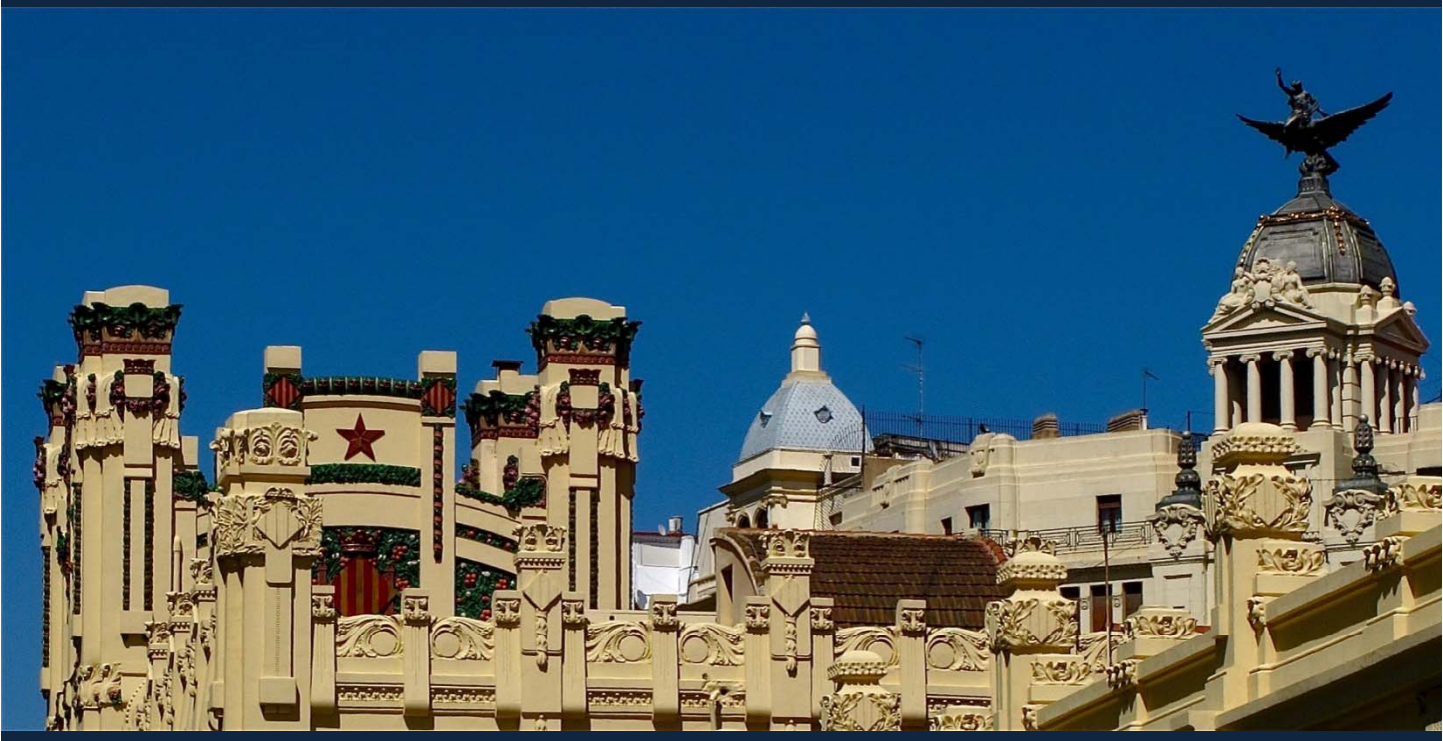
IWSM 2011

Proceedings of the

26th International Workshop on Statistical Modelling

Valencia (Spain), July 11-15, 2011



Editors:
David Conesa
Anabel Forte
Antonio López-Quílez
Facundo Muñoz

Proceedings of the 26th International Workshop on Statistical Modelling

July 11-15, 2011

València

David Conesa, Anabel Forte,
Antonio López-Quílez, Facundo Muñoz
(editors)

Proceedings of the 26th International Workshop on Statistical Modelling.
València, July 11-15, 2011
David Conesa, Anabel Forte, Antonio López-Quílez, Facundo Muñoz, eds.
València 2011.
ISBN 978-84-694-5129-8

Editors:

David Conesa[1], David.V.Conesa@uv.es
Anabel Forte[2], forte@eco.uji.es
Antonio López-Quílez[1], Antonio.Lopez@uv.es
Facundo Muñoz[1], Facundo.Munoz@uv.es

[1] Departament d'Estadística i Investigació Operativa
Universitat de València (Estudi General)
Facultat de Matemàtiques
Dr. Moliner 50, 46100 Burjassot, Spain.

[2] Departamento de Economía
Universitat Jaume I
Facultad de Ciencias Jurídicas y Económicas
Campus del Riu Sec, E-12071 Castelló de la Plana, Spain.

Cover photo: Victor Roda

Printed by Copiformes S.L.

Scientific Programme Committee

- Susie Bayarri (Chair)
  Universitat de València, Spain
- Carmen Armero
  Universitat de València, Spain
- Adrian Bowman
  University of Glasgow, UK
- Charmaine B. Dean
  Simon Fraser University, Canada
- María Durbán
  Universidad Carlos III de Madrid, Spain
- Claire Ferguson
  University of Glasgow, UK
- Herwig Friedl
  Graz University of Technology, Austria
- Gillian Heller
  Macquarie University, Australia
- John Hinde
  University of Galway, Ireland
- Thomas Kneib
  Carl von Ossietzky Universität Oldenburg, Germany
- Arnost Komàrek
  Charles University in Prague, Czech Republic
- Antonio López-Quílez
  Universitat de València, Spain
- Brian Marx
  Louisiana State University, USA
- Pere Puig
  Universitat Autònoma de Barcelona, Spain

Preface

This volume contains all the papers of the 26th International Workshop on Statistical Modelling. Many things have changed since in 1986 an enthusiastic group of statisticians interested in statistical modelling started these series of workshops within a friendly and supportive academic atmosphere. New technologies, more attendants, but always with the same initial spirit: to promote and develop the use of statistical modelling in research and applications.

We are glad to present you these Proceedings, which clearly reflect the aliveness of that spirit. One the one hand, the five invited papers show new advances in theoretical research but always keeping an eye in their applied interest. One the other hand, the great amount of contributions (a total of 140) and their quality demonstrate that the workshop is in good shape. Authors should receive most of the credit for the quality of these Proceedings. Nevertheless, all submissions were carefully reviewed by the members of the Scientific Committee. Their detailed work has been reflected in a big improvement of the preliminary versions jointly with the final selection of contributions.

This 26th edition of the IWSM will be held in Valencia (Spain) in an informal environment (ADEIT- FUNDACIÓ UNIVERSITAT-EMPRESA of the Universitat de València) to encourage discussion and exchange of ideas which could result in future research. Valencia has a great tradition in Statistics and in particular in Bayesian Statistics. This why we are so happy to see that this way of thinking and doing statistics is quite present in these Proceedings reflecting its important role in the Society. We will also like to comment, that many of the contributions in these Proceedings are due to students, which clearly have the future in their hands.

Finally, we wish to acknowledge Carmen Armero, the chair of the local Committee for putting together all the pieces needed in the process of organising this event. Without her interest and passion it would have been impossible.

So welcome to Valencia. Enjoy the city and surroundings and have a great conference.

David Conesa, Anabel Forte, Antonio López-Quílez, Facundo Muñoz
Valencia, June 2011

Contents

Part 1. Invited papers

Part 2. Contributed papers

Part 1. Invited papers

Risk Assessment for Pyroclastic Flows: Combining Deterministic and Statistical Modeling

James O. Berger[1], M. J. Bayarri[2], Eliza S. Calder[3], Keith Dalbey[3], Simon Lunagomez[1], Abani K. Patra[3], E. Bruce Pitman[3], Elaine T. Spiller[4], Robert L. Wolpert[1]

[1] Duke University, Durham, NC 27708-0251, USA
[2] Universitat de València, Av. Dr. Moliner 50, 46100 Burjassot, Valencia, SPAIN
[3] University of Buffalo, Buffalo, N.Y. 14260-2900, USA
[4] Marquette University, Milwaukee, WI 53201-1881, USA

Abstract: Risk assessment of volcanic pyroclastic flows is considered, using a combination of computer modeling, statistical modeling, and extreme-event probability analysis. A computer model of pyroclastic flow is used to allow extrapolation to not-yet-observed pyroclastic flows. Statistical modeling is needed to determine the initial input distributions of the computer model. Direct simulation of rare events using the computer model would be prohibitively expensive. Thus we carry out the analysis using a combination of emulators of the computer model and rare event simulation.

Keywords: Bayesian analysis; Catastrophic events; Emulators; Extreme events; Inverse Problems.

1 Introduction

The talk will discuss the assessment of risk for volcanic pyroclastic flows, as introduced in Bayarri et. al. (2009). Important modifications of the methodology will also be discussed, although the formal implementations of these modifications will not be available until the talk at the conference.
Let $\{E_i\}$ denote individual volcanic pyroclastic flows. We wish to assess the probability that at least one catastrophic event C will occur in the next T years – for instance, the probability that a pyroclastic flow event in the next T years will significantly damage a town.
A modern approach to the problem begins with the development of a deterministic computer model model of pyroclastic flows. The computer model can be run under conditions that have not yet been observed, in order to do risk assessment. The computer model considered here is TITAN2D, developed for modeling the process of volcanic flow. TITAN2D can predict the maximum thickness of a pyroclastic flow at any location (such as the

center of a town) – Patra et.al. (2005). If the flow thickness exceeds one meter, we will call that flow a catastrophic event.
TITAN2D requires inputs to run. The most crucial are the volumes of volcanic flow that can be expected, the initial directions in which the flows proceed down the mountain, and the friction of the flow with the mountain surface. In determining the distributions of these inputs, having available data concerning pyroclastic flows is crucial, as is the development of statistical models of the data.
Computing the probability of rare catastrophic events is also a challenge. Direct simulation is generally impossible, because of the time needed to run the computer model. This challenge is addressed by using emulators (approximations) for the computer model to identify the *threshold* inputs that define the catastrophic event; development of such emulators is another modeling challenge. It is then possible to compute the desired risk probabilities.
The methodology is illustrated on the Soufrière Hills Volcano (to be abbreviated *SHV*) on the island of Montserrat, which has been erupting since 1995 and has generated hundreds of pyroclastic flows with runouts exceeding 2 km. On over 50 occasions, these pyroclastic flows consisted of volumes of material exceeding 10^6m^3.

2 Risk Assessment and Emulation

The inputs to the computer model will be denoted $\mathbf{x} \in \mathcal{X}$ and the computer model prediction of the characteristic of interest by $y^M(\mathbf{x})$. For the SHV in Montserrat, $\mathbf{x} = (V, \varphi, b)$, where V is the volume of the flow, φ is the initialization angle of flow, and b is the basal friction of the flow, i.e., the friction of the flow with the ground. Also, $y^M(\mathbf{x})$ = the maximum height at the center of the target area of a pyroclastic flow from an eruption with characteristics $\mathbf{x}$. A catastrophic event occurs (by definition for this paper) if $\mathbf{x}$ is such that $y^M(\mathbf{x}) \geq 1\text{m}$.
We desire to find the contour that separates catastrophic events from benign events. This contour, which we will call Ψ, can be represented by finding, for each angle $\varphi \in [0, 2\pi)$ and basal friction value b, the minimum volume V that causes catastrophic damage (all larger volumes will cause even worse damage). Thus we will write $\Psi = \Psi(\varphi, b) = \inf\{V : y^M(V, \varphi, b) \geq 1\text{m}\}$.
To find Ψ we ran 256 TITAN2D simulations at design points in a large region of the input space. These design points were chosen according to a Latin hypercube design. Initially the emulator was based on a subset of these design points, which were then augmented in an adaptive way to improve the estimated frontier Ψ. We used the familiar Gaussian process emulator (GaSP) (see Santner et.al. (2003)) to approximate TITAN2D.
Figure 1 shows the mean surfaces of the emulated max-height GaSP's at one of the sites on Montserrat under consideration – the former capitol of

Plymouth – as a function of the volume and angle inputs (and a fixed value of b), along with dots indicating the actual heights at the design-points that were obtained from the computer model runs.

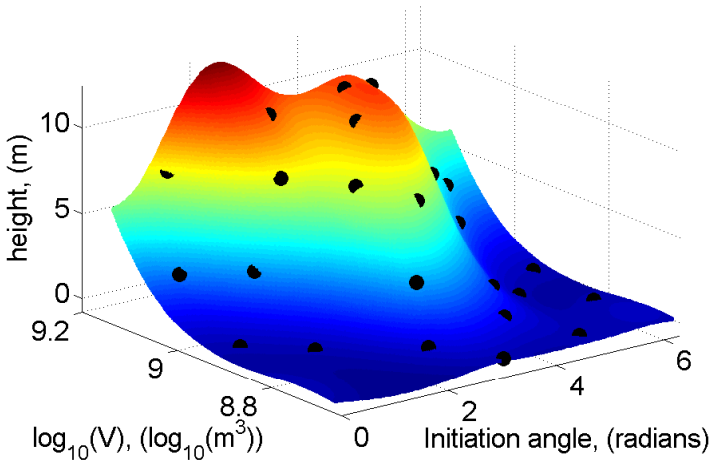


FIGURE 1. Max-height surfaces are the mean of the GaSP emulators at Plymouth. Dark points represent the max-height simulation output at design points.

We approximate the catastrophic event contour Ψ by determining the appropriate contour numerically from the emulated GaSP surface. For Plymouth, $\Psi(\varphi)$ is shown in Figure 2 (at a specified value of b) for the emulator obtained by fitting a GaSP to $\log(y^M(\cdot)+1)$ and transforming back.
The emulators are only approximations to the computer model, so there will be error in the estimation of Ψ. This is reflected in the 90% credible bands given in Figure 2 (found simply by transforming back to meters the 5% and 95% quantiles of the posterior predictive distribution of the Bayesian emulator for the log transformation).
A round of adaptive design was then performed, to obtain new runs of TITAN2D near the critical contour. The results were essentially the same as those obtained in Figure 2.

3 Modeling the Input Distributions

Figure 3 shows an empirical plot of the number of pyroclastic flows exceeding volume $V_j \geq v$ *vs.* v from March, 1996 through July, 2008 for SVH, on a log-log scale, for large volumes $v \geq \epsilon$ (here $\epsilon = 5 \cdot 10^4\,\mathrm{m}^3$).
The near-linear fall-off on a log-log scale suggests that the probability dis-

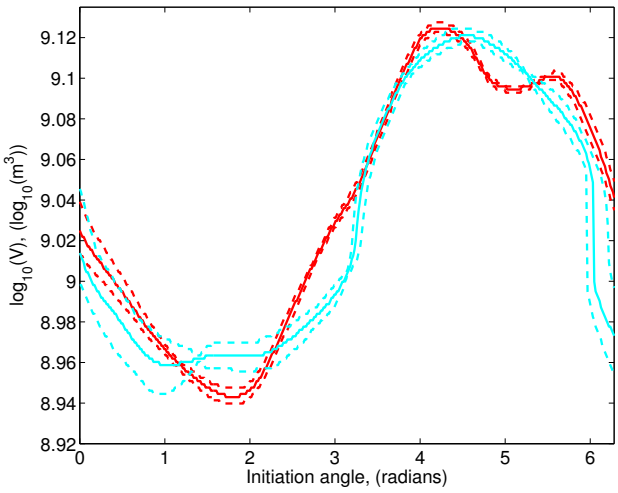


FIGURE 2. Median estimates (solid curves) and 90% credible bands (dashed curves) of frontier $\Psi(\phi)$, based on Gaussian process fit to log-transformed simulation output, at Plymouth.

tribution of flow volumes satisfies

$$\log \mathsf{P}[V \geq v \mid V \geq \epsilon] \approx -\alpha \log(v) + c \tag{1}$$

for some constants $\alpha > 0$ and $c \in \mathbb{R}$, and hence the distribution of the $\{V_j\}$ is approximately Pareto, with

$$\mathsf{P}[V \geq v] \approx (v/\epsilon)^{-\alpha}, \qquad v \geq \epsilon\,. \tag{2}$$

The pyroclastic flows (PFs) whose volume exceed the threshold ϵ are a marked Poisson process with marks as the initial volumes and initiation angle pairs $\{(V_j, \varphi_j)\}$ at times $\tau_j > 0$. We take the Poisson rate to be some constant λ_ϵ; we also assume independence of the Pareto-distributed volumes $\{V_j\}$.
The likelihood function, upon observing $\{(V_j, \tau_j) : \ V_j > \epsilon,\ 0 < \tau_j \leq T\}_{j \leq J_\epsilon}$, is

$$\begin{aligned} L(\alpha, \lambda) \quad &\propto \quad (\alpha\,\lambda)^{J_\epsilon} \, \exp\Big[-\lambda T\, \epsilon^{-\alpha} - \alpha \sum_{j \leq J_\epsilon} \log V_j \Big] \\ &= \quad (\alpha\,\lambda\,\epsilon^{-\alpha})^{J_\epsilon} \, e^{-\lambda T\, \epsilon^{-\alpha} - \alpha S_\epsilon}, \end{aligned}$$

where $S_\epsilon := \sum_{j \leq J_\epsilon} \log(V_j/\epsilon)$.

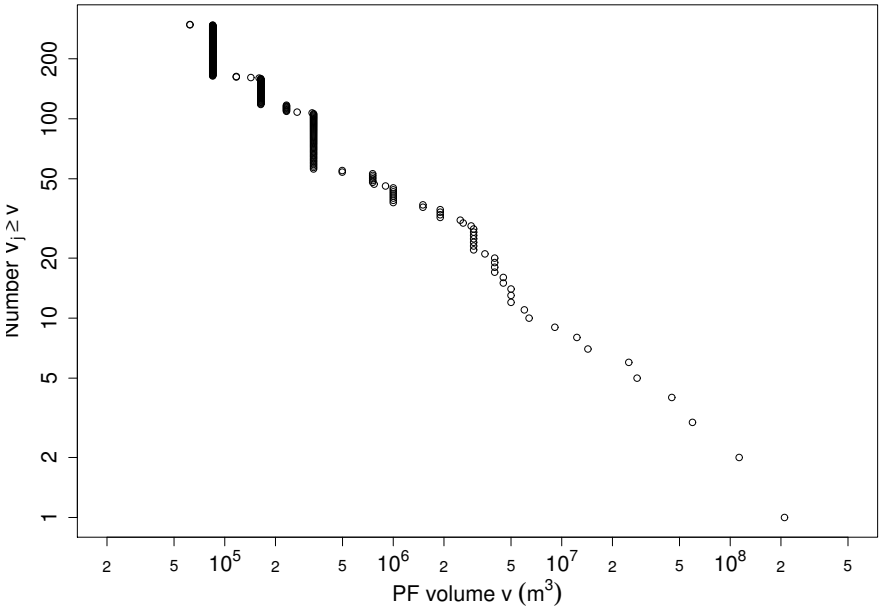


FIGURE 3. Frequency-*vs.*-magnitude plot for pyroclastic flows at SVH.

We analyze this likelihood function from an objective Bayesian perspective. Let $\pi(\alpha, \lambda, \varphi)$ denote an objective prior density function for α and λ and the initiation angle φ. Inferences will be based on the posterior density

$$\pi^*(\alpha, \lambda, \varphi) = \pi(\alpha, \lambda, \varphi \mid \text{data}) \propto L(\alpha, \lambda)\pi(\alpha, \lambda, \varphi)\,;$$

note that the likelihood does not depend on the initiation angle φ.
The obvious objective prior distribution for the angle φ is the uniform distribution on $[0, 2\pi)$, and this was initially deemed reasonable by the geologists in the project; but, later, it was realized that a uniform distribution is not optimal – because of the configuration of the mountain – and a non-uniform distribution will be incorporated in the final analysis. Also, φ was viewed as independent of the other parameters.
A natural objective choice for $\pi(\alpha, \lambda)$ is the reference prior distribution of Berger and Bernardo (1992). There are actually two reference priors, based on declaring first α and then λ to be the parameter of interest:

$$\begin{aligned} \pi_{R\alpha}(\alpha, \lambda) &\propto \lambda^{-1/2}\alpha^{-1}\epsilon^{-\alpha/2}\mathbf{1}_{\{\alpha>0,\lambda>0\}} \\ \pi_{R\lambda}(\alpha, \lambda) &\propto \lambda^{-1/2}[\alpha^{-2} + (\log\epsilon)^2]^{1/2}\epsilon^{-\alpha/2}\mathbf{1}_{\{\alpha>0,\lambda>0\}}\,. \end{aligned}$$

These were both used in the final analysis, and gave virtually identical results.
The final issue is basal friction parameter b. It appears that b and V are strongly related; indeed are nearly linearly related on a log-log-scale over the range of parameters of interest. This relationship is being determined in ongoing work, based on a hierarchical Bayesian analysis of 4 sets of data involving pyroclastic flows at different volcanoes. The analysis then proceeds by simply replacing b by the empirical function of V obtained from this analysis, so that only V remains in the expressions. Hence we henceforth ignore b.

4 Risk Assessment

Combining the previous modeling, we can now compute the probability of a catastrophic event in the next t years at Plymouth.
The number of PFs in a future time interval of length t years whose volume V_i and initiation angle φ_i satisfy $V_i > \Psi(\varphi_i)$ (i.e., the number of catastrophic PFs in t years) will have a Poisson probability distribution with conditional expectation

$$E(\#\text{ catastrophic PFs in } t \text{ yrs} \mid \alpha, \lambda) = \frac{t\,\lambda}{2\pi} \int_0^{2\pi} \Psi(\varphi)^{-\alpha}\, d\varphi, \qquad (3)$$

for given values of the parameters α and λ, so the probability of a catastrophic event is

$$P(\geq \text{one catastrophic PF in } t \text{ yrs} \mid \alpha, \lambda) = 1 - \exp\left[-\frac{t\,\lambda}{2\pi} \int_0^{2\pi} \Psi(\varphi)^{-\alpha}\, d\varphi\right].$$

The posterior probability of catastrophe in t years, using the likelihood function of and the objective prior densities is then given by

$$\begin{aligned} P(t) &= P[\text{At least one PF} > \Psi(\varphi) \text{ in } t \text{ yrs} \mid \text{data}] \\ &= 1 - \int_{\mathbb{R}_+^2} \exp\left[-\frac{t\,\lambda}{2\pi} \int_0^{2\pi} \Psi(\varphi)^{-\alpha}\, d\varphi\right] \pi^*(\alpha, \lambda)\; d\alpha\; d\lambda \qquad (4) \end{aligned}$$

for the posterior density $\pi^*(\alpha, \lambda) = Z^{-1} L(\alpha, \lambda) \lambda^{a-1}\, \alpha^{-1} g(\alpha) \mathbf{1}_{\{\alpha>0, \lambda>0\}}$, with normalizing constant $Z := \iint_{\mathbb{R}_+^2} L(\alpha, \lambda) \lambda^{a-1}\, \alpha^{-1} g(\alpha) d\alpha\, d\lambda$. The λ integral in equation (4) is available in closed form, after computation leaving:

$$P(t) = 1 - \tilde{Z}^{-1} \int_{\mathbb{R}_+} \left[1 + (t/T) I_\epsilon(\alpha)\right]^{-J_\epsilon - a}\; \alpha^{J_\epsilon - 1} e^{-\alpha[S_\epsilon - a \log \epsilon]} g(\alpha)\, d\alpha\,,$$

where $I_\epsilon(\alpha) := \frac{1}{2\pi} \int_0^{2\pi} [\Psi(\varphi)/\epsilon]^{-\alpha}\, d\varphi$ and $\tilde{Z}$ is the normalizing constant.

References

Bayarri, M.J., Berger, J.O., Calder, E.S., Dalbey, K. Lunagomez, S. Patra, A.K., Pitman, E.B., Spiller, E.T., and Wolpert, R.L. (2009). Using statistical and computer models to quantify volcanic hazards. *Technometrics*, **51**, 402-413.

Berger, J. O. and Bernardo, J. M. (1992). On the development of the reference prior method. In: *Bayesian Statistics 4*. 35–49, Oxford, UK: Oxford Univ. Press.

Patra, A. K., Bauer, A. C., Nichita, C. C., Pitman, E. B., Sheridan, M. F., and Bursik, M. I. (2005). Parallel adaptive numerical simulation of dry avalanches over natural terrain. *Journal of Volcanology and Geothermal Research*, **139**, 1–21.

Santner, T. J., Williams, B. J., and Notz, W. (2003), *The Design and Analysis of Computer Experiments*. New York: Springer. NY: Springer-Verlag.

Quasi-variances and extensions

David Firth[1]

[1] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK; d.firth@warwick.ac.uk

Abstract: The notion of quasi-variances, as a device for both simplifying and enhancing the presentation of categorical-predictor effects in statistical models, was developed in Firth and de Menezes (*Biometrika*, 2004, 65–80). The approach generalizes the earlier ideas of Ridout (*GLIM Proceedings*, 1989) and of Easton, Peto and Babiker (*Statistics in Medicine*, 1991 — 'floating absolute risk', which has become rather controversial in epidemiology). In this talk I will outline and exemplify the method to show how it can be useful, and discuss its extension to some other contexts such as parameters that may be arbitrarily scaled and/or rotated.

Keywords: Floating absolute risk; model summary

1 Quasi-variances: The basic idea

When presenting the results of statistical modelling, one very standard summary is a table of parameter estimates and standard errors; in Bayesian analysis, an analogous device is a table of posterior means and standard deviations or — if space permits — a series of marginal views of the posterior density. The device of 'quasi-variances' aims to improve such summaries in situations where at least some of the parameters of interest relate to the effect of a categorical predictor variable. In such situations, *contrasts* among the parameters typically are identified and of interest. Most commonly the standard summary is based on an arbitrarily selected subset of contrasts, for example contrasts with the first or last level of a factor, or with an average over all of the levels. Such a summary works well for those specific contrasts, but does not facilitate valid inference on other contrasts not in the selected subset.

Quasi-variances overcome this difficulty as follows. (The exposition here will be in terms of estimates and standard errors; it could equally well be made in terms of posterior means and standard deviations.) For a set of parameters $\beta_1, \ldots, \beta_p$, we approximate the variance of any contrast $\sum c_r \hat{\beta}_r$ (where $\sum c_r = 0$) by $\sum c_r^2 q_r$, in which the quantities $q_1, \ldots, q_p$ are so-called *quasi-variances*. When good quasi-variances can be found — that is, when the approximation is reasonably accurate for all contrasts of potential interest — this yields a simple summary table from which valid approximate

inference can be drawn about *any* contrast. The simplicity stems from the fact that the $\{q_r\}$ can be read *as if* they were the variances of p uncorrelated estimates. This also allows for simple graphical presentations, for example with a point estimate and error bar for each parameter, whose 'Pythagorean' interpretation is both informative and familiar.
This basic idea was first suggested by Ridout (1989), in the context of estimates from a balanced experimental design. Easton, Peto and Babiker (1991) independently suggested it under the name 'floating absolute risk', with some particular epidemiological applications in mind. A further influential reference is Cox and Reid (2000, p237). In epidemiology the method has proved to be rather controversial (e.g., Easton and Peto, 2000, and references therein); this seems to be partly because the idea of Easton et al. (1991) was not always well enough understood, and partly because the specific approximation recipe used in Easton et al. (1991) was not ideal. Menezes (1999), Firth and Menezes (2004) and Plummer (2004) studied the approximation in detail and suggested methods that are more generally successful. The work of Ridout (1989), whose approximation recipe was indeed one of the 'generally successful' variety, was sadly unknown to the epidemiologists whose arguments about the method's merits spanned several subsequent years.

2 Aims in this talk

In this talk I will review why and when the method of quasi-variances works well, and I will discuss some examples of its fruitful application. The controversy surrounding 'floating absolute risk' will be demystified.
Attention will then turn to extensions of the method:

(i) To some less standard contexts where contrasts are still the identifiable parameter combinations of interest. These contexts include:

- Bradley-Terry models for binary 'tournaments' (Turner and Firth, 2010);
- the homogeneous RC(1) association model of Goodman (1979), for contingency tables;
- multinomial logit regression models for categorical-response data;
- certain other often-used multiplicative interaction models, such as the 'unidiff' model from social mobility studies (Erikson and Goldthorpe, 1992; Xie, 1992).

(ii) To some more general situations, where the contrasts of interest are identified only after fixing some other aspect of parameterization such as *scale* or *angle of rotation*. These include:

- the *non*-homogeneous Goodman RC(1) association model;

- the (homogeneous or non-homogeneous) Goodman RC(2) association models;
- some standard item-response models (Rasch-type scaling models);
- factor analysis of multivariate data.

3 Software

The *R* package qvcalc (Firth, 2003b) implements the basic method efficiently, with direct interfaces to various prominent classes of model object in *R*; summary capabilities include the routine reporting of the accuracy of computed quasi-variances, and facilities for readily interpreted 'error bar' plots of effects of interest. The same package also underlies a simple web-based calculator (originally developed using *Xlisp-Stat*; see Firth, 2000).

Acknowledgments: This work was supported by the Engineering and Physical Sciences Research Council, UK.

References

Cox, D.R. and Reid, N. (2000). *The Theory of the Design of Experiments*. London: Chapman and Hall.

Easton, D., Peto, J. and Babiker, A. (1991). Floating absolute risk: An alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. *Statistics in Medicine*, **10**, 1025–35.

Easton, D. and Peto, J. (2000). Re: 'Presenting statistical uncertainty in trends and dose-response relationships' (letter). *American Journal of Epidemiology*, **152**, 393.

Erikson, R. and Goldthorpe, J.H. (1992). *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford: Clarendon Press.

Firth, D. (2000). Quasi-variances in Xlisp-Stat and on the web. *Journal of Statistical Software*, **5.4**, 1–13.

Firth, D. (2003a). Overcoming the reference category problem in the presentation of statistical models. *Sociological Methodology*, **33**, 1–18.

Firth, D. (2003b). R Package qvcalc. *Comprehensive R Archive Network*, `http://cran.r-project.org/web/packages/qvcalc`.

Firth, D., and Menezes, R.X. de (2004). Quasi-variances. *Biometrika*, **91**, 65–80.

Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, **74**, 537–582.

Menezes, R.X. de (1999). More useful standard errors for group and factor effects in generalized linear models. *D.Phil. thesis*, University of Oxford, UK.

Plummer, M. (2004). Improved estimates of floating absolute risk. *Statistics in Medicine*, **23**, 93–104.

Ridout, M.S. (1989). Summarizing the results of fitting generalized linear models to data from designed experiments. In: *Statistical Modelling: Proceedings of GLIM89 and the 4th International Workshop on Statistical Modelling*, Ed. A. Decarli, B. Francis, R. Gilchrist and G. Seeber, 262–9. New York: Springer Verlag.

Turner, H. and Firth, D. (2010). Bradley-Terry models in R: The BradleyTerry2 package. *Comprehensive R Archive Network*, `http://cran.r-project.org/web/packages/BradleyTerry2`.

Xie, Y. (1992). The log-multiplicative layer effect model for comparing mobility tables. *American Sociological Review*, **57**, 380–95.

Some theoretical thoughts when using a composite endpoint to prove the efficacy of a treatment

Guadalupe Gómez[1]

[1] Statistics and Operations Research Dept., Universitat Politècnica de Catalunya

Abstract: This paper discusses, following Gómez and Lagakos (2011) methodology, to what extent is there a gain in efficiency from adding a component event to a relevant endpoint when the treatment effect on this component is not as strong as on the original relevant endpoint under ideal (independence) circumstances. It presents the bivariate copula model used to overcome the independence assumption and presents the relationship between the components of the asymptotic relative efficiency and a set of interpretable parameters.

Keywords: Asymptotic Relative Efficiency; Composite Endpoints; Composite Outcomes; Copula Model; Logrank tests

1 Introduction and motivating example

In randomized clinical trials it is common to use a composite event as endpoint and to prove the beneficial effects on treatment for this endpoint. A composite event E_* is defined as one of several events $\mathcal{E}_j$ $(j = 1, \cdots, m)$, that is, $E_* = \bigcup_{j=1}^{m} \mathcal{E}_j$. One of the reasons why scientists use composite events is to assure that, for a given sample, enough events are observed during the course of the study, being this especially crucial when one of the events is "rare" or not very frequent. The popular thinking is that "by adding" more events to the composite endpoint, we might have more power to detect treatment differences.

This problem is found in many areas but in particular in cardiovascular studies. For instance, Tardif *et al* (2008) use composite endpoints when studying the addition of succinobucol, a novel anti-oxidant and anti-inflammatory agent, to optimal medical therapy to 6,144 high-risk patients with unstable angina or who had suffered heart attacks. In the double-blind, placebo-controlled clinical trial for succinobucol the following six cardiovascular events are of interest: Cardiovascular death, resuscitated cardiac arrest, myocardial infarction, stroke, hospitalization due to unstable angina or hospitalization due to coronary revascularization. The study shows that succinobucol has no effect on the primary endpoint E_* where all six events are considered, while it has a beneficial effect on the composite secondary

endpoint defined as the union of the first 4 events. In this particular instance, the addition of the hospitalization events (355 (67%) in the succinobucol group and 318 (60%) in the placebo group) to the previous 4 events (207 versus 252) has yielded a non significant result for the primary E_* from a beneficial effect that the treatment has on the composite secondary endpoint.
Gómez and Lagakos' paper (2011) proposes a conceptual framework as an aid to make a decision, when planning a clinical trial, on whether to use a relevant endpoint $\mathcal{E}_1$ or the composite of $\mathcal{E}_1$ and an additional $\mathcal{E}_2$ based on prior information about the disease. The main goal of this paper is to discuss to what extent is there a gain in efficiency from adding a component event to a relevant endpoint when the treatment effect on this component is not as strong as on the original relevant endpoint under ideal (independence) circumstances, to present the copula models used to overcome the independence assumption and to frame them to derive the relative efficiency of $\mathcal{E}_* = \mathcal{E}_1 \cup \mathcal{E}_2$ versus using just the primary endpoint $\mathcal{E}_1$.

2 Notation

We consider two-arm randomized studies involving random assignment to an active treatment ($X = 1$) or to a control treatment ($X = 0$) and we focus on the time from randomization until the first occuring of a specific set of clinical outcomes. We assume that we have two different endpoints of potential interest, $\mathcal{E}_1$ and $\mathcal{E}_2$, where each one can be either single or composite. This paper is restricted to the case where the additional event $\mathcal{E}_2$ cannot include a terminating event, such as death and it corresponds to cases 1 and 3 of Gómez and Lagakos (2011). The individuals are followed until the event of interest, or until the end of the study, whichever occurs first. Denote by $T_1^{(j)}$ and $T_2^{(j)}$ the times to $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively, for patients in group $X = j$ ($j = 0, 1$) and by C the time until the end of the study (assumed equal for both groups). We assume that $T_1^{(j)}$ and $T_2^{(j)}$ are absolutely continuous so that ties cannot occur and that end-of-study censoring is the only noninformative censoring cause. We consider the composite event $\mathcal{E}_* = \mathcal{E}_1 \cup \mathcal{E}_2$ and we measure the effect of treatment on the composite endpoint $T_*^{(j)} = \min\{T_1^{(j)}, T_2^{(j)}\}$ which is the time until the occurrence of $\mathcal{E}_*$ consisting of the earlier occurring of $\mathcal{E}_1$ or $\mathcal{E}_2$.

3 Facts when the independence assumption holds

In this section we show that a beneficial effect on $\mathcal{E}_*$ can occur simultaneously with a beneficial effect on $\mathcal{E}_1$ and a harmful effect on $\mathcal{E}_2$ and that not finding a beneficial effect on the composite event $\mathcal{E}_*$ is no guarantee of not having some effect on the individual events $\mathcal{E}_1$ or $\mathcal{E}_2$.

These facts are shown for the particular case of independence between $T_1^{(j)}$ and $T_2^{(j)}$ and under the assumption that the hazards of $T_1^{(1)}$ versus $T_1^{(0)}$ ($\lambda_1^{(1)}(t)$ and $\lambda_1^{(0)}(t)$) and of $T_2^{(1)}$ versus $T_2^{(0)}$ ($\lambda_2^{(1)}(t)$ and $\lambda_2^{(0)}(t)$) are proportional. Under this assumption, the relative treatment effects on $\mathcal{E}_1$ and on $\mathcal{E}_2$ are the constant hazard ratios $\frac{\lambda_1^{(1)}(t)}{\lambda_1^{(0)}(t)}$ and $\frac{\lambda_2^{(1)}(t)}{\lambda_2^{(0)}(t)}$, respectively, and hazard ratios < 1 (> 1) are indicative of a beneficial (harmful) effect of the treatment.

Proposition For $j = 0, 1$, if $T_1^{(j)}$ and $T_2^{(j)}$ are independent and both $T_1^{(j)}$ and $T_2^{(j)}$ have proportional hazards, then, the hazards of $T_*^{(j)}$ ($\lambda_*^{(1)}(t)$ and $\lambda_*^{(0)}(t)$) are proportional if and only if the baseline hazard functions for the relevant and the additional endpoints, $\lambda_1^{(0)}(t)$ and $\lambda_2^{(0)}(t)$, respectively, are as well proportional. That is, if we have, for given $k_1 > k_2 > 0$, $\lambda_1^{(1)}(t) = k_1\lambda_1^{(0)}(t)$ and $\lambda_2^{(1)}(t) = k_2\lambda_2^{(0)}(t)$ for all t, then there exists k such that $\lambda_*^{(1)}(t) = k\lambda_*^{(0)}(t)$ if and only if $\lambda_2^{(0)}(t) = k_0\lambda_1^{(0)}(t)$ for all t with k and k_o related by $k = \frac{1}{1+k_0}k_1 + \frac{k_0}{1+k_0}k_2$.

Proof *Due to the independence between T_1^j and T_2^j, we have*

$$\lambda_*^{(1)}(t) \quad = \quad k\lambda_*^{(0)}(t) \Leftrightarrow \quad \lambda_1^{(1)}(t) + \lambda_2^{(1)}(t) = k(\lambda_1^{(0)}(t) + \lambda_2^{(0)}(t))$$

hence, since $\lambda_1^{(1)}(t) = k_1\lambda_1^{(0)}(t)$ and $\lambda_2^{(1)}(t) = k_2\lambda_2^{(0)}(t)$, it follows that

$$\begin{aligned} k_1\lambda_1^{(0)}(t) + k_2\lambda_2^{(0)}(t) &= k(\lambda_1^{(0)}(t) + \lambda_2^{(0)}(t)) \Leftrightarrow \\ (k_1 - k)\lambda_1^{(0)}(t) &= (k - k_2)\lambda_2^{(0)}(t) \Leftrightarrow \quad \lambda_2^{(0)}(t) = \frac{(k_1 - k)}{(k - k_2)}\lambda_1^{(0)}(t). \end{aligned}$$

This result establishes that if the baseline hazard functions, $\lambda_1^{(0)}(t)$ and $\lambda_2^{(0)}(t)$ are proportional, then the hazard ratio $\frac{\lambda_*^{(1)}(t)}{\lambda_*^{(0)}(t)}$ is a linear combination of $\frac{\lambda_1^{(1)}(t)}{\lambda_1^{(0)}(t)}$ and $\frac{\lambda_2^{(1)}(t)}{\lambda_2^{(0)}(t)}$, and this has several relevant implications which we summarize in the next Corollary.

Corollary Under the assumptions of the proposition and assuming that $\lambda_2^{(0)}(t) = k_0\lambda_1^{(0)}(t)$,

1. If treatment has no effect on $\mathcal{E}_1$ neither on $\mathcal{E}_2$ ($k_1 = k_2 = 1$), then treatment has no effect on $\mathcal{E}_*$ ($k = 1$).

2. The effect that treatment has on $\mathcal{E}_*$ lies always between the effects that the treatment has on $\mathcal{E}_1$ and $\mathcal{E}_2$. That is, if $k_1 = \frac{\lambda_1^{(1)}(t)}{\lambda_1^{(0)}(t)} < \frac{\lambda_2^{(1)}(t)}{\lambda_2^{(0)}(t)} = k_2$ then $k_1 < \frac{\lambda_*^{(1)}(t)}{\lambda_*^{(0)}(t)} < k_2$ and hence: i) if the treatment effect is

beneficial on $\mathcal{E}_1$ and $\mathcal{E}_2$ ($k_1 < k_2 \leq 1$), the treatment will prove to be beneficial on $\mathcal{E}_*$ and ii) if the treatment effect is harmful on $\mathcal{E}_1$ and $\mathcal{E}_2$ ($1 \leq k_1 < k_2$), the treatment will prove to be harmful on $\mathcal{E}_*$. Analogously if $k_1 > k_2$.

3. If treatment has a beneficial effect for $\mathcal{E}_1$ ($k_1 < 1$) and a harmful effect for $\mathcal{E}_2$ ($k_2 > 1$), you can choose k_0 conveniently to prove either no effect or a beneficial or harmful effect on $\mathcal{E}_*$. For instance, taking $k_1 = 0.5$ and $k_2 = 2$, i) if $k_0 = 1.5$ we have $k = 2$ and treatment has a harmful effect for $\mathcal{E}_*$, ii) if $k_0 = 0.5$ then $k = 1$ and treatment has no effect on $\mathcal{E}_*$ and iii) if $k_0 = 0.25$ then $k = 0.8$ and treatment has a beneficial effect for $\mathcal{E}_*$.

4 Using copulas to model the bivariate survival function

So far we have proved that under the ideal situation of two independent endpoints the beneficial effect on a composite endpoint does not imply the beneficial effect in either component. However, most of the times the two endpoints are correlated and the hazard of the composite cannot be decomposed as the sum of the two marginal hazards. In this situation the joint law of $T_1^{(j)}$ and $T_2^{(j)}$ is needed and we face the challenge of modelling an empirical problem in such a way that is not too complex but still realistic. We can model the joint dependence structure by means of a copula function. A copula is best described, as in Joe (1997), as a multivariate distribution function that is used to bind each marginal distribution function to form the joint. The copula parameterises the dependence between the margins, while the parameters of each marginal distribution function can be estimated separately. The approach via copulas allows much more general types of dependencies to be included than would usually be invoked by a conceptual approach. The approach to formulating a multivariate distribution using a copula is based on the idea that a simple transformation can be made of each marginal variable in such a way that each transformed marginal variable has a uniform distribution. Once this is done, the dependence structure can be expressed as a multivariate distribution on the obtained uniforms, and a copula is precisely a multivariate distribution on uniform random variables. There are many families of copulas which differ in the detail of the dependence they represent. A family will typically have several parameters which relate to the strength and form of the dependence.
Among several classes of copulas the Archimedean copulas are an important family, which have a simple form with properties such as associativity, symmetry and have a variety of dependence structures (Trivedi and Zimmer, 2007). One particularly simple form of an Archimedean bidimensional

copula is given by

$$H(t_1, t_2) = \varphi^{-1}\left(\sum_{i=1}^{2} \varphi(F_i(t_i))\right)$$

where φ is a generator function satisfying $\varphi(1) = 0$, $\lim_{t\to 0}\varphi(t) = \infty$, $\varphi'(t) < 0$ and $\varphi''(t) > 0$, and where F_i $(i = 1, 2)$ are univariate marginal probability distribution functions.
Different choices of the generator function yield as well different copulas with specific features. We are basing our computations in Frank copula's generator defined as $\varphi(t) = -\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$ for dependence parameter θ, $-\infty < \theta < \infty$, because it has the following useful features: it permits negative dependence between the marginals, the dependence is symmetric in both tails, it is comprehensive, that is, it might represent perfect negative dependence, independence and perfect positive dependence between variates. Furthermore, Spearman's ρ linear correlation between $F_1(T_1^{(j)})$ and $F_2(T_2^{(j)})$ is given by $\rho = \rho(\theta) = 1 - \frac{12}{\theta}[\frac{1}{\theta}\int_0^\theta \frac{t}{e^t-1} - \frac{2}{\theta^2}\int_0^\theta \frac{t^2}{e^t-1}dt]$ holding a 1-1 relationship between ρ and θ.
For every group $j = 0, 1$ and given marginal survival (density) functions $S_1^{(j)}(t_1)$ and $S_2^{(j)}(t_2)$ ($f_1^{(j)}(t_1)$ and $f_2^{(j)}(t_2)$) for $T_1^{(j)}$ and $T_2^{(j)}$ and given equal association parameter θ between $T_1^{(j)}$ and $T_2^{(j)}$, the joint survival and density functions based on Frank's copula are as follows:

$$\begin{aligned} S^{(j)}(t_1, t_2; \theta) &= -\theta^{-1}\log\left\{1 + \frac{(e^{-\theta S_1^{(j)}(t_1)} - 1)(e^{-\theta S_2^{(j)}(t_2)} - 1)}{e^{-\theta} - 1}\right\} \\ f_{(1,2)}^{(j)}(t_1, t_2; \theta) &= \frac{\theta e^{-\theta(S_1^{(j)}(t_1) + S_2^{(j)}(t_2)))}}{e^{-2\theta S^{(j)}(t_1,t_2;\theta)}(e^{-\theta} - 1)}[f_1^{(j)}(t_1)][f_2^{(j)}(t_2))] \qquad (1) \end{aligned}$$

For $j = 0, 1$, the survival and density function of $T_*^{(j)} = \min\{T_1^{(j)}, T_2^{(j)}\}$ become equal to

$$\begin{aligned} S_*^{(j)}(t; \theta) &= S^{(j)}(t_1, t_2; \theta) \qquad (2) \\ f_*^{(j)}(t) &= \frac{e^{-\theta S_1^{(j)}(t)}(e^{-\theta S_2^{(j)}(t)} - 1)}{e^{-\theta S_*^{(j)}(t;\theta)}(e^{-\theta} - 1)} f_1^{(j)}(t) + \frac{e^{-\theta S_2^{(j)}(t)}(e^{-\theta S_1^{(j)}(t)} - 1)}{e^{-\theta S_*^{(j)}(t;\theta)}(e^{-\theta} - 1)} f_2^{(j)}(t) \end{aligned}$$

if Frank's copula is used.

5 Log rank test and Asymptotic Relative Efficiency

For the two-arm randomized study described in Section 2, we assume that we have two independent samples, that end-of-study censoring is the only noninformative censoring cause, that end-of-study censoring is identical

across groups and that treatment groups have proportional hazards. To check whether treatment has a beneficial effect, we might use endpoint $\mathcal{E}_1$ carrying the relevant information of the disease process or we might add endpoint $\mathcal{E}_2$ and use the composite $\mathcal{E}_*$. The null hypothesis of no treatment difference is given either by $H_0 : \lambda_1^{(0)}(\cdot) = \lambda_1^{(1)}(\cdot)$ in terms of the marginal hazards of $T_1^{(0)}$ and $T_1^{(1)}$ if $\mathcal{E}_1$ is being used or by $H_0 : \lambda_*^{(0)}(\cdot) = \lambda_*^{(1)}(\cdot)$ in terms of the marginal hazards of $T_*^{(0)}$ and $T_*^{(1)}$ when inferences would be based on $\mathcal{E}_*$. In both cases the logrank test Z (and Z_*) is the chosen statistic on which to base the conclusions.
Following Gómez and Lagakos (2011) we base the strategy in the behaviour of the asymptotic relative efficiency (ARE) of Z_* versus Z given by

$$\mathrm{ARE}(Z_*, Z) = \frac{\left(\int_0^1 \log\left(\frac{\lambda_*^{(1)}(t)}{\lambda_*^{(0)}(t)}\right) f_*^{(0)}(t)dt\right)^2}{\left(\log\left(\frac{\lambda_1^{(1)}(t)}{\lambda_1^{(0)}(t)}\right)\right)^2 (\int_0^1 f_*^{(0)}(t)dt)(\int_0^1 f_1^{(0)}(t)dt)} \tag{3}$$

where $f_1^{(0)}(t)$ and $f_*^{(0)}(t)$ are, respectively, the densities for $T_1^{(0)}$ and $T_*^{(0)}$ in group 0. The method proposes to use the composite endpoint instead of the primary endpoint if $\mathrm{ARE}(Z_*, Z) > 1.25$, to stick to the primary endpoint if $\mathrm{ARE}(Z_*, Z) < 1.1$, and whenever $1.1 < \mathrm{ARE}(Z_*, Z) < 1.25$ balance the benefits of using the composite endpoint over the relevant endpoint on the particular setting.

