\theta_1] \subset \Theta$ for unbiased conventional sample estimator of $\hat{\theta}$ with $E(\hat{\theta})=\theta$ is given by

$$\tilde{\theta}_{sh}=\hat{\theta}+\sqrt{V(\hat{\theta})}\cdot\frac{\theta-\hat{\theta}}{\theta_1-\theta_0}\left[\arctan\left(\frac{\theta_1-\theta}{\sqrt{V(\hat{\theta})}}\right)-\arctan\left(\frac{\theta_0-\theta}{\sqrt{V(\hat{\theta})}}\right)+\frac{V(\hat{\theta})}{2(\theta_1-\theta_0)}\ln\left[\frac{V(\hat{\theta})+(\theta_1-\theta)^2}{V(\hat{\theta})+(\theta_0-\theta)^2}\right]\right]$$

and

$$E(\tilde{\theta}_{sh})=\hat{\theta}+\frac{V(\hat{\theta})}{2(\theta_1-\theta_0)}\ln\left[\frac{V(\hat{\theta})+(\theta_1-\hat{\theta})^2}{V(\hat{\theta})+(\theta_0-\hat{\theta})^2}\right] \quad (8)$$

For $E(\hat{\theta})=\theta$, we have

$$\tilde{\theta}_{sh}=\hat{\theta}+\frac{V(\hat{\theta})}{2(\theta_1-\theta_0)}\ln\left[\frac{V(\hat{\theta})+(\theta_1-\hat{\theta})^2}{V(\hat{\theta})+(\theta_0-\hat{\theta})^2}\right] \quad (9)$$

The estimator obtained in (9) is not linear in $\hat{\theta}$ and becomes identical to the conventional estimator $\hat{\theta}$ as the $V(\hat{\theta})$ tends to zero or as $(\theta_1-\theta_0)$ increases. It also tends to $\hat{\theta}$ as $\hat{\theta}$ moves away from the interval $[\theta_0,\theta_1]$. Furthermore if $\hat{\theta}$ is equal to the midpoint of the interval $\theta_m=(\theta_0-\theta_1)/2$, the estimator $\tilde{\theta}_{sh}$ is equal to $\hat{\theta}$. The exact stochastic properties of the feasible estimator $\hat{\theta}$, including its means and variance, are hard to derive analytically since it is a complex nonlinear function in $\hat{\theta}$. Golosony and Liesenfeld (2011) suggest to approximate $\tilde{\theta}_{sh}$ by a first-order Taylor expansion around the center of the shrinkage interval θ_m. The second derivative of $\tilde{\theta}_{sh}$ equals zero at the point θ_m, so the first-order Taylor expansion in this case is equivalent to second-order one. Since $\tilde{\theta}_{sh}$ has an inflection point at the midpoint of the interval $\hat{\theta}=\theta_m$, this expansion provides a reasonable approximation to $\tilde{\theta}_{sh}$ within the shrinkage interval. So it defined an alternative useful feasible interval shrinkage estimator with mean and variance that can be derived in a closed form. Let the half-length of the shrinkage interval be denoted by $\theta_d=(\theta_1-\theta_0)/2$. The second-order Taylor approximation of the estimator at the point θ_m is linear in $\hat{\theta}$ and is given by

$$\tilde{\tilde{\theta}}_{sh}=\hat{\theta}\left[1-\frac{V(\hat{\theta})}{v(\hat{\theta})+\theta_d^2}\right]+\theta_m\frac{V(\hat{\theta})}{V(\hat{\theta})+\theta_d^2}\theta \quad (10)$$

With mean and variance as

$$E(\tilde{\tilde{\theta}}_{sh})=\theta-(\theta-\theta_m)\frac{V(\hat{\theta})}{V(\hat{\theta})+\theta_d^2}\theta\ ,\ V(\tilde{\tilde{\theta}}_{sh})=V(\hat{\theta})\left(1-\frac{V(\hat{\theta})}{V(\hat{\theta})+\theta_d^2}\right)^2$$

It can be shown that $MSE(\tilde{\tilde{\theta}}_{sh})\leq V(\hat{\theta})$.

4 Numerical Study

To have some idea about mean square error (MSE) of maximum likelihood ($\hat{\theta}_{mle}$) and shrinkage interval ($\breve{\theta}$) estimation, we perform sampling experiments using R software and the results are shown in Tables 1 to Table 3. The MSE of both estimators decrease when sample size increase. Comparing the MSE of two estimators, states that the interval shrinkage estimator works better than maximum likelihood estimator.

References

Asgharzadeh, A. and Fallah, A. (2011). Estimation and prediction for exponentiated family of distributions based on records. *Communication in statistics, theory and methods*, 40: 68-83.

Burr, I. W. (1942). Cumulative frequency distribution, *Annals of Mathematical Statistics*, 13, pp. 214-232.

Nasiri, P. (2016). Interval shrinkage estimators for location parameter of the exponential distribution. *Research Journal of Applied Sciences*, 5(11), 229-231.