If such a method is being used for the design of a given clinical trial, the computation of the $\mathrm{ARE}(Z_*, Z)$ would need to be based on easily interpretable parameters such as the frequencies p_1 and p_2 of observing the endpoints $\mathcal{E}_1$ and $\mathcal{E}_2$ in treatment group 0, the relative treatment effects on $\mathcal{E}_1$ and $\mathcal{E}_2$ given by the hazard ratios $\mathrm{HR}_1 = \frac{\lambda_1^{(1)}(t)}{\lambda_1^{(0)}(t)}$ and $\mathrm{HR}_2 = \frac{\lambda_2^{(1)}(t)}{\lambda_2^{(0)}(t)}$ and to a lesser extent by the dependence degree between the relevant endpoint $T_1^{(0)}$ and the additional endpoint $T_2^{(0)}$ given by Spearman's rank correlation coefficient ρ.
As we see in (3) the $\mathrm{ARE}(Z_*, Z)$ depends on the marginal laws of $T_1^{(0)}$ and $T_*^{(0)}$ in group 0 and on the hazard ratios $\frac{\lambda_1^{(1)}(t)}{\lambda_1^{(0)}(t)}$ and $\frac{\lambda_*^{(1)}(t)}{\lambda_*^{(0)}(t)}$. Assuming Frank's copula for both groups with equal association parameter θ, the density of $T_*^{(j)}$ in group j $(j = 0, 1)$ is given by (2). Hence to derive the $\mathrm{ARE}(Z_*, Z)$ in terms of the above listed interpretable parameters we have to specify marginal parametric laws for $T_1^{(j)}$ and $T_2^{(j)}$ for both treatment groups 0 and 1 and we have to relate their parameters to the frequencies p_1 and p_2, the hazard ratios HR_1 and HR_2 and the Spearman's coefficient ρ.
If for $j = 0, 1$ and $k = 1, 2$, we choose Weibull distributions with scale parameters $b_k^{(j)}$ and shape parameters β_k chosen equal for both groups so

that the proportionality of the hazards holds, the marginal survival function is given by $S_k^{(j)}(t) = \exp\left(-(t/b_k^{(j)})^{\beta_k}\right)$. Then the relationship between $(b_1^{(0)}, b_2^{(0)}, b_1^{(1)}, b_2^{(1)}, \beta_1, \beta_2, \rho)$ and $(p_1, p_2, \mathrm{HR}_1, \mathrm{HR}_2, \beta_1, \beta_2, \rho)$ is given by:

1. The scale parameter $b_1^{(0)}$ is a function of p_1 and β_1 given by $b_1^{(0)} = \frac{1}{(-\log(1-p_1))^{1/\beta_1}}$.
2. (a) If $\mathcal{E}_1$ does not include a terminating event, the scale parameter $b_2^{(0)}$ is a function of p_2 and β_2 given by $b_2^{(0)} = \frac{1}{(-\log(1-p_2))^{1/\beta_2}}$.

   (b) If $\mathcal{E}_1$ includes a terminating event, $T_2^{(j)}$ might be censored by $T_1^{(j)}$ and the probability of observing $\mathcal{E}_2$ will depend on whether $T_1^{(j)} \leq T_2^{(j)}$ or not and hence on the joint density $f_{(1,2)}^{(0)}(t_1, t_2; \theta)$ given in (1). In this case, the scale parameter $b_2^{(0)}$ is a function of $(p_1, p_2, \rho, \beta_1, \beta_2)$ and it is found as the solution of equation $p_2 = \int_0^1 \int_v^\infty f_{(1,2)}^{(0)}(u, v; \theta) du dv$, or equivalently $p_2 = \int_{VL}^1 \left(\int_0^{UL^{(0)}(y)} g(x, y) dx\right) dy$ where $UL^{(0)}(y) = S_1^{(0)}((-\log y)^{1/\beta_2} b_2^{(0)})$, $VL = S_2^{(0)}(1)$ and $g(x, y) = \frac{\theta(1-e^{-\theta})\exp\{-\theta(x+y)\}}{(e^{-\theta}+e^{-\theta(x+y)}-e^{-\theta x}-e^{-\theta y})^2}$.
3. For $k = 1, 2$, the scale parameter $b_k^{(1)}$ is function of the scale parameter $b_k^{(0)}$, the shape parameter β_k and the hazard ratio HR_k as follows: $b_k^{(1)} = \frac{b_k^{(0)}}{\mathrm{HR_k}^{\frac{1}{\beta_k}}}$

Based on the guidelines established in Gómez and Lagakos (2011) they prove that often adding an endpoint to a relevant endpoint can be helpful if the relative effect on treatment on the additional endpoint is larger than on the relevant endpoint, harmful if the effect is smaller and whenever the effect on both endpoints is about the same the frequency of observing the endpoints and their correlation have to be taken into account before reaching a decision.

6 Illustration and conclusion

When studying the addition of succinobucol (Tardif *et al*, 2008) we can split the six components composite event $\mathcal{E}_*$ (cardiovascular death, resuscitated cardiac arrest, non-fatal myocardial infarction, non-fatal stroke, unstable angina, coronary revascularization) into the relevant endpoint $\mathcal{E}_1$ formed by cardiovascular death, resuscitated cardiac arrest, non-fatal myocardial infarction and non-fatal stroke and the additional endpoint $\mathcal{E}_2$ formed by hospitalization for unstable angina and coronary revascularization in order

to assess the best choice as primary endpoint for the analysis under the circumstances of this randomized clinical trial. Based on the published parameters the frequencies of observing $\mathcal{E}_1$ and $\mathcal{E}_2$ are respectively $p_1 = 0.0822$ and $p_2 = 0.0903$ with relative treatment effect on $\mathcal{E}_1$ given by a hazard ratio of $\mathrm{HR}_1 = 0.81$ and on $\mathcal{E}_2$ given by $\mathrm{HR}_2 = 1.05$. For these values the $\mathrm{ARE}(Z_*, Z)$ lies between 0.05 and 0.18 for all the possible degrees of association between $T_1^{(j)}$ and $T_2^{(j)}$ and irrespective of the chosen values for the shape parameters. It is hence clear in this case that adding hospitalization for unstable angina and coronary revascularization is not recommended. As a matter of fact the trial failed to show a statistically significant difference on $\mathcal{E}_*$ (p-value = 0.955) between the succinobucol group and the control group, while it showed a beneficial effect of succinobucol on the relevant endpoint $\mathcal{E}_1$ (p-value = 0.029). Note here that as pointed out in Section 3 composing an event on which treatment has a beneficial effect with an event showing no significant effect we have produced a composite endpoint where the effect has vanished. This clinical trial is extensively discussed in Gómez, Dafni and Gómez (2011) who assess, within the cardiovascular research context, the characteristics of the candidate individual endpoints that should govern the choice of using a composite endpoint as the primary endpoint by means of the asymptotic relative efficiency.
The paper has given more insight into the relationship between the hazard ratios of $T_k^{(1)}$ versus $T_k^{(0)}$ $(k = 1, 2)$ and of $T_*^{(1)}$ versus $T_*^{(0)}$ and has provided a straightforward relationship between the components of the $\mathrm{ARE}(Z_*, Z)$ and a small set of interpretable parameters.

Acknowledgments: This research was partially supported by Grant MTM2008-06747-C02-00 from the Ministerio de Ciencia e Innovación.

References

Gómez, G. and Lagakos, S.W. (2011). Statistical Considerations when Using a Composite Endpoint for Comparing Treatment Groups. *Submitted*.

Gómez, G. , Dafni, U. and Gómez, M. (2011). Informed Choice of Composite Endpoints in Cardiovascular Trials. *Submitted*.

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

Tardif, J-C., McMurray, J.J.V., Klug, E. *et al.* (2008). Effects of succinobucol (AGI-1067) after an acute coronary syndrome: a randomised, double-blind, placebo-controlled trial. *The Lancet*, **371**, 1761-1768.

Trivedi, P.K. and Zimmer, D.M. (2005). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econom.*, **1**, 1-111.

Identifying influential model choices in Bayesian hierarchical models

Peter Green[1], Ida Scheel[2], Jonathan Rougier[1]

[1] School of Mathematics, University of Bristol, Bristol BS8 1TW, UK
[2] Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, 0316 Oslo, Norway

Abstract: Real-world phenomena are frequently modelled by Bayesian hierarchical models. The building-blocks in such models are the distribution of each variable conditional on parent and/or neighbour variables in the graph. The specifications of centre and spread of these conditional distributions may be well-motivated, while the tail specifications are often left to convenience. However, the posterior distribution of a parameter may depend strongly on such arbitrary tail specifications. This is not easily detected in complex models. In this paper we propose a graphical diagnostic which identifies such influential statistical modelling choices at the node level in any chain graph model. Our diagnostic, the local critique plot, examines local conflict between the information coming from the parents and neighbours (local prior) and from the children and co-parents (lifted likelihood). It identifies properties of the local prior and the lifted likelihood that are influential on the posterior density. We illustrate the use of the local critique plot with applications involving models of different levels of complexity. The local critique plot can be derived for all parameters in a chain graph model, and is easy to implement using the output of posterior sampling.

Keywords: Chain graph; graphical diagnostic; hierarchical model; local critique plot; model criticism.

1 Introduction

Bayesian hierarchical models are now widely used to model complex, structured data. Such models are built from a large number of individual factors, representing the conditional distributions of each variable given those higher in the hierarchy, or, in the case of undirected models, potential functions for cliques of variables. Responsible, disciplined model-building requires that specification of all these factors should properly take into account prior information, whether this codifies scientific laws, earlier experiments, or degrees of subjective belief. However, this specification is a very challenging task, and there will often be a concern that it has been done imperfectly. In particular, while it may be relatively easy to specify the location and spread of a marginal or conditional distribution, the shape of the distribution, especially in the tail, is a more taxing question.

Yet the posterior distribution of all unknowns given data may depend on the trading-off of tails of individual model factors. It is important that this phenomenon be detected so that the modeller's attention can be drawn to particular statistical choices that are influential in the analysis, in order to confirm them or to reconsider.
In a simple Bayesian model, conflict between prior and data is easily detected, and this provides a diagnostic for criticising statistical modelling choices. Suppose we have a Bayesian model with a single unknown parameter θ. Then a graphical display of the prior and likelihood functions for θ quickly reveals the extent of any conflict between these two sources of information. In a general hierarchical model, identifying conflict between the sources of information contributing to the posterior distribution of a single node is a more subtle matter. This paper introduces a graphical diagnostic for this purpose.

1.1 Some previous work

Bayesian model criticism is often performed by considering a Bayesian p-value describing the compatibility of the observed data and the model. Such a p-value is typically obtained from some test-statistic or discrepancy measure (possibly depending on parameters as well as data) reflecting important aspects of the model, and a predictive distribution for this discrepancy measure. The type of predictive distribution used varies, e.g. the prior predictive distribution, the posterior predictive distribution (Meng, 1994), and the partial posterior predictive distribution (Bayarri and Berger, 1999, 2000; Bayarri and Castellanos, 2007). The latter approach avoids the need for informative prior distributions, as in the prior predictive approach, as well as the conservatism caused by the double use of data, as in the posterior predictive approach. This conservatism may also be handled by calibration (Hjort et al., 2006). These p-values are usually directed at one specific aspect of a model, not considering model fit at the individual nodes of a hierarchical model. Our idea of looking for conflict between the prior and likelihood information at the node level is not new. O'Hagan (2003) extends the node level residual analysis of Chaloner (1994) to other measures of conflict, to look for conflict between the different sources of information provided for the node in question. In practice, this is done by looking at how much the densities representing two different sources of information overlap, measured by the height of the densities (normalised to have unit maximum height) at the point where the two cross. Marshall and Spiegelhalter (2007) propose a similar p-value for measuring conflict at the node level in hierarchical models, which also avoids specifying a discrepancy measure and acts as an approximation to their cross-validatory, mixed p-value, when it exists.

1.2 Our objective

However, none of the above-mentioned conflict measures really address the nature of the conflict and the impact certain aspects of the prior and the likelihood have on the posterior analysis. The diagnostic we propose examines conflict at the node level by identifying where the posterior samples of a variable are located in what we call the local prior (the information coming from the parents and/or neighbours) and what we call the lifted likelihood (the information coming from children and co-parents).

2 Local critique plots

Our diagnostic technique is defined for a wide class of hierarchical models, including both directed and undirected dependencies (and so in particular handles spatial models in which a Markov random field is one of the model components). For the models we consider, the conditional independence structure can be represented by a *chain graph* (Lauritzen, 1996). In a general chain graph model, the full conditional distribution of any variable has a factorisation of the form

$$p(x_i|x_{-i}) \propto p(x_i|x_{\mathrm{pa}(i)}, x_{\mathrm{ne}(i)}) \times \prod_{c:i\in\mathrm{pa}(V(c))} p(x_{V(c)}|x_{\mathrm{pa}(V(c))})$$

2.1 The local prior and the lifted likelihood

We write $p_i(x) = p(x_i|x_{\mathrm{pa}(i)}, x_{\mathrm{ne}(i)})$ and call it the *local prior* for variable x_i. The other factor $l_i(x) = \prod_{c:i\in\mathrm{pa}(V(c))} p(x_{V(c)}|x_{\mathrm{pa}(V(c))})$ is the *lifted likelihood*.

- local prior measures the influence of *parents* and *neighbours*,
- lifted likelihood, that of *children* and (the possibly many) *co-parents*

2.2 Diagnostic functions

To get a standard 0–1 scale for prior and likelihood 'tension' we use cumulative versions of the local prior and the lifted likelihood:

$$\pi_i(x) = \int_{-\infty}^{x_i} p_i(x^{i\to u})du$$

where $x^{i\to u}$ means x with its ith element replaced by u.

$$\psi_i(x) = \frac{\int_{-\infty}^{x_i} l_i(x^{i\to u})du}{\int_{-\infty}^{\infty} l_i(x^{i\to u})du}$$

(assuming the denominator integral exists – we have a fix if it does not).
$\pi_i(x)$ and $\psi_i(x)$ measure *where* in the effective prior and likelihood the value x_i lies – 0 means the left tail, 1 the right tail. Both can depend on other variables x_j since all unknowns vary dependently.
We use the joint *posterior distribution* of $\pi_i(x)$ and $\psi_i(x)$ as a diagnostic for critically examining model assumptions. We propose to use a plot of this posterior distribution, which we call the *local critique plot* (Scheel, et al, 2011), to examine the degree in conflict between model assumptions at a node level in the graph.
In all but the most trivial applications, this distribution will be intractable, but our method can be implemented by fairly simple post-processing of MCMC output, which can be created by any software. All of our examples are based on output from WinBUGS (Lunn, et al, 2000).

3 An example

In the presentation, a number of motivating illustrations and more substantial examples will be given. In Figure 2, an example of an array of critique plots is shown. Within each frame, the plot shows a sample from the posterior distribution of the cumulative local prior (vertical axis) against the cumulative lifted likelihood (horizontal axis). The two rows present the plots for the 5 group means in a simple normal means hierarchical model, with the upper and lower rows corresponding to two different prior specifications. It is clear, from the way that the distribution is concentrated into the top left corner of the diagram, that for the more informative prior (upper row) there is substantial local prior–lifted likelihood conflict for the 5th parameter.

References

Bayarri,M. J. and Berger, J. O. (1999). Quantifying surprise in the data and model verification. In Bernardo, M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6*, pages 53–82. Oxford University Press.

Bayarri, M. J. and Berger, J. O. (2000). p-values for composite null models (with discussion). *J. Amer. Statist. Assoc.*, **95**:1127–1142.

Bayarri, M. J. and Castellanos, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statist. Sci.*, **22**:322–343.

Chaloner, K. (1994). Residual analysis and outliers in Bayesian hierarchical models. In Freeman, P. R. and Smith, A. F. M., editors, *Aspects of uncertainty: A tribute to D. V. Lindley*, chapter 10, pages 149–157. Wiley.

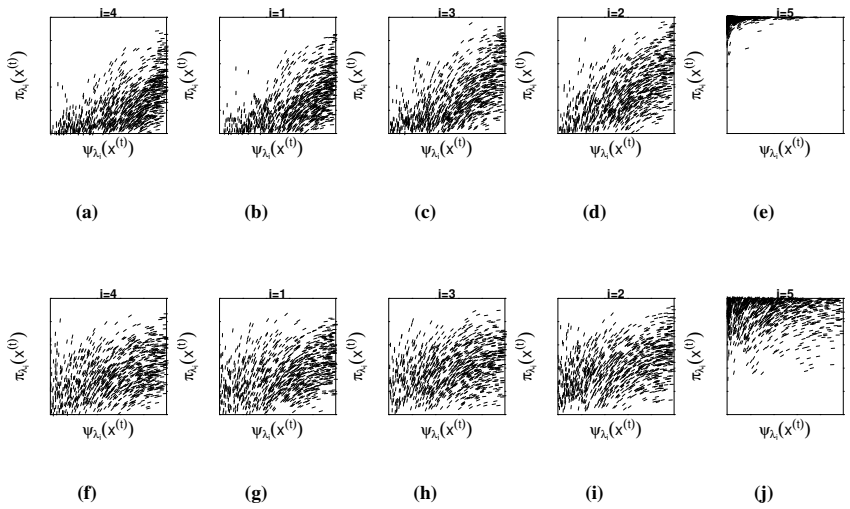


FIGURE 1. Local critique plots for group means in a simple normal means model, under two prior specifications: the less informative prior is used in the lower row of plots.

Hjort, N. L., Dahl, F. A., and Steinbakk, G. H. (2006). Post-processing posterior predictive p values. *J. Am. Statist. Assoc.*, **101**:1157–1174.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling frame-work: concepts, structure, and extensibility. *Statistics and Computing*, **10**:325–337.

Marshall, E. C. and Spiegelhalter, D. J. (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis*, **2**:409–444.

Meng, X. L. (1994). Posterior predictive p-values. *Ann. Statist.*, **22**:1142–1160.

O'Hagan, A. (2003). HSSS model criticism. In Green, P. J., Richardson, S., and Hjort, N. L., editors, *Highly Structured Stochastic Systems*, pages 423–444. Oxford: Oxford University Press.

Scheel, I., Green, P. J. and Rougier, J. C. (2011), A Graphical Diagnostic for Identifying Influential Model Choices in Bayesian Hierarchical Models. *Scandinavian Journal of Statistics*, **38**: doi: 10.1111/j.1467-9469.2010.00717.x

The Ecological Footprint of Taylor's Universal Power Law

Bent Jørgensen[1], Clarice G. B. Demétrio[2], Wayne S. Kendal[3]

[1] University of Southern Denmark, Odense, Denmark
[2] University of São Paulo, Piracicaba, Brazil
[3] The Ottawa Hospital Cancer Centre, Ottawa, Canada

Abstract: Taylor's "universal" power law is an empirical law for the relationship between the mean and variance of population abundance, which over time has been observed for a wide range of different species and ecosystems. Ever since it was proposed 50 years ago, the power law has given rise to discussions, because it seemed to lack a satisfactory theoretical explanation, in spite of its frequent observation in many different ecological, genomic, social science and epidemiological settings. We investigate a possible theoretical explanation for Taylor's power law based on the Tweedie distribution, which is an exponential dispersion model characterized by scale invariance; representing a statistical equilibrium that a system subject to random perturbations will approach over time. By exploring a new self-similarity hypothesis we derive the spatial correlation structure of the population, from which parameters may be estimated by means of estimating functions. These results enable us to investigate the mechanisms that control the spatial structure of the population, including the effects of environmental factors and interaction between species.

Keywords: Power variance function; Scale invariance; Self-similarity; Spatial distribution; Tweedie distribution.

1 Introduction

It is commonly so for species abundance data that sites with higher abundances tend to have higher variability. Let Y_{ij} denote the observed abundances, where $i = 1, \ldots, k$ denotes site and $j = 1, \ldots, n_i$ denotes replicate within site, and let $\mu_i = \mathrm{E}(Y_{ij})$ denote the mean abundance for site i. *Taylor's power law* (Taylor, 1961) stipulates a variance function of the form

$$\mathrm{Var}(Y_{ij}) = a\mu_i^b, \tag{1}$$

where a and b are positive parameters. These two parameters may be estimated by regressing the log empirical variance $\log S_i^2$ on the log empirical mean $\log \overline{Y}_i$, assuming independence both between and within sites. In this way Taylor (1961) confirmed the power law for 24 previously published ecological data sets. Taylor interpreted b as a species-specific index

of aggregation for the population, where $b = 1$ indicates a random dispersion pattern of the individuals and $b > 1$ an aggregated dispersion pattern. Values of b below 1 are rarely observed in practice, and by far the most common values of b are between 1 and 2.
Less than a quarter of a century later, Taylor et al. (1983) reported that the power law had been observed for no less than 444 different species (mainly insects), and subsequently the power law was observed again and again for different species to such an extent that it earned the name "universal". The power law has been observed in an ever expanding variety of different areas such as ecology, epidemiology and genetics, ranging from, say, the number of sexual partners reported by HIV infected individuals (Anderson and May, 1988), to the physical distribution of genes on human chromosome 7 (Kendal, 2004a). Taylor's power law has also been discussed in physics (Eisler et al. , 2008; Fronczak and Fronczak, 2010), where the phenomenon is known as *fluctuation scaling*.
Over the years there have been many attempts at explaining Taylor's power law theoretically, see e.g. Kendal (2004b) and references therein, but no explanation seems to have prevailed. The lack of a definitive theoretical explanation for Taylor's power law has given rise to much confusion in the literature, because the power law as such has little explanatory power, thereby reducing b to be just one out of many possible indices of aggregation (Pedigo and Buntin, 1994, p. 48). Instead we shall investigate the *Tweedie hypothesis* (Kendal, 2004b), namely that Taylor's power law is generated by the Tweedie distribution (Tweedie, 1984), whose variance function coincides with Taylor's power law, and whose shape is governed by the three parameters a, b and μ. We argue that the so-called Tweedie convergence theorem (Jørgensen et al., 1994, 2009) provides a compelling explanation for the ubiquity of Taylor's power law in nature, so that we are in effect observing the direct manifestation of a central limit effect. We discuss some of the historical background and possible ramifications of these ideas.

2 The double power law

A major challenge in ecology is to discover the mechanisms that control the spatial distribution of a population in its habitat. In the conventional approach to Taylor's power law, the replicates within site are obtained by subdividing the site into *quadrats*, where quadrat is the technical term for the frame or sampling device used for isolating an area to be sampled. By holding the quadrat size fixed, and assuming that μ_i varies between sites, we can estimate b, e.g. by the log regression method outlined above. By contrast, the *expanding bin method* of Kendal (2002, 2003, 2007) keeps μ fixed and varies the quadrat size t, which, as it happens, yields a second power law where the variance increases as a power of t.
A possible explanation for these phenomena may be found in the *double*

power law,

$$\mathrm{Var}(Y_{ij}) = at^{2-D}\mu_i^b, \tag{2}$$

where the *factional dimension* D belongs to the interval $(0, 2)$ and t denotes quadrat size. The double power law may be derived from a spatial self-similarity hypothesis, proceeding along similar lines as Jørgensen et al. (2011b). The parameters D, b and μ_i reflect three different aspects of the population distribution, namely the spatial correlation structure (see Eq. (3) below), the social behaviour of the individuals of the species (cf. Jørgensen et al., 2011a), and the combined effects of environmental factors, respectively. The fourth parameter a is a dispersion parameter in the sense of Jørgensen (1997, p. 5).

The special case $D = 1$ corresponds to independence between quadrats, but estimated values of D are usually between 0 and 1, corresponding to a positive correlation between quadrats within site. For example, Fairfield Smith (1938) investigated the heterogeneity of wheat yields and estimated D to be 0.746, whereas Bassingthwaighte et al. (1989) estimated D to be 0.4 in an investigation of regional myocardial blood flow, see also Kendal (2001). Like b, the power $2 - D$ in (2) is usually between 1 and 2, but as already explained, these two powers reflect two different aspects of the spatial distribution.

Inspired by Taylor's regression method for estimating b, one may tentatively estimate the parameters D and b of the double power law by regressing $\log S_i^2$ on $\log \overline{Y}_i$ and $\log t$, for example by adopting the expanding bin method in order to vary the quadrat size t. As pointed out by Perry (1981), Taylor's logarithmic regression method for estimating a and b suffers from a problem of bias. This problem was addressed by Jørgensen et al. (2011a), who proposed a bias-corrected Pearson estimating function for estimating b in Taylor's power law (1). More generally, we may combine the double power law (2) with a regression model for μ_i, for example a log-linear model

$$\log \mu_i = \mathbf{x}_i^\top \beta,$$

where $\mathbf{x}_i$ is a vector of site-specific covariates, and β is the corresponding regression vector. This type of regression model allows us to study the influence of site-specific environmental factors, including the presence of other species at the site, which may lead to a better understanding of factors that e.g. favor invasive species or lead to the extinction of species.

In order to estimate the regression vector β, we need the covariance structure of the data Y_{ij} within each site, which may be derived on the basis of the self-similarity hypothesis. In the present case where the quadrat size t is fixed, the correlation between two quadrats turns out to be constant,

$$\mathrm{Corr}(Y_{ij_1}, Y_{ij_2}) = 2^{1-D} - 1, \tag{3}$$

corresponding to an exchangeable correlation structure. We may hence estimate β by means of the corresponding quasi-score function; essentially

a generalized estimating equation with exchangeable correlation structure. The method of Jørgensen et al. (2011a) may be extended to deal with the double power law (2) by means of a joint bias-corrected Pearson estimating function for estimating a, b and D jointly with β.

3 Discussion

We have outlined a new approach to Taylor's power law based on the Tweedie hypothesis and the spatial self-similarity hypothesis, and we have argued for the plausibility of both hypotheses. This approach enables us to investigate the mechanisms that control the structure and spatial distribution of different biological populations. The question of the influence of environmental factors on the distribution patterns of species, as opposed to internal factors such as behaviour and species interactions, remains a topic of discussion in the ecological literature (Hutchinson, 1953; Jumars and Eckman, 1983; Camazine et al., 2003). Given suitable data, our approach allows a detailed investigation of the influence of environmental conditions and interactions between species, which can potentially revolutionize our understanding of the dispersion and aggregation patterns of biological populations.
We are currently planning a collaboration with ecological colleagues (see Jørgensen et al., 2011a) in order to obtain further data that can verify the hypotheses discussed here in practice. We are also developing the statistical methods required for such an investigation, as outlined above. A further goal is to be able to make predictions and simulations of the spatial distribution of biological populations. We plan to achieve this by developing a multivariate Tweedie distribution, based on the results of Jørgensen (2011).

Acknowledgments: This research was supported by the Danish Natural Science Research Council.

References

Anderson, R.M., and May, R.M. (1988). Epidemiological parameters of HIV transmission. *Nature*, **333**, 514–519.

Bassingthwaighte, J.B., King, R.B., and Roger, S.A. (1989). Fractal nature of regional myocardial blood flow heterogeneity. *Circulation Research*, **65**, 578–590.

Camazine, S., Deneubourg, J.-L., Franks, N.R., Sneyd, J., Theraula, G., and Bonabeau, E. (2003). *Self-Organization in Biological Systems.* Princeton Studies in Complexity. Princeton: Princeton University Press.

Eisler, Z., Bartos, I., and Kertész, J. (2008). Fluctuation scaling in complex systems: Taylor's law and beyond. *Adv. Phys.*, **57**, 89–142.

Fairfield Smith, H. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agr. Science*, **28**, 1–23.

Fronczak, A., and Fronczak, P. (2010). Origins of Taylor's power law for fluctuation scaling in complex systems. *Phys. Rev.* E, **81**, 066112.

Hutchinson, G.E. (1953). The concept of pattern in ecology. *Proceedings of the Academy of Natural Sciences of Philadelphia*, **105**, 1–12.

Jumars, P.A., and Eckman, J.E. (1983). Spatial structure within deep-sea benthic communities. Pages 399–451 of: Rowe, G. (Ed.), *The Sea.* New York: John Wiley and Sons.

Jørgensen, B. (1997). *The Theory of Dispersion Models.* London: Chapman & Hall.

Jørgensen, B. (2011). Construction of multivariate dispersion models. Submitted for publication.

Jørgensen, B., Demétrio, C.G.B., Kristensen, E., Banta, G.T., Petersen, H.C., and Delefosse, M. (2011a). Bias-corrected Pearson estimating functions for Taylor's power law applied to benthic macrofauna data. *Stat. Probab. Lett.*, doi:10.1016/j.spl.2011.01.005.

Jørgensen, B., Martínez, J.R., and Demétrio, C.G.B. (2011b). Self-Similarity and Lamperti convergence for families of stochastic processes. Submitted for publication.

Jørgensen, B., Martínez, J.R., and Tsao, M. (1994). Asymptotic behaviour of the variance function. *Scand. J. Statist.*, **21**, 223–243.

Jørgensen, B., Martínez, J.R., and Vinogradov, V. (2009). Domains of attraction to Tweedie distributions. *Lithuanian Math. J.*, **49**, 399–425.

Kendal, W.S. (2001). A stochastic model for the self-similar heterogeneity of regional organ blood flow. *Proc. Natl. Acad. Sci. USA*, **98**, 837–841.

Kendal, W.S. (2002). A frequency distribution for the number of hematogenous organ metastases. *J. Theor. Biol.*, **217**, 203–218.

Kendal, W.S. (2003). An exponential dispersion model for the distribution of human single nucleotide polymorphisms. *Mol. Biol. Evol.*, **20**, 579–590.

Kendal, W.S. (2004a). A scale invariant clustering of genes on human chromosome 7. *BMC Evolutionary Biology*, **4**(3).

Kendal, W.S. (2004b). Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. *Ecological Complexity*, **1**, 193–209.

Kendal, W.S. (2007). Scale invariant correlations between genes and SNPs on Human chromosome 1 reveal potential evolutionary mechanisms. *J. Theor. Biol.*, **245**, 329–340.

Pedigo, L.P., and Buntin, G.D. (1994). *Handbook of Sampling Methods for Arthropods in Agriculture.* Second Edn. Florida: CRC Press.

Perry, J.N. (1981). Taylor's power law for dependence of variance on mean in animal populations. *Appl. Statist.*, **30**, 254–263.

Taylor, L.R. (1961). Aggregation, variance and the mean. *Nature*, **189**, 732–735.

Taylor, L.R., Taylor, R.A.J., Woiwod, I.P., and Perry, J.N. (1983). Behavioural dynamics. *Nature*, **303**, 801–804.

Tweedie, M.C.K. (1984). An index which distinguishes between some important exponential families. Pages 579–604 of: Ghosh, J.K., and Roy, J. (Eds.), *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference.* Calcutta: Indian Statistical Institute.

Part 2. Contributed papers

Incomplete Clustered Data and Non-Ignorable Cluster Size

M. Aerts[1*], C. Faes[1], N. Hens[12], O. Loquiha[13], G. Molenberghs[12]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics, Universiteit Hasselt. Agoralaan 1, B-3590 Diepenbeek, Belgium
[2] Vaccine and Infectious Disease Institute, University of Antwerp, Campus Drie Eiken, Universiteitsplein 1, 2610 Wilrijk, Antwerp, Belgium
[3] Department of Mathematics and Informatics, Universidade Eduardo Mondlane, Avenida Julius Nyerere, Campus, 3453 ,P.O. Box 257, Maputo, Mozambique
* Corresponding author: marc.aerts@uhasselt.be

Abstract: We consider two different but typical settings of non-ignorable cluster sizes, one of which can be translated into an incomplete data setting. We propose and compare different marginal estimation methods (generalized estimating equations (GEE), weighted GEE, pseudo-likelihood (PL), weighted PL, within-cluster resampling). The two settings are illustrated by a veterinary epidemiology example with herds as clusters, and a development toxicity example with litters as clusters.

Keywords: Cluster-weighted generalized estimating equation, force of infection, prevalence, pseudo-likelihood, pairwise likelihood, missing covariate, inverse probability weight, within-cluster resampling

1 Introduction

Correlated outcome data appear in many applications and there are a number of alternative modelling approaches available nowadays. Here we focus on marginal approaches, not including full likelihood (the specification of which is considered too cumbersome in many situations): generalized estimating equations (GEE) and pseudo-likelihood (PL).
Typically, one considers the following marginal interpretations. A 'non-hierarchical' interpretation refers to the mean parameter as the mean for a typical member of the population of all members (over all clusters). A 'hierarchical' interpretation refers to the mean parameter as the mean for a typical member of a typical cluster of the population of clusters. Both interpretations are marginal but are different in the way they reflect the hierarchical structure represented by the clusters. The second hierarchical

interpretation corresponds to marginal analyses using weighted GEE or PL, with weights equal to the inverse of the cluster size.
This paper focuses on the analogy of non-ignorable cluster size with incomplete data, by studying two similar but different settings. One setting is characterized by the first data example on Bovine Herpesvirus-1 (BHV) in Belgian cattle herds (Boelaert *et al* 2000), with the objective to estimate the age-dependent prevalence of Bovine Herpesvirus-1. The other setting is characterized by data from a developmental toxicity study, investigating the dose-response relationship in mice of the potentially hazardous chemical compound di(2-ethylhexyl)phthalate (DEHP, used as plasticizers for numerous plastic devices made of polyvinyl chloride, see Tyl *et al* 1988).
In Section 2 we formalize the two different settings of non-ignorable cluster sizes. GEE and PL based marginal analyses are applied to the two data examples in Section 5. The paper ends with concluding remarks and topics of future research.

2 Methodology

Sizes of complete and incomplete clusters

Consider a population of clusters and assume we are interested in the estimation of the marginal parameter $\theta = E(y_{ij})$, the mean of the outcome for subject j within cluster $i = 1, ..., N$ (with N the number of clusters). In the first example y_{ij} refers to the indicator whether animal j within herd i has been infected or not, in the second example to the indicator whether fetus j within litter i is malformed or not. Given N, one observes cluster sizes $(m_1, ..., m_N)$.
Given N and m_i, one observes subject-level outcomes y_{ij}, $j = 1, ..., m_i$ and missingness indicators $\delta_i = (\delta_{i1}, ..., \delta_{im_i})$ with $\delta_{ij} = 1$ if y_{ij} is observed and 0 otherwise. Finally, denote $n_i = \sum_{i=1}^{m_i} \delta_{ij}$ the total number of observed outcomes for cluster i. The n_i are referred to as the "incomplete-cluster size" as opposed to the (complete-)cluster sizes m_i.
Different situations (not shown and discussed here) can be distinguished depending on the nonignorability of n_i and/or m_i, and whether m_i and/or n_i is ancillary to θ (not shown in this abstract). In case of nonignorability,

$$f(y_{ij}|x_{ij}, n_i, m_i) \neq f(y_{ij}|x_{ij}),$$

and consequently $E(y_{ij}|x_{ij}, n_i, m_i) \neq E(y_{ij}|x_{ij})$.
For the BHV data m_i is the size of the herd and n_i the number of animals sampled from the herd (in this particular case $m_i = n_i$, as all animals were tested). No outcomes are missing, but one can easily verify that herdsize is non-ignorable. The herdsize can be thought of as a variable summarizes several unobserved variables affecting the infection status of the animal.
For the NTP data m_i is the number of implants of litter i. The m_i is expected to be ignorable for θ, but the incomplete cluster of viable fetuses n_i is non-ignorable.

Marginal Approaches

The most popular marginal approach is undoubtedly generalized estimating (confining all attention to the specification of the first moments of the outcome; Liang and Zeger 1986) and variations such as GEE2 (Liang, Zeger and Qaqish 1991). Next to GEE, PL methods (Arnold and Strauss 1991) have become popular as an alternative to GEE and GEE2. Rather than modifying the 'independent' score equations, the full likelihood is simplified and replaced by a more manageable 'pseudo-likelihood'. Both approaches, GEE and PL, share statistical performance characteristics (e.g. efficiency, robustness) as compared to full likelihood (Geys, Molenberghs and Lipsitz 1988).
If the outcome measured among cluster members is independent of cluster size (i.e., if cluster size is ignorable), then weighted or unweighted GEE analyses produce equivalent results, and the GEE analyses may be optimized by using a more appropriate working correlation than the one corresponding to independence. When cluster size is non-ignorable however, the two marginal analyses are different and GEE should be based on the independence working correlation. Williamson, Datta, and Satten (2003) and Benhin, Rao and Scott (2005) proposed the weighted GEE to deal with non-ignorable cluster size as an alternative to the computation intensive within-cluster resampling approach of Hoffman, Sen, and Weinberg (2001). In this paper we also consider pairwise PL and the conditional version of PL in combination with weighing with the inverse of the cluster size and within-cluster resampling.

3 Analysis of BHV data

Table 1 shows estimated parameters of the model. In this section the age-dependent prevalence of BHV

$$P(\text{animal tests positive}|\text{age}) = \text{expit}\,(\beta_0 + \beta_1 \text{age})$$

based on different GEE and PL based models. The upper part refers to, from left to right: GEE with independence and next with exchangeable working correlation, cluster weighted GEE with independence and next with exchangeable working correlation, within cluster resampling GEE with one animal resampled per herd, within cluster resampling GEE with two animals resampled per herd (if available, otherwise one single animal) with independence and next with exchangeable working correlation. The lower part, first two columns to the left, refers to two repeated analyses based on within cluster resampling full likelihood Dale model (Dale 1986) with two animals resampled per herd (if available, otherwise one single animal). The next columns on the lower part are all PL estimates: pairwise likelihood based on the Dale model using weights $w_{1i} = 1/(n_i - 1)$, next $w_{2i} =$

	GEE-IND	GEE-EXCH	CW-GEE-IND	CW-GEE-EXCH	WCR-GEE (100R)	WC2R-GEE-IND(100R)	WC2R GEE-IND(100R)
$\hat{\beta}_0$	-1.175	-1.362	-1.615	-1.762	-1.600	-1.608	-1.651
se($\hat{\beta}_0$)	0.182	0.134	0.153	0.210	0.139	0.151	0.156
$\hat{\beta}_1$	0.017	0.019	0.017	0.018	0.017	0.017	0.018
se($\hat{\beta}_0$)	0.002	0.002	0.003	0.002	0.002	0.003	0.003
	WCR-D (100R)	WCR-D (100R)	2PLD w_1	2PLD w_2	2PLD w_3	2PLD w_4	WC2R-PL-D(100R)
$\hat{\beta}_0$	-1.646	-1.650	-1.215	-1.599	-0.842	-1.624	-1.664
se($\hat{\beta}_0$)	0.116	0.145	0.178	0.150	0.283	0.135	0.135
$\hat{\beta}_1$	0.018	0.018	0.018	0.017	0.014	0.018	0.019
se($\hat{\beta}_1$)	0.001	0.002	0.002	0.002	0.003	0.002	0.002
$\widehat{\text{OR}}$	11.847	10.533	11.604	10.591	12.376	10.229	8.980
se($\widehat{\text{OR}}$)	2.733	1.704	2.340	2.022	3.810	1.783	1.834

TABLE 1. BHV estimated models according to different GEE based and PL model strategies. Weights: $w_{1i} = 1/(n_i - 1), w_{2i} = 1/\{n_i(n_i - 1)\}, w_{3i} = 1, w_{4i} = 1/\{(1+0.5*(n_i - 1))(n_i - 1)\}$.

$1/\{n_i(n_i - 1)\}$, then $w_{3i} = 1$ (no weights), and finally $w_{4i} = 1/\{(1 + 0.5 * (n_i - 1))(n_i - 1)\}$. The final right column refers to within cluster resampling where two pairs are resampled per herd (if available), and using the Dale model. Within cluster resampling was always combining the results of 100 runs. As expected, the estimates of CW-GEE-IND are in line with the WCR-GEE-IND, WCR2-GEE-IND and WCR-D. It is also clear that PL needs a careful consideration of the different weighting options (combining PL-weights with cluster weights). Using no weights ($w_{3i} = 1$) leads to a biased estimate for the intercept.

4 Analysis of NTP data

The dams were sacrificed, slightly prior to normal delivery, and the status of uterine implantation sites recorded. A total of 1082 live fetuses were examined for malformation, coded as a binary indicator. Fetuses were clustered within mothers; hence the implied association needs to be accommodated in the analysis. Table 2 suggests clear dose-related trends in the malformation rates. The average litter size (number of viable animals) decreases with increased levels of exposure to DEHP. This setting is different in nature as compared to the BHV-example. Indeed, the number of implants m_i of litter i is not related to any outcome, whereas the number of viable fetuses n_i of litter i can be seen as the result of a non-ignorable missing data mechanism. It is well-known in the missing data context that, while ignorability only requires the relatively general missing at random assumption for likelihood and Bayesian inferences, this result cannot be invoked when popular non-likelihood- based method, such as GEE and PL. Molenberghs *et al* (2011)

Exposure	Dose	♯Dams,≥ 1 Impl.	Viab.	Live	Litter Size (mean)	Malformations Ext.	Visc.	Skel.
DEHP	0	30	30	330	13.2	0.0	1.5	1.2
	44	26	26	288	11.1	1.0	0.4	0.4
	91	26	26	277	10.7	5.4	7.2	4.3
	191	24	17	137	8.1	17.5	15.3	18.3
	292	25	9	50	5.6	54.0	50.0	48.0

TABLE 2. NTP summary data for DEHP. The dose is in *mg/kg/day*.

propose a suite of corrections to the standard form of pseudo-likelihood, to ensure its validity under missingness at random. Their corrections follow both single and double robustness ideas, and is relatively simple to apply. Molenberghs *et al* (2011) illustrate that the naive complete case PL-analysis and its weighted version show greatly inflated standard errors, due to the dramatic sample size reduction (only 23 complete litters out of 108 litters with at least one viable fetus). They also proposed a doubly robust version in this setting, which is efficient while it does not even need an explicit model for the missingness probabilities.

5 Discussion

We can conclude that informative cluster sizes do occur more often in the analyses of clustered data than often anticipated. Williamson *et al* (2007) illustrate the use of weighted GEE in a condom use study. In this case, the cluster would be the subject and the individual sex act would be the observation (subunit) within the cluster. Loquiha (2010) studied maternal mortality in Mozambique, and he used weighted GEE to take the numbers of admissions to the health centers as non-ignorable cluster sizes into account. A setting which fits into the missing data context appears in the study of factors associated with periodental disease. Here, data are available on the disease status of each tooth of an individual, but as persons with poor dental health are likely to have fewer teeth than do persons with good dental health, number of teeth is non-ignorable.
Our study indicates that results of GEE and PL approaches are very comparable, at least if appropriate weights are chosen. Pseudo-likelihood is especially appealing in the case of high-dimensional multivariate outcomes, of different nature (mixed continuous and categorical). Simulations are planned to examen their performance further, in different settings.

References

Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya: the Indian Journal of Statistics - Series B*, **53**,

233-243.

Benhin, E., Rao, J.N.K., and Scott, A.J. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika*, **92**, 435–450.

Boelaert, F., Biront, P., Soumare, B., Dispas, M., Vanopdenbosch, E., Vermeersch, J., Raskin, A., Dufey, J., Berkvens, D., and Kerkhofs, P. (2000). Prevalence of bovine herpesvirus-1 in the Belgian cattle population. *Preventive Veterinary Medicine*, **45**, 285–295.

Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 721-727.

Geys, H., Molenberghs, G., and Lipsitz, S.R. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginal odds ratio models. *Journal of Statistical Computation and Simulation*, **62**, 45-72.

Hoffman, E.B., Sen, P.K., and Weinberg, C.R. (2001), Within-cluster resampling. *Biometrika*, **88**, 1121–1134.

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

Liang, K.Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3-40.

Loquiha, O. (2010). Statistical methods for maternal mortality analysis: The case of Mozambique. Master Thesis, University of Hasselt.

Molenberghs, G., Kenward, M.G., Verbeke, G., and Birhanu, T. (2011). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, **21**, 187–206.

Tyl, R.W., Price, C.J., Marr, M.C., and Kimmel, C.A. (1988). Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. *Fundamental and Applied Toxicology*, **10**, 395–412.

Williamson, J.M., Datta, S., and Satten, G.A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics*, **59**, 36–42.

Williamson, J.M., Kim, H-Y, and Warner, L. (2007). Weighting condom use data to account for nonignorable cluster size. *Annals of Epidemiology*, **17**, 603-607.

Bayesian Lee-Carter Model: A Spatio-Temporal Approach.

A. Alvaro-Meca[1], V. Hernandez[1], A. Debón[2], R. Gil[1], A. Gil de Miguel[1]

[1] Dept Preventive Medicine and Public Health and Medics Immunology and Microbiology,
University Rey Juan Carlos, Alcorcón, Madrid, Spain

[2] Corresponding author: andeau@eio.upv.es. Centro de Gestión de la Calidad y del Cambio, Universidad Politécnica de Valencia. Spain.

Abstract: Mortality decreased in all countries in the European Union in the last century, presenting similar patterns in the change in mortality. Despite these similar trends, there are still considerable differences in the levels of mortality in these countries and between men and women. The aim of this article is to adjust and predict mortality and life expectancy at birth for both sexes, in 16 countries in the European Union, modifying the Lee-Carter model with the inclusion of a spatial component. Mortality is decreasing in these countries, and the historical difference between sexes is disappearing, but differences between the studied countries remain.

Keywords: Bayesian Lee-Carter; Europe; Mortality; Spatial.

1 The model

One of the most common models used for the representation of the evolution of mortality and also one of the most used nowadays by actuaries and demographers is the Lee-Carter model (Lee and Carter, 1992). This model and its different extensions have been applied by many authors. The work of Pedroza (2006) in the Bayesian framework is of especial interest. In this paper we propose a Lee-Carter model which modifies mortality specifically for countries that belong to a group by the inclusion of a geographic factor. The Lee-Carter model with a spatial component that we suggest can be formulated as follows,

$$\log\left(\frac{q_{xtr}}{1-q_{xtr}}\right) = a_x + b_x k_t + S_r + \epsilon_{xt}, \tag{1}$$

where x refers to age, t to the year of death and, r to countries to be included in the model, a_x, b_x are vectors of unknown parameters, k_t is an unobserved time series process and S_r is the spatial random effect. ϵ_{xt}

errors are assumed to be independent and identically distributed according to a normal with mean 0 and common variance σ_t^2. This model permits the comparison between countries by means of a simple Index, S_r.

Under the Bayesian paradigm, the researcher can incorporate his knowledge about the matter he is dealing with as a priori information. Afterwards this information is combined with the observed data to obtain the a posteriori distribution of the parameters about which inference is expected to be carried out. Moreover, the Bayesian estimation first requires the likelihood function to be provided, in our study, the Lee-Carter model extended with the spatial component, and the a priori distributions of the parameters of interest. Here, non-informative distributions were chosen for the parameters a_x, b_x, k_t together with non-informative distributions for the variances of parameter σ_t^2. To be precise, we have chosen a distribution $N(0, \sigma_t^2)$ with $\sigma_t^2 \sim Gamma(0, 0.001)$. The initial a priori distributions for the starting point b_0 and k_0 were assumed to be 1 and 0 respectively. In order to study the spatial dependence S_r, we chose a conditional autoregressive model $CAR(\sigma_r^2)$ (Besag, 1974; Clayton, 1993) with $\sigma_r^2 \sim Gamma(0, 0.001)$ as an a priori distribution. This approximation, the most common and the simplest computationally, approximates the spatial dependence as a mean of the spatial effect of its nearby areas.

In order to implement the Bayesian model, it is necessary to obtain the a posteriori distribution of all the parameters of the model. However, the posterior distribution inference is analytically intractable. Instead, several MCMC algorithms have been proposed to obtain the posterior distribution of the parameters. To be precise, we used Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) to draw samples from the joint posterior distribution. This algorithm consists of iteratively sampling from the conditional distribution of each of the parameters given, assigning values to all the other parameters and the data. We used the Winbugs software to fit the model and perform all the posterior inference.

We ran three different chains using 2,500 iterations for the Gibbs sampler, with different over-dispersed starting values. We took the first 500 as burn-in and in the end we obtained a sample for each parameter by selecting the last 2,000 values of each one of the chains. Results presented here are based on the combined 6,000 draws from the posterior distribution. The convergence of the chains was checked by using the Gelman-Rubin statistic (Gelman and Rubin, 1992) implemented in the R-CODA package (Plummer *et al.* 2009). Values lower than 1.1 suggest that convergence has been reached. We have calculated the Gelman-Rubin statistic for the parameters of the model, and in all of them, values lower than 1.01 are reached, indicating convergence.

The last step of the Lee-Carter model consists of fitting a temporal series in the index k_t. In order to carry out predictions of the Lee-Carter model, within the Bayesian framework, the work of Pedroza (2006), has to be considered. This paper carries out predictions for future years by means of

Gibbs sample, following two steps:

1. First drawing the k_t from a normal random distribution with the correct parameters estimated from the data
2. Then, given the k_t, drawing the log mortality rates from a normal distribution with corresponding parameters.

Our proposal consists of fitting a temporal series to the index k_t by using the Box-Jenkins methodology. We decided to use standard adjustment since it provides consistent predictions. However, one disadvantage is that predictions in this model are only based on the variation of one parameter assuming the variabilities provided by the geographic component and age are constants .

2 Application

We analyse mortality in both sexes and for 16 countries in the European Union (Germany, Austria, Belgium, Denmark, Spain, Finland, France, Holland, Ireland, Italy, Luxembourg, Portugal, the United Kingdom and Sweden), using the Bayesian Lee-Carter model, and adding a geographical component for the period 1989-2006. Data have been obtained from Human Mortality Database (2009). These data have mortality details by individual age and year for each country. We used data from 1989 to 2006 in order to adjust. As we wanted to consider Germany as one country we chose 1989 since it set a historic milestone caused by the fall of Berlin Wall and the unification of Germany
Figure 1 shows the exponented spatial effects during the years 1989- 2006, the main conclusion is the convergence of differences in European mortality due to gender.

Acknowledgments: This work was partially supported by a grant from MEyC (Ministerio de Educación y Ciencia), Spain, project MTM2008-05152.

References

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

Clayton, D.G. *et al* (1993). Spatial correlation in ecological analysis. *International Journal Epidemiology* **22**, 193-202.

Debón, A. *et al* (2008). Modelling and forecasting mortality in Spain. *European Journal of Operation Research* **189**, 624-637.

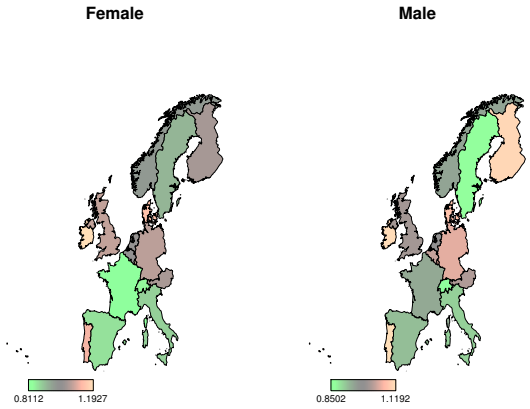


FIGURE 1. Spatial Effects 1989-2006

Gelfand, A.E., and A.F.M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741.

Gelman and Rubin (1992). Inference form iterative simulations using muiltiple sequences (with discussion). *Statistical Science* **7**, 457–472.

Human Mortality Database (2009). Human Mortality Database (2009). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on June 1, 2010).

Lee, R. and Carter, L. (1992). Modelling and forecasting U. S. mortality. *Journal of the American Statistical Association* **87**, 659-671.

Pedroza, C. (2006). A Bayesian forecasting model: predicting U.S. male mortality. *Biostatistics* **7**, 530-550.

Plummer *et al.* (2009). coda: *Output analysis and diagnostics* for MCMC. R package version 0.13-4.

From Empirical Bayes to Leaving-One-Out

Jesús Andrés-Ferrer[1] , Hermann Ney[2]

[1] Universidad Politécnica de Valencia, Valencia, Spain
[2] RWTH Aachen University, Aachen, Germany

Abstract: Empirical Bayes (EB) is a very appealing technique for tasks in which many outcomes of the population do not occur in the sample. In these tasks, it is necessary to estimate sparse probabilities and the EB is known to provide good estimates. However, EB estimates have two main drawbacks: they may be non-monotonic and under some circumstances they cannot be computed. This work presents a framework to constrain EB method by means of its equivalence with the leaving-one-out estimation. Two solutions are derived that amend the previous problems by applying two different sets of constraints: interval and monotonic. The typical application (for us) is language modelling.

Keywords: Leaving-one-out; Empirical Bayes; monotonic constraints, interval constraints.

1 Introduction

We assume a random sample to be drawn from a large population of outcomes. If a particular outcome is observed $r = 0, 1, 2, \ldots$ times in a sample of size N, then the maximum likelihood (ML) estimate r/N is not a good estimate of the population probability, p_r, when r is small. The probability of infrequent outcomes is referred to as *small probabilities* by Good (1953). The small probabilities estimation is a common problem in several tasks such as language modelling, see Ney et al (1997), where the majority of the outcomes are not observed in the sample.
The *Empirical Bayes (EB)* is a Bayesian approach in which the unknown prior distribution is estimated from the sample after making suitable assumptions, see Good (1953) and Robbins (1956). EB uses the posterior mean as an estimator for the population probabilities. First, we assume that the observed counts follow a binomial distribution; and then the population probability for an outcome that has occurred r times is approximated by

$$p_r = \frac{r+1}{N+1}\frac{E_{N+1}(r+1)}{E_N(r)} \approx \frac{r+1}{N}\frac{E_N(r+1)}{E_N(r)} \quad , \tag{1}$$

where $E_N(r)$ is the posterior mean of the count. It is worth noting that the posterior mean is computed over an unknown prior distribution. The posterior means in Eq.(1) are approximated by their observed value n_r as

follows

$$p_r \cong \frac{1}{N}(r+1)\frac{n_{r+1}}{n_r} \quad , \tag{2}$$

where n_r denotes so-called *counts-of-counts (COC)*, which count for the number of outcomes that have been observed r times in the sample, i.e., there are n_r outcomes with a count equal to r.
This variant of EB provides good estimates for small probabilities by shifting probability mass from the outcomes that are frequently observed in the sample to those outcomes that are rarely observed. However, the EB estimates do not require the probabilities p_r to be increasingly monotonic with r. Moreover, the EB estimates in Eq. (2) cannot be computed (or are 0) if n_r (or n_{r+1}) is equal to 0, which specially happens for frequent outcomes. Consequently, the reliability of the probability estimates depend on the observed COC. Ensuring monotonicity while retaining the good properties of EB estimates is very difficult from the EB perspective even making further assumptions. In this work, we will constrain the probabilities, p_r, in order to ensure monotonicity by means of leaving-one-out. In the following sections, two set of constraints will be proposed.

2 Leaving-one-out estimation

The *leaving-one-out (LOO)* estimation as an equivalent derivation of the EB estimates was suggested in Nadas (1985). As in the ML case, we form equivalence classes by grouping all outcomes with the same count r in the class modelled with the probability p_r. The LOO estimation is based on rounds where one outcome observation plays the testing role whereas the remaining observations play the training role. In a given round a sample observation of an outcome, is left out for testing; and a model is trained with the remaining observations in order to predict the probability of the left out observation. The sample counts are then modified accordingly. For instance, if an outcome has been observed r times, then it is moved to the equivalence class $r-1$. The LOO log-likelihood is obtained by repeating this process for all observed counts $r = 1, \ldots, R$ as follows

$$\begin{aligned} F(\{p_0^{R-1}\}) &= \textstyle\sum_{r=1}^{R-1}(r+1)n_{r+1}\log p_r + (n_1-1)\log p_0 & (3)\\ &\approx \textstyle\sum_{r=0}^{R-1}(r+1)n_{r+1}\log p_r & (4)\end{aligned}$$

where the probability of the most frequent outcome, p_R, is assumed to be given, e.g. relative frequency R/N; and where we assumed that $p_0^R = p_0, \cdots, p_r, \cdots, p_R$ are normalised, i.e., they sum up to 1. This constraint can be expressed in the Lagrangian function as follows

$$F(\{p_0^{R-1}\}, \lambda) = \textstyle\sum_{r=0}^{R-1}(r+1)n_{r+1}\log p_r + \lambda(\sum_{r=0}^{R} n_r p_r - 1)\,, \tag{5}$$

where λ is a normalisation constant. The solution to the maximisation of Eq. (4) is obtained by equalling to 0 the gradient of the Lagrangian

$$p_r = (1-\tfrac{R}{N})\tfrac{1}{N}(r+1)\tfrac{n_{r+1}}{n_r} \approx \tfrac{1}{N}(r+1)\tfrac{n_{r+1}}{n_r} \quad , \tag{6}$$

which is equivalent to the EB estimates in Eq.(2). In the remaining, we will show how EB can be constrained under the LOO framework.

2.1 Leaving-one-out with interval constraints

In order to ensure monotonicity EB estimates can be constrained to be in monotonically increasing intervals as follows

$$p_0 \leq \tfrac{1}{N}, \qquad \tfrac{r-1}{N} \leq p_r \leq \tfrac{r}{N} \qquad r = 1, \ldots, R-1 \quad . \tag{7}$$

Conceptually, in order to assure monotonicity an additional constraint would be necessary $p_0 \leq p_1$, even though this constraint is not active in practice. Anyway, the proposed algorithm can be easily extended to include this additional constraint.
In order to find the optimal set of parameters that maximise Eq. (4) constrained by Eq. (7); we consider the estimates p_0^R as a function of the unknown normalisation parameter λ. For doing so, we apply the Karush-Kuhn-Tucker (KKT) conditions and the water-filling method, see Boyd et al (2004), obtaining the following solution

$$p_r(\lambda) = \max\{\tfrac{r-1}{N}, \min\{\tfrac{1}{\lambda}(r+1)\tfrac{n_{r+1}}{n_r}, \tfrac{r}{N}\}\} \quad . \tag{8}$$

Then, the value of λ is obtained by reformulating the normalisation constraint as follows

$$Q(\lambda) = \textstyle\sum_{r=0}^{R} n_r p_r(\lambda) \tag{9}$$

with $p_R(\lambda)$ fixed. Finally, the optimal value of λ must satisfy $Q(\lambda) = 1$. Since the normalisation function $\lambda \mapsto Q(\lambda)$ is monotonically decreasing, it is straightforward to find the optimal value of λ.

2.2 Leaving-one-out with monotonic constraints

Interval constraints force the EB estimates to be in a narrow interval. A more flexible set of constraints that also ensures monotonicity are proposed here

$$p_r \leq p_{r+1} \qquad r = 0, 1, \ldots, R-2 \tag{10}$$

where as usual p_R is kept fixed and not estimated.
The solution to such maximisation problem will result in segments $[r_k, r_{k+1}]$ of constant probabilities q_k for $k = 1, \ldots, K$ such that:

$$\underbrace{\ldots = p_{r_k - 1}}_{q_{k-1}} < \underbrace{p_{r_k} = \ldots = p_{r_{k+1}-1}}_{q_k} < \ldots < \underbrace{p_{r_{k+1}} = \ldots}_{q_{k+1}}, \tag{11}$$

with the following boundary conditions: $r_0 \equiv 0$ and $r_K = R - 1$.
The optimisation problem consists in finding these segment boundaries r_0^K and the probability of each segment q_k. The probability estimates q_k

for each segment $[r_k, r_{k+1}]$, are computed as a function of the unknown segmentation, r_0^K. The optimal values are given by $\hat{q}_k = q(r_k, r_{k+1} - 1)$, where

$$q(r', r) = \frac{1}{\lambda} \frac{A(r', r)}{\sum_{s=r'}^{r} n_s}, \quad 0 \leq r' \leq r < R - 1 \tag{12}$$

with $A(r', r) = \sum_{s=r'}^{r} (s+1) n_{s+1}$, and with λ being a normalisation constant independent of the segmentation, $\lambda = N/(1 - n_R p_R)$. The optimal segmentation is found by the following recurrence

$$F(r) = \max_{r' \leq r \,:\, p_{r'} < p_r} \{F(r'-1) + A(r', r) \log(q(r', r))\} \quad , \tag{13}$$

and tracing back the optimal boundaries used for computing $F(R-1)$.

3 Language modelling

One of the applications of EB is language modelling (LM). LM consists in computing the probability of a given sentence. In this task all the possible sequences of words up to a given length are considered the population and each specific sequence as an outcome. In Andrés-Ferrer and Ney (2009) experiments are reported for the interval constrained estimates applied to a large language modelling task (1.7 million of running words).

Acknowledgments: First author work has been supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project.

References

Nadas, A. (1985) *On Turing's Formula for Word Probabilities.* IEEE Transactions on Acoustics, Speech, and Signal Processing.**33**(6),1414–1416.

Robbins, H. (1955) *An empirical Bayes approach to statistics.* Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. **I**, 157–163.

Ney, H.; Martin,S. and Wessel, F. (1997) *Statistical Language Modelling Using Leaving-One-Outl* In "Corpus-Based Statistical Methods in Speech and Language", 174–207.

Good, I. J. (1953) *The population frequencies of species and the estimation of population parameters.* Biometrika,**40**(3), 237–264.

Boyd, S. and Vandenberghe, L. (2004) *Convex Optimization.* Cambridge University Press, 244–249. ISBN:0521833787

Andrés-Ferrer, J. and Ney, H. (2009) *Extensions of absolute discounting (Kneser-Ney method).* Proceedings of the ICASSP, 4729–4732.

Model Based Estimates of Long-Term Persistence of Induced HPV Antibodies: A Flexible Subject-Specific Approach

M. Aregay[1], Z. Shkedy[2], G. Molenberghs[1,2]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B3000 Leuven, Belgium; *mehreteab.fantahun@student.kuleuven.be*

[2] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, Agoralaan, B3590 Diepenbeek, Belgium

Abstract: To predict the long-term persistence of vaccine-induced anti-HPV-16/18 and to obtain the estimated time points where the individual titers are below the threshold value for protection, a fractional-polynomial model, derived in a data-driven fashion, was used. Initially, model selection was done from among the second-order fractional polynomial, first-order fractional polynomial, and linear mixed model. According to a likelihood ratio-test statistics, the first-order fractional polynomial was selected. Apart from the fractional polynomial model, we also applied a power law model, which is a special case of the fractional polynomial model. Both models were compared with the AIC criterion. Within the observation period, the fractional polynomials fitted the data better than the power-law model; this does not imply that it fits best over the long run. Therefore, we point out that the persistence of the anti-HPV responses induced by these vaccines can only be ascertained empirically by long-term follow-up analysis.

Keywords: Fractional Polynomial Model; AIC; Power-law Model

1 Introduction

During the 1990s, epidemiological studies, supported by molecular technology, provided evidence on the causal role of some human papillomavirus (HPV) infections in the development of cervical cancer (Bosch *et al.* (2002)). In recent years, much attention has been paid to the possibility of vaccination against HPV as a means of preventing cervical pre-cancerous lesions and cancer. David *et al.* (2009) studied the HPV data set, which records information until 75 months. The authors used a conventional power-law model as well as a modified power-law model. However, the first model is limited by the assumption of a progressive decay of antibody and antibody-producing B-cells, while the second model model assumes, in addition, that the proportion of memory B-cells remains stable and identical for all women, which is biologically unlikely.

With these considerations in mind, a modeling endeavor was undertaken to define the long-term duration of vaccine induced anti-HPV. The objective of this paper is to predict the long-term persistence of vaccine-induced anti-HPV-16 and anti-HPV-18 antibodies, to obtain the estimated time point above the threshold value, and also to predict the proportion of subjects above the threshold value with flexible fractional-polynomial model.

2 Motivating Data Set

The data consist of healthy women aged 15–25 years. In the initial phase, blood samples from the 514 women who came from North America (USA and Canada) and Brazil, were evaluated at months 0, 7, 12, and 18 and annually thereafter up to month 90 after first vaccination, for the presence of HPV-16/18 antibodies using a type-specific enzyme-linked immunosorbent assay (ELISA). For the current evaluation, we included women who had received three doses of AS04-adjuvanted HPV-16/18 vaccine and had at least one time point after the third dose with serology results available for at least one vaccine antigen component.

3 Exploratory Data Analysis

The individual profile curves reveal substantial variability between subjects. From the evolution of the mean, we noticed that the decline in antibody level is higher in the first few months, followed by a moderate decrease until the end of the follow up period. The observed variances for each category month indicates that the variance is not constant over time which means that a random intercept model might not be an appropriate model for these studies.

4 Modeling the Mean Antibody Using Subject-specific Fractional Polynomials

Fractional polynomials(Royston and Altman(1994)) were proposed as flexible parametric approach in order to describe the dependency between a response of primary interest and a covariate. In our example, the response of primary interest is log-transformed antibodies and the covariate is time. The mean structure of a fractional polynomial model can be formulated in the following way:

$$\sum_{i=0}^{m} \beta_i H_i(t) + \sum_{i=0}^{m} b_i H_i(t), \tag{1}$$

where m is an integer, $p_1 \leq p_2 \leq \cdots \leq p_m$ is a sequence of powers and $H_i(a)$ is a transformation function given by

$$H_i(t) = \begin{cases} t^{p_i} & \text{if } p_i \neq p_{i-1}, \\ H_{i-1}(t) \times \log(t) & \text{if } p_i = p_{i-1}, \end{cases} \tag{2}$$

with $p_0 = 0$ and $H_0 = 1$. Note that, to take subject heterogeneity into account, we assume two components in the mean structure. The first consists of the fixed parameters β and the latter the subject-specific parameters b_i. For the analyzes presented here, a first order ($m = 1$) and second order ($m = 2$) fractional polynomial was used. Hence, the mean structure for the first order mixed fractional polynomial can be written as

$$f(t_{ij}) = (\beta_0 + b_{0_i}) + (\beta_1 + b_{1_i})t_{ij}^{p_1}$$

Here, b_{0_i} and b_{1_i} are subject specific parameters. Note that for $p_1 = 0$ the factional polynomial model is reduced to the power low model.In this study we used a fractional-polynomial mixed model with serial correlation.

5 Long Term Prediction Using Subject-specific Fractional Poynomials

In the first stage, we selected the model for the serial correlation process and the fractional polynomial model for the mean structure. Four models for the serial correlation process were considered: (1) a model without serial correlation process, (2) a local exponential model, (3) Gaussian serial correlation, and (4) exponential serial correlation. To select the power of the fractional polynomial, powers in the range $\{-3, -2.75, -2.5, \ldots, 2.5, 2.75, 3\}$ were considered. For HPV-16, the model with the smallest AIC obtained for the $p = -1.5$ while for HPV-18 the power is equal to -1.25. For HPV-16, the best serial correlation model is the local exponential model (AIC= 201.5) while for HPV-18 the Gaussian serial correlation model is the model with the best goodness-to-fit (AIC= -317.9). Next, a second-order FP was fitted and the so-called *Function Selection Procedure* (FSP)(Royston and Sauerbrei (2008)) was applied. The first-order FP, reported above, are to be preferred. For each subject in the study, the time to cross a given threshold value, t_τ, can be calculated from the predicted serological result. Three different threshold values(τ) were used. For HPV-16; 1.474,2 and 2.58, and for HPV-18; 1.355, 2, and 2.409. We noticed that for lower threshold (1.47 and 1.355 for HPV-16 and HPV-18, respectively) the proportion of unprotected subjects is 0.002% (1 subject only) for HPV-16 and HPV-18. For $\tau = 2$, 93.2% $(90.7, 95.1)$ and 85.8% $(82.5, 88.6)$ are protected during 25 years for HPV-16 and HPV-18, respectively. Figure 1 shows the long term predicted means for 25 years

6 Discussions and Conclusions

In summary, based on a fractional polynomial model and follow-up data for more than 500 vaccinated women, we are able to predict that vaccination with the AS04-adjuvanted HPV-16/ 18 vaccine induces persistence of both

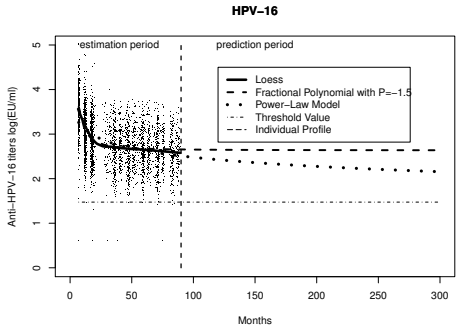


FIGURE 1. Long term(25 years) Prediction for HPV-16

anti-HPV-16 and -18 antibodies for at least 25 years. In this study, the results of long-term prediction using a fractional polynomial model corroborates the findings of previous work which is done on the same data set up to 75 months with a modified power-law model by David *et al.* (2009). Both models feature a long term plateau. The modified power-law model introduces bias towards a plateau in predicting long-term antibody levels. However, the fractional polynomial model is very flexible (Royston and Altman (1994)), being a data-driven method.

References

Bosch, F.X., Lorincz, A., Munoz, N., Meijer, C.J. and Shah, K.V. (2002). The causal relation between human papillomavirus and cervical cancer. *J Clin Pathol* , **55**, 244–65.

David, M., Van Herck, K., Hardt, K., Tibaldi, F., Dubin, G., Descamps, D. and Van Damme, P. (2009). Long-term persistence of anti-HPV-16 and -18 antibodies induced by vaccination with the AS04-adjuvanted cervical cancer vaccine: Modeling of sustained antibody responses. Article in Press.

Royston, P. and Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates :parsimonious parametric modeling, *Applied Statistics*, **43**, 429–467.

Royston, P. and Sauerbrei,W. (2008).*Multivariate Model Building;A pragmatic approach to regression analysis based on fractional polynomials for modeling continuous variables*. John Wiley and Sons: Ltd.

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

Bayesian model selection for assessing the progression of chronic kidney disease in transplanted children.

C. Armero[1], A. Forte[1], H. Perpiñán[1,2], M. J. Sanahuja[3], I. Zamora[3]

[1] Universitat de València
[2] Universidad CEU - Cardenal Herrera
[3] Hospital infantil La Fé de València

Abstract: Bayesian model selection within the regression framework is discussed to assess the progression of chronic kidney disease in transplanted children.

Keywords: Bayes factor; Glomerular filtration rate; Objective prior distribution.

1 Introduction

Chronic kidney disease (*CKD*), is a progressive loss of renal function: kidneys lose their ability to remove wastes, concentrate urine and conserve the electrolytes in the blood. It is irreversible and therapies can only slow down the progression of the disease. *CKD* has five stages of increasing severity which are determined by the glomerular filtration rate (*GFR*). Patients in stage V, also known as *End-stage renal disease*, require replacement therapy, dialysis or kidney transplant, to keep them alive.
Transplantation, if available, is always the best option because the renal function is largely recovered. But it is not a definitive therapy and renal function loss, after transplantation, occurs in a very complex process which is mostly studied in adult but poorly understood in paediatric populations (Areses *et al.*, 2010).
In this paper we discuss the possible relationship between *GFR*, which determines the severity of the disease, and time after transplantation in transplanted children with regard to some patient and donor covariates which could be relevant.

2 Data and variables

Data come from 57 children who have been transplanted for more than one year in the Comunitat Valenciana (CV), an autonomous region in the

east of Spain. They are a cross-section of an observational study aimed to follow-up the evolution of transplanted children in CV.
GFR at the time of the study is recorded together with some patient covariates measured at different periods of time: age of the patient, previous transplants and time in dialysis at transplantation time; *GFR*, presence of antihypertensive medication, rejection episodes, proteinuria and microalbuminuria in urine from a general revision after 12 months of the transplant; and post-transplant time at the cross-sectional time. Information from the donor includes age and her/his, alive or cadaveric, condition. Units of measurement for times are always months.

3 Bayesian model selection

The general framework of this study is multiple linear regression. Bayesian variable selection is firstly considered due to the excessive number of covariates initially in the dataset. Nephrologists in the team consider *GFR* after 12 months of transplantation, post-transplant time, previous transplants and antihypertensive medication as important, necessary and undisputable in the model. These covariates will be considered as fixed in the model and generate the so-called *base model* M_0. The variable selection procedure will only affect to the rest of covariates, a total of seven variables, which generate a total of 2^7 possible different regression models M_i:

$$\mathbf{Y} = X_0\boldsymbol{\beta}_0 + X_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n), \quad i = 0, \ldots, 2^7 - 1$$

where X_0 and $X = [X_0, X_i]$ are, respectively, the design matrix associated to the *base model* M_0 and model M_i, $\boldsymbol{\beta}_0$ and $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i)^T$ the corresponding vectors of regression coefficients and $\boldsymbol{\epsilon}$ the random error with variance-covariance matrix $\sigma^2\mathbf{I}_n$.
Bayesian approach to model selection is based on the posterior distribution for all the candidate models. This information is equivalently expressed in terms of Bayes factor in favor of model M_i and against model M_0, which compares the support of the data for both models, and prior probabilities for all candidate models. Bayes factors depend on the prior distribution for the parameters of the model and for this reason, the elicitation of prior distributions is a key point in this general procedure. Default improper priors generally provides undetermined Bayes factors and subjective elicitation is a colossal task, practically impossible, because of the enormous quantity of prior distributions required.
Objective Bayesian methods for assessing prior distributions for the parameters of all the models are considered. In particular, we take into account the proposal by Forte (2011) based on the so-called Conventional approach (Berger and Pericchi, 2001, and references there) and invariance elements for selecting objective prior distributions for the parameters of all the possible models which provide determinated and closed-form expressions for

the corresponding Bayes factors. It considers the usual Jeffrey's prior for common parameters (the ones in the base model), $\pi_i(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$, and the conditional prior distribution:

$$\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma) = \int_0^1 \mathrm{N}_{k_i}(\boldsymbol{\beta}_i|\mathbf{0}, (\frac{\lambda^{-1}(1+n)}{(k_i+k_0+1)} - 1)\,\Sigma)\,\pi(\lambda)\,\mathrm{d}\lambda$$

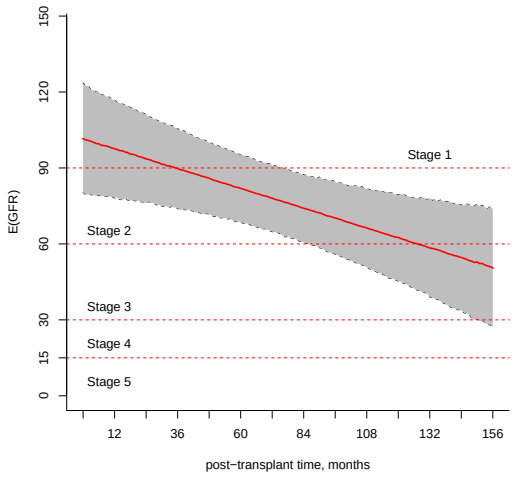
with $\pi(\lambda) = \lambda^{1/2}/2$ and $\lambda \in [0,1]$, for the rest of the β_i parameters of the different models M_i, where k_0 and k_i are, respectively, the dimension of vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_i$ and Σ is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_i$, the MLE of $\boldsymbol{\beta}_i$. Following Scott and Berger (2010), which account for multiplicity control in the election of prior distributions over the model space, the prior distribution for all candidate models has been selected as $p(M_i) = C(7, k_i)^{-1}/8$, where $C(7, k_i)$ is the number of k_i-combinations of 7.

4 Selected model

After computing the posterior probability for all candidate models the selected model will be the one with highest posterior probability. The posterior probability for the chosen model is 0.31 and it includes, jointly with covariates in M_0, the age of the patient at transplantation time. For the second candidate model, which adds proteinuria to that *best* model, the posterior probability turns out 0.09. In addition, the inclusion probability for each one of these covariates, defined as the sum of the posterior probabilities for models which contain it (Barbieri and Berger, 2004), is 0.86 for the age of the patient at transplantation time and 0.32 for proteinuria.
Following the general Bayesian approach to regression models $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, the posterior distribution for the expected *GFR* for patients with covariate vector value X_c can easily computed from the joint posterior distribution (O'Hagan, 1993):

$$\begin{aligned} p(\mathrm{E}(\mathbf{Y}|X_c),\, \sigma^2|data) &= \\ &= p(\mathrm{E}(\mathbf{Y}|X_c) \mid \sigma^2,\, data)\; p(\sigma^2|data) \\ &= \mathrm{N}_k\big(X_c^T\,\hat{\boldsymbol{\beta}}, \sigma^2 X_c^T (X^T X)^{-1} X_c\big)\; \mathrm{IG}\Big(\frac{n-r}{2}, \frac{\hat{\sigma}^2(n-r)}{2}\Big), \end{aligned}$$

where n is the sample size (57 in our case), k the dimension of the model (here 6), $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ the usual MLE estimates of $\boldsymbol{\beta}$ and σ^2 respectively and, $\mathrm{IG}(a, b)$ stands for an inverse Gamma distribution.
Figure above displays the posterior mean of the posterior distribution of the expected *GFR* with regard to post transplant times, in months, for patients with no previous transplants who follow antihypertensive treatment and have covariables values in the model around the mean. Lines in red are the threshold for the different stages of the disease. The expected *GFR* shows



a clear decreasing pattern with post-transplant time and provides direct information about the progression time of the disease.

Acknowledgments: This research has been partially suported by the Ministerio de Ciéncia e Innovación grant MTM2010-19528.

References

Areses, R., Sanahuja, M. and Navarro, M. (2010). Epidemiology of chronic kidney disease in Spanish paediatric population. REPIR II Project. *Nefrologia*, **30**(5): 508-517.

Barbieri, M.M. and Berger, J.O. (2004). Optimal predictive model selection *The Annals of Statistics*, **32**(3): 870-897.

Berger, J.O. and Pericchi, L. (2001). Objetive Bayesian Methods for Model Selection: Introduction and Comparison. *Lecture Notes Monograph Series*, 38(3): 135-207.

Forte, A. (2011). *Objective Bayes Criteria for Variable Selection.* Ph D Thesis. Universitat de València.

Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Association,* **90**(430): 773-795.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics Volume 2B, Bayesian Inference*. Edward Arnold, London.

Scott, J.G. and Berger, J.O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *The Annals of Statistics*, **38**(5): 2587-2619.

Area under the ROC curve using logistic regression with random effects: Estimation and Inference

Llorenç Badiella[1], Emilio Letón[2], Elisa M. Molanes-López[3], Pere Puig[4] , Xavier Sánchez[5]

[1] Servei d'Estadística, UAB, Campus UAB Edifici D, 08193 Cerdanyola (Barcelona), Spain. E-mail: llorenc.badiella@uab.cat
[2] Departamento de Inteligencia Artificial, UNED, C/ Juan del Rosal 16, 28040 Madrid, Spain. E-mail: emilio.leton@dia.uned.es
[3] Departamento de Estadística, UC3M, Avda. de la Universidad 30, 28911 Leganés (Madrid), Spain. E-mail: elisamaria.molanes@uc3m.es
[4] Departament de Matemàtiques, UAB, Campus UAB Edifici D, 08193 Cerdanyola (Barcelona), Spain. E-mail: pere.puig@uab.cat
[5] Departament de Medicina i Cirurgia Animals, UAB, Campus UAB Edifici C, 08193 Cerdanyola (Barcelona), Spain. E-mail: korndraz@hotmail.com

Abstract: Receiver Operating Characteristic (ROC) curves are used to evaluate the accuracy of quantitative tools. The area under the ROC curve (AUC) is a global summary index of the accuracy of a diagnostic test. AUC is commonly estimated using an empirical nonparametric method based on the Mann-Whitney statistic. In this manuscript, we introduce a new approach for constructing nonparametric confidence intervals for the AUC based on logistic regression with random effects. Using several simulated scenarios, this method is compared to other existing methodologies. A real example is used to illustrate the new approach.

Keywords: AUC; GLMM; Mann-Whitney statistic.

1 Introduction

In many scientific areas, ROC curves are used to evaluate the accuracy of quantitative diagnostic tools (biometric markers in medical applications, psychometric tests in psychology, predictive models in machine learning, credit scorings in banking, etc.) in order to distinguish individuals with some trait of interest (a certain disease in medical diagnostic tests) from the rest of individuals.
The ROC curve is built as follows. Let X and Y be the results for a healthy and a diseased subject and let S_X and S_Y be their survival functions. Assuming that higher test values are associated with the diseased population,

for a given cut-off point c, the sensitivity ($Se(c)$) and specificity ($Sp(c)$) are given by:

$$Se(c) = P(Y > c) = S_Y(c) \quad \text{and} \quad Sp(c) = P(X \leq c) = 1 - S_X(c).$$

The ROC curve is a plot of sensitivity versus 1-specificity for any value of c. Based on this curve, there are different indexes to measure the diagnostic accuracy from a local and global point of view. On one hand, the Youden index (see, Le, 2006, Letón and Molanes-López, 2009, and Perkins and Schisterman, 2006, among others), is a local index that at the same time identifies an optimal cut-off point to be used in practice for classifying. On the other hand, the area under the ROC curve, usually denoted as Θ or AUC, is commonly used to summarize the global discriminatory ability of a diagnostic test. This index can be interpreted as the probability that a randomly chosen diseased subject will have a test value (measured in a continuous scale) greater than that of a randomly chosen healthy subject (Bamber, 1975), that is,

$$\Theta = P(Y > X).$$

Moreover, $\alpha = \frac{\Theta}{1-\Theta}$, known as Agresti's α, can be interpreted as a generalization of the Odds Ratio when the response variable is not binary.
In order to evaluate or compare several diagnostic tools, it is of special interest to obtain confidence intervals for their AUC's. In Section 2, several traditional approaches for estimating AUC's variability are summarized, and a direct approach is introduced using logistic regression with random effects. This approach provides confidence intervals within the scope of generalized linear mixed models (GLMM) (Lee et al., 2006). A simulation study is carried out in Section 3, where the new approach is compared to the other methods. Finally, we analyze a motivating and illustrative example in Section 4.

2 New method based on logistic regression with random effects

Suppose that a diagnostic test is measured on m healthy subjects and n diseased individuals. Let X_i and Y_j denote the observations for healthy subjects ($i = 1, ..., m$) and diseased individuals ($j = 1, ..., n$), respectively. The empirical nonparametric estimation of AUC is given by:

$$\widehat{\Theta} = \widehat{AUC} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \Psi(X_i, Y_j)$$

where

$$\Psi(X, Y) = \begin{cases} 1 & X < Y \\ 0 & X > Y. \end{cases}$$

The classical approach to estimate $\mathrm{Var}[\widehat{\Theta}]$ is Bamber's method (Bamber, 1975) which is based on the non-null distribution of the Mann-Whitney statistic. Pepe (2003) suggested applying a logit transformation and use the delta method, to obtain confidence intervals within the unit interval.
Other approaches, different in nature, frequently used to obtain confidence intervals for Θ are based on empirical likelihood (Qin and Zhou, 2006, and Qin and Hotilovac, 2008) and bootstrap resampling methods (Mossman, 1995, and Obuchowski and Lieber, 1998).
We propose here to estimate Θ in a direct way by means of a statistic model using $\Psi(X_i, Y_j)$ as response variable, with $m \times n$ available observations. Although original test values are statistically independent, however, $\Psi(X_i, Y_j)$ values are not, since observations sharing any of both indexes are correlated. To take into account this fact, the model can be stated as follows:

$$W_{ij} = \Psi(X_i, Y_j) = \beta_0 + b_{X_i} + b_{Y_j} + \varepsilon_{ij},$$

where W is the response vector of length $m \times n$, with elements denoted as W_{ij}, $\beta_0 = \Theta$ is the expected value for the response variable, $b_{X_i} \sim N(0, \sigma_X^2)$ is a random term associated to the healthy subjects, $b_{Y_j} \sim N(0, \sigma_Y^2)$ is another random term for the diseased subjects and $\varepsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ models the random error term.
Taking into account the binary nature of the response variable and the presence of random effects, this model belongs to the GLMM class. In our case, the model can be stated as:

$$g(\mu_{ij}) = \beta_0 + b_{X_i} + b_{Y_j},$$

where g is a differentiable monotonic link function (usually a logit function when the response is a binary variable), μ_{ij} is the expected value of W_{ij} and $\beta_0 = \log\left(\frac{\Theta}{1-\Theta}\right)$, which corresponds to the logarithm of Agresti's α index.
This approach can be now evaluated using generic software, for example using SAS PROC GLIMMIX with RMPL (Residual Marginal Pseudo Likelihood) method (Verbeke and Molenberghs, 2000).

3 Simulation study

In this section, we compare six different methods for constructing confidence intervals for Θ: Bamber's method without a logit transformation (CI_B), Bamber's method with a logit transformation (CI_{BLogit}), bootstrap standard normal method (CI_{BootN}), bootstrap percentile method (CI_{BootP}), empirical likelihood approach (CI_{EL}), and our new method (CI_{GLMM}). The comparison is based on coverage probability and length.
Different parametric models are considered for the diagnostic test:

TABLE 1. Averaged coverages and lengths under Scenario 1

		$m = n = 20$		$m = n = 50$		$m = n = 100$	
		Cov(%)	Length	Cov(%)	Length	Cov(%)	Length
$\Theta = 0.5$	CI_B	0.938	0.361	0.944	0.227	0.948	0.160
	CI_{BLogit}	0.956	0.347	0.952	0.223	0.952	0.159
	CI_{BootN}	0.944	0.363	0.942	0.228	0.943	0.161
	CI_{BootP}	0.945	0.365	0.942	0.227	0.943	0.160
	CI_{EL}	0.963	0.361	0.960	0.227	0.955	0.160
	CI_{GLMM}	0.954	0.340	0.950	0.221	0.951	0.158
$\Theta = 0.7$	CI_B	0.928	0.322	0.941	0.203	0.942	0.144
	CI_{BLogit}	0.956	0.315	0.952	0.201	0.948	0.143
	CI_{BootN}	0.929	0.326	0.941	0.204	0.943	0.144
	CI_{BootP}	0.941	0.324	0.948	0.203	0.940	0.144
	CI_{EL}	0.956	0.323	0.958	0.205	0.951	0.144
	CI_{GLMM}	0.949	0.307	0.948	0.199	0.947	0.142
$\Theta = 0.9$	CI_B	0.893	0.185	0.928	0.118	0.941	0.083
	CI_{BLogit}	0.959	0.203	0.955	0.122	0.953	0.084
	CI_{BootN}	0.893	0.188	0.926	0.118	0.941	0.083
	CI_{BootP}	0.914	0.185	0.936	0.118	0.945	0.083
	CI_{EL}	0.917	0.190	0.953	0.124	0.958	0.086
	CI_{GLMM}	0.935	0.191	0.950	0.119	0.952	0.084

TABLE 2. Averaged coverages and lengths under Scenario 2

		$m = n = 20$		$m = n = 50$		$m = n = 100$	
		Cov(%)	Length	Cov(%)	Length	Cov(%)	Length
$\Theta = 0.5$	CI_B	0.931	0.371	0.950	0.235	0.943	0.166
	CI_{BLogit}	0.955	0.356	0.956	0.231	0.946	0.165
	CI_{BootN}	0.933	0.376	0.948	0.235	0.943	0.166
	CI_{BootP}	0.936	0.375	0.948	0.235	0.943	0.166
	CI_{EL}	0.957	0.365	0.956	0.233	0.947	0.166
	CI_{GLMM}	0.951	0.350	0.954	0.229	0.945	0.164
$\Theta = 0.7$	CI_B	0.932	0.337	0.942	0.212	0.940	0.150
	CI_{BLogit}	0.958	0.328	0.953	0.210	0.945	0.149
	CI_{BootN}	0.936	0.341	0.940	0.213	0.941	0.150
	CI_{BootP}	0.939	0.339	0.946	0.212	0.939	0.150
	CI_{EL}	0.959	0.334	0.955	0.212	0.946	0.150
	CI_{GLMM}	0.954	0.322	0.952	0.209	0.944	0.148
$\Theta = 0.9$	CI_B	0.875	0.197	0.919	0.127	0.933	0.090
	CI_{BLogit}	0.946	0.217	0.951	0.132	0.948	0.092
	CI_{BootN}	0.876	0.199	0.918	0.127	0.931	0.090
	CI_{BootP}	0.900	0.196	0.926	0.127	0.936	0.090
	CI_{EL}	0.908	0.202	0.950	0.130	0.951	0.092
	CI_{GLMM}	0.922	0.208	0.947	0.130	0.947	0.091

- Scenario 1: $X \sim N(0,1)$ and $Y \sim N(\sqrt{2}\Phi^{-1}(\Theta), 1)$.
- Scenario 2: $X \sim N(0,1)$ and $Y \sim N(\sqrt{5}\Phi^{-1}(\Theta), 4)$.
- Scenario 3: $X \sim Exp(1)$ and $Y \sim Exp(\frac{1}{\Theta} - 1)$.

We have simulated 2,000 trials for each combination of $\Theta = 0.5$ (null accuracy), 0.7 (moderate) and 0.9 (high), with (m, n)=(20,20), (50,50), (100,100). For the bootstrap-based confidence intervals we have used $B =$ 200 bootstrap resamples.

TABLE 3. Averaged coverages and lengths under Scenario 3

		$m = n = 20$		$m = n = 50$		$m = n = 100$	
		Cov(%)	Length	Cov(%)	Length	Cov(%)	Length
$\Theta = 0.5$	CI_B	0.939	0.361	0.941	0.227	0.951	0.160
	CI_{BLogit}	0.955	0.347	0.945	0.223	0.955	0.159
	CI_{BootN}	0.940	0.366	0.939	0.228	0.948	0.160
	CI_{BootP}	0.946	0.365	0.936	0.227	0.947	0.160
	CI_{EL}	0.963	0.361	0.951	0.227	0.958	0.160
	CI_{GLMM}	0.949	0.340	0.945	0.221	0.955	0.158
$\Theta = 0.7$	CI_B	0.930	0.324	0.941	0.204	0.939	0.144
	CI_{BLogit}	0.964	0.316	0.949	0.202	0.945	0.144
	CI_{BootN}	0.933	0.328	0.938	0.205	0.941	0.144
	CI_{BootP}	0.941	0.326	0.945	0.205	0.942	0.144
	CI_{EL}	0.965	0.323	0.955	0.205	0.950	0.144
	CI_{GLMM}	0.957	0.309	0.948	0.200	0.945	0.143
$\Theta = 0.9$	CI_B	0.880	0.195	0.926	0.125	0.935	0.088
	CI_{BLogit}	0.946	0.214	0.951	0.130	0.949	0.090
	CI_{BootN}	0.880	0.197	0.925	0.125	0.934	0.088
	CI_{BootP}	0.903	0.194	0.933	0.124	0.939	0.088
	CI_{EL}	0.913	0.199	0.951	0.128	0.950	0.089
	CI_{GLMM}	0.924	0.204	0.948	0.127	0.946	0.089

Results of the simulation study are presented in Tables 1-3. They show that the new method based on GLMM provides the smallest interval width and at the same time appropriate coverage probability.

4 Example

In order to illustrate the methodology described above, a real example is used. The study is based on 63 dogs from several races and ages, 35 out of them diagnosed as mitral insufficiency (Sánchez et al., 2010). The diagnosis is usually based on VHS, a quantitative measure from a thoracic radiograph, reflecting heart size and expressed on vertebral bodies units. Confidence intervals of the *AUC* using the approaches described previously are given in Table 4.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12**, 387-415.

TABLE 4. Confidence intervals of *AUC* for the example

Method	$CI(95\%)$
CI_B	0.842 - 0.987
$CI_{B_{Logit}}$	0.809 - 0.964
CI_{BootN}	0.843 - 0.986
CI_{BootP}	0.835 - 0.975
CI_{EL}	0.812 - 0.964
CI_{GLMM}	0.813 - 0.963

Le, C.T. (2006). A solution for the most basic optimization problem associated with an ROC curve. *Statistical Methods in Medical Research*, **15**, 571-584.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*, Chapman & Hall/CRC Boca Raton.

Letón, E. and Molanes-López, E.M. (2009). Adjusted empirical likelihood estimation of the Youden index and associated threshold for the bigamma model. *Statistics and Econometrics Series*, 07, Working Paper 09-19.

Mossman, D. (1995). Resampling techniques in the analysis of non-binormal ROC data, *Medical Decision Making*, **15**, 358-366.

Obuchowski, N.A. and Lieber, M.L. (1998). Confidence intervals for the receiver operating characteristic area in studies with small samples. *Academic Radiology*, **5**, 561-571.

Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction.* New York: Oxford University Press.

Perkins, N.J. and Schisterman, E.F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, **163**, 670-675.

Qin, G.S. and Hotilovac, L. (2008). Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Statistical Methods in Medical Research*, **17**, 207-221.

Qin, G.S., and Zhou, X.H. (2006). Empirical likelihood inference for the area under the ROC curve. *Biometrics*, **62**, 613-622.

Sánchez, X., Prandi, D., Domènech, O. (2010). A new radiologic measurement to study left atrial enlargement in dogs with mitral insufficiency: preliminary study. *ACVIM Annual Meeting, Anaheim, CA*.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, New York: Springer.

Optical properties of fresh date palm in different stages of maturity

X. Barber[1], A. M. Martín[2], A. Mayoral[1], J. Morales[1], J. A. Pérez-Álvarez[2]

[1] Centro de Investigación Operativa. Universidad Miguel Hernández de Elche.
[2] Dpto. de Tecnología Agroalimentaria. Universidad Miguel Hernández de Elche.

Abstract: The date palm color is the main external quality feature which allows the evaluation of its ripeness (kimri, khalal or rutab), thereby influencing the decision of the consumer to purchase. This study characterizes spectrophotometrically the various stages of maturation of fresh dates range Medjoul, using a spectrophotometer Minolta CM-2600. Regarding the reflectance spectra (360-740 nm), we can differentiate the three states. The one we can differentiate clearer is kimri state, since the spectrum shape is modified, mainly in wavelengths ranging between 650 and 700 nm.