Pandey, B. N. (1983). Shrinkage estimation of the exponential scale parameter. *IEEE Trans. Reliability* 32.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *J. Amer. States. Assoc.* 63, 113-123.

Fitting binomial mixtures to RNA-seq data to find discrete gene regulation mechanisms

Adriaan van der Graaf[1]

[1] Universitair Medisch Centrum Groningen, Netherlands

E-mail for correspondence: adriaan.vd.graaf@gmail.com

Abstract: Gene expression measurements are becoming more and more relevant to detect how cellular processes operate. In this paper binomial mixtures are fit to RNA-seq data, searching for certain discrete gene regulation mechanisms. My results suggest that gene expression regulation is performed through a continuous process and not by a small number of discrete ones. One other option could be that the data used was of insufficient quality to provide accurate results.

Keywords: RNA-seq; binomial mixtures

1 Introduction

Next generation RNA sequencing (RNA-seq) is the current gold standard of gene expression measurements. Unlike microarray gene expression measurements, RNA-seq is able to detect novel transcripts. Counts of detected transcripts in RNA-seq are a proxy of cellular gene expression, but because of the count nature of the data, statistical techniques that deal with microarray expression are not always suitable because of their assumption of Gaussian distributed data.

Variance of gene expression can only be partially explained by an organisms genetics, Lloyd-Jones et al. (2017), implying that there are multiple factors driving gene expression. Modelling possibly unknown factors that influence gene expression can be attempted through a mixture of expression distributions, each of which could represent some biological process.

This short paper describes the steps taken to pre-process Geuvadis RNA-seq data from Lappalainen et al. (2013), fitting a binomial mixture distribution using a custom EM algorithm and finally doing model selection to find the best performing model and discussing biological implications.

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 Methods

2.1 Data

The Geuvadis mRNA-seq dataset from Lappalainen et al. (2014) was used. This consisted of 660 human individuals of which RNA-seq data was available from lymphoblastoid cell lines, the total sequencing depth had a median of 4.97 million transcripts per individual. For every gene on the first chromosome, transcripts were counted and used as input data for the EM algorithm.

2.2 EM algorithm

The rationale for modeling RNA-seq using the binomial distribution is the following: If there are n_i RNA-seq transcripts detected in individual i, then, finding a number k_i RNA-seq transcripts from some gene, can be modeled through a binomial process, with some probability π. Where a success is considered as finding a transcript on the gene and a failure if the transcript is not detected on the gene. The biological expectation in this study is that total expression can be considered a mixture of binomial distributions, each of which may represent some biological process.

The EM algorithm that is often used to estimate Gaussian mixtures can be adapted to fit other probability functions, in our case the binomial. I consider some vector of i.i.d. transcripts $\mathbf{k} = (k_1, ..., k_N)$ for some gene, and a vector of total transcripts for the individual $\mathbf{n} = (n_1, ..., n_N)$ where N is the number of individuals of which RNA-seq data is available. Analyzing each gene separately, the log likelihood is maximized for a mixture of $M \in \{1, 2, ..., 50\}$ binomial distributions.

$$l(\pi|\mathbf{p}, \mathbf{k}, \mathbf{n}, \mathbf{\Delta}) = \sum_{m=1}^{M} p_m \sum_{i=1}^{N} \Delta_{i,m} \cdot \binom{n_i}{k_i} \cdot \pi_m^{k_i} \cdot (1 - \pi_m)^{n_i - k_i} \tag{1}$$

Where $\mathbf{k}$ and $\mathbf{n}$ are N dimensional vectors with entries k_i and n_i respectively. π and $\mathbf{p}$ are an M dimensional vector with entries $\pi_m \in [0, 1]$ and p_m respectively, where $\sum_{m=1}^{M} p_m = 1$, represents the relative proportion of a distribution in the mixture. $\mathbf{\Delta}$ is an N-by-M matrix with each element $\Delta_{i,m} \in \{0, 1\}$, indicating if individual i is part of distribution m.

An EM algorithm for binomial mixtures was used to estimate the distribution parameters, the EM algorithm is initialized four times, and after convergence, the model with the maximum log likelihood is retained. Initialization parameters were: $\pi_m \sim U(ARGMIN_i(k_i/n_i), ARGMAX_i(k_i/n_i))$ and $p_m = 1/M$. In the E step, membership probability of an individual is determined m: $\gamma_{i,m} = \frac{Bin(n_i,k_i,\hat{\pi}_m)}{\sum_{m=1}^{M} Bin(n_i,k_i,\hat{\pi}_m)}$. In the M step, the maximum likelihood estimators are $\hat{p}_m = \frac{\sum_{i=1}^{N} \gamma_{i,m}}{N}$ for probability of membership and $\hat{\pi}_m = \frac{\sum_{i=1}^{N} \gamma_{i,m} k_i}{\sum_{i=1}^{N} \gamma_{i,m} n_i}$ for the binomial proportion. The log likelihood and BIC is determined, by taking the highest $\gamma_{i,m}$ for each individual, and setting it's respective $\Delta_{i,m} = 1$. The algorithm is repeated until there is less than a 10^{-10} difference between new and old $\hat{\pi}_m$ parameters, or if the maximum number of iterations (1000) has been reached.

3 Results

Some genes showed non-convergence, which was mostly attributable to numerical precision problems, when there were a low number of transcripts. If only genes were considered median($\mathbf{k}$) > 10, 89.2 % of initializations reached convergence. The BIC of four representative genes are shown in figure 1.

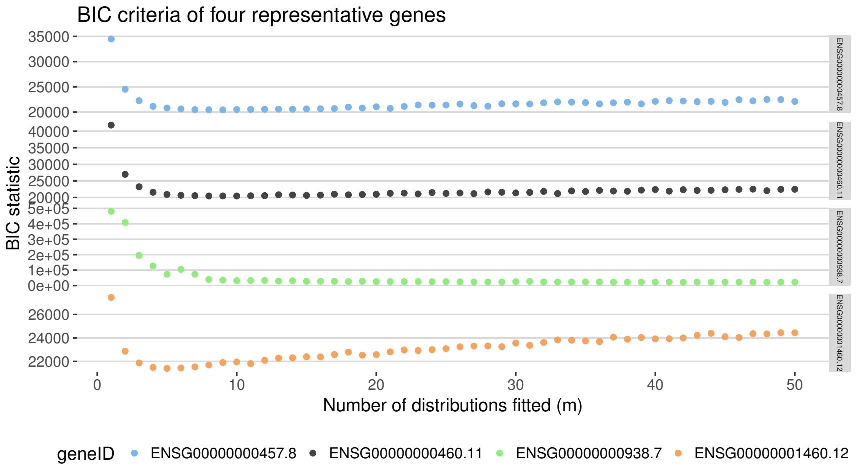


FIGURE 1. The BIC statistics of a number of genes, over a different number of mixture distributions. The Genes are identified by their ENSEMBL ID

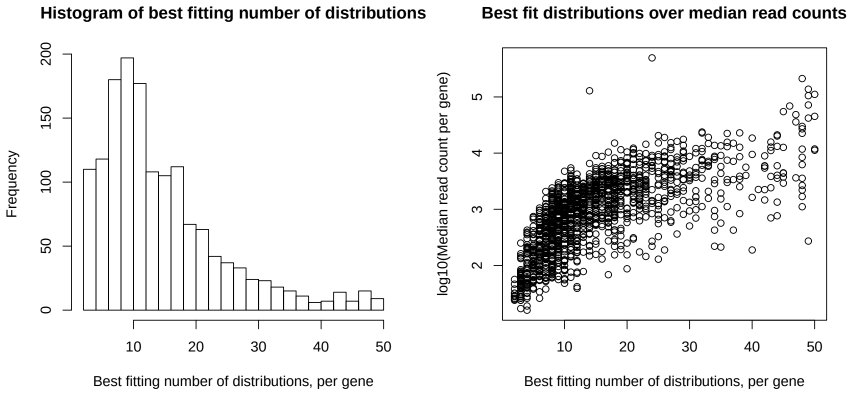


FIGURE 2. Left: A histogram of the optimal number of mixtures, based on minimized BIC. Right: A scatter plot of optimal number of mixtures compared to median transcript count of the gene.

A histogram showing the best fitting number of mixtures is shown in figure 2

(left). This would indicate that most genes are under control of a small number of regulatory processes.
However, when plotted against the median transript count of a gene, there is a clear correlation between the number of best fitting mixtures and the median transcript count of a gene (figure 2, right). This indicates that the accurate detection of binomial mixtures is dependent on the number of reads $\mathbf{k}_j$ from a gene.
Simulations were also done to validate the EM algorithm. Reads were simulated according to convergence parameters from the algorithm, based on parameters found in converging EM algorithms. In all cases, the algorithm converged to the same solutions

4 Conclusion and discussion

This short paper details the steps taken to fit up to 50 different binomial mixtures to RNA seq data. Fitted using an EM algorithm. Evidence for a small and discrete number of distributions which regulate gene expression was not found. Perhaps because of sample size issues, but it could also indicate that gene expression is regulated by more continuous mechanisms.
One thing to note is that RNA-seq data can be overdispersed, so using a binomial distribution for modelling is not always correct, Sun (2011) therefore used the negative binomial distribution for regression of expression, using an organisms genotype as a dependent variable. Adapting this to the negative binomial could provide more accurate results.

Acknowledgments: Special Thanks to Marco Grzegorczyk, for much needed guidance, in this unfamiliar field.

References

Lloyd-Jones, L. R. et al. (2017). The genetic architecture of gene expression in peripheral blood *American Journal of Human Genetics* **100** 228 – 237.

Lappalainen, T. et al. (2014). Transcriptome and genome sequencing uncovers functional variation in humans *Nature* **501** 506 – 511.

Sun, W (2011). A Statistical Framework for eQTL Mapping Using RNA-seq Data *Biometrics* **68** 1 – 11.