Keywords: Smoothness, P-splines, spectrometry , color, maturity

1 Introduction

The appearance of food is of great importance in the food industry. Color is the quality parameter with the largest influence on consumer purchasing criteria and color is also related to technological treatments or degradation processes (Maroulis and Saravacos 2003). Therefore, an objective measurement of color can be an instrument to assess the main characteristics of many foods (Francis 1995).
Color, as seen by the human eye, is the result of complex series of physiological and psychological responses to electromagnetic radiation of wavelengths in the range of 400 to 700 nm. The use of instrumental methods are necessary for reproducible, precise, accurate and fast results allowing a better interpretation of color differences (Pérez-Álvarez 1996).
The color of a food, plant or animal, can be described by different color coordinate systems. In particular, using the color space CIE L*a*b*, the color differences are similar to those perceived by the human eye (Abbott 1999).
In the date palm, as in most fruits, the outer color is one of the most important quality parameters. It experiences major changes during the different stages of their growth and maturation (kimri, khalal and rutab). Therefore, the analysis of the optical properties of date palm at different

stages of maturity, using non-destructive techniques may be of interest to the date palm industry (Vilella-Espla, Pérez-Álvarez et al. 2004).

2 Material and Methods

Samples: date palms of the Medjoul variety harvested in different stages of maturity (kimri, khalal and rutab) were supplied by Phoenix Station of Elche (Alicante).
Determination of optical properties: reflection spectra and color coordinates, L * (lightness), a * (red-green coordinate), b * (yellow-blue coordinate), C * (color intensity) and h (Hue angle = arctang b / a), were determined using a Minolta CM-2600 spectrophotometer with illuminant D65, observer 10, as SCI, opening for the illumination of 11 mm and 8 mm for measurement. The colorimeter was calibrated according to manufacturer's instructions before taking action with a white tile (CM-A145) and a zero calibration box (CM-A32). Nine replicates of each sample were analyzed on whole fruit.
Statistical Analysis: The study of the refection spectra shows a curve for each of the maturity states. The idea is to test the differences among curves for each stage of maturity. A Penalized Spline Mixed Model with interaction between reflection and maturation status (Durban and Lee, 2008) has been used for the study of these curves:

$$y_{ij} = f_{z_i}(x_{ij}) + a_{i1} + a_{i2}x_{ij} + \varepsilon_{ij};$$

i.e.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + Z_i u_k \sum_{l=2}^{L} tr_{il}(\gamma_{0l} + \gamma_{1l}x_{ij}) +$$
$$+ \sum_{l=2}^{L} Z_i w_k^l + a_{i1} + a_{i2}x_{ij} + \varepsilon_{ij}.$$

where

$$w_k^l \sim N(0, \sigma_{wl}^2), \qquad (a_{i1}, a_{i2})^T \sim N(0, \Sigma), \qquad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

and y_{ij} represents the reflectance, x_{ij} represents the wavelength applied, and finally $\gamma_{0l} + \gamma_{1l}x_{ij} + Z_i w_k^l$ shows the difference in the fitted curves between the different states of maturation (kimri, khalal and rutab).

3 Results

Figure 2 compares the reflection spectra between the different stages of maturity. An isobestic point, (ie, a point which, in the same wavelength

match the different stages of maturity, point of intersection) was found in the area near the 550 nm. In this area (400-500 nm), where both chlorophylls and carotenoids absorb the reflection of the greenest date is lower. Therefore, this could be a characteristic of the date palms. The biggest difference between the samples is observed on the 678 nm, where the date kimri state, has a very pronounced minimum. This minimum corresponds to the absorption band of chlorophyll (Merzlyak, Solovchenko et al. 2003). In addition, looking at the bands, it would be possible to differentiate between rutab and khalal states at wavelengths between 400 and 430 nm, areas without overlapping.

4 Conclussion

The reflectance spectra provide more information about the characteristics of each state of ripeness. Our recommendation is to use the reflection spectra as a tool for differentiation among stages by analyzing the percentage of reflection at wavelengths of 678 nm (the minimum value in the kimri state).

Acknowledgments: Special Thanks to Agencia de Cooperación Internacional para el Desarrollo AECID (A/030696/10), Ministerio de Ciencia e Innovación, Plan Nacional de I+D+I 2008-2011, MTM2010-20540 and Maria Durban

References

Abbott, J. A. (1999). Quality measurement of fruits and vegetables. *Postharvest Biology and Technology* **15**(3):207-225.

Francis, F. J. (1995) Quality as influenced by color. *Food Quality and Preference* **6**(3):149-155.

Maroulis, Z B. and Saravacos, G.D. (2003) *Food process design.* New York, Marcel Dekker, Inc.

Pérez-Álvarez, J. A. (1996) *Contribución al estudio objetivo del color en productos cárnicos crudo-curados.* España, Universidad Politécnica de Valencia.

Vilella-Espla, J. M., Pérez-Álvarez, J. A. et al. (2004). Inexpensive technique to measure colour in date palm fruit (Phoenix dactylifera, L.) at different stages of maturity. *Journal of Food Techonology* **2**(4): 246-252.

Durban, M., Lee, D-J (2008). *Splines con penalizaciones (P-splines).* Universidad Publica de Navarra.

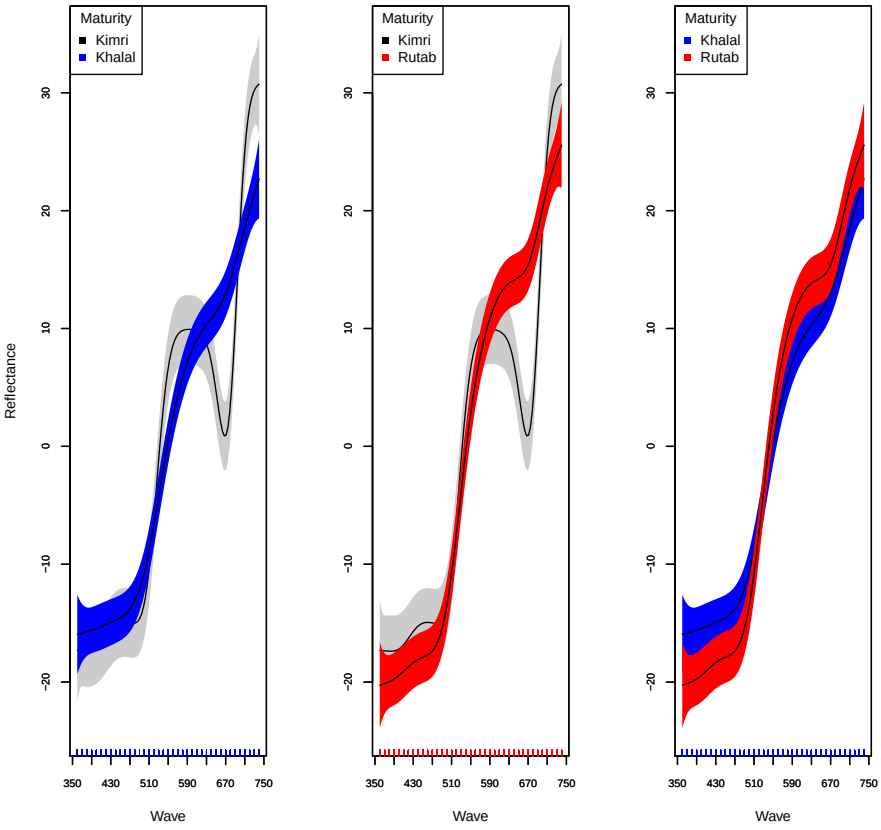


FIGURE 1. Spectometry Date Palm and Maturity.

Measuring the real estate bubble: a house price index for Bilbao.

M. J. Bárcena[1], P. Menéndez[2], M. B. Palacios[2], F. Tusell[1]

[1] Departamento de Econometría y Estadística, Universidad del País Vasco, Avenida Lehendakari Aguirre, 83, 48015 Bilbao, Spain
[2] Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra, Campus de Arrosadía, 31006 Pamplona, Spain
email: blanca.palacios@unavarra.es *(corresponding author)*

Abstract: The aim of this work is to calculate house price indexes taking into account the location of the houses. Two approaches were used to fit the regression model. The first one includes the postal code as a categorical variable. The second one takes into account the house coordinates and, the model is fitted using geographically weighted regression. We estimate index for housing prices in the city of Bilbao (Spain) over the period 2005-2010.

Keywords: hedonic models; geographically weighted regression; price index; spatio-temporal data

1 The real estate bubble

Since the accession of Spain to the EU in 1986 and, in particular during the first years of the present century, housing prices have experienced substantial increases. This phenomenon has been observed in a large part of the western world, but in Spain has been exacerbated by a number of circumstances: monetary stability, with low or even negative real interest rates, easy borrowing, high economic growth and fiscal allowances for home buyers. Since the first years of this century it became commonplace to refer to the *real state bubble*.
As the situation began to deteriorate after the onset of the financial crisis of 2008, widely different figures have been given on how much housing prices have already dropped. Previously, widely different figures were given on the extent of the price increase while the bubble lasted.
Part of the discrepancy can be traced to the fact that different sources sometimes speak of different markets (and it is known, for instance, that second residences in coastal areas have been hit harder than urban properties in or near large cities). But even when speaking of roughly the same market, figures are so widely disagreeing so as to raise the issue of how they were obtained and what they really measure.

Compiling a housing price index is particularly difficult, due to the opacity of the market, the incentives to understate transaction prices, and the fact that any single house is traded only very infrequently, and has no exact replicates. In this paper we obtain an index as the estimate of a non-observable component, of the price level, in a semi-parametric model of readily available offered prices in the city of Bilbao (Spain).

2 Methodology

We address the problem of opacity making resort to publicly available data of the Spanish house prices in one of the leading housing market portals on the web (`www.idealista.com`). It is open to question whether these prices even approach final transaction prices, however for our purposes, it is enough to assume that they overstate transaction prices by a factor relatively constant over time.
We fit a model to the response $\log(Price/m^2)$ to capture the influence on house prices of attributes such as number of total surface, type of dwelling, number of bedrooms, bathrooms, type of heating, age of the building, orientation, availability of services such as garage, elevator, etc. In general our model is as follows,

$$\log(Price/m^2) = \beta_0 + \sum_k \beta_k x_k + s(t) + \epsilon \tag{1}$$

where the aforementioned attributes or explanatory variables are denoted by x_k and the effect of time t is captured using as a regressor a cubic smoothing spline s, whose suitably normalized profile provides an estimate of the price index. To capture the influence of the location, two approaches have been used: including the postal code as a categorical regressor and using geographically weighted regression (Fotheringham et al. 2002), with each house geocoded to UTM coordinates.
While the first approach, using areal postal code information, can be implemented easily using available software (we have used Simon Wood's `mgcv` package, available in R; see Wood, 2004), mixing geographically weighted regression and a non-parametric trend requires what is in essence a back-fitting algorithm (see, e.g. Hastie and Tibshirani, 1990), which alternates the fitting of the different effects present in the model.
A little over five thousand observations have been used, the exact number varying across models on account of missing values in several regressors.

3 Results and Discussion

The results obtained include an estimated temporal profile of the price indexes which are shown in Figure 1 (left), the different curves were computed

via a backfitting algorithm. The black line is the initial estimate whereas the green one is the final estimate. This estimated temporal profile, clearly shows a drop of about 15% from market heights which bottomed around mid 2009. After that period, the temporal index exhibits some small fluctuations with a slightly increasing trend towards the second semester of 2010. Ever since, and contrary to common belief, prices have shown an upturn.
The estimated prices per squared meter, for the different locations investigated in the city of Bilbao, are presented in Figure 1(right). Likewise, the estimates for the effect of the considered attributes on house prices, broken down by geographical location, were computed (Figure 1, left). For illustration, we display the impact of attributes such as the number of bathrooms and bedrooms, availability of garage, as well as the contribution of elevator, into the house price in Figure 2. In particular, when focused on those attributes, it is observed that the availability of elevator and garage have a greater contribution to the price (per squared meter) than the number of bedrooms and bathrooms. The different contribution between elevator-garage and bedrooms-bathrooms might be explained by the fact that the price per squared meter does not discriminate between the type of room in the dwelling.
In general, the results presented in this paper agree quite well with perceived trends in the Spanish market, in particular, in the city of Bilbao. Furthermore, given the opacity of the real transactions in the market, our approach provides a way to compute a price index curve that can be used to estimate the non-observable transaction prices. Moreover, the model presented here allows to quantify the contribution of different attributes in the housing prices. As a concluding remark, we would like to stress the fact that the methodology affords easy, cheap and almost real time monitoring of the market as new information accumulates which might be very appealing when the data are updating continously.

Acknowledgments: Partial support from grants ECO2008-05622 (MCyT) and IT-347-10 (Basque Government) is gratefully acknowledged.

References

Fotheringham, S., Charlton, M., and Brunsdon, C. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: John Wiley & Sons.

Hastie, T.J., and Tibshirani, R.J. (1991) *Generalized Additive Models. 2nd. edition*. London: Chapman & Hall.

Wood, S.N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99**, 673-686.

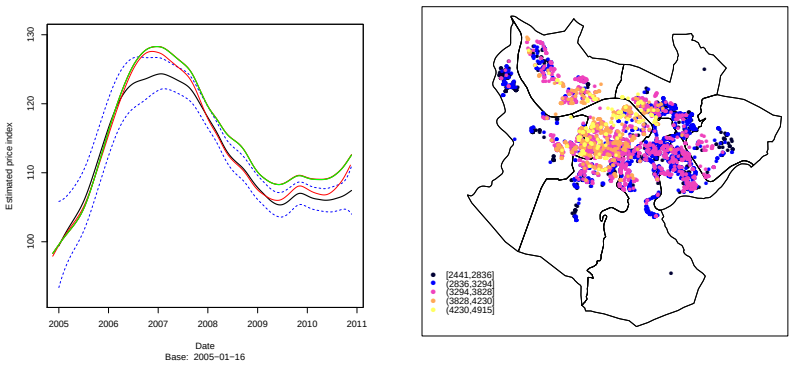


FIGURE 1. (Left) Temporal estimated curve of the price index. (Right) Estimated prices per squared meter for housing market in Bilbao.

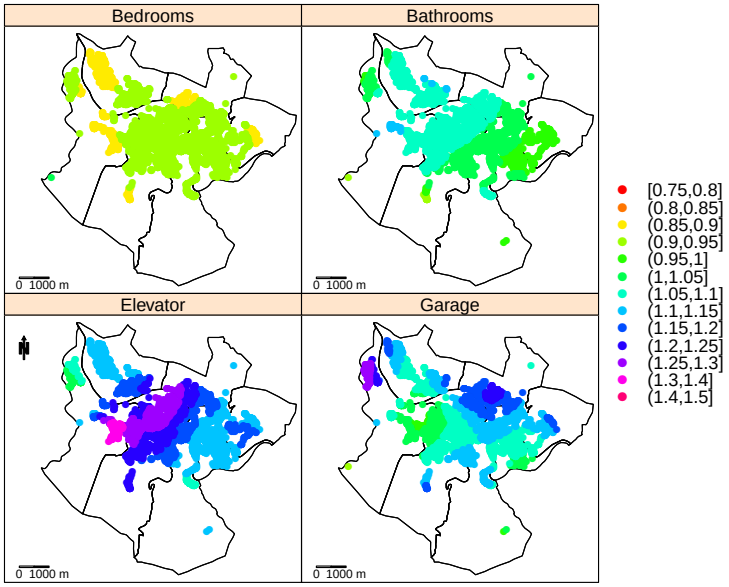


FIGURE 2. Relative contributions to the estimated price per squared meter for the number of bathrooms, bedrooms, availability of garage and elevator.

Missing data, multiple imputation and the UK National Vascular Database

P D Baxter[1], B A Cattle[1], C P Gale[1], M S Gilthorpe[1], D J A Scott[2]

[1] Centre for Epidemiology and Biostatistics, School of Medicine, University of Leeds, LS2 9JT, UK, p.d.baxter@leeds.ac.uk.

[2] Division of Cardiovascular and Diabetes Research, Leeds Institute of Genetics Health and Therapeutics, University of Leeds, LS2 9JT, UK.

Abstract: The National Vascular Database (NVD) collects information of the quality of care and outcomes of patients admitted to acute hospitals in England, Wales, Scotland and Northern Ireland with (i) Abdominal Aortic Aneurysms (AAA), (ii) lower limb ischaemia requiring bypass, (iii) Carotid Endarterectomy and (iv) Amputation. The NVD has proved to be an important resource for clinical audit (Prytherch et al., 2001), by contrast its potential as a valuable research tool remains underexploited. Use for research is dependent on the ability to adjust for case-mix, which in turn is dependent on the completeness and quality of data collected. In this work we present an illustration of Multiple Imputation by Chained Equations (MICE) (van Buuren & Groothius-Oudshoorn, 2010) to address the problems of missing data. We follow the analysis protocol of (Sterne et al., 2009) and compare the VBHOM model (Tang et al., 2007) based on imputed data, with a complete cases analysis.

Keywords: Abdominal Aortic Aneurysms; Missing Data; Multiple Imputation by Chained Equations.

1 Selection of variables to impute

As a general rule using every bit of available information yields multiple imputations that have minimal bias (Collins et al., 2001; Meng, 1994). This principle suggests that the number of predictors should be as large as possible. Practically however, the imputation scheme should be at least as rich as the models that the analyst intends to use for their statistical modelling after the imputations: a property referred to as congeniality (Meng, 1994). As well as including predictor variables of the VBHOM model (Tang et al., 2007) we have also included auxiliary variables that can improve prediction of the missing values in the variables of interest. When selecting auxiliary variables it is important to include both clinical judgment on which variables might usefully predictor those that are missing and also statistical judgment to avoid variables that are highly collinear (i.e. do not contribute

Variable	Type	% missing	Imputation method	Clinically plausible limits
AAA Surgery	Categorical	23.2	Polytomous regression	
Stroke	Categorical	66.7	Polytomous regression	
Mode of admission	Binary	0.8	Polytomous regression	
Gender	Binary	0.0	Polytomous regression	
Diabetes	Binary	5.9	Polytomous regression	
Current Smoker	Binary	21.5	Polytomous regression	
Renal Dialysis	Binary	13.7	Polytomous regression	
Renal Transplant	Binary	15.5	Polytomous regression	
Previous Aortic Stent Surgery	Binary	11.2	Polytomous regression	
Haemorrhage	Binary	10.5	Polytomous regression	
Myocardial Infarction	Binary	62.6	Polytomous regression	
Cardiac Failure	Binary	65.4	Polytomous regression	
Hypotension	Binary	65.4	Polytomous regression	
Discharge Status	Binary	3.2	Predictor only	
Age	Continuous	1.3	Predictive mean matching	(18,100)
Haemoglobin	Continuous	10.5	Predictive mean matching	(2,20)
White Cell Count	Continuous	12.3	Predictive mean matching	(2,50)
Urea	Continuous	14.1	Predictive mean matching	(0.1,800)
Sodium	Continuous	10.9	Predictive mean matching	(105,165)
Potassium	Continuous	26.3	Predictive mean matching	(2,40)
Lowest Systolic BP	Continuous	17.4	Predictive mean matching	(20,250)
Highest Pulse	Continuous	17.4	Predictive mean matching	(20,200)

FIGURE 1. Key variables and summary of missing data. For a full description of the variables, see http://www.vascularsociety.org.uk/library/audit.html.

any additional information relative to the variables that have already been selected) (Collins et al., 2001). The subset of variables to be imputed and their missingness characteristics are summarised in Fig. 1. Note that for continuous variables any data values that lie outside clinically plausible limits have been declared as missing data.

2 Choice of imputation methods

In this work we have opted to use predictive mean matching (PMM) to impute continuous predictor variables and polytomous regression for binary and categorical predictors. PMM is a general purpose imputation method (Little, 1988) in which the imputations are confined to the observed distribution. An advantage of PMM is that it can preserve non-linear relations between predictors. A possible disadvantage of PMM is that it may fail to produce enough between imputation variation when the number of predictors is small (van Buuren, & Groothius-Oudshoorn, 2010). As the sample sizes of the NVD AAA data is large and the number of predictors is also large, we believe that PMM offers a useful method of imputing continuous variables and preserving non-linear relationships. Moreover, partly to mitigate concerns regarding insufficient between imputation variation, we have

elected to use 20 imputations in our work rather than the 'standard' five imputations routinely suggested.
Multiple imputation assumes normality of the variables being imputed, and it is important to check that this assumption will be approximately satisfied. For those variables that are found to have a non-normal distribution a transformation to approximate normality is required. A logarithmic transformation will often suffice. In this work we have opted for the logarithmic transformation for the variables White Cell Count, Urea, Sodium and Potassium, all of which are sufficiently non-normal to cause concern about the validity of the normality assumption.
It is important to include the outcome variable (in this case mortality status at discharge) as a predictor in the imputation model. Failing to include the outcome will severely dilute the associations between the outcome and the other variables (Moons et al., 2006). Missing outcomes will also be imputed, but the results of the imputations are excluded in the final analyses.

3 Evaluating robustness of the imputation scheme

There is no definitive method for checking the imputations or the within imputation iterations.
The chain mean and standard deviation at each iteration can be plotted and on convergence the different streams should freely intermingle without showing any definite trends. Although the default setting of five iterations is often sufficient, in this work we used 20 within imputation iterations. Successful convergence of chain means and standard deviations is exhibited for this data set.
In general, a good imputed value is one which could have been observed had it not been missing. The missing at random assumption can never be tested on the observed data, but we can check that the imputations are plausible by comparing the distributions of the observed and imputed values for each imputed data set. In this work, distributions of observed and imputed values appear similar.

4 Comparison of imputed data with complete case analyses

Fig. 2 shows the performance of the VBHOM model for predicting status at discharge (dead / alive) using only data with complete cases and a full data set with missing values imputed and pooled using the MICE scheme explained above. The magnitudes of the coefficients in the model are broadly similar both with and without imputation. However, notice that, for all of the variables in the VBHOM model, the confidence intervals are narrower for the MICE imputed data. A narrower confidence interval represents reduced uncertainty in the model coefficients and hence greater confidence in the validity of the model.

	(95% confidence interval for odds ratio of death on discharge) [width of confidence interval]	
Variable	Complete cases only	MICE imputed data
Gender: Male	(0.664, 1.008) [0.344]	(0.787, 1.069) [0.282]
Admission mode: non-elective	(0.184,15.530) [15.345]	(0.184, 4.450) [4.266]
Age	(1.036, 1.061) [0.025]	(1.042, 1.060) [0.018]
Urea	(1.006, 1.026) [0.020]	(1.002, 1.018) [0.016]
Sodium	(0.994, 1.007) [0.013]	(0.994, 1.004) [0.010]
Potassium	(0.992, 1.135) [0.142]	(0.986, 1.117) [0.131]
Haemoglobin	(0.743, 0.806) [0.063]	(0.723, 0.771) [0.047]
White Cell Count	(1.126, 1.165) [0.038]	(1.102, 1.131) [0.030]

FIGURE 2. Comparison of VBHOM model complete case analysis versus MICE.

Acknowledgments: We grateful acknowledge the Vascular Society of Great Britain and Ireland for providing access to the National Vascular Database.

Collins, L.M., Schafer, J.L., and Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* **6**(3), 330-351.

Little, R.J.A. (1988). Missing data adjustments in large surveys (with discussion). *Journal of Business Economics and Statistics* **6**, 287-301.

Meng, X.L. (1994). Multiple imputation inferences with uncongenial sources of input. *Statistical Science* **9**, 538-558.

Moons, K.G, Donders, R.A., Stijnen, T., and Harrel Jr., F.E. (2006). Using the Outcome for Imputation of Missing predictor Values was Preferred. *Journal of Clinical Epidemiology* **59**, 1092-1101.

Prytherch, D.R., et al. (2001). A model for national outcome audit in vascular surgery. *European Journal of Vascular and Endovascular Surgery*, **21**(6), 477-483.

Sterne, J.A.C., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal* **338**.

Tang, T., Walsh, S.R., Prytherch, D.R., Lees, T., Varty, K., and Boyle, J.R. (2007). VBHOM, a data economic model for predicting the outcome after open abdominal aortic aneurysm surgery. *British Journal of Surgery* **94**, 717-721.

van Buuren, S., and Groothuis-Oudshoorn, K. (2010). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. In Press.

A Comparison of Frequentist and Bayesian Approaches to Latent Class Modelling of Susceptibility to Asthma and Patterns of Antibiotic Prescriptions in Early Life

Danielle Belgrave[1], Christopher Bishop[2], Adnan Custovic[3], Angela Simpson[3], Aida Semic-Jusufagic[3], Andrew Pickles[4], Iain Buchan[1]

[1] North-West Institute for Bio-Health Informatics, The University of Manchester, Oxford Road, Manchester, United Kingdom, M13 9PL.
[2] Microsoft Research Cambridge, Roger Needham Building, 7 JJ Thomson Avenue, Cambridge, United Kingdom, CB3 0FB
[3] Respiratory Department, The University of Manchester, University Hospital South Manchester, Southmoor Road, United Kingdom, M23 9LT
[4] Institute of Psychiatry, Kings' College London, Box P, De Crespigny Park, London, United Kingdom, SE5 8AF
*Corresponding author: danielle.belgrave@manchester.ac.uk

Abstract: The assessment of patterns of antibiotic use in early life may have major implications for understanding the development of asthma. This paper compares a classical generalized latent variable modelling framework and a Bayesian machine learning approach to define latent classes of susceptibility to asthma based on patterns of antibiotic use in early life. We compare the potential advantages of each method for elucidating clinically meaningful phenotypes or classes.

Keywords: Longitudinal Latent Class Analysis; Bayesian Inference; Infer.NET.

1 Introduction

The assessment of patterns of antibiotic use in early life may have major implications for our understanding of the development of asthma. Within the medical literature, antibiotic use has been investigated as having a causal association with asthma. We hypothesise that antibiotic use in early life, rather than being causally related to asthma gives an indication of a child's susceptibility to infection with a heightened response to exposure since such children are more likely to receive antibiotics early on in life due to their immunodeficiency. Thus antibiotic use in early life can be used as a marker in order to identify children who are more susceptible to outcomes of exacerbation of wheeze and asthma symptoms. The aim of this project is to identify latent classes of susceptibility to characterize children

according to susceptibility based on patterns of early-life antibiotic use and investigate whether this latent class is predictive of contemporaneous and future asthma and wheeze symptoms. We infer that antibiotic use picks up a signal of something that occurs very early in life, and which is completed by 24 months of age.

2 Methods

The Manchester Asthma and Allergy Study (MAAS) is an unselected, prospective population-based birth cohort study designed to determine early life factors for the development of asthma and allergic disease. Subjects were recruited prenatally and followed prospectively. A trained physician extracted the information on antibiotic prescription receipt and symptoms of asthma/wheezing from the primary care medical records (n=916). Based on a longitudinal model for antibiotic use within the first 2 years of life, latent class analysis was carried out to obtain a phenotypic definition of susceptibility. We then investigate whether these latent classes of susceptibility are predictive of contemporaneous and future asthma and wheeze symptoms. We describe and use two different statistical approaches for defining latent classes of susceptibility to asthma: a classical generalized linear latent and mixed models framework using the gllamm package in STATA and a Bayesian machine learning approach using Infer.NET.

Using the classical approach to latent class analysis we formulated a longitudinal trajectory model which allows us to hypothesize that there may be subgroups of children who, because of differing maturity of their immune response, have changing levels of susceptibility over time. We specified this as a two-level random-coefficient logistic regression model for antibiotic use with level-1 units as the monthly measurement occasions and the level-2 units as children. This model characterises the child's susceptibility through their age and exposure to two particular known risk factors (older siblings and day-care) and by membership of different possible classes defined by the intercept and slope in the regression equation for antibiotic use yij of child i at time j which was specified as:

$$Logit\{Pr(y_{ij}) = 1|x_{ij}, c_i = k\} = \beta_{0k} + \beta_{1k}x_{1i} + \beta_{2k}x_{2ij} + \beta_{3k}x_{3ij} \quad (1)$$

where x_{1i} is the time point for a specified child i. x_1 represents monthly time periods and $x_1 = 1,\ldots,24$ months; x_{2ij} is child i's day-care attendance at time j; and x_{3i} is the number of older siblings child i has which remains constant at all time points j. We also introduce a prior distribution $Pr(c_i = k)$ over the classes given by a multinomial distribution.

We assume that each child belongs to one of a set of N latent classes, with the number of classes and their size not known a priori. Other than random temporal fluctuation, each child's pattern of antibiotic prescription is to be explained by their belonging to a particular class of susceptibility. Children belonging to the same class are similar with respect to the observed variables in the sense that their observed scores are assumed to come from the same probability distributions, whose parameters are, however, unknown quantities to be estimated. Using empirical Bayes' techniques, children are assigned to the latent class with the largest posterior probability. We also consider a restricted random-intercept form of this model in which the classes are allowed to differ in their intercepts, but not in their slopes i.e. in which the relative susceptibility remained constant over time. These models were fit using `gllamm`, a program implemented in Stata (www.stata.com) to fit generalized linear latent and mixed models.

We then investigated parallel models using a machine learning approach with the Bayesian inference software Infer.NET. The Bayesian machine learning method provides a unified framework for modelling and quantifying uncertainty–employing probabilistic modelling strategies based on defining priors in such a way that probabilities can be associated with unknown parameters. The three steps for defining a model in Infer.NET are: i) the definition of a probabilistic model; ii) the creation of an inference engine for performing inference; and iii) the execution of an inference query. Since the Bayesian approach to statistical modelling enables us to quantify model uncertainty through the incorporation of priors, we assumed uninformative priors for y_{ij}. Variables x_{1i}, x_{2ij} and x_{3ij} are specified as a vector array X with a vector of coefficients β. The k unobserved latent classes are accompanied by a random temporal fluctuation or noise, ξ_k and c_i is assumed to be multinomial over k classes with a prior uniform Dirichlet distribution (Dirichlet (1,1)) and the random noise has a prior Gaussian distribution(0,1).

We compare models that assume varying numbers of latent classes using the Bayesian Information Criterion as a measure of goodness-of-fit. We then investigate whether the inferred phenotypes of susceptibility predict current or future asthma and wheeze symptoms using conventional time-to-event analyses.

3 Results

Using the classical generalized linear latent and mixed models framework, we identified a model with three distinct latent classes of susceptibility

based on patterns of antibiotic use in the first two years of life. Based on our interpretation of the model, Class 1 were children resilient to infection (31.1%), Class 2 showed a normal immune response (55.7%) and Class 3 were susceptible to infection (13.2%). Compared to Class 1 and Class 2, children in Class 3 had a significantly higher hazard of reported asthma or wheeze symptoms in the first three years of life (HR=3.72 [95% CI 2.72 – 5.10, $p < 0.01$] and 1.61 [95% CI 1.25 – 2.09, $p < 0.01$] respectively. Class 2 had a greater hazard of experiencing exacerbations of asthma and wheeze symptoms than Class 1 (HR=1.90 [95% CI 1.21 – 2.98, $p < 0.01$]) however, after the third year of life, this hazard ratio ceased to be statistically significant (HR=1.39 [95% CI 0.79 – 2.45, $p = 0.25$]). Similar results were obtained using a Bayesian machine learning framework. We demonstrate the potential advantages of Bayesian models for elucidating clinically meaningful phenotypes.

4 Conclusion

By analysing trajectories of antibiotic use in early life, we identified a group of children with high susceptibility to the development of asthma. Our results suggest that antibiotic use in early life indicates a child's susceptibility to infections. Since the Bayesian and frequentist approaches provided different perspectives for identifying the latent classes, with concordant results, the combination of methodologies was complementary – Bayesian extensions to classical epidemiology have the potential to shape hypotheses with more complete use of the data and current knowledge.

References

Bishop, C.M (2006). *Pattern Recognition and Machine Learning*. USA: Springer.

Custovic, A., Simpson, B.M., Murray, C.S., Lowe, L., and Woodcock, A. (1994). The National Asthma Campaign Manchester Asthma and Allergy Study. *Paediatric Allergy Immunology, Supplement.*, **15**, 32-7.

Goodman, L.A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I – A modified latent structure approach *American Journal of Sociology*, **79**, 1179-1259

Rabe-Hesket, S., Skrondal, A., Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, **128**, 301-23.

Who uses Complementary and Alternative Medicine? An analysis for cancer patients

Ester Boixadera[1], Anna Espinal[1], Cristina Pallí[2], Jun Lluch[2]

[1] Servei d'Estadística Aplicada
[2] Departament de Psico-oncologia, Institut Català d'Oncologia-Girona. Hospital Universitari Josep Trueta de Girona

Abstract: We present an analysis of cancer patients which use Complementary and Alternative Medicine (CAM). We find out variables coming from their own characteristics, their disease and their own perception of health and quality of life as factors that influencing the use of CAM as a complement to the conventional healthcare.

1 Introduction

The controversial term *Complementary and Alternative Medicine* (CAM) includes any practice of care or healing "that does not fall within the realm of conventional medicine", or maybe "that still has not been shown consistently to be effective".
However, it is common that people choose complementary therapies after or during having experienced some limitations in the conventional medicine for treating certain diseases. That's the case, for instance, of patients with chronic diseases, or with a psycho-somatic component, or to treat side effects caused by conventional treatments. As far as we know, there are few studies about the use of CAM as a complement of the hospital treatment. Some related papers are Evans et al (2007), Albert & Shen (2005).
In this study, we are interested in characterising which kind of patients uses some of those therapies coming from alternative medicine, even though we don't have information on whether it helps to the conventional treatment.

2 The Data

The initial sample was collected in three hospitals. However in this study we restrict to the oncologic patients of Hospital Universitari Josep Trueta de Girona. The database comes from the answers of a questionnaire that patients respond with the help of a qualified member of the clinical team. The sample contains 205 cancer (Breast, Lung, Gastric, Colon, Gynecological, Hematologic) patients with available variables given in four groups:

- Socio-demographic variables of the patient: Sex, age, marital status, studies and nationality.
- Disease's characteristics: Cancer types, months in treatment, months from diagnostic, metastasis and, past and current treatments (chemotherapy, radiation therapy, surgery).
- CAM's characteristics: a dummy variable of whether or not the patient uses CAM. For those receiving CAM it is included:
  - type of CAM (Traditional Chinese Medicine, Homeopathy, Reiki, Bach flower therapy, physiotherapy, diet and others).
  - How they knew of the therapy and what it is used for
  - About the therapist.
  - Since when he is using CAM (in weeks)
  - If the oncologist knows that patient uses CAM and his reaction.
  - Their belief that CAM helps on different facets.
- Quality of life questionnaire: Moreover, patients also answered the EORTC QLQ-C30 questionnaire. The content areas covered by the questionnaire reflect the multi-dimensionality of the Quality of Life (QoL) construct (see Aaronson et al., 1993). Quality of life scales has been used as in Fayers et al (2001). This incorporate five functional scales (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, and nausea and vomiting), a global health status / QoL scale, and a number of single items assessing additional symptoms commonly reported by cancer patients (dyspnoea, loss of appetite, insomnia, constipation and diarrhea) and perceived financial impact of the disease. Scales were used as a binary variables defined by: More than 75% *vs* Equal to or less than 75% in Global health status (QL2cat) and Functional scales; Equal to or greater than 25% *vs* Less than 25% in Symptom scales/items (FIcat).

3 Statistical Analysis

Due to the huge amount of variables, firstly we performed a multivariate correspondence analysis (Lebart et al., 2004) using the disease characteristics as active variables.
After multivariate analysis to characterize patients we establish a logistic regression model (Hosmer and Lemeshow, 2002) for the binary indicator of using CAM. The main goal was finding the main factors that motivate the use of CAM for cancer patients. Covariates included in the model have been selected from this previous multivariate analysis. Moreover, since men and women present different types of cancer, results has been obtained stratifying by sex.

4 Results

The database of 205 patients was analysed for these previous analysis. We analized 85 men (41%) and 120 women (59%) with average age of 57.85 years old (stdev=12.54). The percentage of patients using CAM was 33% and the distribution among type of cancer was: Breast (33%), Colon (19%), Gynaecological (6%), Gastric (9%), Hematologic (16%), Lung (17%).
No statistically significant factor was obtained for the men subsample.
For women, statistically significant factors for the logistic model were: Education, months in treatment, QL2cat (Global health status/QoL, see Data section) and FIcat (Perceived financial impact of the disease, see Data section). Even though type of cancer is not statistically significant but it is clinically relevant; that is why we introduced the breast cancer indicator (1=breast cancer; 0=other cancer).

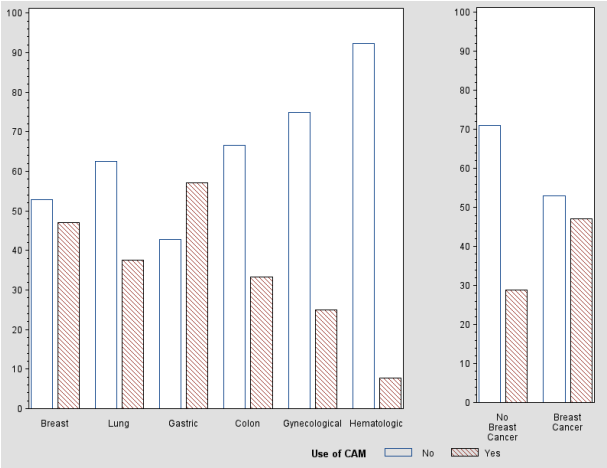


FIGURE 1. Use of CAM by Cancer type.

Results obtained from the model are in Table 1. Women with breast cancer have more than twice the risk of using CAM than others. About the duration of treatment, women with less than 6 month of treatment have two and a half the risk of using CAM. Women who feel better have 2.14 the risk of using CAM. And, woman who didn't perceive financial impact of the disease is 2.83 times the odds of woman who perceived financial impact of the disease, this is, women who do no perceive financial impact of the disease use CAM.
Hosmer and Lemeshow Goodness-of-Fit Test is calculated Statistic Chi-Square=1.2361, DF=7 and p_value=0.9901.
All results were obtained using the software SAS v9.2 (SAS Institute Inc., Cary, NC, USA).

TABLE 1. Odds Ratio Estimates

Variable	Effect	OR	95% Wald CL	
Education	Elementary vs None	4.92	0.54	45.22
	Universitary vs None	4.80	0.48	48.26
	Middle vs None	13.69	1.50	125.45
Breast Cancer Indicator	Yes vs No	2.56	1.02	6.44
Months in treatment	< 6 vs ≥ 6	2.45	1.00	6.00
QL2cat	$> 75\%$ vs $\leq 75\%$	2.14	0.87	5.26
FIcat	$\geq 25\%$ vs $< 25\%$	2.83	1.05	7.61

5 Conclusions

In this study, the main part of patients would like to receive information about complementary therapies from the hospital (85%). For this reason, we encourage to do studies on the effectiveness of these therapies, as a complement of some conventional treatments which may be too severe for some patients.
The main goal of the analysis was to find out patient characteristics which may lead an individual to use complementary therapy. From the model, we got that statistically significant factors were education, breast cancer, few months in treatment, high global health status, and lesser perception of financial impact of the disease.

References

Evans M, Shaw A, Thompson EA, Falk S, Turton P, Thompson T, Sharp D. (2007) *Decisions to use complementary and alternative medicine (CAM) by male cancer patients: information-seeking roles and types of evidence used*, BMC Complementary and Alternative Medicine 2007, **7**:25

Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A, on behalf of the EORTC Quality of Live Group (2001). *The EORTC QLQ-C30 Scoring Manual (3rd Edition)*, Published by: European Organization for Research and Treatment of Cancer, Brussels 2001.

Shen, J. Wenger, N. Glaspy, J. Hays, R. Albert, P. Choi, C. and Shekelle, P. (2000) *Electroacupunture for control of myeloablative chermotherapy-induced emesis: a randomized controlled trial*, J. Am. Med. Ass, **284**: 258-263.

Assessing isotropy with the variogram

Adrian W. Bowman[1] , Rosa M. Crujeiras[2]

[1] School of Mathematics and Statistics, The University of Glasgow, U.K.,
[2] Department of Statistics and Operations Research, University of Santiago de Compostela, Spain

Abstract: Spatial isotropy implies that the dependence between two observations of a spatial process is a function only of the distance between the two sample locations, and not the direction. Although it is a common assumption in spatial data analysis, it may not be realistic in practice. In this work, we present a variogram-based testing technique for assessing isotropy. The method is illustrated with a real data example.

Keywords: Isotropy; Smooth surface; Variogram.

1 Introduction

A common assumption in the analysis of spatial data is isotropy which means direction invariance of the dependence structure. However, in some practical contexts, this assumption may not be reasonable. For instance, when monitoring pollutants coming from a certain emission source, such as an industrial site, wind directions may play a role in the evolution of the process.
Isotropy is a simplifying assumption in data analysis, and it is usually explored by drawing the variogram for different directions, although some formal tests have been introduced by Guan et al. (2004), based on subsampling estimator of the covariance matrix. In this paper, we present a testing procedure for assessing isotropy based on the variogram.
Denote by Z a spatial process defined on a spatial domain $\mathcal{D} \subset \mathcal{R}^2$, $\{Z(s),\ s \in \mathcal{D}\}$. In order to characterize the spatial dependence structure of Z, the variogram is a useful and well-known tool. Specifically, the variogram is denoted by 2γ and it is defined as:

$$2\,\gamma(h) = \mathrm{Var}(Z(s) - Z(s+h)), \quad s, s+h \in \mathcal{D}.$$

Consider n observations of Z at locations $s_1, \ldots, s_n$, denoted by $Z(s_i)$, for $i = 1, \ldots, n$. The variogram is usually estimated by its empirical version based on binning the data, as

$$2\,\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} \{Z(s_i) - Z(s_j)\}^2,$$

where $N(h) = \{(i,j) : \ (s_i, s_j),\ s_i - s_j = h\}$ and $|N(h)|$ denotes the cardinality of $N(h)$. A more robust estimator of the variogram is obtained in the square-root-absolute-value (srav) scale (see Cressie (1993)), as follows:

$$2\,\hat{\gamma}^*(h) = \frac{1}{|N(h)|} \sum_{(i,j)\in N(h)} \sqrt{|Z(s_i) - Z(s_j)|}. \tag{1}$$

In addition, when the original data are normally distributed, the srav of the differences between observations are also well described by a normal distribution. 'The estimate $\hat{\gamma}^*$ can be converted into an estimate of γ by a suitable transformation.

If the spatial process Z is isotropic, then the variogram depends on the difference vector h only through its size. In Section 2, we will briefly describe the testing procedure for assessing isotropy based on the srav-scale variogram, that is, the variogram information is given by the square-root absolute values of the differences between observations. An illustration with real data will be provided in Section 3.

2 Assessing isotropy

Consider the binned-variogram estimator in the srav scale given by (1), obtained from the sample $Z(s_1), \ldots, Z(s_n)$. The difference vectors between sample locations can be expressed in polar form as: $s_i - s_j = h_{ij} e^{iv_{ij}}$, where $h_{ij} = \|s_i - s_j\|$ and $v_{ij} = \angle(s_i, s_j)$. An estimation of the variogram surface in (1) can be obtained by smoothing the data cloud (h_{ij}, v_{ij}, d_{ij}) where $d_{ij} = \sqrt{Z(s_i) - Z(s_j)}$ denote the observed differences in the srav scale, which follow a normal distribution. It is computationally more convenient to work with data which has been binned across both distance and angles. Under the assumption of isotropy, the marginal effect of the distance will be the same for all angles $v \in [0, 2\pi]$ and the estimated surface is obtained as a one-dimensional smooth curve, which is the same for all possible angles v_{ij}. That is, the variogram surface under isotropy varies with h but not with the angle v.

The test statistic is given by:

$$T = \sum_i \left(M_0(h_i, v_i) - M_1(h_i, v_i)\right)^2, \tag{2}$$

where M_0 denotes the smooth variogram surface under isotropy and M_1 is the smooth variogram surface estimator, and i indexes the binned data. Both M_0 and M_1 are obtained by local linear smoothing in practice. Therefore, the testing problem reduces to the comparison of nonparametric surfaces, similar to Bowman (2007).

The test statistic in (2) can be written as a quadratic form in normal random variables, $T = d'Qd$, where d denotes the vector of binned srav

differences and $'$ denotes the transpose. The matrix in the quadratic form is given by $Q = (S_1 - S_0)'\Sigma(S_1 - S_0)$, where S_0 and S_1 are the smoothing matrices in the nonparametric estimators of the variogram surfaces, under the assumption of isotropy and in the general case, respectively. Computation of a p-value can be done using moment matching techniques (see Bowman, 2007) which involves the estimation of the covariance matrix Σ under the null hypothesis of isotropy.

3 Real data analysis

Mosses have been used for decades as biomonitors in order to determine levels of heavy metal concentrations in the atmosphere, since the uptake of metals in mosses comes mainly from the air. A sampling network using this technique has been established in Galicia (NW Spain) since 1995. In 2006, measurements of mercury (Hg, in parts per billion) jointly with other heavy metals were collected on a grid with 148 points covering Galicia and nearby locations. This dataset has been analyzed in order to assess for isotropy. Applying the test statistic for these data, the p-value obtained is 0.018, giving evidence of lack of isotropy.

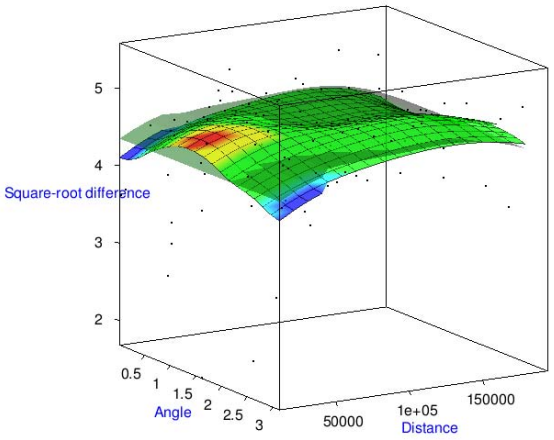


FIGURE 1. Variogram surfaces for Mercury concentrations in March. Isotropic variogram: dark green surface.

Figure 1 shows the fitted variogram surfaces on the srav scale, with colour shading indicating the distance between the isotropic variogram (dark green surface) and the smooth variogram. Red and blue areas indicate that the two surfaces are more than (respectively, less than) two standard errors apart. Hence, the red and blue streaks in the picture indicate higher and lower variance in these directions, suggesting non-isotropy in the blue (lower variance) direction.

Acknowledgments: Work of Rosa M. Crujeiras has been supported by the MTM2008-03010 Project from the Spanish Ministry of Science.

References

Bowman, A.W. (2007). Comparing nonparametric surfaces. *Statistical Modelling*, **6**, 1-21.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics

Guan, Y., Sherman, M. and Calvin, J.A. (2004). A nonparametric test for spatial isotropy using subsampling *Journal of the American Statistical Association*, **99**, 810-821.

Simplified regular vines for modeling high-dimensional financial risk data

Eike Christian Brechmann[1], Claudia Czado[1], Kjersti Aas[2]

[1] Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, 85748 Garching b. München; corresponding e-mail: *brechmann@ma.tum.de.*
[2] Department of Statistical Analysis, Image Analysis and Pattern Recognition, Norwegian Computing Center, Gaustadalléen 23, 0314 Oslo.