Modelling weekly temperatures in Eelde using Bayesian Structural Time Series

Hung Chu [1]

[1] Johann Bernoulli Institute for Mathematics and Computer Science, the Netherlands

E-mail for correspondence: hung_c@hotmail.com

Abstract: In this paper we study the weather temperature estimation performance of a Bayesian approach based on a structural time series model. The approach makes use of Monte Carlo Markov Chain (MCMC) sampling to approximate the joint distribution of the model parameters. For our empirical analysis we consider data with a strong seasonal pattern, namely weekly weather data.

Keywords: Bayesian Structural Time Series Model; MCMC; Estimation.

1 Introduction

Traditional methods to estimate time series data are the Multiple Linear Regression (MLR) and the autoregressive integrated moving-average (ARIMA) model. We apply the *Bayesian Structural Time Series* (BSTS) model, which was introduced by Scott et al. (2013), for the estimation of weather temperature using contemporaneous predictors. The BSTS allows for 1. decomposing the time series data into several latent components that can describe the underlying dynamics of the data, such as trend, seasonality and regression, 2. variable selection and 3. Bayesian model averaging. The structural time series model of an observation y_t at time t is

$$\begin{aligned} y_t &= \mu_t + \gamma_t + \beta^T x_t + \varepsilon_t, \\ \mu_t &= \mu_{t-1} + \delta_{t-1} + \eta_{1t} \\ \delta_t &= \delta_{t-1} + \eta_{2t} \\ \gamma_t &= -\sum_{s=1}^{S-1} \gamma_{t-13s} + \eta_{3t}, \end{aligned} \tag{1}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ for some positive constant σ_ε^2, $\eta_t \equiv (\eta_{1t}, \eta_{2t}, \eta_{3t})^T \sim \mathcal{N}(0, Q)$ for some diagonal variance matrix Q, x_t a vector of contemporaneous or lagged

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

predictors and S the number of seasons. Furthermore, each season is assumed to have a duration of exactly 13 weeks. The model is decomposed into four components: μ_t, γ_t, $\beta^T x_t$ and ε_t respectively describe the general trend, seasonal pattern, regression effect and error. The model parameters are σ_ε^2, β and Q.

1.1 Spike-and-slab regression

The BSTS model incorporates the spike-and-slab regression for variable selection. Let K denote the number of regressors. For $k = 1, \ldots, K$, let

$$\gamma_k = \begin{cases} 1 & \text{if } \beta_k \neq 0 \\ 0 & \text{if } \beta_k = 0 \end{cases},$$

and assume that the prior $p(\gamma)$ is an independent Bernoulli:

$$\gamma \sim \prod_{k=1}^{K} \pi_k^{\gamma_k} (1 - \pi_k)^{1-\gamma_k}.$$

Intuitively, π_i is the probability of inclusion of β_i.
Let β_γ denote the subset of β for which $\beta_k \neq 0$. The spike-and-slab prior on the regression coefficients is then given by

$$p(\beta, \sigma_\varepsilon^2, \gamma) = p(\beta_\gamma | \gamma, \sigma_\varepsilon^2) p(\sigma_\varepsilon^2 | \gamma) p(\gamma).$$

See Scott et al. (2013) for the conditional priors $p(\sigma_\varepsilon^2|\gamma)$ and $p(\beta_\gamma|\gamma, \sigma_\varepsilon^2)$.

1.2 Conditional posterior distributions

Let $y = (y_1, \ldots, y_T)^T$ and $\alpha = (\alpha_1, \ldots, \alpha_T)^T$, where T is the size of the data. Observe that the model in (1) can be represented as a state space time series model:

$$\begin{aligned} y_t &= Z^T \alpha_t + \varepsilon_t \\ \alpha_t &= T\alpha_{t-1} + R\eta_{t-1}. \end{aligned}$$

Therefore we can use the Kalman filter and smoother to obtain the posterior distribution of the states given the data, i.e. $p(\alpha|y, \theta)$, where θ is the set of model parameters. However, we cannot just generate each α_t from $p(\alpha_t|y, \theta)$, because we have to take into account the autocorrelation between α_t and α_{t+1}. For this we use the simulation smoother of Durbin et al. (2002) to generate samples from $p(\alpha|y, \theta)$.
The conditional posterior $p(\gamma|y, \alpha)$ is obtained by a Gibbs sampling algorithm, where each γ_i is drawn conditional on every other element of γ. The conditional posteriors $p(Q|\alpha)$, $p(\sigma_\varepsilon^{-2}|y, \alpha, \gamma)$ and $p(\beta|y, \alpha, \sigma_\varepsilon^{-2}, \gamma)$ are closed-form, from which we can generate samples directly. See Scott et al. (2013) for the details.

2 Model estimation

Let $\phi \equiv (\theta, \alpha)^T$ and let ψ denote the set of parameters other than β and σ_ε^2. Considering an initial $\theta = \theta^{(0)}$ generated from the prior distributions, we obtain a stationary distribution $p(\phi|y)$ using MCMC as follows.

1. Simulate α from $p(\alpha|y, \theta)$ using the simulation smoother.
2. Simulate ψ from $p(\psi|y, \alpha, \beta, \sigma_\varepsilon^2)$.
3. Simulate β and σ_ε^2 from $p(\beta, \sigma_\varepsilon^2|y, \alpha, \psi)$.

By cycling through Steps $1 - 3$ for M times, we obtain a sequence of MCMC draws $\phi^{(1)}, \ldots, \phi^{(M)}$. The first m samples (also known as *burn-in* samples) may not be representative for the target posterior distribution and hence will be discarded. The remaining sequence of MCMC draws is used to estimate the posterior distribution of y_t by means of (1) and the fit by:

$$\hat{y}_t = \frac{1}{M - m} \sum_{i=m+1}^{M} \hat{y}_t^{(i)}.$$

3 Empirical analysis

For the empirical analysis we use weekly weather data from Eelde (the Netherlands) from 20–12–2011 until 20–12–2016, provided by Koninklijk Nederlands Meteorologisch Instituut (KNMI). We have y_t denoting the average weekly temperature and x_t containing nineteen candidate predictors, such as average wind speed, relative humidity, global radiation, but also last week's temperature. Furthermore, we perform $M = 5,000$ MCMC iterations, where the first $m = 1,000$ samples are considered for burn-in.

TABLE 1. R^2 and R^2_{adj} of the fitted models.

Approach	BSTS	BSTS$_{sub}$	MLR	MLR$_{sub}$	ARIMA(1,1,1)
R^2	0.9375	0.9338	0.9191	0.9110	0.7678
R^2_{adj}	0.9323	0.9319	0.9127	0.9088	0.7669

The set of predictors with inclusion probability greater than 0.500 are *reference evapotranspiration* (1.000), *global radiation* (0.999), *last week's temperature* (0.995), *minimum relative humidity* (0.655) and *average relative humidity* (0.531). We compare the fit of the BSTS model with all nineteen predictors to the BSTS model with the set of five predictors above (BSTS$_{sub}$), MLR with all nineteen predictors, MLR with the set of five predictors above (MLR$_{sub}$) and the ARIMA(1,1,1) model. The estimation performance of the models in terms of the (adjusted) coefficient of determination (R^2 and R^2_{adj}) can be found in Table 1. The result in the table indicates that BSTS and even BSTS$_{sub}$ provide a better fit than the traditional methods. The fitted temperature and individual components by BSTS (with all predictors) are depicted in Figure 1. The true temperature is included in the first subplot (red dashed line).

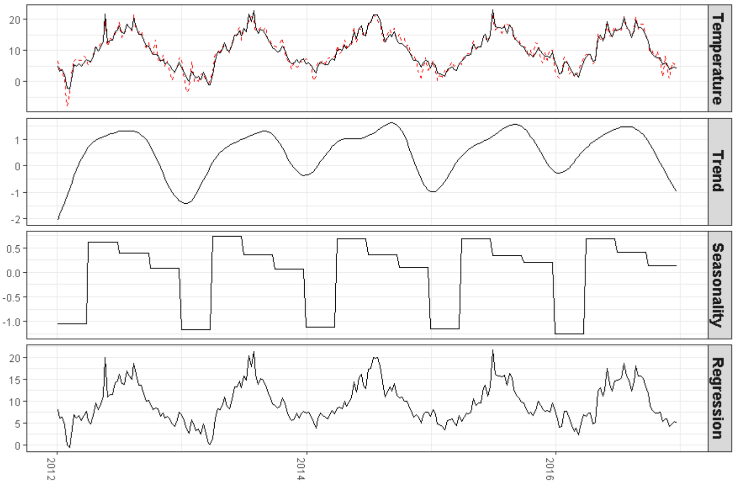


FIGURE 1. Plot of the true (red dashed line) and fitted temperature by the BSTS model as well as the individual fitted components (black solid lines).

4 Conclusion

Considering a dataset with a strong seasonal pattern, we empirically showed that the BSTS model (even with the selected subset of predictors) provides a better fit than the MLR and ARIMA(1,1,1) model, based on R^2 and R^2_{adj}. The BSTS model has the advantage that it allows to visualize and study the underlying components of the time series data as well as variable selection. Moreover, this study demonstrates that the BSTS model performs the variable selection and the fitting of weather temperature very well.

References

Durbin, J. and Koopman, S.J. (2002). A Simple and Efficient Simulation Smoother for State Space Time Series Analysis). *Biometrika*, **89**, 603 – 613.