Abstract: Regular vines constitute a flexible class of high-dimensional dependency models which use only bivariate copulas as building blocks. The flexibility however comes along with a strongly increasing complexity in higher dimensions. In order to counteract this problem, we propose using efficient statistical model selection techniques to simplify a regular vine. The newly proposed approaches were evaluated in extensive simulation studies and used to investigate a 19-dimensional financial data set of Norwegian and international market variables.

Keywords: multivariate copula; regular vines; simplified vines.

1 Introduction

Introduced by Bedford and Cooke (2001, 2002) and discussed in detail in Kurowicka and Cooke (2006) *regular vines* (R-vines) are a flexible class of high-dimensional dependency models which use only bivariate copulas as building blocks. Each so-called *pair copula* can be chosen arbitrarily and the full model exhibit complex dependence patterns such as asymmetry and tail dependence.
The flexibility however comes along with a strongly increasing complexity in higher dimensions: the number of pair copulas increases quadratically and the number of different R-vines even exponentially. Very recently, there has been considerable progress in constructing R-vines (Dißmann et al. 2011). Nevertheless, for R-vines to be really useful in practice, one needs to be able to fit such structures to data with more than 20 dimensions for which efficient methods were lacking so far. Hence, we treat the problem of determining whether an R-vine can be *simplified* or even *truncated* using simple Gaussian and independence copulas, respectively, after explicitly modeling a certain number of dependency levels.

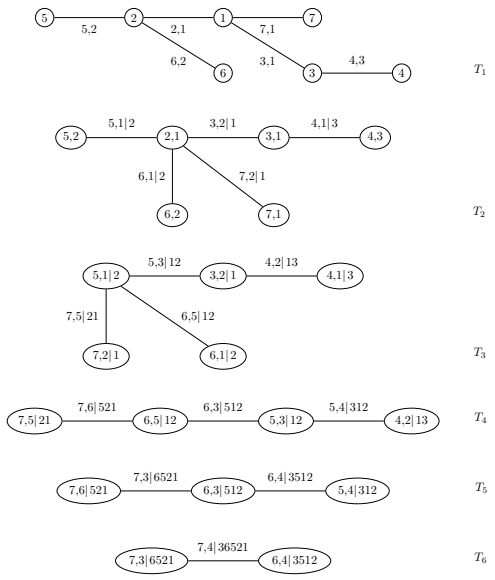


FIGURE 1. An R-vine tree specification on seven variables with edge indices.

2 Regular vines and their simplification

An R-vine on d variables is a sequence of trees $T_1, ..., T_{d-1}$ with nodes N_i and edges E_i, $i = 1, ..., d-1$, which satisfies the following: T_1 has nodes $N_1 = \{1, ..., d\}$ and edges E_1. For $i = 2, ..., d-1$, T_i has nodes $N_i = E_{i-1}$, where two edges in T_i must share a common node if they are to be joined by an edge in tree T_{i+1}. An example of a seven-dimensional R-vine tree specification is given in Figure 1.

The statistical model based on R-vine trees is then obtained by associating each edge $e = j(e), k(e)|D(e)$ in E_i, $i = 1, ..., d-1$, with a bivariate copula density $c_{j(e),k(e)|D(e)}$, a pair copula. These pair copulas constitute the building blocks of the R-vine distribution. Kurowicka and Cooke (2006) prove that the joint density of the d-dimensional random vector $(X_1, ..., X_d)$ is uniquely determined and given by

$$f(\boldsymbol{x}) = \left[\prod_{k=1}^{d} f_k(x_k)\right] \times \left[\prod_{i=1}^{d-1} \prod_{e \in E_i} c_{j(e),k(e)|D(e)}\right],$$

where $f_k, k = 1, ..., d$, are the marginal densities of X_k, $k = 1, ..., d$, and $c_{j(e),k(e)|D(e)} := c_{j(e),k(e)|D(e)}(F(x_{j(e)}|\boldsymbol{x}_{D(e)}), F(x_{k(e)}|\boldsymbol{x}_{D(e)}))$. $\boldsymbol{x}_{D(e)}$ denotes the subvector of $\boldsymbol{x} = (x_1, ..., x_d)'$ determined by the indices $D(e)$. Conditional distribution functions such as $F(x_{j(e)}|\boldsymbol{x}_{D(e)})$ can be obtained recursively using the copula specifications of previous trees (Dißmann et al. 2011).

The number of different possible R-vine tree specifications is very large (Morales-Napoles et al. 2009), e.g., in seven dimensions (cp. Figure 1) there are already more than 2.5 million different R-vines. Following the idea that we want to model the strongest dependencies in the first trees, we therefore construct R-vine trees heuristically by capturing as much dependence as possible in each tree using a maximum spanning tree algorithm. If for example Kendall's τ is used as dependence measure, we select the spanning tree that maximizes the sum of pairwise absolute empirical Kendall's taus $\hat{\tau}_{i,j}$, i.e.,

$$\max \sum_{e=\{i,j\} \text{ in spanning tree}} |\hat{\tau}_{i,j}|.$$

See Dißmann et al. (2011) for more details.
Also the number of pair copulas and hence the computational effort needed to estimate all R-vine parameters strongly increase with the dimension: there are $d(d-1)/2$ pair copulas in a d-dimensional R-vine. In high dimensions and under limited time and computational resources, we therefore want to find the best possible specification of the first K trees in the R-vine, while higher order trees should only involve simple pair copulas.
Specifically, we denote an R-vine a *pairwisely simplified* K level one, $\text{sRV}(K)$, if we replace all pair copulas in trees higher than K by Gaussian copulas which are easier to specify than other copulas and straightforward to interpret in terms of the correlation parameter. Further, we speak of a *pairwisely truncated* R-vine at level K, $\text{tRV}(K)$, if all pair copulas in trees higher than K are set to independence copulas. Truncation may also be regarded as a special case of simplification, using Gaussian pair copulas with correlation parameter equal to zero. Hence, it constitutes the greatest possible simplification.
The density of a pairwisely simplified K level R-vine distribution is given by

$$f(\boldsymbol{x}) = \left[\prod_{k=1}^{d} f_k(x_k)\right] \times \left[\prod_{i=1}^{K} \prod_{e \in E_i} c_{j(e),k(e)|D(e)}\right] \times \left[\prod_{i=K+1}^{d-1} \prod_{e \in E_i} c^{\rho}_{j(e),k(e)|D(e)}\right],$$

where $c^{\rho}_{j(e),k(e)|D(e)}$ denote Gaussian pair copulas and arguments have been omitted for reasons of readability.
The density of a pairwisely truncated R-vine at level K is given similarly with the rightmost part of the above equation collapsing to 1, since the density of the independence copula is simply 1.
For canonical vines, a special case of R-vines (Aas et al. 2009), the product of all pair copulas involved in trees higher than K gives a $(d-K)$-variate copula. We call this *joint simplification*. It has previously been treated by Valdesogo (2009) and Heinen and Valdesogo (2009) and more details on it can also be found in Brechmann et al. (2010).

3 Selection criteria

We will now consider the selection of simplification levels. Truncation level selection follows as a special case and can exploit the nestedness of $\mathrm{tRV}(K)$ and $\mathrm{tRV}(K+1)$. $\mathrm{sRV}(K)$ and $\mathrm{sRV}(K+1)$ are not nested in general.
We will start with $K = 1$ and fit a simplified R-vine. We thereafter increase K by one and assess the gain by fitting the extra tree. If the gain is negligible we stop and use the resulting specification. If the gain is large enough, we increase K by one again, and proceed in this way until we have reached a simplification level K_0, which either gives a sufficient fit, or we have reached the computational time frame we allowed for the estimation process.
To assess whether there is gain to move from model $\mathrm{sRV}(K)$ to $\mathrm{sRV}(K+1)$, we now consider two kinds of statistical model selection techniques. First, we choose the one of $\mathrm{sRV}(K)$ and $\mathrm{sRV}(K+1)$ with the smaller AIC or BIC value. If for some K_0 the smaller model is chosen, we stop, and use the model $\mathrm{sRV}(K_0)$. AIC/BIC comparisons for non-nested models however induce an increased variability. Second, we therefore consider the likelihood-ratio based test for non-nested model comparisons by Vuong (1989). It determines the simplification level as the level K_0 for which $\mathrm{sRV}(K_0+1)$ does not provide a *statistically significant* gain. The Vuong test statistic may also be corrected for the number of model parameters using the Akaike and the Schwarz corrections, which correspond to the AIC and BIC penalty terms, respectively.
In the case of joint simplification of canonical vines one may in addition to the above-mentioned model selection methods, use copula goodness-of-fit tests to determine the truncation/simplification level (Brechmann et al. 2010).
To sum it up, at each level four steps have to performed:

1. Tree construction using maximum spanning trees.
2. Pair copula type selection (e.g., using copula goodness-of-fit tests or AIC comparisons).
3. Pair copula parameter estimation.
4. Investigating whether truncation and/or simplification are possible.

In extensive simulation studies we validated this heuristic approach and, in particular, all five procedures for the selection of simplification and truncation levels: AIC, BIC, Vuong test with and without Akaike and Schwarz correction. It turned out that the Vuong tests are superior to AIC/BIC and perform quite well. More parsimonious models are typically found using the BIC and, especially, the Vuong test with Schwarz correction. Details can be found in Brechmann et al. (2010).

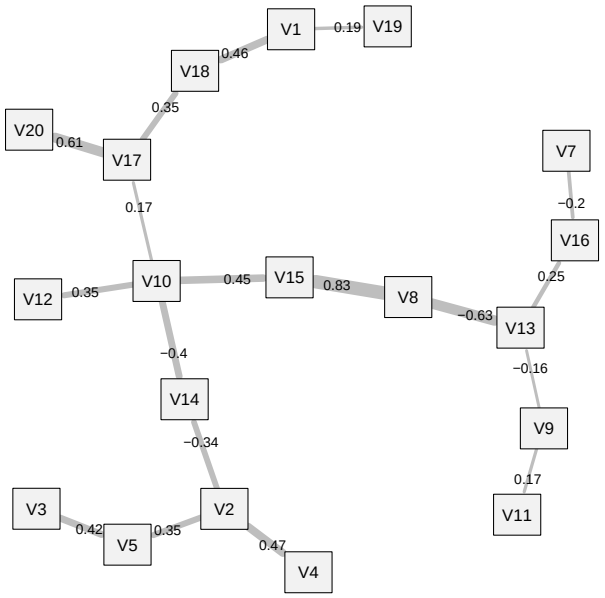


FIGURE 2. First R-vine tree for the financial data set. Edge labels indicate empirical Kendall's τ's between the respective variables.

4 Application

We analyzed a 19-dimensional data set consisting of Norwegian and international financial variables with 1107 daily observations from March 2005 to March 2008. The variables constitute the market portfolio of a large Norwegian financial institution and hence, it is crucial to correctly model the dependencies between them.

When investigating possible simplification of adequate R-vine specifications with marginal ARMA-GARCH models and a range of ten different copula types (allowing, amongst others, for tail dependence and asymmetric dependence), simplification at level 2 and truncation at level 4 or 6 turned out to be appropriate. Further, in comparison to the multivariate t-copula, currently the state-of-the-art approach for modeling financial return data, all truncated and simplified models were statistically equivalent or superior. In economical terms, our model has an evident interpretation. It is constituted of three clusters of economically similar variables (see the first R-vine tree in Figure 2). The first cluster consists of stock, hedge fond and real estate indices (V1, V17-V20). The second cluster consists of interest rates and bond indices (V7-V16), and finally, exchange rates (V2-V5) constitute the third cluster. The identified simplification and truncation levels indicate that dependencies within these clusters are the most important ones to model accurately.

5 Conclusion

The methods discussed here allow for the first time to efficiently construct flexible R-vine models even in higher dimensions and under time and resource restrictions. As such, R-vine models constitute a powerful class of high-dimensional dependency models, available for a wide range of applications (see Brechmann and Czado (2011) for a large scale financial application).

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, **44**, 182-198.

Bedford, T., and Cooke, R. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, **32**, 245-268.

Bedford, T., and Cooke, R. (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics*, **30**, 1031-1068.

Brechmann, E.C., and Czado, C. (2011). Extending the CAPM using pair copulas: The regular vine market sector model. *Submitted*.

Brechmann, E.C., Czado, C., and Aas, K. (2010). Truncated regular vines in high dimensions with application to financial data. *Submitted*.

Dißmann, J., Brechmann, E.C., Czado, C., and Kurowicka, D. (2011). Selecting and estimating regular vine copulae and application to financial returns. *In preparation.*

Heinen, A., and Valdesogo, A. (2009). Asymmetric CAPM dependence for large dimensions: the Canonical Vine Autoregressive Model. CORE Discussion Papers 2009069, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Kurowicka, D., and Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Chichester: Wiley.

Morales-Napoles, O., Cooke, R.M., and Kurowicka, D. (2009). About the number of vines and regular vines on n nodes. *Submitted*.

Valdesogo, A. (2009). *Multivariate volatility models using copulas*. Ph. D. thesis, Université Catholique de Louvain, Belgium.

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307-333.

Climate Envelopes for Species Distribution Models

Mark J Brewer[1], Robert B O'Hara[2], Barbara J Anderson[3], Ralf Ohlemüller[4]

[1] Biomathematics and Statistics Scotland, The James Hutton Institute, Craigiebuckler, Aberdeen, Scotland, AB15 8QH, UK, **Email:** M.Brewer@bioss.ac.uk
[2] Biodiversity and Climate Research Centre - Bik-F, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany
[3] Dept of Biology, University of York, PO Box 373, York, YO10 5YW, UK
[4] School of Biological and Biomedical Sciences, Durham University, South Road, Durham, DH1 3LE, UK

Abstract: Spatial models of species distribution often include attempts to describe relationships with climate variables via low-degree spline curves; these are commonly termed "climate envelopes". Such curves should typically be either unimodal or monotonic. We propose a simple parametric alternative to spline curves which appeals to biological plausibility and can capture common expected features of species' presence/climate relationships. Furthermore, the methodology can be extended to the multivariate case in a straightforward manner.

Keywords: Climate envelope; niche modelling; species distribution.

1 Introduction

As an important part of understanding the likely effects of future climate change, it is vital we understand the relationships between species distributions and climate conditions. Historically, methods for modelling these relationships have proved inadequate: Austin (2002) criticised the then-popular use of Normal distribution curves to model responses as being too restrictive and unrealistic, instead recommending smooth spline ("GAM") terms. Alternatively, Oksanen and Minchin (2002) suggested using the suite of five curves (ranging from completely flat to skew-unimodal) from Huisman *et al.* (1993); these, however, can suffer from discontinuities. Recent papers (e.g. Heikkinen and Mäkipää, 2010) have tended to use smooth spline terms with few degrees of freedom. In practice, species' responses to climate variables (if present) are expected to be unimodal or monotonic. Here we present a parametric form for modelling species/climate relationships that is biologically plausible, is efficient in terms of the number of parameters involved, and can be readily extended to a multivariate setting to account for interactions between climate variables. We study data on

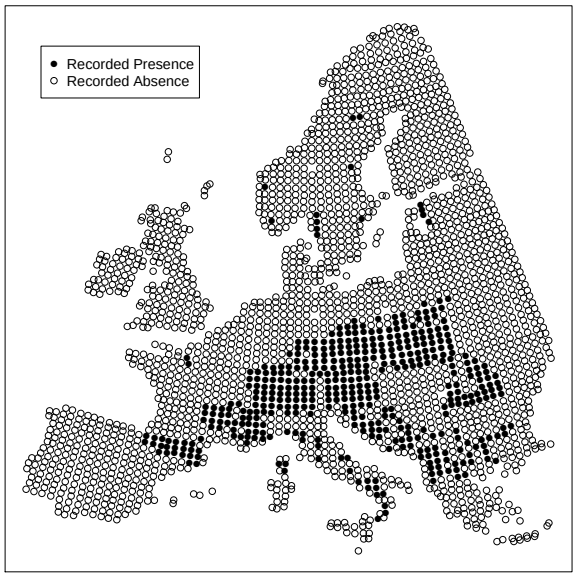


FIGURE 1. Distribution of *Abies alba* in Europe.

tree species from the Europe-wide data set Atlas Florae Europaeae. For the species *Abies alba* (European Silver Fir) recorded presences are shown in Figure 1. Four potential climate variables have been identified: an index of drought (DRO); the number of "growing degree days" (GDD); the mean temperature of the coldest month (MTCO); and the mean temperature of the warmest month (MTWA).

2 A Species Distribution Model

We fit a Bayesian logistic spatial regression model to the presence/absence data including a spatially-structured random effect (Besag *et al.*, 1991) in the WinBUGS software (Lunn *et al.*, 2000). With presence/absence response Y_i for observation cell $i = 1, \ldots, 2606$ we have a Binomial error model with logit link and linear predictor

$$\mu_i = \beta_0 + g(D_i) + g(G_i) + g(C_i) + g(W_i) + u_i$$

for: intercept β_0; covariates D_i (DRO), G_i (GDD), C_i (MTCO) and W_i (MTWA); random effect u_i where

$$f\left(u_i \mid \tau\right) \propto \tau^{n/2} \exp\left\{-\frac{\tau}{2} \sum_{i \sim i'} \left(u_i - u_{i'}\right)^2\right\}$$

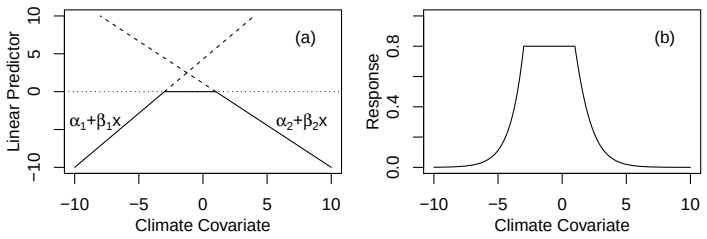


FIGURE 2. Envelopes: (a) linear predictor scale; (b) response (probability) scale.

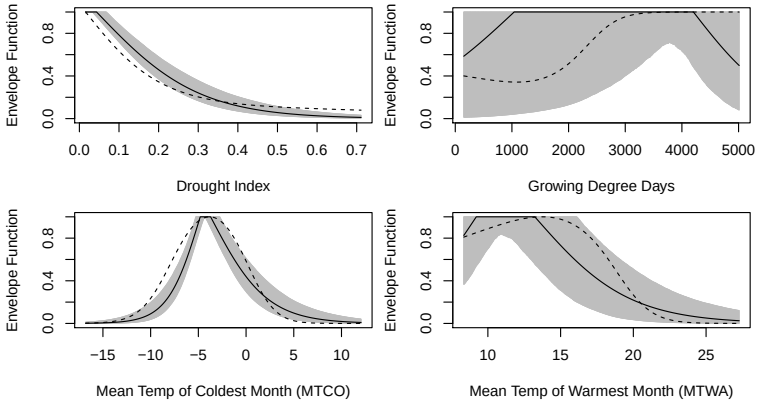


FIGURE 3. Estimated univariate climate envelopes for *Abies alba*.

with smoothing parameter τ; and (initially) univariate envelope functions $g()$ of the form

$$g(x) = \min\{\alpha_1 + \beta_1 x, 0, \alpha_2 + \beta_2 x\} \tag{1}$$

with $\beta_1 > 0$ and $\beta_2 < 0$. The form of functions $g()$ is piecewise linear on the linear predictor scale, as seen by the solid line of Figure 2(a). This maps onto the curve shown in Figure 2(b) on the response scale; the key here is that this flexible yet efficient parametric form allows for a "plateau", and is guaranteed to be either unimodal or monotonic.

Univariate climate response functions for the *Abies alba* data are shown in Figure 3 as solid curves; the shaded regions represent the associated uncertainty, while the broken curves show alternative fits from a GAM, where the $g()$ functions are replaced by splines of low degree. While the envelopes are similar, the spline term for growing degree days does have an unlikely feature—the rise towards zero days.

3 Multivariate Envelope Functions

The piecewise polynomial function of (1) can be generalised to a multivariate setting by thinking of the climate envelope as a "top-sliced warped cone", allowing for different slopes either side of the apex for each covariate; pairwise interactions can also be included. A general multivariate envelope cone in this form can be defined by the cartesian equation, in M dimensions, for apex $\mathbf{a}$ and covariates x:

$$\begin{aligned}(z-a_z)^2 &= \sum_{i=1}^{M} \beta_{i,1}\left(x_i - a_{x_i}\right)^2 I\left[x_i < a_{x_i}\right] + \beta_{i,2}\left(x_i - a_{x_i}\right)^2 I\left[x_i \geq a_{x_i}\right] \\ &\quad + \sum_{i>j} \beta_{i,j}\left(x_i - a_{x_i}\right)\left(x_j - a_{x_j}\right)\end{aligned}$$

and where the apex height is constrained positive (i.e. $a_z > 0$), where $\beta_{i,1} > 0, \beta_{i,2} < 0\, \forall i$, and the top-slicing is enforced by $z = \min(z, 0)$. Our model therefore becomes

$$\mu_i = \beta_0 + g(D_i, G_i, C_i, W_i) + u_i$$

for multivariate envelope function $g()$. In terms of implementation, the multivariate cone can be coded efficiently via max/min operations, and note that despite a longer per-iteration run time, in practice convergence has proved to be much faster than with our univariate envelopes.

We illustrate the multivariate envelope on a second species, *Castanea sativa* (Sweet Chestnut), whose distribution is shown in Figure 4. Using a GAM, fitting a smooth to all four variables at once proved highly unsatisfactory, as the results were too "wiggly" and not at all realistic; what is shown in Figure 5 is a model having a bivariate smooth for MTCO and MTWA but separate additive univariate terms for the other two variables. Using the `gam` function in the `mgcv` package (Wood, 2006) in R (R Development Core Team, 2010), we needed to fix the degrees of freedom at 4 for the bivariate spline smooth, as any other value gave an envelope function that was very different and did not seem at all realistic; it seems, then, that using "default" GAM software in this way has a considerable lack of robustness. Note also the slight rise for very low values of minimum temperature, itself an unlikely feature. The "top-sliced warped cone" model, however, has no such issues, and the plot shown in the right-hand panel of Figure 5 is a 2-D projection for the temperature variables from a model with a full 4-D fit with all four variables and all pairwise interaction terms; to generate the plot, the values of DRO and GDD chosen corresponded to the apex $\mathbf{a}$. The fitting process required no tuning, and generic priors were appropriate for all parameters; convergence was achieved around 500 iterations, and the envelope shown was obtained from a further 500. The plot suggests a much clearer picture, showing that mean temperature of the coldest month is very important for *Castanea sativa*, as it is generally not found in regions with colder winters.

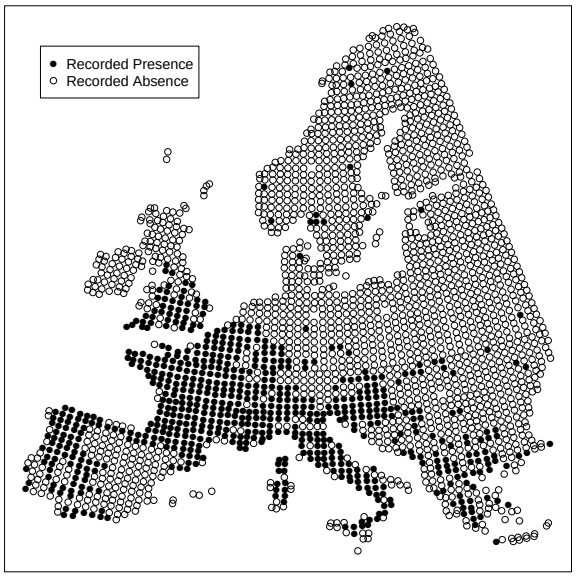


FIGURE 4. Distribution of *Castanea sativa* in Europe.

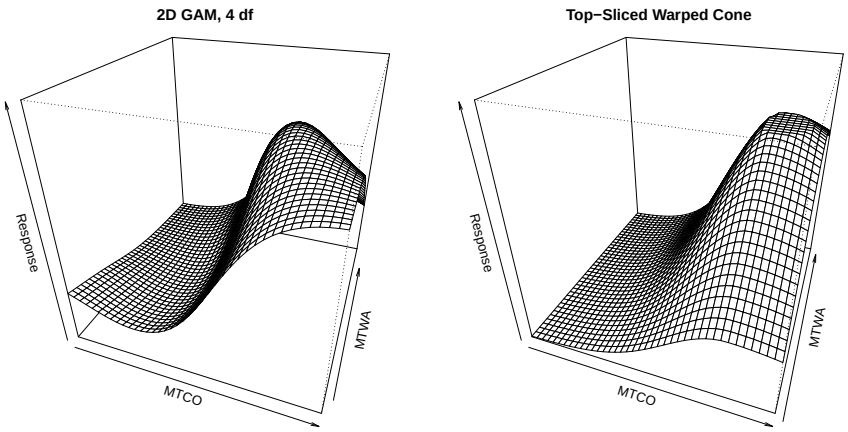


FIGURE 5. Multivariate climate envelopes for *Castanea sativa*.

4 Discussion

Potential correlations between climate variables suggest that envelope functions should be multivariate, but using standard GAM/spline implementa-

tions in R for example, we found fitting 4-D surfaces problematic with our species presence/absence data; even 2-D surfaces were highly sensitive to choice of degrees of freedom. It is possible that realistic envelopes could be obtained by setting further constraints on GAM smooths—for example, by forcing spline curves to be either monotonic or a combination of two monotonic segments (having a maximum at the join). However, the top-sliced warped cone function introduced here is simple, fits quickly, is intrinsically realistic and is easily interpretable.

Acknowledgments: This work received funds from the Rural and Environment Research and Analysis Directorate (RERAD) of the Scottish Government, and from the UKPopNet project "Developing Bayesian hierarchical models of UK butterfly distributions".

References

Austin, M.P. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101-118.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.

Heikkinen, J. and Mäkipää, R. (2010). Testing hypotheses on shape and distribution of ecological response curves. *Ecological Modelling*, **221**, 388-399.

Huisman, J., Olff, H., and Fresco, L.F.M. (1993). A hierarchical set of models for species response analysis. *Journal of Vegetation Science*, **4**, 37-46.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325-337.

Oksanen, J., and Minchin, P.R. (2002). Continuum theory revisited: what shape are species responses along ecological gradients? *Ecological Modelling*, **157**, 119-129.

R Development Core Team (2010). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Wood, S.N. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

XD survival regression models with frailty

Kevin Burke[1] , Gilbert MacKenzie[1,2]

[1] Centre for Biostatistics, University of Limerick, Ireland.
[2] ENSAI, Rennes, France.

Email: kevin.burke@ul.ie gilbert.mackenzie@ul.ie

Abstract: We aim to explore the survival distributions based on the extreme dispersion, *XD*, models proposed by Jørgensen (2010). Survival times can be modelled within the *XD* framework by taking $\log T = Y^* \sim XD(\mu, \lambda)$, where T is the survival time and Y^* is the extreme dispersion random variable. We will show how these survival models can be generated within *XD* and investigate the distribution and properties of the survival time $T = e^{Y^*}$. We will also introduce a frailty generalization of *XD*.

Keywords: Extreme dispersion; Morris class; Quadratic slope; Survival regression model; Frailty.

1 Introduction

Jørgensen (2010) proposed an extreme value analogue of exponential dispersion models, *ED*, and generalized linear regression models. The slope function is introduced as an analogue of the variance function. Therefore the slope function characterizes the extreme dispersion model, *XD* in much the same way that the variance function characterizes the exponential dispersion model.
We start with a basic no-parameter survivor function $\Pr(Y > y) = G(y)$, where Y is contained in the interval $\mathcal{C} \subseteq \mathrm{R}$. Typically $\mathcal{C} = [a, b)$, however the interval can be open or closed at either end point. The density function corresponding to $G(y)$ integrates to 1 over this interval, $\int_a^b f(y)\mathrm{d}y = 1$.
In the *XD* framework $G(y)$ is analogous to the moment generating function of *ED*. Therefore, apart from a sign change, the cumulative hazard function $H = -\log G$ is analogous to the cumulant generating function. As we know, $h = H'$ is the hazard function which is therefore analogous to the mean value mapping, τ, of *ED*. So we have

$$r(Y) = h(0), \qquad s(Y) = h'(0), \tag{1}$$

where r is the *rate* and s is the *slope*, being analogous to the mean and variance. Unlike variance, the slope can be negative as well as positive. In *ED*, τ must be monotone increasing. However, h can be monotone increasing or

decreasing. Interestingly, Jørgensen (2010) shows that the hazard function for $T = e^Y$ does not need to be monotone, even if it does for Y. Therefore we are not limited in this sense when considering the survival extension of the *XD* model.
Going back to the random variable Y, the *unit slope function* is defined as

$$v(\mu) = h'(h^{-1}(\mu)), \tag{2}$$

where $\mu \in \Psi = h(\mathcal{C}) = [h(a), h(b))$ is the *rate* parameter (more on this in the next section). The slope function, $v(\mu)$, maps Ψ onto $\mathcal{R}_+$ when h is increasing and onto $\mathcal{R}_-$ when h is decreasing. Here $h^{-1}(\cdot)$ is the inverse hazard function so that $h(h^{-1}(\mu)) = \mu$.
It can be shown that the inverse hazard function satisfies the differential equation

$$\frac{\mathrm{d}h^{-1}(\mu)}{\mathrm{d}\mu} = \frac{1}{v(\mu)}. \tag{3}$$

Replacing h with τ and v with V in Equations (2) and (3) gives us the equations in the exponential dispersion framework which show the relationship between the mean value mapping and the unit variance function.

2 The extreme dispersion model

The extreme dispersion model, $XD(\mu, \lambda)$, generated by G has survivor function

$$G(y^*; \mu, \lambda) = G^\lambda \left(\frac{y^*}{\lambda} + h^{-1}(\mu) \right). \tag{4}$$

The *XD* model has support on $\mathcal{C}^* = \lambda(\mathcal{C} - h^{-1}(\mu))$, i.e. this is the support of the *XD* random variable Y^*, where we have included the superscript "$*$" to differentiate from $Y \in \mathcal{C}$, which is the random variable corresponding to $G(y)$; the function we're using to generate *XD*. Note that the survivor function $G(y)$ has no parameters. It is used merely to generate $G(y^*; \mu, \lambda)$ which has two parameters.

Here $\mu \in \Psi = h(\mathcal{C})$ and $\lambda > 0$. So to generate the *XD* survivor function we replace y by $\frac{y^*}{\lambda} + h^{-1}(\mu)$ in $G(y)$ and raise it to the power of λ.
The reason for carrying out this operation is based on properties of the moment generating function and the fact that $G(y)$ is analogous to this function.
We can see that the cumulative hazard function for the *XD* model is

$$H(y^*; \mu, \lambda) = \lambda H \left(\frac{y^*}{\lambda} + h^{-1}(\mu) \right), \tag{5}$$

where $H(\cdot) = -\ln G(\cdot)$. And the hazard function is

$$h(y^*;\mu,\lambda) = h\left(\frac{y^*}{\lambda} + h^{-1}(\mu)\right) \tag{6}$$

where $h(\cdot) = H'(\cdot)$.
The density function for the *XD* model is therefore

$$f(y^*;\mu,\lambda) = h\left(\frac{y^*}{\lambda} + h^{-1}(\mu)\right)\exp\left(-\lambda H\left(\frac{y^*}{\lambda} + h^{-1}(\mu)\right)\right). \tag{7}$$

It can be shown that $h(0;\mu,\lambda) = \mu$ and $h'(0;\mu,\lambda) = \frac{1}{\lambda}v(\mu)$ for the *XD* model. So we see that μ is the *rate* and $\frac{1}{\lambda}v(\mu)$ is the *slope* of Y^*. We also see why $v(\mu)$ is called the *unit slope function* as it corresponds to $\lambda = 1$.
Jørgensen (2010) puts special emphasis on *XD* models with quadratic unit slope function. These are analogous to the exponential dispersion models with quadratic variance functions known as the Morris class.

3 *XD* survival model

To obtain the survival model based on the *XD* model, we let $Y^* = \log T$ where T is the positive survival time. This gives us the following

$$G_T(t;\mu,\lambda) = G^{\lambda}\left(\frac{\ln t}{\lambda} + h^{-1}(\mu)\right), \tag{8}$$

$$H_T(t;\mu,\lambda) = \lambda H\left(\frac{\ln t}{\lambda} + h^{-1}(\mu)\right), \tag{9}$$

$$h_T(t;\mu,\lambda) = \frac{1}{t}h\left(\frac{\ln t}{\lambda} + h^{-1}(\mu)\right), \tag{10}$$

and,

$$f_T(t;\mu,\lambda) = \frac{1}{t}h\left(\frac{\ln t}{\lambda} + h^{-1}(\mu)\right)\exp\left(-\lambda H\left(\frac{\ln t}{\lambda} + h^{-1}(\mu)\right)\right), \tag{11}$$

which has support

$$\mathcal{C}_T = \exp(\mathcal{C}^*) \tag{12}$$

where the subscript T here indicates that these functions correspond to the survival time T.

4 Estimation

Jørgensen (2010) suggests both a quasi-likelihood method and maximum likelihood for fitting these models (in the regression setting). He discusses some potential problems with both methods. We will consider the latter method here.
Looking only at the survival *XD* model, the log-likelihood function is given by:

$$\ell_T(\mu,\lambda) = \sum_{i=1}^{n} \delta_i \log\left[\frac{1}{t_i} h\left(\frac{\ln t_i}{\lambda} + h^{-1}(\mu)\right)\right] - \lambda H\left(\frac{\ln t_i}{\lambda} + h^{-1}(\mu)\right). \tag{13}$$

where δ_i is the censoring indicator, $\delta_i = 0$ if the survival time is censored.

5 Example: Rayleigh-*XD*

We now look at an example; the 'Rayleigh-*XD*', so called because it is the $XD(\mu,\lambda)$ model based on the Rayleigh generator, $G(y) = \exp(-y^2/2)$. Using $G(y)$ (the generator) and the equations in Sections 2 and 3, we can obtain the *XD* model and it's survival counterpart.

5.1 *XD* Model

$$G(y^*;\mu,\lambda) = \exp\left(-\frac{\lambda}{2}\left(\frac{y^*}{\lambda} + \mu\right)^2\right) \tag{14}$$

$$H(y^*;\mu,\lambda) = \frac{\lambda}{2}\left(\frac{y^*}{\lambda} + \mu\right)^2 \tag{15}$$

$$h(y^*;\mu,\lambda) = \frac{y^*}{\lambda} + \mu \tag{16}$$

$$Support : \mathcal{C}^* = \lambda \times ([0,\infty) - \mu) = [-\lambda\mu, \infty) \tag{17}$$

5.2 Survival

$$G_T(t;\mu,\lambda) = \exp\left(-\frac{\lambda}{2}\left(\frac{\log t}{\lambda} + \mu\right)^2\right) \tag{18}$$

$$H_T(t;\mu,\lambda) = \frac{\lambda}{2}\left(\frac{\log t}{\lambda} + \mu\right)^2 \tag{19}$$

$$h_T(t;\mu,\lambda) = \frac{1}{t}\left(\frac{\log t}{\lambda} + \mu\right) \tag{20}$$

$$Support : \mathcal{C}_T = \exp(\mathcal{C}^*) = [\exp(-\lambda\mu), \infty) \tag{21}$$

6 Simulation

We simulated survival times from the survival Rayleigh-*XD*. We did this for all combinations of $\mu = (2, 0.5)$, $\lambda = (2, 0.5)$ and $n = (100, 1000)$, where n is the sample size. So in total there were $2 \times 2 \times 2 = 8$ different settings. There was no censoring for the purposes of this simulation. Survival times were generated using the inverse function,

$$t = F^{-1}(u) = \exp\left(\lambda\left\{\left[-\frac{2}{\lambda}\log(1-u)\right]^{1/2} - \mu\right\}\right), \tag{22}$$

where $u \sim Uniform(0, 1)$ is generated using a random number generator. Maximum likelihood was then used to fit the survival Rayleigh-*XD* model to the simulated data. Within each setting this was done 1000 times. The average of the 1000 MLEs for each setting is shown in Table 1 below, along with the % bias given by $pbias = 100(\tilde{\theta} - \theta)/\theta$, where $\tilde{\theta}$ represents the average of the 1000 MLEs.

TABLE 1. Simulation Results

	μ	λ	n	$\tilde{\mu}$	$pbias_\mu$	$\tilde{\lambda}$	$pbias_\lambda$
1	2.0	2.0	100	2.07	3.54	1.94	-2.80
2	0.5	2.0	100	0.50	0.58	1.95	-2.53
3	2.0	0.5	100	2.05	2.69	0.49	-2.42
4	0.5	0.5	100	0.49	-2.86	0.49	-2.40
5	2.0	2.0	1000	2.01	0.51	1.99	-0.53
6	0.5	2.0	1000	0.50	-0.02	1.99	-0.54
7	2.0	0.5	1000	2.01	0.34	0.50	-0.46
8	0.5	0.5	1000	0.50	-0.51	0.50	-0.44

We can see from Table 1 that there does not seem to be an issue with the MLEs in this simple setting for the survival Rayleigh-*XD*.

7 Extreme generalized linear models

Extreme generalized linear models are formed by regression of the rate $r(Y_i^*) = \mu(x_i'\beta)$ where $Y_1^*, \ldots, Y_n^*$ are independent random variables. Here $\mu(x_i'\beta)$ denotes the inverse link function, x_i is a vector of covariates for the ith case and β is a vector of unknown regression parameters. Therefore

$$Y_i^* \sim XD(\mu(x_i'\beta), \lambda). \tag{23}$$

We can then model T_i, the survival time for the ith individual, by letting $Y_i^* = \log T_i$.

8 Frailty

We can further extend this survival *XD* regression by multiplying the hazard function, $h_T(t;\beta,\lambda) = h_T(t;\mu(x_i'\beta),\lambda)$, by a *random effects* term. Thus the *conditional* hazard and survivor functions are

$$h_T(t;\beta,\lambda\,|\,z) = zh_T(t;\beta,\lambda), \tag{24}$$

and

$$G_T(t;\beta,\lambda\,|\,z) = e^{-zH_T(t;\beta,\lambda)}. \tag{25}$$

If we assume that z has a gamma distribution, $g(z)$, with $E(Z) = 1$ and $Var(Z) = \sigma^2$, then we can integrate out the random effect to obtain the *marginalized* survivor function,

$$G_T(t;\beta,\lambda,\sigma^2) = \int_0^\infty e^{-zH_T(t;\beta,\lambda)}g(z)\mathrm{d}z = \left[1+\sigma^2 H_T(t;\beta,\lambda)\right]^{\frac{-1}{\sigma^2}}, \tag{26}$$

and the corresponding marginalized hazard function is

$$h_T(t;\beta,\lambda,\sigma^2) = \frac{h_T(t;\beta,\lambda)}{1+\sigma^2 H_T(t;\beta,\lambda)}. \tag{27}$$

This now depends on σ^2 which is another parameter that must be estimated.

9 Discussion

The *XD* class of extreme dispersion models offers the prospect of developing a new class of survival models with novel properties. Our early work confirms that this is indeed the case. However, the resulting survival models appear to embody a latent period before which failure becomes operational. Whilst this is always a testable hypothesis, in the group setting, the extent to which these models can be usefully applied in practice remains to be ascertained.

Acknowledgments: This work was supported, in part, by the SFI's (www.sfi.ie) BIO-SI research programme, grant number, **07MI012**. The first author is an IRCSET Scholar (www.ircset.ie) and the second is the Principal Investigator of BIO-SI (www.ul.ie/bio-si).

References

Jørgensen, B. (1997) *The Theory of Dispersion Models*. Chapman & Hall.

Jørgensen, B., Goegebeur, Y. and Martinez, J.R. (2010) Dispersion models for extremes. *Extremes*, **13**, 399-437.

Least-squares signal estimation using correlated delayed observations transmitted by different sensors

R. Caballero-Águila[1], A. Hermoso-Carazo[2] , J. Linares-Pérez[2]

[1] Dpto. de Estadística e I.O., Universidad de Jaén, 23071 Jaén, Spain (raguila@ujaen.es)
[2] Dpto. de Estadística e I.O., Universidad de Granada, 18071 Granada, Spain (ahermoso@ugr.es, jlinares@ugr.es)

Abstract: This paper is concerned with the least-squares (LS) linear filtering problem of discrete-time signals from noisy measurements coming from multiple randomly delayed sensors with different delay characteristics. It is assumed that the Bernoulli random variables characterizing the measurement delay are correlated at consecutive sampling times. Using an innovation approach, a recursive linear filtering algorithm is obtained without requiring the state-space model generating the signal, but only the covariance functions of the signal and the noise, the delay probabilities and the correlation function of the Bernoulli variables.

Keywords: Least-squares estimation; Randomly delayed observations; Covariance information; Innovation approach; Multiple sensors.

1 Introduction

In many practical situations, for example in networked systems with a heavy network traffic, data packets may suffer transmission delays due to numerous causes, such as network congestion, random failures in the transmission device, accidental loss of some measurements, or data inaccessibility at certain times. Moreover, these time-delays are often random in nature.
Standard estimation algorithms are not applicable in such situations where measurements are randomly delayed, thus being necessary to modify these algorithms incorporating the effects of random delays. The signal estimation problem for models with random delays has been widely investigated assuming full knowledge of the signal state-space model (see e.g. Su and Lu (2001)) and using only the covariance functions of the processes involved in the observation model (see e.g. Nakamori et al. (2005)).
However, most papers concerning systems with randomly delayed sensors assume that all the sensors have the same delay characteristics. In the last years, Hounkpevi and Yaz (2007) (using the state-space model) and Caballero-Águila et al. (2010) (using covariance information) have generalized this situation considering multiple delayed sensors with different delay

characteristics. The main assumption in these papers is that the delays are mutually independent. In the current paper, this restriction is weakened; specifically, we consider different sequences of Bernoulli variables correlated at consecutive sampling times to characterize the measurement delay of each sensor. This correlation model covers situations where consecutive observations cannot be delayed; for example, signal transmission problems with stand-by sensors where any transmission failure in a sensor is immediately detected and the failed sensor is replaced.

2 Delayed observation model

Consider m scalar sensors whose real measurements, $\widetilde{y}_k^i$, of the n-dimensional signal, z_k, are perturbed by additive noise vectors v_k^i; that is,

$$\widetilde{y}_k^i = H_k^i z_k + v_k^i, \quad k \geq 1, \quad i = 1, \ldots, m. \tag{1}$$

Assume that at time $k = 1$ the real measurements, $\widetilde{y}_1^i$, are always available for the estimation, but at any time $k > 1$, the available measurement coming from each sensor may be randomly delayed by one sampling time according to different delay characteristics. Therefore, if $\{\gamma_k^i;\ k > 1\}$, $i = 1, \ldots, m$, denote sequences of Bernoulli random variables, the available measurement of the ith sensor, y_k^i, is described by

$$y_k^i = (1 - \gamma_k^i)\widetilde{y}_k^i + \gamma_k^i \widetilde{y}_{k-1}^i, \quad k > 1; \quad y_1^i = \widetilde{y}_1^i, \quad i = 1, \ldots, m. \tag{2}$$

From (2) it is clear that, if $\gamma_k^i = 1$, which occurs with probability p_k^i, then $y_k^i = \widetilde{y}_{k-1}^i$ and the measurement of the ith sensor is delayed by one sampling period; otherwise, $\gamma_k^i = 0$ and $y_k^i = \widetilde{y}_k^i$, which means that the measurement is up-to-date with probability $1 - p_k^i$. Therefore, the variables $\{\gamma_k^i;\ k > 1\}$ model the random delay of the ith sensor and the values $\left\{p_k^i,\ k > 1\right\}$ represent the delay probabilities in the measurements coming from the ith sensor. It is also assumed that a delay in the observation at time k depends on a delay at time $k - 1$, but it is independent of delays at times previous to $k - 1$; this is formulated by imposing the stochastic independence of the Bernoulli variables γ_k^i and γ_s^j when $|k - s| \geq 2$.

For simplicity, (1) and (2) are rewritten as follows:

$$\begin{array}{c} \widetilde{y}_k = H_k z_k + v_k, \quad k \geq 1, \\ y_k = (I_m - \Gamma_k)\widetilde{y}_k + \Gamma_k \widetilde{y}_{k-1}, \quad k > 1; \quad y_1 = \widetilde{y}_1, \end{array} \tag{3}$$

where $\widetilde{y}_k = \left(\widetilde{y}_k^1, \ldots, \widetilde{y}_k^m\right)^T$, $H_k = \left(H_k^{1T}, \ldots, H_k^{mT}\right)^T$, $v_k = \left(v_k^1, \ldots, v_k^m\right)^T$, $\Gamma_k = Diag\left(\gamma_k^1, \ldots, \gamma_k^m\right)$ and I_m is the $m \times m$ identity matrix.

To address the LS linear estimation problem of the signal based on the randomly delayed observations (2), the following hypotheses are assumed:

(H.1) $\{z_k;\ k \geq 1\}$ has zero mean and factorizable covariance function $K^z_{k,s} = E[z_k z_s^T] = A_k B_s^T,\ s \leq k$, with A_k and B_s known $n \times M$ matrices.

(H.2) The noise, $\{v_k;\ k \geq 1\}$, is a zero-mean white sequence with known covariances $Cov[v_k] = R_k,\ \forall k \geq 1$.

(H.3) For $i = 1, \ldots, m$, the noises $\left\{\gamma_k^i;\ k > 1\right\}$ are sequences of Bernoulli variables with $P\left[\gamma_k^i = 1\right] = p_k^i$. For $i, j = 1, \ldots, m$ the variables γ_k^i and γ_s^j are independent for $|k - s| \geq 2$, and $\mathrm{Cov}[\gamma_k^i, \gamma_{k-1}^j]$ are known.

(H.4) The signal process, $\{z_k;\ k \geq 1\}$, and the noises, $\{\gamma_k;\ k > 1\}$ and $\{v_k;\ k \geq 1\}$, where $\gamma_k = (\gamma_k^1, \ldots, \gamma_k^m)^T$, are mutually independent.

Clearly from (H.3), the following properties hold:

- The mean of the random matrix Γ_k is $\Gamma_k^p = Diag\left(p_k^1, \ldots, p_k^m\right)$ and, for any random matrix $G_{m \times m}$ independent of $\{\Gamma_k,\ k \geq 1\}$, it is satisfied $E[\Gamma_k G_{m \times m} \Gamma_s] = E[\gamma_k \gamma_s^T] \circ E[G_{m \times m}]$ ($\circ$ denotes the Hadamard product).

- The random vectors γ_k and γ_s are independent for $|k - s| \geq 2$ and the covariance matrices of γ_k and γ_s for $s = k, k-1$, which will be denoted by $K^{\gamma}_{k,s}$, are known.

- The mean of the random vector γ_k is $E\left[\gamma_k\right] = p_k = \left(p_k^1, \ldots, p_k^m\right)^T$ and the correlation functions of γ_k and $\mathbf{1} - \gamma_k$ ($\mathbf{1} = (1, \ldots, 1)^T$ is the $m \times 1$ ones vector) are denoted by

$$E[\gamma_k \gamma_k^T] = P_k^p,\ E[(\mathbf{1} - \gamma_k)(\mathbf{1} - \gamma_k)^T] = P_k^{\mathbf{1}-p},\ E[\gamma_k (\mathbf{1} - \gamma_k)^T] = P_k^{p,\mathbf{1}-p}.$$

3 Linear filtering algorithm

Using an innovation approach and the Orthogonal Projection Lemma, the following recursive algorithm, for the LS linear filter of the signal z_k based on the randomly delayed observations $\{y_1, \ldots, y_k\}$ given in (3), is derived.