Scott, S.L. and Varian, H.R. (2013). Bayesian Variable Selection for Nowcasting Economic Time Series. *NBER Working Paper 19567.*

A comparison between GMLVQ and Random Forests

Thomas Nijman[1]

[1] Rijksuniversiteit Groningen, Netherlands

E-mail for correspondence: thomasrutger@gmail.com

Abstract: This paper presents a comparison of two classification algorithms, the first one is an advanced prototype based classifier known as General Matrix Learning Vector Quantization. Which is an extension of Learning vector Quantization and is a relatively new algorithm. The second algorithm is the Random Forest algorithm which is a very popular classification method that has proven to be a very reliable classifier. It was found that GMLVQ gave slightly better results than the Random Forest algorithm for the first two datasets. On the third dataset however, the Random Forest algorithm performed much better (13 % of test examples were misclassified by GMLVQ as opposed to only 7 % by the Random Forest algorithm).

Keywords: GMLVQ; Random Forest; Comparison.

1 Introduction

This section gives some theoretical background on GMLVQ and Random Forests. As the Random Forest classifier is already well known, only a brief description is given.

1.1 GMLVQ

GMLVQ (General Matrix LVQ) [Schneider et. al., 2009] is an extension of the prototype based classification method Learning Vector Quantization, in which a set of labeled prototypes are fitted to the data. A new data point with an unknown class label is then assigned to the class of the closest prototype. Any distance measure can be used, but usually Euclidean distance is used. GMLVQ extends LVQ by using an advanced distance measure (eq. 1), using a matrix Λ of adaptive relevances, depicting the relevance for each feature as well as correlations between features.

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

$$d^{\Lambda}(\boldsymbol{\xi}, \boldsymbol{w}) = (\boldsymbol{\xi} - \boldsymbol{w})^{\mathrm{T}} \Lambda (\boldsymbol{\xi} - \boldsymbol{w}) \tag{1}$$

The matrix Λ transforms each feature vector to an alternative feature space which provides more discriminative power. Consider the closest 2 prototypes to a training sample $(\boldsymbol{\xi}, y)$ with $c(\boldsymbol{w}_J) = y$ and $c(\boldsymbol{w}_J) \neq y$. The updating of the prototypes is done by minimising a cost function:

$$E_{GMLVQ} = \sum_{i=1}^{P} \Phi(\mu_i^{\Lambda}) \text{ with } \mu_i^{\Lambda} = \frac{d_J^{\Lambda}(\boldsymbol{\xi}_i) - d_K^{\Lambda}(\boldsymbol{\xi}_i)}{d_J^{\Lambda}(\boldsymbol{\xi}_i) + d_K^{\Lambda}(\boldsymbol{\xi}_i)} \tag{2}$$

The model is trained by minimising E_{GLVQ} with respect to the model parameters [Schneider et. al., 2009].

The diagonal entries of Λ reflect the importance of each feature, while off-diagonal elements account for correlations between features. Feature vectors can be projected onto the 2 eigenvectors of Λ corresponding to the 2 largest eigenvectors of Λ, Giving a 2D visualisation of the dataset. This makes a GMLVQ classifier easy to interpret.

1.2 Random Forest

Random Forest [Breiman, 2001] is an ensemble classification method that builds many decision trees and combines them to form a new classifier. Each decision tree is trained on a bootstrap sample of the training data. It eliminates some of the key issues of decision trees. In particular, decision trees tend to be non-robust and may lack accuracy when compared to other approaches [Gareth et. al., 2015]. Classification can be done by outputting the class that is most frequently given as output by the individual trees.

2 Method

Comparing two classification methods is not easy as classification methods can have many different parameters. Some parameters such as the learning rates for GMLVQ have been heuristically chosen. The number of prototypes per class was set to 1 for all the experiments.

For GMLVQ the intitial prototype learning rate used was 1 and the initial relevance Matrix learning rate used was 2. Furthermore, the learning rates are reduced by a factor of 1.5 after every training epoch (1 ”sweep” through the training set). 100 training epochs were used, this seemed to be a sufficient amount of time for the prototypes to converge in all the experiments without showing overfitting effects.

For Random Forests, the number of trees used was 128. 128 was selected heuristically as it was found that using more trees did not do much for the classification performance in the experiments. It has also been shown that the error does not decrease linearly as the number of trees increases [Oshiro et. al., 2012].

10-fold cross validation was used for the experiments. Perhaps not a necessity, but it could help in avoiding random splits that are beneficial for one classifier.

3 Results

The experiments were done using three different datasets. They were all obtained from the "UCI Machine Learning Repository" [M. Lichman, 2013]. The first dataset is the "Wine" dataset. Which is a 2 class problem, containing 178 12 dimensional feature vectors. The second dataset is the "Wisconsin Diagnostic Breast Cancer" dataset. A 2 class problem with 569 30 dimensional feature vectors. The third dataset is the "Image Segmentation" dataset. This is a 7 class problem with 2310 19 dimensional feature vectors. The third feature has been removed as it is constant over all the feature vectors to obtain 18 dimensional feature vectors.

The experiment described in "Method" was repeated 10 times for all the datasets for both GMLVQ and Random Forests. The test errors over all the validation runs were averaged and can be seen in Table 1. These numbers were obtained by dividing the number of incorrect classifications by the total number of test examples.

4 Discussion

When looking at Table 1 it can be seen that GMLVQ shows robust performance when compared to Random Forests. The error for the first 2 datasets is relatively low. It is even a little bit lower than the error for Random forests. When looking at the Image Segmentation dataset, Random Forest does perform significantly better, however. This problem is also somewhat harder due to the fact that it has 7 classes. Performance here for GMLVQ might still be optimized by tweaking the learning parameters. Especially increasing the number of prototypes might increase classification performance as some classes may be spread out over a subspace that is too large for 1 prototype.

5 Conclusion

In this short study it can be seen that GMLVQ is a suitable classifier for certain problems. Rivaling Random Forests in the Wine and Breast Cancer datasets. An advantage of GMLVQ is that it is sophisticated, yet easy to interpret. One obvious downside is that there are plenty of parameters to tweak that may affect classification performance. Random Forest clearly wins when it comes to the Image Segmentation dataset, but little improvements could perhaps still be made by tweaking the parameters of GMLVQ.

TABLE 1. Test Errors

	Wine	Breast Cancer	Image Segmentation
GMLVQ	0.0107	0.0309	0.1305
Random Forest	0.0181	0.0414	0.0724

References

Breiman, L. (2001).
Random Forests.
Public Library of Science
In: Machine Learning, 45(1), 5 – 32.

Gareth, J., Witten, D., Hastie, T. and Tibshirani, R. (2015).
An Introduction to Statistical Learning.
New York: Springer.

Lichman, M. (2013).
UCI Machine Learning Repository.
url: http://archive.ics.uci.edu/ml
University of California, Irvine, School of Information
and Computer Sciences

Oshiro T.M., Perez P.S., Baranauskas J.A. (2012).
How Many Trees in a Random Forest?
Machine Learning and Data Mining in Pattern Recognition:
8th International Conference, MLDM 2012, Berlin, Germany,
July 13-20, 2012. Proceedings.
Springer Berlin Heidelberg Berlin, Heidelberg, 154 – 168.

Schneider, P., Biehl, M. and Hammer, B. (December 2009).
Adaptive Relevance Matrices in Learning Vector Quantization.
MIT Press.
In: Neural Comput., 21(12), Cambridge, MA, USA, 3532 – 3561.

Polynomial and spline regression

D.A. Langbroek[1]

[1] RUG, Groningen

E-mail for correspondence: davidlangbroek@hotmail.com

Abstract: In this article I compare polynomial regression and cubic spline regression methods on some functions. I summarise some advantages and disadvantages of both regression methods, and give insight when to apply either method.

Keywords: Polynomial; Spline; Regression

1 Introduction

In this paper I take a look at polynomial regression and cubic spline regression. Polynomial regression fits a function of the form $c_{k+1}x^k + c_kx^{k-1} + \cdots + c_2x + c_1$ to the data, where k is the degree and each c_i to be determined constant. The flexibility in the model comes from the degree. Spline regression fits a polynomial as well but with the addition of dividing the range of X into regions, on which each separately polynomial regression takes place. Cubic spline regression specifies that the to be fit spline polynomials have degree 3 and connect on the section knots in a continuous and second order smooth manner. The flexibility of splines is mostly generated by the number of regions. In this paper I first look at the Runge function, to illustrate some weaknesses of polynomial regression and how splines can help. Then I will show in two cases that splines can be applied to more functions than polynomial regression. Finally I show some downsides of splines.

2 Statistical methods

All the figures, plots, table and simulation have been made in R with build in regression functions. The simulations have been made by taking 100 random points uniform distribution (1000 for the cosine) over the domain of the underlying function as data points to fit the regression methods with. These points are the same for all figures. The error in Figure 4 is created by adding a random value between $[-0.2, 0.2]$ taken form uniform distribution. The adjusted R^2 values are from the corresponding built in function aswell. Table 1 has an overview

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

of indication for all figures how well each model fits. A sample R code can be found in the appendix.

3 Runge function

In this section I will use both polynomial and cubic spline regression methods on the Runge function. The Runge function is defined as: $y = \frac{1}{1+25x^2}$. In Figure 1 I have plotted the regression polynomials of degree 4,9,16 over the original Runge function and cubic spline with 3 and 7 knots at respective data quantiles. The

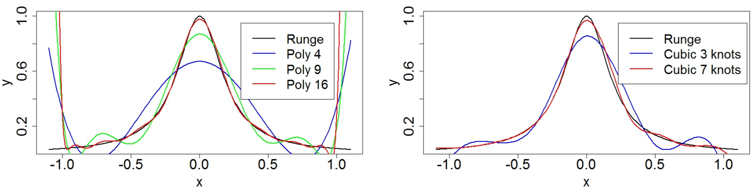


FIGURE 1. Runge function polynomial/spline regression

Figure and Table 1 show that a more flexible polynomial leads to a better fit. However, at the boundaries, the polynomials are extremely poor. The cubic spline with 7 knots is more flexible then the 3 knots spline, and fits the function better. Note that the left boundary is approximated well but that the right boundary still has a bit of fluctuation. Indicating that splines can alleviate the boundary problems, but not necessarily outright solve them.