The linear filter, $\widehat{z}_{k/k}$, of the signal z_k is obtained as

$$\widehat{z}_{k/k} = A_k O_k, \quad k \geq 1,$$

where the vectors O_k are recursively calculated from

$$O_k = O_{k-1} + J_k \Pi_k^{-1} \nu_k, \quad k \geq 1; \quad O_0 = 0,$$

and the matrix J_k is given by

$$J_k = G_{B_k}^T - r_{k-1} G_{A_k}^T - J_{k-1} \Xi_k^T, \quad k \geq 2; \quad J_1 = B_1^T H_1^T,$$

with $r_k = E[O_k O_k^T]$ recursively obtained from

$$r_k = r_{k-1} + J_k \Pi_k^{-1} J_k^T, \quad k \geq 1; \quad r_0 = 0.$$

The innovation, ν_k, satisfies

$$\nu_k = y_k - G_{A_k} O_{k-1} - \Xi_k \nu_{k-1}, \quad k \geq 2; \quad \nu_1 = y_1.$$

and Π_k, the innovation covariance matrix, is given by

$$\begin{aligned}\Pi_k &= P_k^{\mathbf{1}-p} \circ \left[H_k A_k B_k^T H_k^T + R_k\right] + P_k^p \circ \left[H_{k-1} A_{k-1} B_{k-1}^T H_{k-1}^T + R_{k-1}\right] \\ &\quad + P_k^{\mathbf{1}-p,p} \circ \left[H_k A_k B_{k-1}^T H_{k-1}^T\right] + P_k^{p,\mathbf{1}-p} \circ \left[H_{k-1} B_{k-1} A_k^T H_k^T\right] \\ &\quad - G_{A_k} r_{k-1} G_{A_k}^T - \Xi_k \Pi_k \Xi_k^T - G_{A_k} J_{k-1} \Xi_k^T - \Xi_k J_{k-1}^T G_{A_k}^T, \qquad k \geq 2, \\ \Pi_1 &= H_1 A_1 B_1^T H_1^T + R_1.\end{aligned}$$

The matrices G_{A_k}, G_{B_k} and Ξ_k are given by

$$G_{\Psi_k} = (I - \Gamma_k^p) H_k \Psi_k + \Gamma_k^p H_{k-1} \Psi_{k-1}, \qquad \Psi = A,\ B,$$

$$\begin{aligned}\Xi_k &= \Big[K_{k,k-1}^{\gamma} \circ \left((H_k A_k - H_{k-1} A_{k-1})(H_{k-1} B_{k-1} - H_{k-2} B_{k-2})^T - R_{k-1}\right) \\ &\quad + \left(p_k (\mathbf{1} - p_{k-1})^T\right) \circ R_{k-1}\Big] \Pi_{k-1}^{-1}, \ k > 2, \\ \Xi_2 &= \Gamma_2^p R_1 \Pi_1^{-1}.\end{aligned}$$

The accuracy of the LS linear filter is measured by the filtering error covariance matrices $\Sigma_{k/k} = E\left[(z_k - \widehat{z}_{k/k})(z_k - \widehat{z}_{k/k})^T\right]$ which, using hypothesis (H.1) and the filter expression, are given by

$$\Sigma_{k/k} = A_k \left[B_k^T - r_k A_k^T\right], \quad k \geq 1.$$

Acknowledgments: This research is supported by Ministerio de Educación y Ciencia (grant No. MTM2008-05567) and Junta de Andalucía (grant No. P07-FQM-02701).

References

Caballero-Águila, R., Hermoso-Carazo, A., Jiménez-López, J.D., Linares-Pérez, J., and Nakamori, S. (2010). Signal estimation with multiple delayed sensors using covariance information. *Digital Signal Processing*, **20**, 528-540.

Hounkpevi, F.O., and Yaz, E.E. (2007). Minimum variance generalized state estimators for multiple sensors with different delay rates. *Signal Processing* , **87**, 602-613.

Nakamori, S., Caballero-Águila, R., Hermoso-Carazo, A., and Linares-Pérez, J. (2005). Recursive estimators of signals from measurements with stochastic delays using covariance information. *Applied Mathematics and Computation*, **162**, 65-79.

Su, C.L., and Lu, C.N. (2001). Interconnected network state estimation using randomly delayed measurements. *IEEE Transactions on Power Systems*, **16**, 870-878.

Filtering algorithm for fractional order discrete systems with uncertain observations

R. Caballero-Águila[1], A. Hermoso-Carazo[2] , J. Linares-Pérez[2]

[1] Dpto. de Estadística e I.O., Universidad de Jaén, 23071 Jaén, Spain (raguila@ujaen.es)
[2] Dpto. de Estadística e I.O., Universidad de Granada, 18071 Granada, Spain (ahermoso@ugr.es, jlinares@ugr.es)

Abstract: This paper considers the least-squares linear estimation problem in linear fractional order discrete state-space systems with uncertain observations. An extension of the fractional Kalman filter is obtained for this class of systems whose observations may not contain the signal, and this uncertainty is described by introducing independent Bernoulli variables in the observation model.

Keywords: Least-squares estimation; Discrete fractional state-space systems; Uncertain observations; Fractional Kalman filter.

1 Introduction

The least-squares estimation problem of stochastic signals from noisy observations has been widely treated when the observation sequence contains the signal to be estimated with probability one. Although the Kalman filter has played an important role because of its wide applicability in many fields, for fractional order models the Kalman filter is not directly applicable and new estimation algorithms are needed. In Sierociuk and Dzieliński (2006) and Sierociuk et al. (2011) generalizations of the Kalman filter for linear fractional order discrete state-space systems, called *fractional Kalman filter* and *improved fractional Kalman filter*, respectively, are proposed.
On the other hand, in many practical situations, the signal vector enters in the observation equation randomly. This can occurs, for example, in problems where there exist intermittent failures in the observation device, fading phenomena in propagation channels, target tracking, accidental loss of some measurements, or inaccessibility of the data during certain times; that is, problems where, due to different reasons, the transmitted data packet can contain observations which are only noise. These situations are described by an observation equation which includes not only an additive noise, but also a multiplicative noise component, modelled by a sequence of Bernoulli random variables whose values, one or zero, indicate the presence or absence of the signal in the observation. The state estimation problem

from uncertain observations has been widely studied in linear systems under different hypotheses on the variables describing the uncertainty (see e.g. Hermoso-Carazo et al. (2008), Caballero-Águila et al. (2011) and references therein); our aim in this paper is to obtain a filtering algorithm from uncertain observations, when the state evolution is described by a linear fractional order discrete equation and the uncertainty is modelled by independent variables.

2 System description

A linear fractional order discrete state equation (Sierociuk et al. (2011)) is given by

$$\begin{aligned} &\Delta^{\alpha} x_{k+1} = Ax_k + Bu_k + w_k, \quad k \geq 0 \\ &x_{k+1} = \Delta^{\alpha} x_{k+1} - \sum_{j=1}^{k+1} (-1)^j \begin{pmatrix} \alpha \\ j \end{pmatrix} x_{k+1-j}, \quad k \geq 0 \end{aligned}$$

where x_k is the n-dimensional state vector, u_k is a d-dimensional system input, w_k represents the system noise and A, B are known matrices of appropriate dimensions.

We assume that the equation orders are not identical, and the following generalized definition is considered:

$$\begin{aligned} &\Delta^{\Upsilon} x_{k+1} = Ax_k + Bu_k + w_k, \quad k \geq 0 \\ &x_{k+1} = \Delta^{\Upsilon} x_{k+1} - \sum_{j=1}^{k+1} (-1)^j \Upsilon_j x_{k+1-j}, \quad k \geq 0 \end{aligned} \tag{1}$$

with

$$\Upsilon_j = diag\left[\begin{pmatrix} \alpha_1 \\ j \end{pmatrix} \cdots \begin{pmatrix} \alpha_n \\ j \end{pmatrix} \right], \qquad \Delta^{\Upsilon} x_{k+1} = \begin{bmatrix} \Delta^{\alpha_1} x_{1,k+1} \\ \vdots \\ \Delta^{\alpha_n} x_{n,k+1} \end{bmatrix}$$

where $\alpha_1, \ldots, \alpha_n$ are the system equation orders.

The aim of this paper is to determine the least-squares (LS) linear estimator of x_k from noisy measurements which may not contain the signal with different probabilities. Specifically, assume that the observation at each sampling time, k, denoted by y_k, may either contain the state to be estimated, x_k, or be only noise, v_k; this uncertainty about the state being present or missing in the observation is modelled by Bernoulli random variables, γ_k. The observation model is thus described as follows:

$$y_k = \gamma_k H_k x_k + v_k, \quad k \geq 1. \tag{2}$$

If $\gamma_k = 1$, then $y_k = H_k x_k + v_k$ and the measurement contains the signal; otherwise, $\gamma_k = 0$ and $y_k = v_k$, which means that such measurement is only noise.

To address the LS linear estimation problem of the state (1) based on the observations (2), the following hypotheses are assumed:

- The initial state, x_0, is a zero-mean random vector with known covariance matrix $Cov[x_0] = P_0$.
- The noises, $\{w_k;\ k \geq 0\}$ and $\{v_k;\ k \geq 1\}$, are zero-mean white sequences with known covariance matrices $Cov[w_k] = Q_k$, $Cov[v_k] = R_k,\ \forall k$.
- The multiplicative noise $\{\gamma_k;\ k \geq 1\}$, which describes the uncertainty in the observations, is a sequence of independent Bernoulli random variables with $P[\gamma_k = 1] = p_k$.
- The initial state, x_0, and the noises, $\{w_k;\ k \geq 0\}$, $\{\gamma_k;\ k \geq 1\}$ and $\{v_k;\ k \geq 1\}$ are mutually independent.

3 Linear filtering algorithm

The LS estimator of the state x_k given the observations $Y^k = \{y_1, \ldots, y_k\}$ is $E[x_k/Y^k] = \int x_k g(x_k/Y^k)dx_k$ and, hence, its determination requires knowledge of the conditional density $g(x_k/Y^k)$. The uncertainty in the observations produces that the density of each observation y_j is a mixture, or weighted sum, of two densities (corresponding to $\gamma_j = 0$ and $\gamma_j = 1$) and hence, the computation of the conditional density $g(x_k/Y^k)$, mixture of 2^k densities, requires an exponentially growing memory. For this reason, the estimation problem in systems with uncertain observations has usually been focused on searching for suboptimal, basically linear, estimators.
In this paper, we propose a filtering algorithm that, as usual in the LS filtering problem, performs in two steps: first, approximations of the mean and covariance of the state x_k given the observations Y^{k-1} ($\widehat{x}_{k/k-1}$ and $P_{k/k-1}$, respectively) are obtained and, from them, the conditional mean and covariance given Y^k are approximated by the following expressions, with a similar structure to those of the Kalman filter:

$$E[x_k/Y^k] \simeq \widehat{x}_{k/k} = \widehat{x}_{k/k-1} + P^{x\nu}_{k/k-1}\Pi^{-1}_{k/k-1}\nu_{k/k-1},\ k \geq 1; \quad \widehat{x}_{0/0} = 0,$$
$$Cov[x_k/Y^k] \simeq P_{k/k} = P_{k/k-1} - P^{x\nu}_{k/k-1}\Pi^{-1}_{k/k-1}P^{\nu x}_{k/k-1},\ k \geq 1;\ P_{0/0} = P_0.$$

In these expressions, $\nu_{k/k-1}$ denotes the innovation at time k (difference between the new measurement, y_k, and its prediction from the previous ones), $\Pi_{k/k-1}$ is the conditional covariance matrix of $\nu_{k/k-1}$ given Y^{k-1}, and $P^{x\nu}_{k/k-1}$ denotes the conditional cross-covariance matrix of x_k and $\nu_{k/k-1}$.

One-stage state predictor. Since the observation model is not used in the prediction step, the prediction estimates and error covariance matrices are approximated by the following expressions proposed in Sierociuk and Dzieliński (2006) for fractional system estimation when there is no uncertainty in the observations:

$$\widehat{x}_{k/k-1} = A\widehat{x}_{k-1/k-1} + Bu_{k-1} - \sum_{j=1}^{k}(-1)^j \Upsilon_j \widehat{x}_{k-j/k-j}, \quad k \geq 1.$$

$$P_{k/k-1} = (A+\Upsilon_1)P_{k-1/k-1}(A+\Upsilon_1)^T + Q_{k-1} + \sum_{j=2}^{k} \Upsilon_j P_{k-j/k-j} \Upsilon_j^T, \quad k \geq 2,$$

$$P_{1/0} = (A+\Upsilon_1)P_{0/0}(A+\Upsilon_1)^T + Q_0.$$

Innovation $\nu_{k/k-1}$ and covariance matrix $\Pi_{k/k-1}$. From (2), using the model hypotheses and the conditional expectation properties, we have that

$$\nu_{k/k-1} = y_k - p_k H_k \widehat{x}_{k/k-1}, \quad k \geq 1,$$

and, rewriting $\nu_{k/k-1} = (\gamma_k - p_k)H_k x_k + v_k + p_k H_k (x_k - \widehat{x}_{k/k-1})$, we obtain

$$\Pi_{k/k-1} = p_k(1-p_k)H_k \widehat{x}_{k/k-1}\widehat{x}_{k/k-1}^T H_k^T + p_k H_k P_{k/k-1} H_k^T + R_k, \quad k \geq 1.$$

Conditional cross-covariance $P_{k/k-1}^{x\nu}$. Using the above innovation expression, we have

$$P_{k/k-1}^{x\nu} = p_k P_{k/k-1} H_k^T, \quad k \geq 1.$$

Acknowledgments: This research is supported by Ministerio de Educación y Ciencia (grant No. MTM2008-05567) and Junta de Andalucía (grant No. P07-FQM-02701).

References

Caballero-Águila, R., Hermoso-Carazo, A. and Linares-Pérez, J. (2011). Linear and quadratic estimation using uncertain observations from multiple sensors with correlated uncertainty. *Signal Processing*, **91**, 330-337.

Hermoso-Carazo, A., Linares-Pérez, J., Jiménez-López, J.D., Caballero-Águila, R. and Nakamori, S. (2008). Recursive fixed-point smoothing algorithm from covariances based on uncertain observations with correlation in the uncertainty. *Applied Mathematics and Computation*, **203**, 243-251.

Sierociuk, D. and Dzieliński, A. (2006). Fractional Kalman filter algorithm for the states, parameters and order of fractional system estimation. *International Journal of Applied Mathematics and Computer Science*, **16**, 129-140.

Sierociuk, D., Tejado, I. and Vinagre B. M. (2011). Improved fractional Kalman filter and its application to estimation over lossy networks. *Signal Processing*, **91**, 542-552.

The Log-Generalized Modified Weibull Regression Model

Jalmar M. F. Carrasco[1], Edwin M. M. Ortega[1], Gauss M. Cordeiro[2]

[1] University of São Paulo
[2] Federal Rural University of Pernambuco

Abstract: We introduce the log-generalized modified Weibull regression model based on the modified Weibull distribution (Carrasco et al., 2008a). This distribution can accommodate increasing, decreasing, bathtub and unimodal shaped hazard functions. Other advantage is that it includes classical distributions reported in lifetime literature as special cases. We obtain maximum likelihood estimates for the model parameters by considering censored data and evaluate local influence on the estimates of the parameters by taking different perturbation schemes. In addition, we define martingale and deviance residuals to detect outliers and evaluate the model assumptions. We demonstrate that our extended regression model is very useful to the analysis of real data and may give more realistic fits than other special regression models.

Keywords: Generalized modified Weibull distribution; Log-Weibull regression; Residual analysis; Sensitivity analysis; Survival function.

1 The Model

Most generalized Weibull distributions have been proposed in reliability literature to provide a better fitting of certain data sets than the traditional two and three parameter Weibull models. The GMW distribution with four parameters $\alpha > 0$, $\gamma \geq 0$, $\lambda \geq 0$ and $\varphi > 0$, introduced by Carrasco et al. (2008a), extends the MW distribution (Lai et al., 2003) and should be able to fit various types of data. Its density function for $t > 0$ is given by

$$f(t) = \frac{\alpha\varphi\big(\gamma + \lambda t\big)t^{\gamma-1}\exp\big[\lambda t - \alpha t^{\gamma}\exp(\lambda t)\big]}{\big\{1 - \exp\big[-\alpha t^{\gamma}\exp(\lambda t)\big]\big\}^{1-\varphi}}. \tag{1}$$

Henceforth, T is a random variable following the GMW density function (1) and Y is defined by $Y = \log(T)$. It is easy to verify that the density function of Y obtained by replacing $\gamma = 1/\sigma$ and $\alpha = \exp(-\mu/\sigma)$ reduces to

$$f(y) \quad = \quad \varphi\big[\sigma^{-1} + \lambda\exp(y)\big]\exp\left\{\Big(\frac{y-\mu}{\sigma}\Big) + \lambda\exp(y) - \exp\Big[\Big(\frac{y-\mu}{\sigma}\Big) +\right.$$

$$\lambda \exp(y)\Big]\Big\}\Big\{1-\exp\Big[-\exp\Big\{\Big(\frac{y-\mu}{\sigma}\Big)+\lambda\exp(y)\Big\}\Big]\Big\}^{\varphi-1}, \quad (2)$$

where $-\infty < y, \mu < \infty$, $\sigma > 0$, $\lambda \geq 0$ and $\varphi > 0$. We refer to equation (2) as the LGMW distribution, say $Y \sim \text{LGMW}(\lambda, \varphi, \sigma, \mu)$, where $\mu \in \Re$ is the location parameter, $\sigma > 0$ is the scale parameter and λ and φ are shape parameters. The random variable $Z = (Y - \mu)/\sigma$ has density function

$$f(z) = \varphi\sigma(\sigma^{-1} + v)\exp\big[\omega - \exp(\omega)\big]\big\{1 - \exp\big[-\exp(\omega)\big]\big\}^{\varphi-1}, \quad (3)$$

where $\omega = v + z$ and $v = \lambda\exp(\mu + \sigma z)$.

In many practical applications, the lifetimes are affected by explanatory variables such as the cholesterol level, blood pressure, weight and many others. Based on the log-generalized modified Weibull (LGMW) density, we propose a linear location-scale regression model linking the response variable y_i and the explanatory variable vector $\mathbf{x}_i^\top = (x_{i1}, \ldots, x_{ip})$ as follows $y_i = \mathbf{x}_i^\top\boldsymbol{\beta} + \sigma z_i$, $i = 1, \ldots, n$, where z_i the random error, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\sigma > 0$, $\lambda \geq 0$ and $\varphi > 0$ are unknown parameters. The parameter $\mu_i = \mathbf{x}_i^\top\boldsymbol{\beta}$ is the location of y_i. The location parameter vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ is represented by a linear model $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is a known model matrix.

Consider a sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ of n independent observations, where each random response is defined by $y_i = \min\{\log(t_i), \log(c_i)\}$. We assume non-informative censoring such that the observed lifetimes and censoring times are independent. Let F and C be the sets of individuals for which y_i is the log-lifetime or log-censoring, respectively. Conventional likelihood estimation techniques can be applied here. The log-likelihood function for the vector of parameters $\boldsymbol{\theta} = (\lambda, \varphi, \sigma, \boldsymbol{\beta}^\top)^\top$ has the form $l(\boldsymbol{\theta}) = \sum_{i\in F} l_i(\boldsymbol{\theta}) + \sum_{i\in C} l_i^{(c)}(\boldsymbol{\theta})$, where $l_i(\boldsymbol{\theta}) = \log[f(y_i)]$, $l_i^{(c)}(\boldsymbol{\theta}) = \log[S(y_i)]$, $f(y_i)$ is the density and $S(y_i)$ is survival function of the generalized modified Weibull of Y_i. The total log-likelihood function for $\boldsymbol{\theta}$ reduces to

$$l(\boldsymbol{\theta}) = \sum_{i\in F} l_1(\lambda, \varphi, z_i, u_i) + \sum_{i\in C} l_2(\lambda, \varphi, z_i, u_i), \quad (4)$$

where

$$\begin{aligned} l_1(\lambda, \varphi, z_i, u_i) &= \log\big[\varphi(\sigma^{-1} + u_i)\big] + \big[z_i + u_i - \exp(z_i + u_i)\big] + \\ &\quad (\varphi - 1)\log\big\{1 - \exp\big[-\exp(z_i + u_i)\big]\big\}, \\ l_2(\lambda, \varphi, z_i, u_i) &= \log\Big\{1 - \big[1 - \exp\big\{-\exp(z_i + u_i)\big\}\big]^\varphi\Big\}, \end{aligned}$$

$u_i = \lambda\exp(\sigma z_i + \mathbf{x}_i^\top\boldsymbol{\beta})$, $z_i = (y_i - \mathbf{x}_i^\top\boldsymbol{\beta})/\sigma$ and r is the number of uncensored observations (failures). The maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\theta}}$ of the vector of unknown parameters can be calculated by maximizing the log-likelihood (4). We can use the likelihood ratio (LR) statistic for comparing some special sub-models with the LGMW model.

2 Application

Survival times for the Golden shiner data, Notemigonus crysoleucas, were obtained from field experiments conducted in Lake Saint Pierre, Quebec, in 2005 (Laplante-Albert, 2008). Each individual fish was attached by means of a monofilament chord to a chronographic tethering device that allowed the fish to swim in midwater. A timer in the device was set off when the tethered fish was captured by a predator. The device was retrieved approximately 24 hours after the onset of the experiment and survival time was then obtained from the difference: time elapsed between onset of the experiment and retrieval-time elapsed in device timer since predation event. The Golden shiner data have been analyzed by Carrasco et al. (2008b) using the LMW regression model. We now reanalyzed these data using the LGMW regression model. First, we consider the equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \sigma z_i, \qquad (5)$$

where the random variable y_i has the LGMW distribution. The MLEs (p-values in parentheses) are: $\hat{\lambda} = 0.001$, $\hat{\varphi} = 12.855$, $\hat{\sigma} = 5.086$, $\hat{\beta}_0 = -1.894(0.748)$, $\hat{\beta}_1 = 2.197(0.001)$, $\hat{\beta}_2 = 0.097(0.008)$, $\hat{\beta}_3 = -0.125(0.001)$, $\hat{\beta}_4 = 0.035(0.001)$, $\hat{\beta}_5 = 0.022(0.202)$ and $\hat{\beta}_6 = 0.222(0.278)$. Further, we calculate the maximum unrestricted and restricted log-likelihoods and the LR statistics for testing some sub-models. An analysis under the LGMW regression model provides a check on the appropriateness of the LW, LEW and LMW sub-models and indicates the extent for which inferences depend upon the model. For example, the LR statistic for testing the hypotheses H_0:$\varphi = 1$ versus H_1:H_0 *is not true*, i.e. to compare the LMW and LGMW regression models, is $w = 2\{-201.142 - (-204.577)\} = 6.87$ (p-value < 0.05) which yields favorable indications toward to the LGMW regression model. A summary of the values of the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Consistent Akaike Information Criterion (CAIC) to compare the LGMW and LMW regression models is given in Table 1. The LGMW regression model outperforms the LMW model irrespective of the criteria and can be used effectively in the analysis of these data. The explanatory variables x_1, x_2, x_3 and x_4 are marginally significant for the LGMW model at the significance level of 5%.

TABLE 1. Statistics AIC, BIC and CAIC for comparing the LGMW and LMW models.

Model	AIC	BIC	CAIC
LGMW	422.3	424.6	448.9
LWM	427.2	429.0	451.1

3 Concluding Remarks

We introduce the so-called log-generalized modified Weibull (LGMW) distribution whose hazard rate function accommodates four types of shape forms, namely increasing, decreasing, bathtub and unimodal. We derive an expansion for its moments. Based on this new distribution, we propose a LGMW regression model very suitable for modeling censored and uncensored lifetime data. The new regression model permits testing the goodness of fit of some known regression models as special sub-models. Hence, the proposed regression model serves as a good alternative for lifetime data analysis. Further, the new regression model is much more flexible than the exponentiated Weibull, modified Weibull and generalized Rayleigh sub-models. We use the matrix programming language Ox (MaxBFGS function) to obtain the maximum likelihood estimates and perform asymptotic tests for the parameters based on the asymptotic distribution of these estimates. We examine a simulation study. We discuss influence diagnostics and model checking analysis in the LGMW regression models fitted to censored data. We also discuss the sensitivity of the maximum likelihood estimates from the fitted model via deviance component residuals and sensitivity analysis. We demonstrate in one application to real data that the LGMW model can produce better fit than its sub-models.

Acknowledgments: Special Thanks to CNPq and CAPES.

References

Carrasco, J. M. F., Ortega, E. M. M. and Cordeiro, M. G. (2008a). A generalized modified Weibull distribution for lifetime modeling. Computational Statistics and Data Analysis, **53**, 450-462.

Carrasco, J. M. F., Ortega, E. M. M. and Paula, G. A. (2008b). Log-Modified Weibull Regression Models with Censored Data: Sensitivity and Residual Analysis. Computational Statistics and Data Analysis, **52**, 4021-4029.

Doornik, J. A. (2007). *An Object-Oriented Matrix Language Ox 5*. Timberlake Consultants Press: London.

Lai, C. D., Xie, M. and Murthy, D. N. P. (2003). A modified Weibull distribution. *IEEE Transactions on Reliability,* **52**, 33-37.

Laplante-Albert, K.A., 2008. Habitat-dependent mortality risk in lacustrine fish. M.Sc. thesis. Université du Québec à Trois-Rivières, Canada.

An exponential dispersion family to modelling critical phenomenon

Joan del Castillo[1] , Isabel Serra[1]

[1] Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain (castillo@mat.uab.cat, iserra@mat.uab.cat).

Abstract: We propose to model the power dissipation index of tropical cyclones by a truncated gamma distribution, using maximum likelihood estimation (MLE). This significantly improves the fit obtained with the Pareto distribution. Numerical procedures for this three parameter gamma model often show instabilities that are solved with new parameters in the framework of exponential dispersion models.

Keywords: Exponential models, Domain of the means, Critical phenomenon, Extreme value theory.

1 Introduction

Often, we observe a disaster when it becomes important, for example, the proximity of an important city. That is a main reason to use the Pareto distribution to modelling in this scope. However, for some cases, the model have no heavy tails; that is the main motivation to consider the gamma distributions. On the other hand, the possible lack of data for low values of the measure leads to work with truncated models, where to compute MLE becomes more difficult. We propose to consider gamma truncated models with a new parameter of truncation, we provide an algorithm for determining MLE and we show the improvement in the fit. Remark, that all the probability densities are on $(0, \infty)$. Finally, we apply it to Tropical Cyclone data in the North Atlantic occurred between 1966 and 2009. To measure the importance, we consider the power dissipation index (PDI) from the work of Corral, A. *et al.* (2010).

2 Exponential dispersion models

Let X be a continuous non negative variable, for any threshold, $u > 0$, the threshold excedances are the values of $(X - u)$ conditional to $X > u$, $X_u = (X - u \mid X > u)$. If X has distribution function $F(x)$ and density function $f(x)$ the density function of X_u is

$$f_u(x) = f(x + u)/(1 - F(u)). \tag{1}$$

Let $\mathcal{P}$ be an exponential model, generated by the *Lebesgue* measure on $[0, \infty)$, with canonical statistic $T(x)$. The model $\mathcal{P}$ corresponds to the set of densities

$$\exp(\theta \cdot T(x)) / C(\theta) \tag{2}$$

for each $\theta \in D$, where D denotes the set such that the Laplace transform

$$C(\theta) = \int_0^\infty \exp(\theta \cdot T(x))\, dx. \tag{3}$$

converges. D is called the *natural domain of parameters*. If D is an open set, then the likelihood equations have one and only one solution provided the observation is in the domain of the means, Barndorff-Nielsen (1978). The *domain of the means* is the image of the interior of D by the map

$$\theta \mapsto \nabla k(\theta)$$

where k is defined by $\log C(\theta)$. Given a sample $\mathbf{x} = \{x_1, \cdots, x_n\}$ such that the sample value of the statistic T, $t(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n T(x_i)$, is in the interior of the domain of the means, then the likelihood estimator for the sample is in the interior of the natural domain of parameters.
In the same way as in (1), given $u > 0$ be a fixed threshold, then we can consider a new exponential model with statistic $T(x + u)$, or equivalently $T(1+x/u)$. Therefore, we can extend the model to an exponential dispersion model with statistic $T(1 + x)$. This is a closed model by truncation and scale.

2.1 Maximum Likelihood for exponential dispersion model

Let $\mathcal{P}$ be a model as described above, the exponential model associated to $\mathcal{P}$ is the set of densities of the form $f(x; \theta, \sigma) = \exp(\theta \cdot T(x/\sigma) - k(\theta))/\sigma$. Given a sample $\mathbf{x} = \{x_i\}$ of size n, the log-likelihood function is

$$l(\mathbf{x}; \theta, \sigma) = \theta \cdot t(\mathbf{x}/\sigma) - k(\theta) - \log(\sigma)).$$

We take the derivatives in respect θ and σ, and we can describe the likelihood equations by

$$t(\mathbf{x}/\sigma) - \nabla k(\theta) = 0 \tag{4}$$

$$\theta \cdot \nabla t(\mathbf{x}/\sigma) \cdot \mathbf{x}/\sigma - 1 = 0 \tag{5}$$

Define $\psi = (\nabla k)^{-1}$ and use (4) to simplify (5) to get

$$\psi(t(\mathbf{x}/\sigma)) \cdot \nabla t(\mathbf{x}/\sigma) \cdot \mathbf{x}/\sigma = 1 \tag{6}$$

This is the equation in σ which has to be solved.

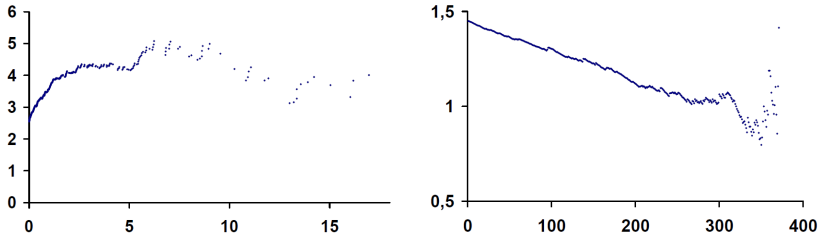


FIGURE 1. In the left, the ME-plot shows increasing line tendence for a small thresholds. In the right, the CV-plot shows the tendence to 1 for near behavior samples.

2.2 The Full Truncated Gamma model

The truncated gamma distribution is a three-parameter model of continuous probability distributions with support on $(0, \infty)$ which probability density function is given by

$$f(x; \alpha, \sigma, \rho) = \rho^{\alpha} (1 + x/\sigma)^{\alpha - 1} \exp(-\rho (1 + x/\sigma)) / (\sigma \Gamma(\alpha, \rho)) \quad (7)$$

where $\Gamma(\alpha, \rho)$ is the incomplete gamma function. Indeed, this corresponds to the dispersion exponential model associate to $T(x) = (x, \log(1+x))$. Remark, this model contains more distributions in addition to the truncated gamma, for this reason we call *full truncated gamma* (FTG). In particular, for $\rho = 0$ is the Pareto model.
To compute the MLE we can use the algorithm: solve the equation (6) to determine σ and, in the case that σ gives us a value of the statistic t in the interior of the domain of the means, use it to solve (4). Therefore, we have to determine the domain of the means. From Castillo, J. *et al.*, we can prove that the domain of the means for the exponential model with statistic T is

$$\{(x, y) \; ; \; x > 0, \log(1 + x) > y > x/(1 + x)\}.$$

3 Tropical Cyclones

The measure considered to fit the Tropical Cyclones described on the introduction is the PDI. This is defined as $\sum_t v_t^3 \Delta t$, where t denotes time and runs over the entire lifetime of the storm and v_t is the maximum sustained surface wind velocity at time t. The unit that we will use is $10^{10} m^3/s^2$. From the evidence of lack of data for low values, we are going to consider the TC with PDI bigger than 0.3, that is a sample of size 372 (75% of the original data). We can see in Figure 1 that for differen thresholds don't get stability by the tail index of Pareto model. In fact, the tail is liked to an exponential tail.

TABLE 1. Likelihood ratio test

	lv FTG	α	σ	ρ	lv Pareto	α	σ	p-value
MLE	-667.58	0.28	0.09	0.02	-680.06	-1.63	2.01	5.8e-7
s.e.		0.15	0.11	0.02		0.22	0.41	

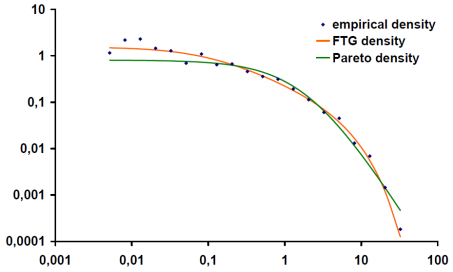


FIGURE 2. The fit of the empirical density using Pareto and FTG model in logarithm scale for both axes.

The solution of (5) is $\sigma = 0.09$, then the value of t is $(28.34, 2.55)$ and it is in the interior of the domain of the means. Therefore, the MLE is in the interior of the domain of parameters. Using the likelihood ratio test we can conclude that the difference with the Pareto model is significative, it is reject with p-value $5.8\,10^{-7}$, we refer to Table 1. In Figure 2 we show the fit using the methodology of Corral, *et al.* (2010). In fact, in both cases we obtain goodness of fit, but the FTG model gives us stability in the tail.

Acknowledgments: We thank A. Corral for generously sharing their data sets and spending his time with us. We thank to G. Letac that shared with us his vast culture.

References

Barndorff-Nielsen (1978). *Information and exponential families in statistical theory.* Wiley, New York.

Castillo, J. del and Puig, P. (1999). Invariant Exponential Models Applied to Reliability Theory and Survival Analysis. *JASA*, 94, 522-528.

Corral, A., Osso, A. and Llebot, J.E. (2010). Scaling of tropical-cyclone dissipation. *Nature Physics*, 6, 693-696.

Hierarchical Bayesian modelling to assess divergence in disease mapping

Dolores Catelan[12], Annibale Biggeri[12]

[1] Department of Statistics "G. Parenti", Viale Morgagni 59, 501234, Florence
[2] Biostatistics Unit, ISPO Cancer Prevention and Research Institute,Via Cosimo il Vecchio 2, Florence

Abstract: In Disease mapping several areas, often spatially close each other, can have relative risk of a given disease that diverges from the reference. Use of cross-validation posterior predictive distributions to detect outlying observation has been suggested. We propose a hierarchical modelling approach to the problem and show how to specify a full range of informative "null" priors.

Keywords: Cross-validation predictive distributions; hierarchical Bayesian model; Disease Mapping.

Corresponding author: Dolores Catelan, email: catelan@ds.unifi.it

1 Introduction

Disease mapping focused on relative risk surface estimation. This explain why great emphasis was put on spatial patterns. The main goal was on investigating the geographical distribution of the risk. Since the seminal paper of Clayton and Kaldor (1987) spatially-structured priors were considered in almost all the proposed models in the literature. The Besag, York and Mollié model (1991) is a benchmark because it combines spatially structured and un-structured random effects, gaining in flexibility. However, inference on area-specific relative risks received little attention in the literature despite of the need to identify areas (or regions) at unusual (high or low) risk. Stern and Cressie (2000) used cross-validation posterior predictive distributions to explore model fitting and identify outlying areas in disease mapping. The idea of cross-validation is to re-fit the model removing one observation in turn. The model is thus fitted to a subset of data Y_{-i} from which the i-th observation is dropped. The posterior predictive distribution $P(Y_i^{rep}|Y_{-i})$ for a replicate (Y_i^{rep}) of the i-th observation conditional to the remaining data Y_{-i} is then used to evaluation purposes. The extremeness is usually measured by some summaries over $P(Y_i^{rep}|Y_{-i})$, for example the posterior predicted p-values, $P(Y_i^{rep} \leq y_i|Y_{-i})$, or the conditional predictive ordinate, $p(Y_i^{rep} = y_i|Y_{-i})$. Marshall and Spiegelhalter (2003) noted that "...There are essentially two reasons why observations/regions may

be divergent. First, the statistical assumptions underlying the model may be incorrect...[second], these regions could represent genuine 'hot-spots' of disease requiring further investigation." Poor model fit is a reasonable explanation when a relevant number of observations/areas are identified as divergent while the presence of real hot-spots or outliers is the usual interpretation of few divergent ones. Stern and Cressie (2000) did not fully exploited the potentiality of this approach: they stay essentially on model checking. This is because the posterior predictive distribution was obtained under the alternative hypothesis and any discrepancy detected was naturally interpreted as a symptom of lack of model fit. Their approach is computationally intensive and time consuming. Marshall and Spiegelhalter (2007) proposed a mixed approach to perform cross-validatory checks in disease mapping which can be easily implemented in WinBugs (Lunn et al., 2000).
The aim of this work is to develop a hierarchical model to detect divergent areas under hierarchical null models in the context of disease mapping. This is pursued by specifying appropriate hyperpriors and obtaining cross-validation posterior predictive distributions (and related quantities like posterior p-values). We take advantage of a the real example on the distribution of Lung cancer in Tuscany.

2 Motivating example

Lung cancer death certificates were considered for males resident in the 287 municipalities of the Tuscany Region (Italy) for the period 1995-1999. Data were made available by the Regional Mortality Register. A set of reference rates (Tuscany, 1971-1999) have been used to compute the expected number of cases for each municipality, following indirect standardization and classifying the population by 18 age classes (0-5, ..., 85 or more). The goal is to identify municipalities with a divergent risk from the general mean.

3 Methods

3.1 Models for disease mapping

Let Y_i be the number of observed cases in the i-th area ($i = 1, \ldots, 287$) which follows a Poisson distribution with mean $E_i\theta_i$, where E_i is the expected number of cases under indirect standardization and θ_i the relative risk.
Clayton and Kaldor (1987), assumed a Gamma(κ, ν) prior distribution for θ_i. We specify a full Bayesian model where the hyperparameters κ and, ν are assumed to be exponentially distributed. In this model, Poisson random variability is filtered out and relative risk estimates are shrunken toward the general mean. Besag et al. (1991) specified a random effect log linear

model for the relative risk $log(\theta_i) = u_i + v_i$. The heterogeneity random term u_i represents an unstructured spatial variability component assumed a priori distributed as Normal $(0, \lambda_u)$ where λ_u is the precision parameter modelled as Gamma. The clustering term v_i represents the structured spatial variability component assumed to follow a priori an intrinsic conditional autoregressive (ICAR) model. In other words, denoting S_i as the set of the areas adjacent to the i-th area, $v_i|v_{j\in S_i}$ is assumed distributed as Normal($\bar{v}_i$,$\lambda_v n_i$) where $\bar{v}_i$ is the mean of the terms of adjacent areas to the i-th one (Besag and Kooperberg, 1995) and $\lambda_v n_i$ is the precision, which is dependent on n_i, the cardinality of S_i. Through these two random terms the BYM model shrinks the relative risk estimates both toward the local and the general mean.
We now look at these two models as hierarchical models for the null. The problem becomes how to specify a portfolio of suitable prior distributions.

3.2 Priors specification

The choice of a suitable combination of hyperparameters leads to different degree of prior vagueness on the extent relative risk heterogeneity among areas.
The exploration is facilitated in the case of the conjugate Poisson-Gamma model since we have a close solution and the problem reduces to parameters specification of the predictive negative binomial distribution.
For the Besag et al. (1991) model we took advantage of the proposal of Bernardinelli et al. (1995). The hyperpriors for the precision parameters were parameterized in terms of the ratio between the 95th percentile and the 5th percentile of the relative risk distribution.
In particular, the 90 per cent range of variation of RR mapped as a ratio of RRs, is approximately $\frac{\theta_{0.95}}{\theta_{0.05}} = exp(2z_{1-\alpha/2}\sqrt{\sigma_u + c\sigma_v})$ where the constant c depends on the observed adjacency structure and the neighbour weighting matrix, with $1/\sigma_u \propto \chi^2_{\nu_u}/s_u$ and $1/\sigma_v \propto \chi^2_{\nu_v}/s_v$. Such distributions depends on the prior scale parameters s_u and s_v and the prior degrees of freedom parameters ν_v and ν_u.

3.3 Cross-validation predicted p-values

Divergence from the hierarchical null models is assessed via posterior predictive distribution.
The posterior predictive distribution is:
$P(Y^{rep}|Y) = \int P(Y^{rep}|Y,\theta)P(\theta|Y)d\theta = \int P(Y^{rep}|\theta)P(\theta|Y)d\theta$
assuming conditional independence of Y^{rep} and Y given the parameters. This is too confident since the data are used twice, for deriving posteriors and for obtaining replicates (Plummer 2008). To control for excess in optimism the posterior predictive distribution is replaced by the cross-validation (leave-one-out) posterior predictive distributions:

$P(Y^{rep}|Y_{-i}) = \int P(Y^{rep}|\theta)P(\theta|Y_{-i})d\theta$
Cross validation posterior predicted distributions are computationally prohibitive. Several approximations have been proposed. A mixed approach was given by Marshall and Spiegelhalter (2007) and it is particularly convenient under WinBugs. At each Montecarlo iteration a replicate value for the random parameters for the i-th observation is generated and then used to generate a replicate observation Y_i^{rep}. This approach is called mixed because random effects are drawn from their predictive distribution and not from the posterior.
A measure of divergence can be the cross validation posterior predicted p values defined, using mid-p for a discrete response, as:

- if $Y_i > E_i$: $Pr(Y_i^{rep} > Y_i^{obs}|Y_{-i}) + \frac{1}{2}Pr(Y_i^{rep} = Y_i^{obs}|Y_{-i})$
- if $Y_i < E_i$: $Pr(Y_i^{rep} < Y_i^{obs}|Y_{-i}) + \frac{1}{2}Pr(Y_i^{rep} = Y_i^{obs}|Y_{-i})$

where Y_i is the observed and E_i the expected number of cases in the i-th area.

4 Results

Table 1 and table 2 reports some possible choices of hyperprior parameters for the Poisson-Gamma and Besag et al. models. Table 1 shows also the prior 90% centile range of relative risk. These ranges represent different reference beliefs about the background variability of disease risk among areas. Each choice will produce a different set of divergent observations (see Figure 1). Notice that the priors defined by the hyperparameters values in the tables are very informative. We deliberately specify a bad-fitting models on the basis of prior null expectation. A simple leave-one out cross-validation has a very little effect on the posterior distributions of the model parameters in the Disease mapping context (data not shown here).

5 Conclusion and Discussion

Similar approaches to hierarchical modelling of the null is described in Ohlssen et al. (2009). The authors argued that fitting null model by leave-one out cross-validation may be sufficient to detect divergent observations. We disagree to this point, as we show in the results section. In Disease mapping hierarchical modelling of the null can be reached by specifying a full range of informative null priors.

References

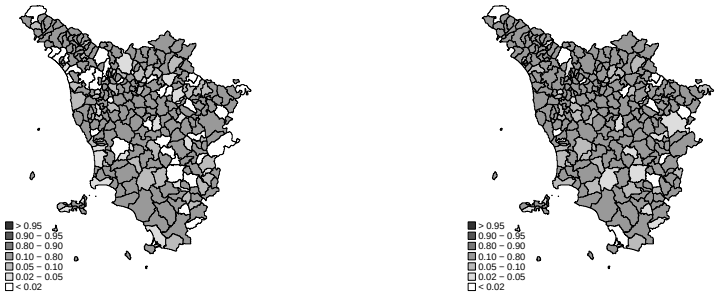
Bernardinelli, L., Clayton, D. and Montomoli, C. (1995). Bayesian Estimates of disease maps : how important are priors? *Statistics in Medicine*, **14**, 2411-2431.

TABLE 1. Possible choices of priors for the Poisson Gamma model.

	ν	α	μ	σ	5%	95%	RATIO
EB	34.6	35.3	0.98	0.17	0.72	1.26	1.75
large	95.0	99.0	0.98	0.10	0.82	1.14	1.39
small	266.8	272.2	0.98	0.06	0.88	1.07	1.22

TABLE 2. Possible choices of priors for the Besag et al. (1991) model and number of divergent areas at different probability thresholds (5% and 1%).

	DF	5%	1%
Prior 1	15	26	2
Prior 2	20	28	4
Prior 3	25	63	19



Prior 3 Prior 2

FIGURE 1. Cross validation posterior predicted p values under two different null priors (prior 3 and prior 2 of Table 2). Lung cancer, males, 1995-1999, Tuscany.

Besag, J., York,J., and Mollié, A. (1991). Bayesian Image Restoration, with Two Applications in Spatial Statistics (with discussion). *Annals of the Institute of Statistical Mathematic*, **43**, 1-59.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disea-se Mapping. *Biometrics*, **43**, 671-681.

Lunn, D.J., Thomas, A., Best, N,. and Spiegelhalter D. (2000):WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**, 4, 325-337.

Marshall, C.A. and Spiegelhalter, D.J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, **22**, 1649-1660.

Marshall, C.A. and Spiegelhalter, D.J (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis*, **2**, 409-444.

Ohlssen, D.I., Sharples, L.D. and Spiegelhalter, D.J. (2007). A Hierarchical Modelling Framework for Identifying Unusual Performance in Health Care Providers. *Journal of the Royal Statistical Society, Series A*,**170**, 865-890.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523-539.

Stern, H.S. and Cressie, N. (2000). Posterior predictive model checks for disease mapping models *Statistics in Medicine*, **19**, 2377–2397.

LASSO Penalised Likelihood in High-Dimensional Contingency Tables

Susana Conde[1,2] , Gilbert MacKenzie[1,3]

[1] Centre of Biostatistics, Department of Mathematics & Statistics, The University of Limerick, Ireland, susana.conde@ul.ie & gilbert.mackenzie@ul.ie
[2] School of Mathematical Sciences, Room 1.65, Western Gateway Complex, University College Cork, Ireland
[3] ENSAI, Rennes, France

Abstract: We consider several least absolute shrinkage and selection operator (LASSO) penalized likelihood approaches in high dimensional contingency tables and with hierarchical log-linear models. These include the proposal of a parametric, analytic, convex, approximation to the LASSO. We compare them with "classical" stepwise search algorithms. The results show that both backwards elimination and forward selection algorithms select more parsimonious (i.e. sparser) models which are always hierarchical, unlike the competing LASSO techniques.

Keywords: high dimensional contingency tables; LASSO; model selection; penalized likelihood; stepwise search algorithms.