4 Spline and polynomials fit

In this section I will show that a spline can approximate some functions that polynomial regression can not while not performing worse on polynomials. First I will approximate the polynomial: $y = x^6 + x^5 - 3x^4 - 3x^3 - 2x^2 + x$ with spline and polynomial regression. The original function is a sixth degree polynomial, and as a result polynomial regression of degree equal or more then 6 give an (almost) perfect fit. In Figure 2 I have plotted the original function, and polynomial and spline approximations. The cubic splines are capable of approximating this polynomial as well and are better fits as indicated in Table 1. The other function that I approximate is the periodic cosine function on the domain $x \in (-4\pi, \pi)$. In this domain the cosine function achieves 9 extreme values. In Figure 3 the polynomials of degree 9 and 16 have been plotted over the cosine. I have not included a figure where a cubic spline is fit, as the cubic spline with appropriately chosen knots (knots at multiples of π) fits the underlying function near perfect.

5 Weakness of splines

So far splines outperform polynomials. However splines are not necessarily better, and in the section I will show two explicit downsides of splines: *knot placements*

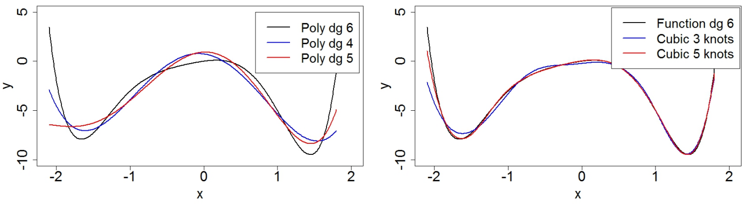


FIGURE 2. Polynomial degree 6 polynomial/spline regression

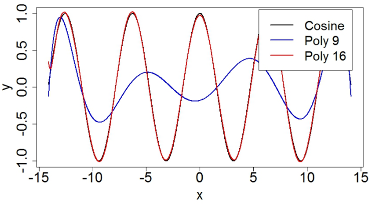


FIGURE 3. Cosine polynomial regression

and variance. Previously I have claimed that a cubic spline with appropriately chosen knots can perfectly approximate a cosine function. However if the knots are chosen at a different location or another amount of knots is chosen, then the result is not necessarily as good a fit. In the following Figure 4 I have plotted a cubic spline with 9 knots, positioned at multiples of 10% of the regression data points. As well as a cubic spline with 7 knots, at the respective data quantiles. Table1 indicates that both are worse fits than the polynomial of degree 16. To indicate the variance when using spline I will try to approximate the Runge function again, but this time the data points have an random error of up to ±0.2. The results are shown in Figure 4, including the data points used for the approximation. The spline with 20 knots is very flexible and wobbly, typical overfit.

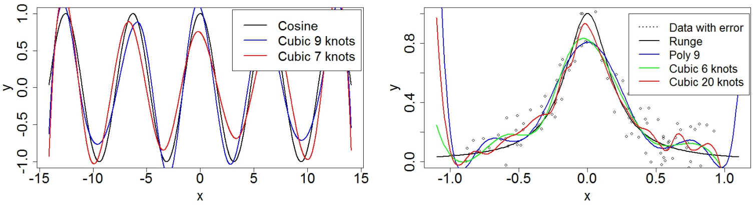


FIGURE 4. Cosine Spline regression/ Runge function with error regression

TABLE 1. Overvieuw adjusted R^2 error of fits

	Fits of first sub-Figure			Fits of second sub-Figure		
Plots	Fit 1	Fit 2	Fit 3	Fit 1	Fit 2	Fit 3
Figure 1	0.7655	0.9523	0.9985	0.9399	0.9963	
Figure 2	0.9081	0.9263		0.9913	0.9993	
Figure 3	0.2579	0.9991				
Figure 4	0.8841	0.9960		0.8156	0.8339	0.8418

6 Conclusion

Polynomial regression is a simple method to approximate functions. However at boundary values it may become poor, and increasing the degree does not always improve the results. Moreover there are functions that can not be well approximated using polynomials. In all these cases, spline regression can yield better results. All while not performing worse in situations that polynomial regression suffices. It seems that in general spline performs better. However it also comes with its own downsides. It is overly complex and has the tendency to overfit the data. Also the position and number of knots matters for the results, but finding the proper values can be difficult. Lastly, the boundary problems can persist when using splines.

Acknowledgments: Special Thanks to dr. M.A. Grzegorczyk for his feedback on the writing of this paper

References

Hastie, T., James, G., Tibshirani, R., and Witten, D. (1994). *An Introduction to Statistical Learning, with apllication in R*. New York: Springer.

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**
- **Center for Language and Cognition, RUG**
- **Behavioural and Cognitive Neurosciences, RUG**
- **R Studio**
- **Statistical Analysis Software (SAS)**
- **The welcome reception is sponsored by:**
  - **The Province of Groningen**
  - **The Muncipality of Groningen**
  - **The Rijksuniversiteit Groningen (RUG)**