1 Introduction

Conde and MacKenzie (2008) have recently developed new stepwise search algorithms for binary variables, in the context of high dimensional contingency tables and hierarchical log-linear models. They introduced the idea of measuring dependence between binary comorbidities using interactions in a hierarchical log-linear setting. The algorithms can work with any number of binary variables, and constitute one approach to the problem of model selection in this context.
In this paper we consider a different approach which has been more recently developed in the literature, based on Penalized Likelihood. The idea is to attach a penalty to the usual likelihood function. Different penalties may be adopted to achieve various desirable properties: e.g., sparsity (Friedman, 2008) or smoothness of solutions (Eilers and Marx, 1996), etc. Here we are primarily interested in encouraging sparse solutions in order to identify a more parsimonious model in high-dimensional contingency tables.
We compare our methods with the penalized likelihood approach given by Dahinden *et al.* (2007), who provided just such an extension for contingency tables using the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) penalty. We consider a multinomial likelihood,

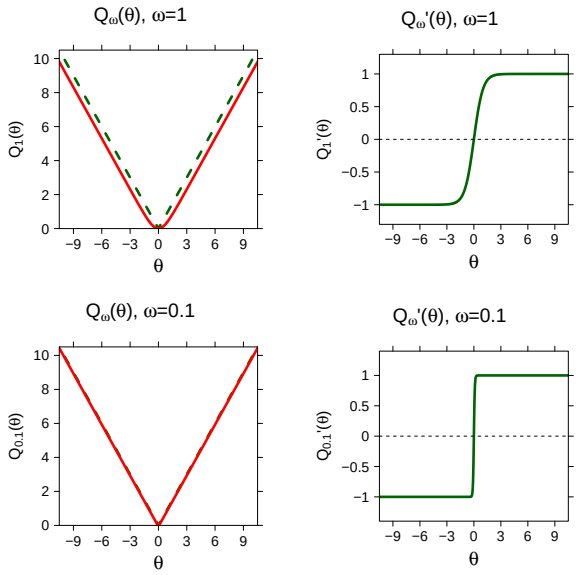


FIGURE 1. Graphs of Q_ω and Q'_ω for $\omega = 1, 0.1$.

with penalties: (a) the LASSO; (b) the LASSO only in the interactions; (c) a parametric, convex, analytic approximation to the LASSO (Lee, 2010). The latter two developments are novel.

2 Penalized Likelihood Inference

Given p binary variables, let consider the p-dimensional contingency table with $q = 2^p$ cells. If we define $\mu_i = E(Y_i)$, the expected value in the ith cell, $i = 1, \ldots, q$ let consider a log-linear regression model with k parameters (with $k \leq q$):

$$\ln(\mu_i) = \sum_{j=1}^{k} a_{ij}\, \theta_j. \tag{1}$$

where $A = (a_{ij})$ is a $(q \times k)$ design matrix, k the number of linearly independent parameters; and θ, the vector of unknown parameters measuring the influence of the constant, main effects and interactions on the response. The dimension of θ is that from the vector space spanned by the columns of A. Thus, A has full rank $= k$. A log-linear model is a generalized linear model.

For inference, we consider that the penalized negative log-likelihood is:

$$-\ell^{\mathcal{P}}(\theta, \lambda) = -\ell_{\text{mult}}(\theta) + \text{pen}_\lambda$$

where $\ell_{\text{mult}}(\theta)$ is the log-likelihood of a multinomial random variable, and pen_λ, the penalty term, is, for $\lambda > 0$

$$(a): \quad \lambda \sum_{j=2}^{k} |\,\theta_j\,|, \qquad\qquad (b): \quad \lambda \sum_{j=2+p}^{k} |\,\theta_j\,|,$$

and where for the case of the smooth approximation, we have that $\sum_j |\,\theta_j| \approx \sum_j Q_\omega(\theta_j)$ with $Q_\omega(\theta_j) = \omega \ln \left[\cosh\left(\frac{\theta_j}{\omega}\right)\right]$ for a certain constant ω that regulates the approximation of the function to that of the absolute value, see Figure 1, whence the penalty term is

$$(c): \quad \lambda \sum_{j=l}^{k} \omega \ln \left[\cosh\left(\frac{\theta_j}{\omega}\right)\right]$$

where $l = 2$ or $p + 2$. Note that $Q_\omega(\theta_j) \in \mathcal{C}^\infty$, the set of functions that are infinitely differentiable, and is convex. We define then the maximum penalised likelihood estimates (MPLEs), according to the terminology of Green and Silverman (1994) as

$$\hat{\theta}^{\mathcal{P}}(\lambda) := \arg\min_{\theta \in \Theta} \left\{-\ell_{\text{mult}}(\theta) + \text{pen}_\lambda\right\}. \tag{2}$$

For a large λ, all the estimates have gone to 0; and for $\lambda = 0$, there is no constraint whence $\hat{\theta}^{\mathcal{P}}(0) \equiv \hat{\theta}$, the maximum likelihood estimates. We estimated the regularization parameter using cross-validation with different folds (5-, 10-, 20-) as required.

We also note that the LASSO penalty is a non-differentiable function, which can complicate optimization. Muggeo (2010) proposes a penalty which is a smoother parametric approximation to the LASSO; nevertheless, that approximation is only once differentiable (Conde, 2011) and standard Newton algorithms require that the function is at least twice differentiable.

2.1 Hierarchical Log-Linear Models

We note too that the models derived from penalties encouraging sparsity are not necessarily hierarchical: the penalty and the estimation procedure do not take the hierarchical rules into account. Overall, this is a major disadvantage of the methodology since non-hierarchical models are not invariant to the choice of design matrix (Conde, 2011) and, accordingly, are of no scientific interest. This remark does *not* apply to simple main effects analysis.

2.2 Computation

To compute the solutions of (2), we used the `logilasso` package, contributed by Dahinden (2007). The package fits log-linear models in sparse

contingency tables using penalized likelihoods. The penalties supported are the LASSO, the group-L_1, and the L_2. The `logilasso` procedure calculates the estimates of the parameters along a path of λs, estimating λ by cross-validation. The functions in the package use a path following algorithm (Dahinden *et al.*, 2007). They re-scale $\lambda^* = 0$ to 1 so that the latter value corresponds to $\lambda = +\infty$. The algorithm starts with $\lambda^* = 1$, for which all the parameters are 0. Then, in each step, it tries to add, to the active set, which is the set of non-zero parameters, the parameters for which the condition for a minimum in the previous inactive set, has been violated. The estimates of the parameters are calculated using a Newton formula with the current λ and the previous estimate. Our penalty (b) is not included in the `logilasso` package and we used `nlm` in R and as (c) is an analytic penalty we again used `nlm`, obtaining the standard errors directly.

3 Results

In this paper, merely as an illustrative example of the methodology, we analyze a simulated contingency table, corresponding to $p = 5$, $n = 2000$, sampled from the model with all two-way interactions present. We used a design matrix up to and including all the 3-ways. Then, the model has the constant, 5 main effects, 10 2-way interactions, and 10 3-way interactions, a total of 25 parameters (without including the constant). The table is in vector format and Fortran standard order:

$$\begin{aligned}\mathcal{Y}^* = (&39, 23, 21, 27, 42, 7, 37, 21, 75, 70, 21, 56, 50, 21, 14, 28,\\ &87, 55, 9, 21, 46, 13, 12, 4, 325, 520, 28, 129, 103, 61, 10, 25)^{\mathrm{T}};\end{aligned}$$

In Figure 2 we present the graph showing the MPLEs corresponding to the 10 3-way interactions along the path of λs, with a LASSO penalty: panel (a) using the `logilasso` package (Dahinden, 2007). Irrespectively of the fold of the cross-validation, the final model found is the same (i.e. 6 three-ways went to zero); panel (b) using the `nlm` function in R with the LASSO penalty; panels (c) and (d), using the `nlm` function in R with the smooth LASSO for $w = 1$. The paths of the 3-ways are not stabilized until λ is very large (panel (c)) and they go to 0 much slower than in the previous cases. Note that the range of λs in panel (d) corresponds to those in panels (a) and (b). Finally, panels (a) and (b) compare the use of the `logilasso` package with `nlm`.

The final models found using each penalty are: for the LASSO, six three-way interactions out of the ten went to zero; for the LASSO only in the interactions, five three-ways went to zero; for the smooth LASSO, if we use the 95% confidence interval includes 0 as a cut-off criterion, the method found the correct model, i.e. the all two-way interactions model. Our step-wise search algorithms (Conde and MacKenzie, 2008; Conde, 2011) found either the true model (the backwards elimination algorithms), or a model

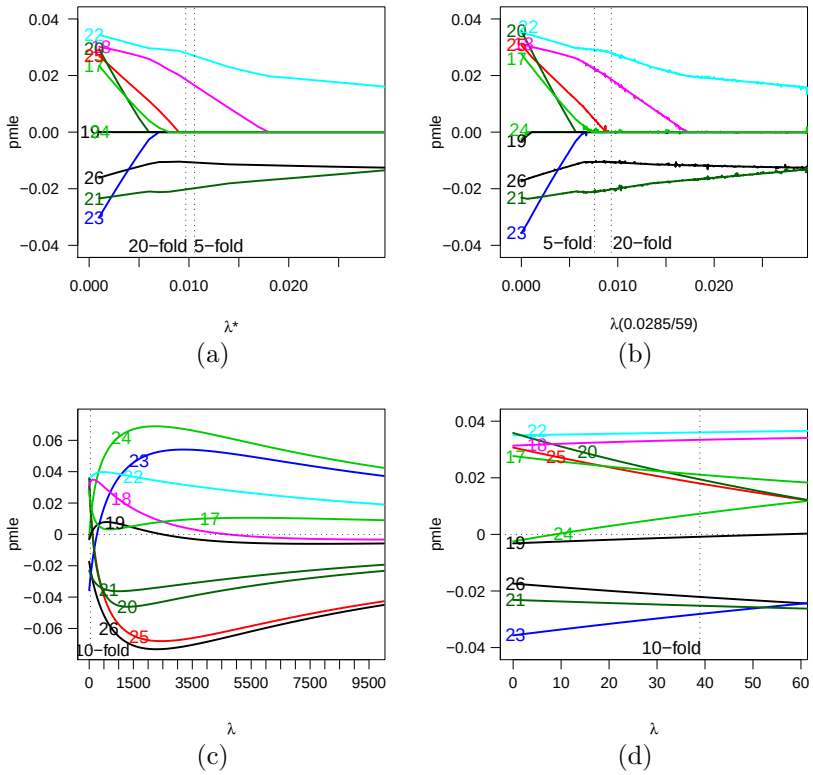


FIGURE 2. MPLEs of the 3-way interactions, with a LASSO penalty. (a) Using the package `logilasso`, $\lambda^* \in [0, 1]$; (b) Using the `nlm` function; (c) and (d) Using the smooth approximation for $\omega = 1$. The range of λ in (a), (b), and (d) coincide.

with nine two-ways (the Forward Selection algorithm), i.e., in all cases a more parsimonious (and hierarchical) model.

For this example table, the LASSO penalized likelihood method found the least parsimonious or least sparse model, in "regularization" terminology. This is just one table, but it illustrates a direct contradiction to the view that penalized likelihood methods produce sparse(st) solutions. We have many more examples including simulated and real data, and including cases from the $q \gg n$ scenario, with similar conclusions (Conde, 2011).

4 Final Remarks

As far as we know, this is the first time that penalized likelihood approaches have been compared with some "classical" stepwise search algorithms in contingency tables. In the light of the results, we recommend the use of the stepwise algorithms which outperform the LASSO penalized likelihood approaches. We will present more detailed finding at the workshop.

Acknowledgments: The work for this paper was supported Glaxosmithkline (GSK) who funded the first author from 2006-2010. It was also supported by Science Foundation Ireland (SFI, www.sfi.ie) under their Mathematics Initiative, II, via the BIO-SI (www.ul.ie/bio-si) research programme in the Centre of Biostatistics, University of Limerick, Ireland (grant number 07/MI/012). The second author is the PI of BIO-SI project.

References

Conde Llinares, S. (2011). *Interactions: Log-Linear Models in Sparse Contingency Tables.* PhD Thesis. The University of Limerick, Ireland. Submitted.

Conde Llinares, S. and MacKenzie, G. (2007). Modelling High Dimensional Sets of Binary Co-morbidities. In: *Proceedings of the 22nd International Workshop on Statistical Modelling*, Barcelona, 177-180. Eds.: del Castillo J., Espinal A. and Puig, P.

Conde Llinares, S. and MacKenzie, G. (2008). Search Algorithms for Log-Linear Models in Contingency Tables: Comorbidity Data. In: *Proceedings of the 23rd International Workshop on Statistical Modelling*, Utrecht, 184-187. Ed.: Eilers, P.H.C.

Dahinden, C., Parmigiani, G., Emerick, M.C. and Bühlmann, P. (2007). Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics*, **8:476**.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science*, **11(2)** 89-121.

Friedman, J.H. (2008). Fast Sparse Regression and Classification. In: *Proceedings of the 23rd International Workshop on Statistical Modelling*, Utrecht, 27-57. Ed.: Eilers, P.H.C.

Lee, W. (2010). *Personal communication.*

Muggeo, V.M.R. (2010). LASSO regression via smooth L_1-norm approximation. In: *Proceedings of the 25th International Workshop on Statistical Modelling*, Glasgow, 391-396. Ed.: Bowman, A.W.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, **58(1)** 267-288.

Describing the geography of Spanish bank branching.

David Conesa[1], Anabel Forte[2], Luisa Alamá[2], Emili Tortosa-Ausina[2,3]

[1] Universitat de València
[2] Universitat Jaume I
[3] Instituto Valenciano de Investigaciones Económicas

Abstract: In this work we undertake an analysis of the association between socio-economic variables and the spatial distribution of bank branches in Spain in 2008. In particular, a Poisson regression with random effects is used to model to number of bank branches in each Spanish municipality. Bayesian approach is used to make inference about parameters. The main result is that we can determine whether a given municipality (or province) is *under-branched*, which would suggest the existence of financial exclusion (in terms of bank service accessibility), or *over-branched*, which could be indicative of misguiding expansion policies. .

Keywords: Banking; Generalized linear mixed models; Hierarchical Bayesian modeling.

1 Introduction

During the last few years the geography of bank branching has been changing in several countries around the world. In the U.S., recent laws have ultimately removed branching restrictions at both intra- and inter-state levels (Jayaratne and Strahan 1996, 1999). In Spain, one of the five largest banking systems in Europe, these laws have allowed savings banks to enter other markets different to their traditional markets, since they could set offices in regions different from their regions of origin. This bank deregulatory initiative triggered off the morphing of the geography of Spanish banking, in which, simultaneously to the Latin America forays of some large commercial banks, savings banks expanded geographically throughout the country, becoming the main actors in dimensions as important as the total number of branches (Fuentelsaz et al. 2002).
However, the recent economic and financial crisis has questioned the validity of the geographic expansion policies set by most savings banks, and not only the largest ones. Many of these firms based their expansions in financing the housing bubble, whose burst is closely related to the difficulties they are going through nowadays. As a result of such difficulties, the 46 savings banks existing by the end of 2009 have virtually reduced to 17 due to the

merger process enforced by the Bank of Spain, whose principal aim was to strengthen the Spanish financial system. In addition, prior to the start of the merger process, some savings banks had already initiated a back off policy, by closing some offices—the total number of bank branches in Spain decreased for the first time ever in 2008, and the decline has intensified in 2009 and 2010.

One can therefore forecast the total number of bank branch offices to decrease further in the next few years, not only for savings banks but also for commercial banks and credit unions—the other two types of firms operating in the Spanish banking system. However, since this pattern is not expected to reverse, some concerns might be raised on its likely negative effects. In the particular case of Spain, although financial exclusion has not been high in the political agenda, at least in comparison with other countries, various institutions have pursued objectives aimed at helping to reduce financial exclusion. In particular, Spanish savings banks offer banking products that are designed specifically for vulnerable groups. The concentration process in the Spanish banking industry has led to a much lower number of savings banks which will ultimately operate as commercial banks and, consequently, their contribution to financial inclusion could be thwarted (Bernard et al. 2008).

Therefore, taking into account the relevance of financial exclusion and the recent changes in the banking industry in general and the geography of bank branching in particular, this work undertakes an analysis of the association between socio-economic variables and the spatial distribution of bank branches in Spain in 2008. Since the access to bank services is unlikely to be improved simply by an increase in the number of bank branches, the spatial distribution of branches need to address points of actual and growing unmet demand—which could also be points of declining demand. We tackle these issues using a Generalized Linear Mixed Model (in particular a Poisson regression with random effects). In order to make inference about parameters, we use the Bayesian approach. The main result is that we can determine whether a given municipality (or province) is *under-branched*, which would suggest the existence of financial exclusion (in terms of bank service accessibility), or *over-branched*, which could be indicative of misguiding expansion policies.

2 Modeling the number of bank offices

Taking into account that our interest is describing the number of branches in each municipality O_i, we model these counts using a Poisson regression model. This is a particular case of Generalized Linear Models where each observation follows a Poisson distribution centered in the expected number of cases by a multiplier λ_i:

$$O_i \sim \mathrm{Po}(E_i \times \lambda_i)\,,$$

where the expected number of branches in each municipality is computed taking into account the corresponding population:

$$E_i = pob08_i \times \frac{\text{Total number of offices in Spain}}{\text{Total population in Spain in 2008}}.$$

The quantity λ_i is a factor which modifies the expected number of offices. As a first step of the study, this parameter was modeled taking into account several covariates. But we found that the variance of our observations is quite larger than their mean (a variance of 2709.89 with a mean of 6.04 offices), which is far away from expected (same mean and variance for Poisson data). To solve this issue we introduced the extra variability through a Generalized linear mixed model.

For this study we considered three different covariates. The population density, the unemployment rate and the foreigners rate. In particular, the population density is considered through its logarithm. This is because of the magnitude of this variable. Also, we considered reasonable that a variation of the density when it is small should have a larger effect than its variation when the density is large.

The other effect affecting the number of bank offices is the geographical region (the province or the community of the municipality). This geographical effect can considered in three different ways: as fixed effects, as an independent random effect or as random effects with a dependence structure. Although we have implemented all of them, we only present here the one which better fits our data. In particular, the best geographical effect is a fixed effect per province, taking Burgos to be the base province.

The resulting linear predictor is linked with λ_i in the usual way as:

$$\begin{aligned} \log(\lambda_i) &= \alpha_0 + \alpha_1 * \log(\text{density}_i) + \alpha_2 * \text{unemployment}_i \\ &+ \alpha_3 * \text{foreing}_i + \beta_1 * \text{prov1}_i + \cdots + \beta_{51} * \text{prov51}_i + U_i, \end{aligned}$$

with $U_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, 8109$ municipalities in Spain.

Once the model is determined, the next step is to estimate its parameters. As we are using the Bayesian paradigm, we have to specify the (hyper) prior distributions of each parameter involved in the model. We have considered rather noninformative prior distributions, with the aim of expressing our initial vague knowledge about the parameters. Expressions above jointly with the priors of all the parameters contain all our knowledge of the system but they do not yield to analytical estimates. Therefore, we have to resort to numerical methods in order to obtain the posterior distributions of all the parameters and also to make prediction about the presence/absence in a series of unsampled locations. In particular, MCMC inference have been carried out using WinBUGS (Spiegelhalter et al., 1999).

It is worth noting that this modeling is the one resulting after a model selection process among the possible geographical effects. This process has been done using the Deviance information criterion (DIC) (Spiegelhalter et

al., 2002), the more useful criterion when comparing models whose posterior distribution has been approximated by MCMC.

Acknowledgments: David Conesa and Anabel Forte would like to thank the financial support of the Ministerio de Educación y Ciencia (jointly financed with European Regional Development Fund) via the research Grant MTM2010-19528 and of the Generalitat Valenciana via the research Grant ACOMP11/218.

References

Bernad C.,Â Fuentelsaz L., and Gómez J. (2008). Deregulation and its long-run effects on the availability of banking services in low-income communities. *Environment and Planning A*, 40(7), 1681–1696.

Fuentelsaz L., Gómez J., and Polo Y. (2002). Followers' entry timing: Evidence from the Spanish banking sector after deregulation. *Strategic Management Journal*, 23(3), 245–264.

Jayaratne J. and Strahan P.E. (1996). The finance-growth nexus: Evidence from bank branch deregulation. *The Quarterly Journal of Economics*, 111(3), 639–670.

Jayaratne J. and Strahan P.E. (1999). Entry restrictions, industry evolution, and dynamic efficiency: evidence from commercial banking. *The Journal of Law and Economics*, 16, 239–273.

Spiegelhalter, D.J., Myles, J.P., Jones, D.R., Abrams, K.R. (1999). Methods in health service research. An introduction to Bayesian methods in health technology assessment. *British Medical Journal*, 319, 508–612

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

Spatio-temporal disease modeling and surveillance with Bayesian hierarchical Poisson models

Ana Corberán-Vallet[1], Andrew B. Lawson[1]

[1] Division of Biostatistics and Epidemiology,
Medical University of South Carolina,
135 Cannon St. Suite 303, 29425 Charleston, South Carolina, United States,
corberan@musc.edu, lawsonab@musc.edu

Abstract: This study deals with the development of statistical methodology for prospective spatio-temporal disease surveillance. Within the framework of Bayesian hierarchical Poisson count models, we show how the conditional predictive ordinate, a general Bayesian diagnostic which detects observations discrepant from a given model, can be adapted in a surveillance context to detect small areas of unusual aggregation of disease as quickly as possible. As a local measure, different alarms will be sounded for those areas of increased disease incidence. In order to address the problem of multiple comparisons, a common prior probability that a given area signals an alarm when no change in risk takes place is introduced into the model specification. Once an incident cluster is identified, our model formulation allows us to determine the change in the relative risk pattern.

Keywords: Public health surveillance; On-line surveillance; Areal data; Lagged loss function; Conditional predictive ordinate; Multiple comparisons.

1 Disease modeling and surveillance

The ability to rapidly detect any substantial change in disease incidence is of critical importance to public health practitioners, thus facilitating timely public health interventions. Unlike testing methods, modeling for spatio-temporal disease surveillance is a relatively undeveloped arena of statistical methodology. Most spatio-temporal models have been developed for retrospective analyses of complete data sets. However, data in public health registries accumulate over time and sequential analyses of all the data collected so far is a key concept to early detection of changes in disease risk over space and time.

When small area disease data in the form of counts are available, Bayesian hierarchical Poisson models are commonly used to describe the behavior of diseases. At the first level of the model, the Poisson distribution with a mean which is a function of the expected counts of disease and the unknown area-specific relative risks is considered for modeling the within area

variability of the counts. At the second level of the model, the logarithm of the relative risk is usually decomposed in additive components representing spatial, temporal, and space-time interaction effects (Lawson, 2009, ch 11). In this study we build on Poisson count models for prospective spatio-temporal disease surveillance. In particular, we use the convolution model to describe the behavior of disease under endemic conditions. Those are simple and robust models where the relative risks are assumed to be constant over time (Besag et al., 1991). Each time new observations become available, we show how the conditional predictive ordinate (CPO, Geisser, 1980) can then be adapted in a surveillance context to detect small areas of unusual disease aggregation. In particular, for each small area, we define the surveillance conditional predictive ordinate (SCPO) as the conditional predictive density of the new observation given the data from previous time periods, values close to zero indicating that the new observation is not representative of the data expected under the previously fitted model. As a local measure, different alarms will be sounded for those areas of increased incidence. Hence, the proposed surveillance technique can be used to detect multiple clusters of varying size and magnitude simultaneously.

From a Bayesian viewpoint, there is no need to introduce a penalty term for performing multiple comparisons simultaneously (Scott and Berger, 2006). However, the multiple comparisons problem has to be carefully addressed to assure a good performance of the surveillance procedure. We propose to model the number of alarms at each time period under the null by the Binomial distribution with parameters the number of small areas under surveillance and the probability of each area signaling an alarm. We can then evaluate, at each time period, the probability of observing at least the same number of alarms. An alarm for an out-of-control system will be triggered if this probability is below a critical level. All the alarms associated with small areas of unusual aggregation of disease will then be reported.

Once an incident cluster of disease has been detected, an additional effect representing the expected additive increase of disease counts due to the outbreak is added to the mean of the Poisson distribution. For non-infectious diseases, we assume that the outbreak component follows a Gaussian random walk. For infectious disease, the epidemic component is modeled as a function of the previous numbers of cases in the area as well as in the neighboring areas, which allows us to explain the spread of the disease.

It is important to emphasize that at each time period a new set of data are included in the model. This implies restarting the MCMC simulation process from scratch at each time period, which can be time consuming. Here we propose to use a sliding window with fixed time units (Lawson, 2004) to estimate the convolution model describing the endemic state. That is, only the most recent observations are used to estimate the model at each time period. The observations corresponding to outbreaks of disease are only used to model the epidemic state, and so they will be assumed to be missing in the estimation of the convolution model.

2 Numerical example

We analyze culture positive (C+) laboratory notifications of influenza in South Carolina from October 2004 to April 2005, with a total of 13 biweekly time periods. As we can see in Figure 1, a slightly high count can be observed at time period 4. However, it is not until time period 6 when we declare an epidemic state, with a length of 6 time periods. Figure 2 displays a selection of cumulative crude count maps for 4 time periods: week beginning 15th December 2004, 15th January, 1st March, and 1st April 2005. Differential timing of incipient epidemic waves and also differences in duration can be observed across the state.

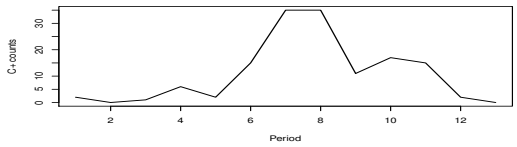


FIGURE 1. Biweekly C+ laboratory notifications of influenza in South Carolina.

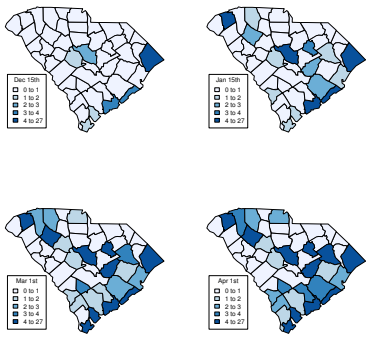


FIGURE 2. A selection of four county maps during the flu season of 2004/2005: cumulative counts of C+ notifications.

Figure 3 displays the posterior average mean C+ notifications temporal profiles for Charleston and Richland, 2 major urban regions of the state. Because of the limited amount of historical data, we start the surveillance exercise at time period 4, using the first 3 time periods to estimate the endemic behavior of disease. Alarms for an epidemic state (solid points) were properly sounded at those time periods of increased disease incidence.

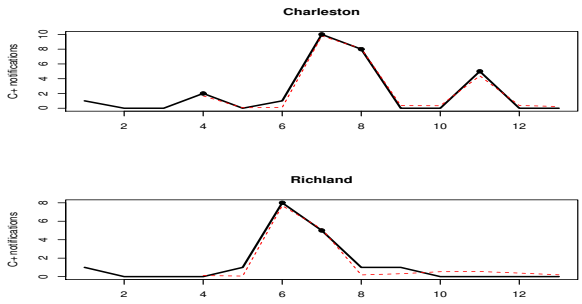


FIGURE 3. Real (solid line) and posterior mean (dashed line) C+ notifications for Charleston and Richland. Solid points represent detected unusual observations.

3 Conclusions

In this study, we show how the conditional predictive ordinate can be adapted in a surveillance context to detect areas of unusual aggregation of disease. The results obtained in the analysis of influenza data at county level are encouraging. Our surveillance procedure allows us to detect influenza epidemics at the very moment of their onset. In addition, our model formulation provides a good description of the epidemic behavior.

References

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.

Geisser, S. (1980). Comment on: Sampling and Bayes' inference in scientific modelling and robustness. By G.E.P. Box, *Journal of the Royal Statistical Society, Series A*, **143**, 416-417.

Lawson, A.B. (2004). Some issues in the spatio-temporal analysis of public health surveillance data. In: *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*, Brookmeyer, R., and and Stroup, D.F. (eds), Chapter 11. Oxford: Oxford University Press.

Lawson, A.B. (2009). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology.* Boca Raton: Chapman & Hall.

Scott, J.G., and Berger, J.O. (2006). An Exploration of Aspects of Bayesian Multiple Testing. *Journal of Statistical Planning and Inference*, **136**, 2144-2162.

Time series modeling and Bayesian forecasting with exponential smoothing models

Ana Corberán-Vallet[1], José D. Bermúdez[2], Enriqueta Vercher[2]

[1] Division of Biostatistics and Epidemiology,
Medical University of South Carolina,
135 Cannon St. Suite 303, 29425 Charleston, South Carolina, United States,
corberan@musc.edu

[2] Department of Statistics and Operational Research,
University of Valencia,
Dr Moliner 50, 46100 Burjassot, Valencia, Spain,
Jose.D.Bermudez@uv.es, Enriqueta.Vercher@uv.es

Abstract: This study deals with the prediction of time series using a Bayesian forecast approach based on exponential smoothing models. In particular, we describe the Bayesian analysis of the Holt-Winters model formulated as a linear heteroscedastic model. This alternative formulation simplifies the Bayesian analysis of the model, since it provides the joint distribution of the data vector. As a consequence, any conditional distribution is also known, which allows us to develop a straightforward approach for dealing with missing data problems. In addition, we show how the linear formulation generalizes in a natural manner to the multivariate case, which allows us to jointly forecast correlated time series. The Bayesian analysis of the multivariate Holt-Winters model formulated as a seemingly unrelated regression model is straightforward. MCMC simulation techniques and Monte Carlo integration are used in order to approach the posterior and predictive distributions, which are not analytically tractable.

Keywords: Holt-Winters model; Positive time series; Missing data; Multivariate time series; Monte Carlo methods.

1 Exponential smoothing models

Exponential smoothing methods, due to their simplicity and robustness, are widely used forecasting techniques (Gardner, 2006). Statistical foundation for exponential smoothing was provided by the introduction of a class of innovations state space models underlying exponential smoothing methods (Hyndman et al., 2008). Within this framework, Bermúdez et al. (2007) formulated the additive Holt-Winters model as a heteroscedastic linear model. This formulation simplifies the Bayesian analysis of the model, since it shows the joint distribution of the data vector: a multivariate Normal distribution with mean vector and covariance matrix depending on the

initial conditions and the smoothing parameters (Bermúdez et al., 2010). In this study we present two practical applications of this linear formulation. First, we develop a Bayesian forecast procedure that allows us to analyze positive demand time series with a proportion of zero values and a high variability for the non-zero data. The proposed procedure relies on the analysis of an unconstrained latent demand time series underlying the observed data, which can take negative values but those can only be observed by the value zero. Given that knowing the joint distribution of the data vector implies that any conditional distribution is also known, the linear formulation for the Holt-winters model provides a suitable framework for the analysis of time series with missing and censored observations and, consequently, for the analysis of positive time-series data with zero values.

On the other hand, it is common in practice to find sets of time series subject to correlated random disturbances or where the observations of a time series are related to past and present values of other series. On those occasions, the use of a multivariate time series model accommodating the existing interrelationship is necessary to improve the fit and forecast accuracy with respect to the univariate analyses of the series. Here we show how the linear formulation of the Holt-Winters model can be easily extended to the multivariate case. Assuming that each of the individual time series comes from the univariate Holt-Winters model and that there is a contemporaneous correlation between corresponding errors in the different equations, a multivariate general model is obtained, which can be formulated as a seemingly unrelated regression (SUR) model (Zellner, 1971). From conventional non-informative prior distributions we derive the posterior distribution of all the unknowns. This posterior distribution is not analytically tractable but can be approached by MCMC simulation techniques. In particular, we propose a Metropolis-within-Gibbs algorithm that allows us to simulate from the full conditional posterior distributions of the model parameters. The predictive distribution, which encapsulates all the information concerning the future values of the time series and allows us to calculate both point forecasts and prediction intervals, is finally estimated using Monte Carlo integration (Corberán-Vallet, 2009).

The general multivariate Holt-Winters model includes previously studied exponential smoothing models as particular cases. Of special importance from a practical viewpoint is the homogeneous multivariate model, resulting from assuming that the univariate series share a common structure. The multivariate model can then be formulated as a traditional multivariate regression model (Zellner, 1971), which simplifies its Bayesian analysis (Bermúdez et al., 2009). To decide between the general and the homogeneous multivariate Holt-Winters models we propose to use the deviance information criterion (DIC, Spiegelhalter et al., 2002), which can be easily calculated from samples generated by MCMC simulation techniques.

2 Numerical examples

We first analyze the time series corresponding to the number of buses manufactured in Spain from January 1998 to December 2004, with changing local level and seasonal pattern over time. Figure 1 depicts the time plot of the series together with the forecast obtained when we consider as historical data the observations for the first six years and forecast the last one, 2004, to measure the post-sample accuracy of our forecasts. The corresponding forecast SMAPE is 16.13.

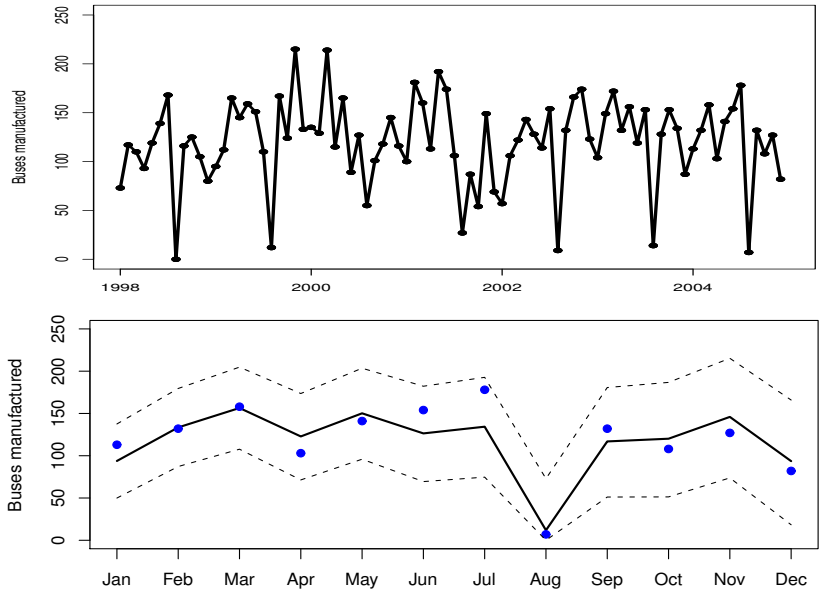


FIGURE 1. Top half: Time plot of the series corresponding to the number of buses manufactured in Spain from January 1998 to December 2004. Bottom half: Monthly point forecasts (solid line) and 80% prediction intervals (dashed lines) for year 2004. Real data are represented by solid points.

Second, we study the data bank that contains monthly hotel occupancy in Castellón, Valencia and Alicante, the three provinces that make up the Valencian Community, from January 2001 to December 2006. Figure 2 shows the data series with a regular growth and additive seasonality. In addition, it is reasonable to assume that the three series are correlated, so the joint analysis of the series with the multivariate Holt-Winters model is justified. In order to illustrate the performance of the multivariate model, we consider as historical data the observations for the first 5 years, 2001-2005, and forecast the last one, 2006. The first step in the analysis of the time series is to select the most adequate multivariate model for describing their

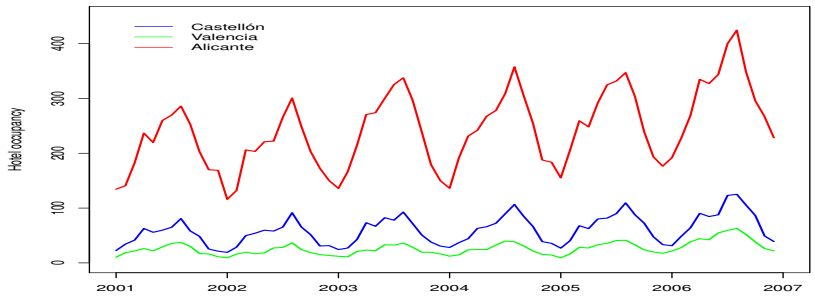


FIGURE 2. Monthly hotel occupancy in Castellón, Valencia and Alicante, in thousands of travelers.

behavior. The values of the DIC criterion are 868.11 and 884.08 respectively for the general and the homogeneous model. Thus, the assumption of a common structure for the time series is not appropriate in this example and the general multivariate model is advisable for the joint analysis of the series. Table 1 shows the SMAPE forecast errors obtained, for each time series, with the general multivariate Holt-Winters model. For comparative purposes, we also include the corresponding errors obtained from the homogeneous model and the univariate analyses.

TABLE 1. Forecast SMAPE obtained, for each time series, from both the general and homogeneous multivariate Holt-Winters models and the univariate models.

	Castellón	Valencia	Alicante	Mean
General multivariate H-W	10.57	28.58	17.78	**18.98**
Homogeneous multivariate H-W	12.95	37.04	18.77	22.92
Univariate Holt-Winters	10.13	35.00	17.58	20.90

References

Bermúdez, J.D., Corberán-Vallet, A., and Vercher, E. (2009). Multivariate exponential smoothing: A Bayesian forecast approach based on simulation. *Mathematics and Computers in Simulation*, **79**, 1761-1769.

Bermúdez, J.D., Segura, J.V., and Vercher, E. (2007). Holt-Winters forecasting: an alternative formulation applied to UK air passenger data. *Journal of Applied Statistics*, **34**, 1075-1090.

Bermúdez, J.D., Segura, J.V., and Vercher, E. (2010). Bayesian forecasting with the Holt-Winters model. *Journal of the Operational Research Society*, **61**, 164-171.

Corberán-Vallet, A. (2009). *Un análisis Bayesiano de modelos multivariantes de suavizado exponencial*. Doctoral thesis.

Gardner Jr., E.S. (2006). Exponential smoothing: The state of the art - Part II. *International Journal of Forecasting*, **22**, 637-666.

Hyndman, R.J., Koehler, A.B., Ord, J.K., and Snyder, R.D. (2008). *Forecasting with exponential smoothing: the state space approach*. Berlin: Springer.

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

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

Assessment of e-government maturity in Portuguese municipalities using regression and clustering approaches

Marco Costa[12], Gonçalo Paiva Dias[13]

[1] Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, Apartado 473, 3754-909 Águeda, Portugal
[2] Centro de Matemática e Aplicações Fundamentais da Universidade de Lisboa
marco@ua.pt
[3] Unidade de Investigação em Governança, Competitividade e Políticas Públicas da Universidade de Aveiro
gpd@ua.pt

Abstract: In order to evaluate the development of websites of the 308 Portuguese municipalities in this work it was performed an analysis using regression models and clustering techniques. That analysis allowed recognizing a group of socioeconomic variables that are significant to characterize homogenous groups of municipalities in what concerns e-government maturity.

Keywords: regression model; clustering; e-government; public administration.

1 Introduction

Public attention to performance analysis in the public sector has grown considerably in recent decades (Heinrich, 2008) and, particularly, in recent years, in the area of e-government. Statistical analysis on the performance assessment of e-government issues is also very recently. Most of these works apply linear regression models and correlation analysis as Mitra and Gupta (2008) or Kumar and Best (2006). To investigate the 'demand' side of e-government, Gauld et al. (2010) applied the multiple logistic regression. Principal components analysis (PCA) was applied in an study about citizens' attitudes towards e-government and e-governance in United Kingdom by Kolsaker and Lee-Kelley (2008). This paper focuses on the evaluation of website maturity of the 308 Portugal's municipalities regarding the features they offer. A combination of multivariate techniques, as regression models and clustering procedures, allows recognizing a group of variables that are significant to characterize some homogeneous groups of municipalities identified by cluster analysis.

2 Methodology

The websites of the 308 municipalities were classified according to the features available, namely, into three dimensions: information online (*Info*), online services (*Serv*) and online participation (*Particip*). Each component was classified according to an evaluation grid translating into a score from an ordinal scale (0-4 points). Thus, each municipality is characterized by a vector with three scores (*Info*, *Serv*, *Partic*). Moreover, a large set of variables was collected (19 variables almost all the National Institute of Statistics of Portugal, INE) that includes variables related with demographic characteristics, economic development, education levels, participation in government modernization programs, etc. Firstly, it was performed an exploratory analysis of variables and outliers were identified for some of them. From a global point of view, the analysis focuses on the sum of the three variables collected, i.e., in a new variable *Total* that indicates a global measure of the maturity of a website. The preliminary analysis indicated possible quadratic relations between some independent variables and *Total* that were considered in modelling procedure. An ordinary least squares (OLS) multiple regression model with backward procedure was fitted to identify a restrict group of exogenous variables that describes significantly the global indicator *Total* as a dependent variable, removing the independent variable with largest p-value (since more than 5%). The final model was validated by verification of the usual assumptions. In a supplementary analysis, a clustering procedure was performed to identify homogenous groups taking into account the websites' scores (*Info*, *Serv*, *Partic*). The clusters analysis considered the squared euclidian distance as disparity measure and the average linkage to hierarchical clustering process. The choice of number of clusters is performed analyzing clusters distance and R-squared criterion. Finally, the solution obtained in the clusters analysis was interpreted through an analysis of the variable's statistics.

3 Regression analysis

Table 1 summarizes the results of the final regression model

$$Total_i = \beta_0 + \beta_1 IRS_i + \beta_2 S1EdInitial_i + \beta_3 Digital_i + \beta_4 Population100_i +$$

$$+\beta_5 PercUrbPop_i + \beta_6 MTT_i \beta_7 MTT_i^2 + \beta_8 MTV_i + \beta_9 EHR_i^2 + \epsilon_i$$

with $i = 1, 2, ..., 308$. It is possible to identify that the demand of digital services, as the online submission of tax forms (*IRS*), is significant to the municipality's score as well as the participation of the municipalities in the modernization program *Simplex* 2008/09 (*S1EdInitial*). Variables *Population100* (number of residents, unity=100 000) and *%UrbPop* (% of population residing in an urban area) represent a demographic characterization. The economic development of municipality is represented by tax

TABLE 1. Regression results.

variable	$\hat{\beta}$	std. error	p-value
Intercept	3.024	0.526	.000
IRS	.021	.009	.015
S1EdInitial	1.326	0.579	.023
Digital	.413	.179	.021
Population100	−1.764	.613	.004
%UrbPop	1.036	.377	.006
MTT	.049	.022	.025
MTT^2	−.001	.000	.011
MTV	.938	.442	.035
EHR^2	−.001	.000	.019
$R^2 = .224$			

variables as MTT (municipal tax on transfers of property, in millions of euros) and MTV (municipal tax on vehicles, in millions of euros). Last variable ERH (expenditure on human resources) incorporates one factor related with staff dimension, namely the expenditure on human resources. The assumptions of normality of errors was verified with the Kolmogorov-Smirnov test (Statistic=.038; p-value=.200).

4 Clustering analysis

A clustering procedure was performed to identify homogenous groups of municipalities considering the vectors of the initial three variables. It was considered the square euclidian distance as measure of disparity and the average linkage approach in the hierarchical clustering procedure. As there are 308 objects, it is very difficult to choose the number of clusters based on dendogram because the graphic is huge. Therefore, two approaches were implemented to support this choice:

- R-squared criterion (greater than 80%), $R^2 = \frac{\sum\sum n_{ij}(X_{ij}-\bar{X}_{ij})^2}{\sum\sum\sum(X_{ijk}-\bar{X})^2}$ performed with the support of the usual ANOVA one-way;
- distance between clusters obtained in the agglomeration process.

Combining these approaches, seven clusters were adopted. Figure 1 shows, in a geographical view, the classification of the 308 municipalities according to the solution of clusters procedure. Attending to statistics of variables in each cluster, presented in Table 2, the clusters analysis allowed the identification ofseven main profiles of websites with different median scores in the three assessment components.

TABLE 2. Characterization of clusters solution (medians for ordinal variables and averages to quantitatives.

Cluster	Info	Serv	Partic	Total	IRS	%UrbPop	MTV
1 (6.2%)	3	3	2	8	69.2	59.5	.81
2 (6.8%)	3	3	1	7	67.1	40.7	.50
3 (11.6%)	3	1	2	6	62.7	32.1	.30
4 (32.1%)	3	1	1	5	61.0	32.0	.34
5 (14.9%)	2	1	1	4	60.3	32.7	.28
6 (17.9%)	3	1	0	4	61.9	23.8	.20
7 (10.4%)	2	0	0	2	56.6	12.5	.09

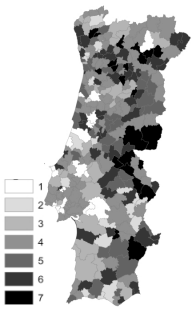


FIGURE 1. Representation of clusters analysis solution in Portugal's map.

References

Gauld R., Gold, S., Horsburgh S. (2010). Do they want it? Do they use it? The ?Demand-Side? of e-Government in Australia and New Zealand. *Gov Inform Q*, **27**, 177-186.

Heinrich, C.J. (2008). Advancing public sector performance analysis *Appl Stoch Model Bus Ind*, **24**, 373-389.

Kolsaker, A., Lee-Kelley, L. (2008). Citizens' attitudes towards e-government and e-governance: A UK study. *International Journal of Public Sector Management*, **21(7)**, 723-738.

Kumar, R., Best, M.L. (2006). Impact and Sustainability of E-Government Services in Developing Countries: Lessons Learned from Tamil Nadu, India. *Inform Soc*, **22**, 1-12.

Mitra, R.K., Gupta, M.P. (2008). A contextual perspective of performance assessment in eGovernment: A study of Indian Police Administration. *Gov Inform Q*, **25**, 278-302.

Joint Modeling Longitudinal Health Care Costs and Time-to-Event Data in Matched Pairs

An Creemers[14], Marc Aerts[1], Niel Hens[13], Frank De Smet[2], Philippe Beutels[3]

[1] Interuniversity Institute for Biostatistics, Hasselt University and Catholic University of Leuven, Belgium
[2] Medical Direction, National Alliance of Christian Mutualities, Belgium
[3] Center for Health Economics Research and Modeling Infectious Diseases, Center for the Evaluation of Vaccination, and Vaccine and Infectious Disease Institute, University of Antwerp, Belgium
[4] Communicating Author: Universiteit Hasselt gebouw D, B-3590 Diepenbeek, an.creemers@uhasselt.be

Abstract: The aim of the study is to investigate the excess health care expenditures for persons with pneumococcal disease, not only at the moment of diagnosis, but also long before and after diagnosis. The dataset contains health care costs and the occurrence time for patients diagnosed with the disease and for a matched control. Joint modeling of costs and the gap times is performed using mixed models. Exponential, Weibull and Gamma distributions with a different link functions to model the gap times are compared. The costs themselves are modeled conditional on the time-to-recurrent-event.

Keywords: Joint Modeling, Longitudinal Data, Health Care Costs

1 Introduction

Streptococcus pneumoniae (or "pneumococcus") is a bacterial pathogen that affects children and adults worldwide. It can cause disseminated invasive disease (including meningitis, bacteraemia and pneumonia) as well as non-invasive disease, including otitis media, non-invasive pneumonia and sinusitis. Anyone can acquire pneumococcal infection, but the invasive pneumococcal disease mostly affects children, the elderly and immunocompromised individuals. Here, we will focus on pneumococcal infections, which are so severe or persistent that they warrant diagnosis by a positive isolate (these are predominantly, though not necessarily, cases of invasive pneumococcal disease).
We focus on the medical costs incurred by people who acquire pneumococcal infections. In the health economic literature there is much theoretical

debate about the inclusion of future unrelated costs in economic evaluations of interventions. Future unrelated costs are often taken to be the discounted population-averaged accumulated health care costs after the age at intervention (i.e. future costs accumulated during years of life that would not have been lived or would have been lived differently without the intervention). However, heterogeneity in susceptibility to illness implies that persons who suffer from particular diseases (especially those who die) are more likely to suffer from other diseases during their hypothesized remaining life span than the average person of the same age. In an attempt to contextualize the costs associated with pneumococcal infections in this manner, we aim to study the overall health care costs after the infection was cleared (i.e. for patients who are vulnerable to the more severe expressions of the disease, but survive the episode), as compared to costs incurred by undiagnosed persons. Furthermore, in addition to analyzing these unrelated costs for the future (i.e. after the time at diagnosis), here we also analyse these for the past (i.e. before the time at diagnosis).

2 The Dataset

The dataset that will be considered in this paper was obtained by merging two databases, one from the National Reference Laboratory, containing all positive pneumococcal isolates in Belgium, the other from the *National Alliance of Christian Sickness Funds* (NACSF), containing all resource use information of members of the largest sickness fund in Belgium.
Merging the two databases described above, resulted in a dataset of resource use by cost category of 876 people who have had a pneumococcal infection at a known point in time and could be matched with 876 patients in terms of municipality, age, gender and social category to unrelated NACSF members in other aspects. Thus, the final dataset contains all medical costs incurred by 1752 NACSF members.
The considered NACSF members were divided into four age groups based on expected differences in levels of severity of experienced pneumococcal disease. Age group 1 contains all diagnosed members and matched members younger than 5 years (in total 2×316 patients), age group 2 the members between 5 and 49 years (2×253 patients), age group 3 all members between 50 and 64 years (2×113 patients) and age group 4 the members aged 65 years or more (2×194 patients).
Not only the size of the costs are considered, also the times at which these costs took place are taken into account. One might expect that diagnosed patients on average not only have higher costs, but also have more frequent costs, i.e. when a member performs a cost at a certain time, it will take less time for a diagnosed member to have the next cost, compared to an undiagnosed member. Creemers et al. (2011) analyzed these data for the first time and they did take this concept into account by analyzing the

cumulative costs rather than the original costs. By doing so, the time of the costs is (implicitly) included in the analysis. Here, we will explicitly include the time at which the costs are made in analysis, by modeling jointly the time to the next cost and the magnitude of the cost.

3 The Joint Model

Denote t_{ij}^k the j^{th} month for the diagnosed (k=1) or the undiagnosed (k=2) member of pair i, $i = 1, \ldots, 876$ and $j = -n_{bi}^k, \ldots, n_{ai}^k$. Negative months coincide with timepoints before the diagnosis, positive months with timepoints after diagnosis. The moment of diagnosis (or in case of a matched member: the moment of diagnosis of the corresponding diagnosed member) takes place at time 0. In stead of looking at the time the costs take place, one can define a variable that describes the *time to the next cost* (when the considered timepoint is after diagnosis) or the *time to the previous cost* (when the considered timepoint is before the diagnosis):

$$\begin{aligned} s_{ij}^k &= t_{ij+1}^k - t_{ij}^k; \qquad k = 1,2 \text{ and } j = 0, \ldots, n_{ai}^k, \\ s_{ij}^k &= t_{ij}^k - t_{ij-1}^k; \qquad k = 1,2 \text{ and } j = -n_{bi}^k, \ldots, -1. \end{aligned}$$

s_{ij}^k is referred to as the *gap time*, i.e. the time between two successive costs. The gap times for an individual i are grouped into a vector $\mathbf{s}_i^k$.
It might be interesting to model jointly these gap times and the costs. One then would like to consider and estimate the joint density $f(\mathbf{y}_i^1, \mathbf{y}_i^2)$, where $\mathbf{y}_i^k = (\mathbf{s}_i^k, \mathbf{c}_i^k)$, with $\mathbf{s}_i^k$ and $\mathbf{c}_i^k$ the vector of respectively gap times and costs for the diagnosed (k=1) or undiagnosed (k=2) patient of pair i. Several modelling assumptions can be made and in most cases, some factorization of the joint density is applied. One possible factorization is as follows:

$$\begin{aligned} f(\mathbf{y}_i^1, \mathbf{y}_i^2) &= \int\int f(\mathbf{y}_i^1|u_i) \times f(\mathbf{y}_i^2|v_i) \times f(u_i, v_i) du_i dv_i \\ &= \int\int f(\mathbf{s}_i^1, \mathbf{c}_i^1|u_i) \times f(\mathbf{s}_i^2, \mathbf{c}_i^2|v_i) \times f(u_i, v_i) du_i dv_i \\ &= \int\int \underbrace{\left[f(\mathbf{c}_i^1|\mathbf{s}_i^1, u_i) \times f(\mathbf{s}_i^1|u_i)\right]}_{F_1} \times \underbrace{\left[f(\mathbf{c}_i^2|\mathbf{s}_i^2, v_i) \times f(\mathbf{s}_i^2|v_i)\right]}_{F_2} \\ &\qquad \times f(u_i, v_i) du_i dv_i. \end{aligned} \tag{1}$$

The first factor F_1 in (1) refers to the diagnosed patients, while the second factor F_2 refers to the undiagnosed patients. Diagnosed and undiagnosed patients are linked by the joint random-effects distribution $f(u_i, v_i)$. For

this joint distribution, a bivariate distribution (for example a bivariate normal density) can be assumed. Another option is to assume that the random effects are independent, and thus $f(u_i, v_i) = f(u_i) \times f(v_i)$, which simplifies the model drastically.

4 Results and Discussion

Firstly, we only focus on the gap times (i.e. the second terms of F_1 and F_2). Within each age group, a covariate effect of age is included and the mean structure is described in the following ways:

$$\begin{aligned}
\text{(a)} \quad G\left(\mu_{ij}^k\right) &= \text{before} \times \left(\alpha_1 + \beta_1 age_{ij}^k\right) + \text{after} \times \left(\alpha_2 + \beta_2 age_{ij}^k\right) + b_i^k \\
b_i^1, b_i^2 &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{b^1}^2 & \sigma_{b^1 b^2} \\ \sigma_{b^1 b^2} & \sigma_{b^2}^2 \end{pmatrix}\right], \qquad \text{and} \\
\text{(b)} \quad G\left(\mu_{ij}^k\right) &= \text{before} \times \left(\alpha_1 + \beta_1 age_{ij}^k + \gamma_1 s_{ij-1}^k\right) + \\
&\quad \text{after} \times \left(\alpha_2 + \beta_2 age_{ij}^k + \gamma_2 s_{ij-1}^k\right) + b_i^k \\
b_i^1, b_i^2 &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{b^1}^2 & \sigma_{b^1 b^2} \\ \sigma_{b^1 b^2} & \sigma_{b^2}^2 \end{pmatrix}\right],
\end{aligned} \tag{2}$$

with G the considered link function. (identity or log link). The Gamma distribution did not lead to convergency in any of the cases and therefor will not be discussed here. Until now, results for the full structures as described in (2) could be obtained only for the exponential distribution with identity link and without effect of the previous gap time. However, results for all models could be obtained when assuming independency between random effects b_i^1 and b_i^2. Comparing the situation with a general covariance and independency under an exponential distribution, identity link and effect of only age, suggested that this independency assumption might be a good approximation. However, when more results for the general covariance case are available, it should be checked if this indeed is valid.

The Weibull distribution gives smaller AIC values compared to the exponential distribution, a log link behaves better compared to the identity link and including an effect of the previous gap time results in seriously reduced AIC values. In Table 1 parameter estimates and standard errors for the Weibull distribution with a log link and with a mean structure including the previous gap time are summarized. The effect of the covariate age (within an age group) was significant only in age group 4 and in age group 1 before the diagnosis. In the oldest age group, the effect of the age of the member is negative, inducing that in this group, the older patients will make more frequent costs. In the youngest age group, before diagnosis, the effect of age is positive, inducing that in age group 1, the younger members will make more frequent costs. The effect of the previous gap time was

TABLE 1. Parameter estimates and standard errors for the Weibull distribution with a log link and with a mean structure including the previous gap time.

Group	Before			After				
	α_1	β_1	γ_1	α_2	β_2	γ_2	σ_b^2	k^*
P1	0.23	0.06	0.06	0.67	-0.02	0.05	0.13	1.57
	(0.05)	(0.02)	(0.01)	(0.04)	(0.02)	(0.01)	(0.01)	(0.37)
P2	0.66	0.004	0.03	0.75	-0.002	0.05	0.20	1.43
	(0.07)	(0.002)	(0.003)	(0.08)	(0.002)	(0.005)	(0.02)	(0.009)
P3	-0.64	0.02	0.10	-0.58	0.02	0.07	0.22	1.75
	(0.67)	(0.01)	(0.005)	(0.67)	(0.01)	(0.01)	(0.03)	(0.03)
P4	1.95	-0.02	0.07	1.03	-0.01	0.06	0.25	1.98
	(0.41)	(0.005)	(0.004)	(0.41)	(0.005)	(0.01)	(0.03)	(0.01)
M1	0.41	0.05	0.05	0.86	-0.01	0.04	0.15	1.46
	(0.05)	(0.02)	(0.01)	(0.04)	(0.02)	(0.01)	(0.01)	(0.01)
M2	0.85	-0.001	0.03	0.82	-0.001	0.03	0.20	1.43
	(0.07)	(0.002)	(0.004)	(0.08)	(0.002)	(0.004)	(0.02)	(0.009)
M3	1.79	-0.02	0.03	1.16	-0.01	0.03	0.23	1.58
	(0.69)	(0.01)	(0.004)	(0.69)	(0.01)	(0.006)	(0.03)	(0.03)
M4	2.22	-0.02	0.08	1.78	-0.02	0.05	0.19	1.85
	(0.35)	(0.005)	(0.004)	(0.36)	(0.005)	(0.007)	(0.02)	(0.01)

highly significant in all cases ($p < 0.0001$). Estimates for this effect were positive, meaning that a large gap time between the previous cost and the current cost will result in a large mean gap time between the current cost and the next cost. The random effects were found to be significant in all cases, with p values all < 0.0001 using a mixture of χ^2 distributions.
Figure 1 shows the average observed and average predicted profiles in function of the previous gap times for age group 1. Predictions are made using the model that results from Table 1: the covariate age is included only before diagnosis. The average predicted profiles approximate the average observed profiles well in most cases. For longer previous gap times, predicted averages can deviate from observed averages. This might suggest there is still room for improvement in the model. Possible generalizations include different random effects before and after diagnosis, and the inclusion of random slopes.
Secondly, the full joint model as described in (1) is fitted. However, models are still running and results could not be obtained yet. More details will be given in the presentation.

References

Creemers, A., Aerts, M., Hens, N., Shkedy, Z., De Smet, F., and Beutels, P. (2011). Revealing age-specific past and future unrelated costs of pneumococcal infections by flexible generalized estimating equations. *Journal of Applied Statistics* **00**, 000-000.

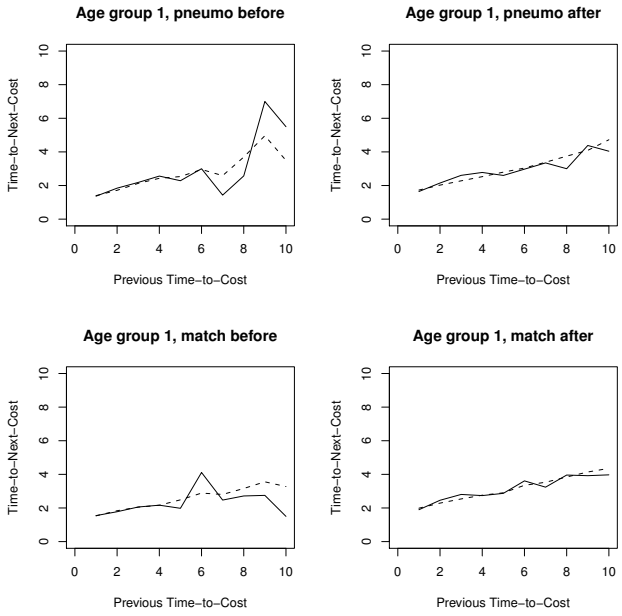


FIGURE 1. Average observed (full lines) and average predicted (dotted lines) profiles in the diagnosed group and the undiagnosed group for age groups 1, before and after the diagnosis.

Bartlett–type Correction in Heteroscedastic Symmetric Nonlinear Models

Audrey H. M. A. Cysneiros[1]

[1] Departamento de Estatística – Universidade Federal de Pernambuco, Recife – PE, Brazil – e-mail: audrey@de.ufpe.br

Abstract: We present simple matrix formulae for corrected score statistics in heteroscedastic symmetric nonlinear regression models, with link functions for both mean and dispersion parameter. We compare the sizes and the powers of the corrected score tests with original score test.

Keywords: Bartlett-type correction; Heteroscedastic model; Nonlinear model; Score statistic.

1 Heteroscedastic Symmetric Nonlinear Models

We consider an heteroscedastic symmetric nonlinear model where both mean and dispersion parameters vary across observations through nonlinear regression structures. Homoscedasticity of the dispersion parameter is a common assumption in nonlinear models. However, this may not be appropriate in some situations and for others may not show the dependence of the dispersion parameter on covariates available in the data. This type of heteroscedastic regression has been discussed in many areas of applied statistics. The random variables $Y_1, \ldots, Y_n$ are assumed to be independent, and each observation Y_ℓ has a symmetric density with mean parameter $\mu_\ell \in I\!R$ and dispersion parameter $\phi_\ell > 0$ given by $\pi(y; \mu_\ell, \phi_\ell) = \frac{1}{\sqrt{\phi_\ell}}\, g(u_\ell)$, $y \in I\!R$, where $g : I\!R \to [0, \infty)$ is such that $\int_0^\infty g(u)du < \infty$ and $u_\ell = \phi_\ell^{-1}(y_\ell - \mu_\ell)^2$. The function $g(\cdot)$ is typically known as the density generator. We will denote $Y_\ell \sim \mathrm{S}(\mu_\ell, \phi_\ell, g), \ell = 1, \ldots, n$. The symmetrical class includes all symmetrical continuous distributions with heavier and lighter tails than the normal ones. First, we assume that the mean response is $\mu = (\mu_1, \ldots, \mu_n)^\top$ with $\mu_\ell = f(x_\ell; \beta)$, where $x_\ell = (x_{\ell 1}, \ldots, x_{\ell m})^\top$ is an $m \times 1$ vector of known explanatory variables associated with the ℓth response, $f(\cdot; \cdot)$ is a twice continuously differentiable function in β and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown regression parameters to be estimated. We also assume that β is defined in a subset $\in \Omega_\beta$ of $I\!R^p$ $(p < n)$. Furthermore, the $n \times p$ matrix of derivatives of μ with respect to β, denoted by $\widetilde{X} = \partial\mu/\partial\beta$, is assumed to be of full rank, i.e., $\mathrm{rank}(\widetilde{X}) = p$ for all β. Second, we introduce a systematic component for the dispersion parameter vector $\phi = (\phi_1, \ldots, \phi_n)^\top$

given by $\phi_\ell = h(\tau_\ell)$, where $h(\cdot)$ is a known one-to-one continuously differentiable function of the dispersion linear predictor defined by $\tau_\ell = z_\ell^\top \gamma$, where $z_\ell = (z_{\ell 1}, \cdots, z_{\ell q})^\top$ is a $q \times 1$ vector of explanatory variables that may have components in common with x_ℓ and $\gamma = (\gamma_1, \cdots, \gamma_q)^\top$ is a vector of unknown parameters to be estimated. The function $h(\cdot)$ is usually called dispersion link function and should be a positive-value function. One possible choice for $h(\cdot)$ is $h(\tau) = \exp(\tau)$. We introduce the following notation: $\delta_{abcde} = E\{t^{(1)^a} t^{(2)^b} t^{(3)^c} t^{(4)^d} z^e\}$ for $a, b, c, d, e = 0, 1, 2, 3, 4$, where $t^{(r)} = \mathrm{d}^r t(z)/\mathrm{d}z^r$ and $t(z) = \log h(z^2)$. Fisher's information matrix for (β, γ) is block diagonal and is given by $K = \mathrm{diag}\{K_\beta, K_\gamma\}$, where $K_\beta = \delta_{(2,0,0,0,0)} \widetilde{X}^\top \Lambda \widetilde{X}$ with $\Lambda = \mathrm{diag}\{\phi_1^{-1}, \ldots, \phi_n^{-1}\}$ and $K_\gamma = \widetilde{P}^\top V \widetilde{P}$, $V = \mathrm{diag}\{v_1, \ldots, v_n\}$ with $v_i = \frac{(\alpha_{2,0,0,0,2}-1)h_i'^2}{4\phi_i^2}$ and $\widetilde{P}$ is the $n \times q$ matrix of derivatives of ϕ with respect to γ, denoted by $\widetilde{P} = \partial\phi/\partial\gamma$. The parameters β and γ are globally orthogonal and then the MLEs $\hat{\beta}$ and $\hat{\gamma}$ are asymptotically independent. A nonlinear optimization method, such as Fisher's scoring algorithm, is needed for obtaining $\hat{\beta}$ and $\hat{\gamma}$; see Cysneiros et al. (2010).

2 Improved score tests

The basic idea of transforming the score test statistic in such a way that it becomes better approximated by the reference chi-squared distribution is due to Cordeiro and Ferrari (1991). The corrected score statistic S_R^* proposed by these authors is given by $S_R^* = S_R\{1 - (c + bS_R + aS_R^2)\}$, where the coefficients a, b and c are of order n^{-1} and come from the expansion of the distribution function of S_R under the null hypothesis given by Harris (1985). Also, the coefficients a, b and c depend on the functions of joint cumulants of log-likelihood derivatives up to the fourth order. The null distribution of S_R^* is chi-squared with approximation error reduced from order $O(n^{-1})$ to $O(n^{-3/2})$. The improved statistic S_R^* is not always a monotone transformation. To overcome this, Kakizawa (1996) suggested the monotone transformation $K(S_R) = S_R^* + P(S_R)$, where $P(S_R) = \frac{1}{4}\left\{c^2 S_R + 2bcS_R^2 + \left(2ac + \frac{4}{3}b^2\right) S_R^3 + 3abS_R^4 + \frac{9}{5}a^2 S_R^5\right\}$. Also, Cordeiro et al. (1998) found an alternative formula to the modified score statistic S_R^*, which is a monotone transformation of S_R. If $a = 0$ and $b \neq 0$, the alternative statistic, $\widetilde{S}_R$, is given by $\widetilde{S}_R = \frac{1}{2b}\exp(-c)\{1 - \exp(-2bS_R)\}$. If $a = b = 0$, S_R^* is a monotone transformation of S_R and there is no need to define an alternative corrected statistic. The three statistics S_R^*, $K(S_R)$ and $\widetilde{S}_R$ are equivalent up to order n^{-1}, i.e., they typically differ by $O_p(n^{-3/2})$. Partitioned the parameters vectors γ as $\gamma = ({\gamma_1}^\top, {\gamma_2}^\top)^\top$, where $\gamma_1 = (\gamma_1, \cdots, \gamma_{q_1})^\top$ is a vector of parameters of interest, $\gamma_2 = (\gamma_{q_1+1}, \cdots, \gamma_q)^\top$ and $\beta = (\beta_1, \cdots, \beta_p)^\top$ are nuisance parameters. We are interested in testing the null hypothesis $\mathcal{H}_0 : \gamma_1 = \gamma_1^{(0)}$ against the al-

ternative hypothesis $\mathcal{H}_1 : \gamma_1 \neq \gamma_1^{(0)}$ where $\gamma_1^{(0)}$ is a specified vector of dimension q_1 ($q_1 \leq q$). Corresponding to this partition, we write $\widetilde{P} = (\widetilde{P}_1, \widetilde{P}_2)$ where $\widetilde{P}_1$ and $\widetilde{P}_2$ are full rank matrices with dimensions given by $n \times q_1$, $n \times (q - q_1)$, respectively. We can now express the score statistic S_R for testing $\mathcal{H}_0$ in the heterocedastic symmetric nonlinear model as $S_R = \widetilde{\zeta}^\top \widetilde{P}_1 \left(\widetilde{P}_1^\top V \widetilde{P}_1\right)^{-1} \widetilde{P}_1^\top \widetilde{\zeta}$, with the functions being evaluated at $(\gamma_1^{(0)\top}, \tilde{\gamma}_2^\top, \tilde{\beta}^\top)$. Here, the quantities above are $S = \mathrm{diag}\{\mathrm{s}_1, \cdots, \mathrm{s}_\mathrm{n}\}$, $s_l = \frac{-2g'(u_l)}{g(u_l)}$, $\zeta = (SF_1u - F_1\iota)(\delta_{(2,0,0,0,2)} - 1)^{-1/2}\Lambda$, $u = (u_1, \cdots, u_n)^\top$, $u_l = \frac{(y_l - \mu_l)^2}{\phi_l}$, $F_1 = \mathrm{diag}\{\mathrm{h}_1', \cdots, \mathrm{h}_\mathrm{n}'\}$ where the primes denote differentiation with respect to τ and ι is an $n \times$ vector of the ones, $l = 1, \cdots, n$. The general expressions for the $A's$ consider a test on all elements of γ, i.e., $\gamma = \gamma^{(0)}$ become $A_1 = b_1 a_0\, \iota^\top \Lambda_4 Z_{\beta d} Z_\gamma Z_{\beta d} \Lambda_4 \iota + 3a_1\, \iota^\top \Lambda_4 Z_\beta \odot Z_\gamma \odot Z_\beta \Lambda_4 \iota + b_2\, \mathrm{tr}\{\Lambda_7 Z_{\beta d} Z_{\gamma d}\}$, $A_2 = b_3 a_3\, \iota^\top \Lambda_1 Z_{\gamma d} Z_\gamma Z_{\gamma d} \Lambda_1 \iota - a_4\, \iota^\top \Lambda_4 Z_{\beta d} Z_\gamma Z_{\gamma d} \Lambda_1 \iota + a_6\, \mathrm{tr}\{\Lambda_9 Z_{\gamma d}^{(2)}\}$, and $A_3 = -b_9 a_3 \iota^\top \Lambda_1 Z_{\gamma d} Z_\gamma Z_{\gamma d} \Lambda_1 \iota - \frac{2b_9}{3} a_3 \iota^\top \Lambda_1 Z_\gamma^{(3)} \iota$, where $Z_\beta = \delta_{(2,0,0,0,0)}^{-1} X(X^\top \Lambda X)^{-1} X^\top$, $Z_\gamma = \delta_{(2,0,0,0,0)}^{-1} \tilde{P}(\tilde{P}^\top V \tilde{P})^{-1} \tilde{P}^\top$, $a_0 = -\delta_{(0,0,1,0,1)} + 2\delta_{(0,1,0,0,0)}$, $a_1 = (\delta_{(3,0,0,0,1)} - \delta_{(1,0,0,0,1)})$, $a_2 = \delta_{(0,1,0,0,2)} - \delta_{(0,0,1,0,3)}$, $a_3 = (1 - 3\delta_{(0,1,0,0,2)} - \delta_{(0,0,1,0,3)})$, $a_4 = \frac{4\, b_1\, \delta_{(2,0,0,0,0)}}{(\delta_{(2,0,0,0,2)} - 1)} a_3$, $a_6 = \frac{3}{(\delta_{(2,0,0,0,2)} - 1)^2} a_5$ with $a_5 = -6 + \delta_{(2,0,0,0,2)}(12 - 3\delta_{(2,0,0,0,2)}) + 4\delta_{(3,0,0,0,3)} + \delta_{(4,0,0,0,4)}$, $b_1 = \frac{3(\delta_{(1,1,0,0,1)} - \delta_{(0,1,0,0,1)})}{(\delta_{(2,0,0,0,2)} - 1)\delta_{(2,0,0,0,0)^2}}$, $b_3 = \frac{-12(1 - 3\delta_{(0,1,0,0,2)} - \delta_{(0,0,1,0,3)})}{(\delta_{(2,0,0,0,2)} - 1)^2}$ and $b_2 = \frac{-6(\delta_{(2,0,0,0,2)}\delta_{(0,1,0,0,2)} + 2\delta_{(3,0,0,0,1)} + \delta_{(4,0,0,0,2)})}{(\delta_{(2,0,0,0,2)} - 1)\delta_{(2,0,0,0,0)^2}}$.

3 Numerical Evidence

In Table 1, we report some simulation results in order to compare the sizes of the usual score test and of the tests based on the following modified score statistics: S_R^*, $K(S_R)$ and $\widetilde{S}_R$. We use the following nonlinear regression model that assumes the predictor: $\eta_\ell = \beta_0 + \beta_1 x_{1\ell} + \exp(\beta_2 x_{2\ell})$ with $\phi_\ell = \exp(z_i^\top \gamma)$ being $\tau = \gamma_1 + \gamma_2 z_{2i\ell} + \gamma_3 z_{3\ell}$, $\ell = 1, \cdots, n$. The null hypothesis that we consider is $\gamma_1 = 0$ and the response was generated from a type–I logistic distribution. The independent variable x_1, x_2, z_2 and z_2 were chosen as random draws from a uniform U(0,1) distribution and their values were held fixed throughout the simulations with equal sample sizes. Ten thousand samples of 30, 35, 40, 45 and 100 observations were generated with $\beta_1 = 5$, $\beta_2 = 2$, $\beta_3 = 1$, $\gamma_2 = 0.3$ and $\gamma_3 = 0.5$. Table 1 displays the null rejection rates of the three tests for 10% and 5% nominal levels (α). The Table 1 reveal important information. The score test is largely conservative for small samples and the corrections are really necessary for small and moderate sample sizes. Simulation studies (omitted here) they show that the powers of the three corrected tests are similar and larger than the power of the original score test.

TABLE 1. Size simulations: rejection rates of the score and three corrected score; entries are percentages.

n	α	Type I logistic model			
		S_R	S_R^*	$K(S_R)$	$\tilde{S}_R$
30	5	2.3	2.7	2.8	2.8
	10	5.2	6.2	6.4	6.3
35	5	2.5	3.2	3.3	3.3
	10	5.2	7.2	7.4	7.3
40	5	2.7	3.5	4.1	4.0
	10	6.4	7.7	7.9	7.8
45	5	2.8	3.9	4.5	4.4
	10	6.5	8.0	8.1	8.1
100	5	4.1	4.5	4.9	4.9
	10	8.9	9.9	10.0	9.9

Acknowledgments: Financial support from L'oréal/ABC/UNESCO, CNPq and FACEPE - Brazil is gratefully acknowledged.

References

Cordeiro, G.M., Ferrari, S.L.P., and Cysneiros, A.H.M.A. (1998). A formula to improve score test statistics. Journal of Statistical Computation and Simulation, **61**, p. 123-136.

Cordeiro, G.M., and Ferrari, S.L.P. (1991). A modified score test statistic having chi-squared distribution to order n^{-1}. *Biometrika*, **78**, 573–582.

Cysneiros, F.J.A., Cordeiro, G.M. and Cysneiros, A.H.M.A. (2010). Bias-Corrected Maximum Likelihood Estimators in Nonlinear Heteroscedastic Models. *Journal of Statistical Computation and Simulation*, 80, **4**, 451–461.

Harris, P. (1985). An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika*, **72**, 653–659.

Kakizawa, Y. (1996). Higher order monotone Bartlett–type adjustment for some multivariate test statistics. *Biometrika*, **83**,923–927.

A Symbolic Robust Regression Model

Francisco José A. Cysneiros[1], Roberta A. A. Fagundes[2 3] , Renata M. C. R. de Souza[3]

[1] Departamento de Estatística, Universidade Federal de Pernambuco, Recife PE-Brazil e-mail: cysneiros@de.ufpe.br
[2] Universidade de Pernambuco, Recife PE-Brazil, e-mail:raaf@cin.ufpe.br
[3] Centro de Informática, Universidade Federal de Pernambuco, Recife PE-Brazil email:rmcrs@cin.ufpe.br

Abstract: This paper proposes a robust regression method for large data sets using symbolic data analysis. Large standard data sets are transformed into symbolic interval data sets based on a generalization process. Each interval of the input data is described by range and mid-point variables. To validate this model, experiments to software size estimation using two large data projects from the NASA repository are considered. The prediction quality is assessed by a mean magnitude of relative errors calculated from test data sets.

Keywords: symbolic data analysis; robust regression; software size estimation.

1 Introduction

Due to the explosive growth in the use of databases, new approaches have been proposed for discovering regularities and summarizing information stored in large data sets. In real-world applications of decision making is usual that inaccuracy, uncertainty or variability must be taken into account to represent available information. In these cases, classical data are not able to represent these nuances and other kinds of data, such interval-valued data are required. *Symbolic Data Analysis* (SDA) has been introduced as a new domain related to multivariate analysis, pattern recognition and artificial intelligence for extending classical exploratory data analysis and statistical methods to symbolic data. SDA (Diday and Noirhomme-Fraiture (2008)) aims allows multiple (sometimes weighted) values for each variable and new variable types (interval, categorical multi-valued and modal variables) have been introduced. In particular, SDA is a powerful tool to represent units that may be derived through a definition or an aggregation data and can be applied to situation where inaccuracy and uncertainly must be by considering to faithfully represent the real world.
This paper introduces a regression approach for large data sets using a robust model for interval data. Here, this regression model is over a statistical view of learning in an application with software size estimation. An

accurate estimate of software size is an essential element in the calculation of estimated project costs and schedules. Two large projects of the NASA data base containing software modules are considered in this application. Initially, each large data set is preprocessed in order to generate interval data from of standard data. The intervals are formed through an aggregation way using a discrete variable. In the following, the robust regression method is applied to these interval data sets. A comparative study between (linear and robust) regression methods for interval data is carried out. The performance of the methods is measured by the prediction accuracy that is assessed based on the mean magnitude of relative errors ($MMRE$).

2 Generation of symbolic data

Software estimation (Bielak (2000)) is responsive to the widespread problems the software industry has experienced in creating meaningful cost and schedule estimates. Two traditional size measures for estimating are source lines of code and function points. The lines of code that a project generates are strongly influenced by the software languages used, individual coding style, and organizational standards. The NASA repository (*http://mdp.ivv.nasa.gov/*) contains 13 projects. Each project is formed by software modules described by a set of variables. Here, we consider the projects $MC1$ and $PC2$ of sizes: 9466 and 5586 respectively. The project $MC1$ is a combustion experiment that is designed to fly on the space shuttle. This project consists of more than 63 $KLOC$ of C and $C++$ codes. The project $PC2$ is a dynamic simulator for attitude control systems. It consists of 26 $KLOC$ of C code. In this work, the software size estimation for these projects is based on symbolic regression model in which the number of code lines NL is used as the response variable and the predictor variables are: number of operators NO, number of operands NA and branch count BC. Interval data can be generated from standard data according to a generalization process based on the difficulty level (LD) variable. This variable has a variability that allows us to represent maximum and minimum values and to obtain intervals. Therefore, the numerical variables describing project modules lead to interval variables describing groups of project modules. Consider the intervals $[a, b]$, $[c, d]$, $[e, f]$, $[\alpha, \beta]$ and a value v of the variable LD. The symbolic description of a group of project modules is accomplished in the following way: given a value v of the variable LD compute $a = min\{NO\}$, $b = max\{NO\}$, $c = min\{NA\}$, $d = max\{NA\}$, $e = min\{BC\}$, $f = max\{BC\}$ $\alpha = min\{NL\}$ and $\lambda = max\{NL\}$ $\forall i$ such that $LD = v$. In the situations in which the frequency of $LD = v$ is equal 1, we aggregate the data by groups of values of the variable LD. After generalization process, the projects $MC1$ and $PC2$ concerns of 112 and 61 groups of software modules, respectively. For example, the project $PC2$ is a description of: $(NO = [5, 17]; NA = [3, 9]; BC = [3, 6]; NL = [18, 81])$ for $LD = 10$.

3 Robust regression model for interval-valued data

Let $\Omega = 1, \ldots, n$ be a data set of n objects described by the response interval-valued variable Y and p predictor interval-valued variable (X_1, $\ldots, X_p$). Each object i of Ω is represented as an interval feature vector $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ where $x_{ij} = [a_{ij}, b_{ij}] \in \Im = \{[a, b] : a, b \in \Re, a \leq b\}$ $(j = 1, \ldots, p)$ and $y_i = [\alpha_i, \lambda_i] \in \Im$.
Let Y^c and X_j^c and Y^r and X_j^r be, respectively, quantitative variables that describe the midpoints and the ranges of the intervals $y_i \in Y$ and $x_{ij} \in X_j$ $(j = 1, 2, \ldots, p)$. This means that each example $\Omega = 1, \ldots, n$ is represented by pairs $\mathbf{v}_i = (\mathbf{x}_i^c, y_i^c)$ and $\mathbf{r}_i = (\mathbf{x}_i^r, y_i^r)$ with $\mathbf{x}_i^c = (x_{i1}^c, \ldots, x_{ip}^c)$ and $\mathbf{x}_i^r = (x_{i1}^r, \ldots, x_{ip}^r)$ where $x_{ij}^c = [a_{ij} + b_{ij}]/2$, $x_{ij}^r = b_{ij} - a_{ij}$, $y_i^c = [\alpha_i + \lambda_i]/2$ and $y_i^r = \lambda_i - \alpha_i$ are respectively, the observed values of X_j^c, X_j^r, Y^c and Y^r.
Consider $\boldsymbol{\beta}^c = (\beta_0^c, \beta_1^c, \ldots, \beta_p^c)'$ and $\boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, \ldots, \beta_p^r)'$ as being two vectors of $p + 1$ parameters and $\boldsymbol{\varepsilon^c} = (\varepsilon_1^c, \varepsilon_2^c, \ldots, \varepsilon_n^c)'$ and $\boldsymbol{\varepsilon^r} = (\varepsilon_1^r, \varepsilon_2^r, \ldots, \varepsilon_n^r)'$ as being two vectors of n unknown errors on the midpoint and range of the intervals. Two linear regression equations, respectively, on midpoint and range values are given by: $y_i^c = \mathbf{x}_i^{'c}\boldsymbol{\beta}^c + \varepsilon_i^c$ and $y_i^r = \mathbf{x}_i^{'r}\boldsymbol{\beta}^r + \varepsilon_i^r$.
The vectors $\boldsymbol{\beta}^c$ and $\boldsymbol{\beta}^r$ are estimated minimizing a criterion function based on a function ρ for both the residuals $e_i^c = y_i^c - \mathbf{x}_i^{'c}\hat{\boldsymbol{\beta}}^c$ and $e_i^r = y_i^r - \mathbf{x}_i^{'r}\hat{\boldsymbol{\beta}}^r$.
The function ρ is related to the likelihood function for an appropriate choice of the error distribution. Here, both the errors ε_i^c and ε_i^c are independent and identically distributed according to a distribution $L(\cdot/\sigma)$ where σ is a scale parameter (usually unknown). The criterion function is given by $\sum_{i=1}^{n} \rho\left(\frac{\epsilon_i^c}{s}\right) + \rho\left(\frac{\epsilon_i^r}{s}\right)$ where s is a robust estimate of scale and ρ is particular function. The Fisher scoring method can be easily applied to get $\hat{\boldsymbol{\beta}}^c$ and $\hat{\boldsymbol{\beta}}^r$ that one can be interpreted as a modified least square.
The i-th prediction of the lower and upper bounds $\hat{y}_i = [\hat{\alpha}_i, \hat{\lambda}_i]$ of a new example is based on the prediction of $\hat{y}^c$ and $\hat{y}^r$. Given a interval vector $\mathbf{x}_i = ([a_{i1}, b_{i1}], \ldots, [a_{ip}, b_{ip}])$ with $x_{ij}^c = (a_{ij} + b_{ij})/2$ and $x_{ij}^r = b_{ij} - a_{ij}$ $(i = 1, \ldots, n)$ $(j = 1, \ldots, p)$, the interval $\hat{y}_i = [\hat{\alpha}_i, \hat{\lambda}_i]$ is obtained as follows: $\hat{\alpha}_i = \hat{y}_i^c - \hat{y}_i^r/2$ and $\hat{\lambda}_i = \hat{y}_i^c + \hat{y}_i^r/2$ where $\hat{y}_i^c = \mathbf{x}_i^{'c}\hat{\boldsymbol{\beta}}^c$ and $\hat{y}_i^r = \mathbf{x}_i^{'r}\hat{\boldsymbol{\beta}}^r$.
An experimental evaluation of the robust regression method for interval-valued data developed in this work using two NASA projects that were preprocessed in order to transform quantitative data into interval data. Moreover, a comparative study regarding this robust regression method and the linear regression one introduced in Lima Neto and De Carvalho (2008) is also discussed.
The accuracy prediction of the method is measured by the mean magnitude of relative error ($MMRE$) that is estimated by the hold-out method in the framework of a Monte Carlo simulation with 200 replications. According to the robust regression method using Tukey's biweight criterion function and the linear regression method introduced in Lima Neto and De Carvalho

TABLE 1. $MMRE$ for interval data sets

Project	Linear Regression	Robust Regression
$MC1$	0.16 ± 0.08	0.08 ± 0.04
$PC2$	0.18 ± 0.22	0.12 ± 0.14

(2008) using least squares criterion function, the $MMRE$ is given by

$$MMRE = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left\{\left|\frac{\alpha_i - \hat{\alpha}_i}{\alpha_i}\right| + \left|\frac{\lambda_i - \hat{\lambda}_i}{\lambda_i}\right|\right\}.$$

Table 1 shows the average and the standard deviation of the $MMRE$ for projects $MC1$ and $PC2$ considering the linear and robust regression methods for test data sets (25% of the interval data set). The results in this table points out that, the robust method is the best option in terms of $MMRE$.

4 Conclusion

In this paper, we have introduced a robust regression model for large data sets using symbolic data analysis. Here, a generalization processing is applied to large point data sets in order to obtain symbolic interval data sets. Experiments using two large data projects of the well-known NASA data set for software size estimation were carried out. We have compared the proposed regression method with a linear regression one and the results showed that the robust regression model is better than the linear regression one in terms of prediction quality. In addition, the use of symbolic interval data allowed to describe projects modules taking into account variability.

Acknowledgments: This study was supported by CNPq and Facepe, Brazil.

References

Bielak, J. (2000). Improving size estimates using historical data. *IEEE Computer Society*, **17**, 27-35.

Diday, E., Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley-Interscience.

Lima Neto, E.A., De Carvalho, F.A.T. (2008). Centre and Range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis*, **52**, 1500-1515.

Bayesian inference for copula based GARCH models

Claudia Czado[1], Andreas Dill[1] , Mathias Hofmann[1]

[1] Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, 85748 Garching b. München; corresponding e-mail: *cczado@ma.tum.de.*

Abstract: Pair copula constructions (PCCs) allow for the construction of flexible multivariate copulas. These multivariate copulas are formulated only using bivariate copula terms. One such PCC class is the class of D-vines, which allows to incorporate asymmetric and tail dependencies for different pairs of variables. In this paper we combine D-vines for modelling the residual dependency among stock indices after marginal time dependencies are captured by univariate GARCH margins. We follow a Bayesian approach, where marginal and copula parameters are estimated jointly in a MCMC setup and show how they can be used to quantitify value-at-risk. Model selection of the D-vine structure as well as the family of bivariate copula families is discussed and illustrated by the analysis of four major stock indices. Comparison to corresponding standard GARCH models show the superiority of the discussed models.

Keywords: multivariate copula; D vines; GARCH, value-at-risk

1 Introduction

Pair copula constructions (PCCs) (see Kurowicka and Cooke (2006), Aas et. al. (2009), Czado (2010) and Kurowicka and Joe (2011) and references therein) have become quite popular choices for multivariate copulas, since they allow to construct very flexible dependency models and thus extending standard multivariate copula models such as elliptical copulas. In particular they can be used, when different pairs of variables exhibit different asymmetric and tail dependencies. They are defined by a sequence of trees, which determine pairs of variables together with a set of variables. The distribution of these variables conditioned by the indicated set of variables are subsequently modeled by a bivariate copula. These bivariate copula terms can be chosen arbitrarily from a large catalogue of bivariate copulas. The resulting joint density can be written as a product of the corresponding bivariate copula densities. The structure of the allowable trees are quite general allowing for many different PCC's. Both the normal as well as the multivariate Student t-copulas with common degree of freedom are special cases. One particular simple structure are the ones corresponding to D-vine

copulas. The resulting d dimensional D-vine copula density is given by

$$f(u_1,\ldots,u_d) = \prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{i,(i+j)|(i+1)\cdots,(i+j-1)}(F(u_i|u_{i+1},\cdots,u_{i+j-1}),F(u_{i+j}|u_{i+1},\cdots,u_{i+j-1})) \quad (1)$$

for $c_{r,s|i_1,\cdots,i_l}(\cdot,\cdot)$ an arbritrary chosen bivariate copula density. The arguments in (1) are conditional cdf's, which can be recursively determined in PCC models. In the application we utilize as bivariate copulas normal, Student t-, BB1 and BB7 copulas, respectively. The BB1 and BB7 copulas of Joe (1997) have the advantage that they allow for different tail dependence in the upper and lower tail in contrast to symmetric tail dependence for the Student t-copula or no tail dependence of the Gauss copula. For some bivariate copula components also the independence copula is chosen.

2 D-vine copula based GARCH models

To utilize copulas for the analysis of financial time series data, we first have to remove the marginal time dependencies before we can construct an i.i.d distribution to be modelled by a copula model. For modeling the time dependence within each margin we use separate GARCH(1,1) models with t innovations. The corresponding d dimensional vector of standardized innovations at time t are now i.i.d distributed over different time points. The margins are t distributed with the marginal degree of freedom determined by the corresponding univariate GARCH model. The dependency among the components of the standardized innovation vector is modeled by a D-vine copula.

For the complete specification of a D-vine copula we have to specify the D-vine tree structure and the family of bivariate copulas to be chosen for each bivariate copula term. From (1) we see that for $j = 1$ we model unconditional dependencies between pairs of variables and that the computational complexity of the likelihood increases as j increases, since the recursive calculations of the conditional cdf's increases. Therefore it is desirable to model those pairs of variables with highest dependence (as measured for example with Kendall's τ) directly in (1) for $j = 1$. Therefore we find an order of the variables from 1 to d, such that the dependency between (U_i, U_{i+1}) is large for many $i = 1,\cdots,d$. This specifies the bivariate copula terms to be modeled. For the unconditional $c_{i,i+1}$ we use for example empirical contour plots with standard normal margins based on data $(u_{i,t}, u_{i+1,t}, t = 1,\cdots,T$. Once these families are chosen then the corresponding parameter value is estimated using for example the inversion of the empirical Kendall's τ value. Once all unconditional bivariate copula families are determined (denoted by $S_{i,i+1}$ and their parameter values estimated (denoted by $\hat{\theta}_{i,i+1}$), we create pseudo

data given by $(F(u_{i,t}|u_{i+1,t}, S_{i,i+1}, \hat{\theta}_{i,i+1}$ and $(F(u_{i+1,t}|u_{i,t}, S_{i,i+1}, \hat{\theta}_{i,i+1}$ for $i = 1, \cdots, d-1$ and use this data in a similar way to choose the copula family and its parameter estimate for $c_{i,i+2|i+1}$. We continue in this way until all bivariate copula families are determined. This sequential proceeding gives a first set of parameter estimates, which are used as starting values in maximum likelihood estimation (see e.g. Aas et.al (2009)). We could also use these to construct prior distributions for the copula parameters in a Bayesian set up.
A major advantage of the D-vine copula based GARCH model over the standard multivariate CCC GARCH model is that asymmetric dependency effects can be captured and the model specification allows for independent choices of the copula terms, while in a CCC GARCH model extra care is needed to achieve the a positive definite correlation matrix.

3 Bayesian inference for D-vine copula based GARCH models

For statistical inference D-vine copula based GARCH models have to estimate marginal and copula parameters. Commonly in classical statistics a two step approach is taken, i.e. first the marginal parameters are estimated separately ignoring the dependency between the margins. In a second step the likelihood of the copula parameters is considered, where the marginal parameters are set to their estimated values to reduce the dimension of the optimisation. This might introduce bias. To facilitate joint estimation we follow a Bayesian approach. This allows us to construct credible intervals, while confidence intervals are difficult to obtain due to the possible non positive definiteness of the Hessian matrix. A further advantage of a Bayesian approach is that we can assess the uncertainty of derived quantities such as the value-at-risk.
We developed a Markov Chain Monte Carlo (MCMC) algorithm for joint estimation of all parameters. Performance was improved by using a joint update for marginal GARCH parameters. More details can be found in Hofmann and Czado (2010).

4 Application

For daily stock indices (DAX, S&P500, Nikkei 225, MSCI) from March 31, 1999 until December 15, 2009 were chosen for our analysis. For determining appropriate D-vine copula based GARCH models we fitted 4 separate GARCH(1,1) models with Student t-innovations. Then we applied the probability integral transform to standardized residuals based on a univariate t-distribution with fitted degree of freedom for each margin. The resulting four dimensional copula data is now used to allow for determine 5 plausible D-vine copula models involving 2 D-vine tree structures,

bivariate t-copulas for copula terms and copula terms with different bivariate copula specifications using the considerations for model selection given before. As benchmark models we considered three CCC GARCH models, each with GARCH(1,1) margins and t-innovations together with a multivariate Gauss, t-copula with common degree of freedom and the independence copula, respectively. For these benchmark models we also developed a Bayesian MCMC algorithm to jointly estimate all marginal and copula parameters. To allow for time dependence, we selected 5 low and 5 high volatility periods and fitted separate models for each period.
The Bayesian estimation results for all investigated models and sub periods were compared based on the deviance information criterion showing that the D-vine based models are superior to the benchmark models especially in periods of high volatility. Bivariate t-copulas are often sufficient as choice for the copula families. Extensive backtests for the value-at-risk shows that the D-vine based models are clearly preferred over the benchmark models especially in high volatility periods. More detailed results can be found in Dill (2010).

5 Conclusions

This work extends the Bayesian D-vine copula estimation of Min and Czado (2010) to joint estimation of marginal and copula parameters and considers more than 2 dimensions as was done by Ausin and Lopes (2010). The application shows that they are useful extensions in the context of multivariate financial time series data. Extensions to include time varying copula effects will be pursued in the future.

References

Aas, K., Czado, C. , Frigessi, A. and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance Mathematics and Economics*,**44 (2)**, 182-198.

Ausin, A. and Lopes, H. (2010). Time-varying joint distributions through copulas, *Computational Statistics and Data Analysis*, **54**, 2383-99.

Czado, C. (2010). Pair-copula constructions of multivariate copulas, P. Jaworki, F. Durante, W. Härdle and W. Rychlik, (Ed.) *Workshop on Copula Theory and its Applications*, Springer, Dordrecht, 93-103.

Dill, A. (2010). *Bayesian value-at-risk calculations in copula based GARCH models*. Diploma Thesis, Technische Universität München, Germany.

Hofmann, M., and Czado, C. (2010). *Assessing the VaR of a portfolio using D-vine copula based multivariate GARCH models*. preprint, Technische Universität München, Germany.

Joe, H. (1997). *Multivariate models and dependence concepts*. Chapman & Hall, London

Kurowicka, D., and Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Chichester: Wiley.

Kurowicka, D., and Joe, H.(Ed.) (2011). *Dependence Modeling - Handbook on Vine Copulae*, World Scientific, Singapore.

Min, A. and Czado, C. (2010). Bayesian Inference for Multivariate Copulas using Pair-copula Constructions. *Journal of Financial Econometrics*, **8(4), 511-546.**

Bayesian Dose Escalation in phase I studies of Combinations of Drugs with Control

David Dejardin[1], Paul Hamberg[2], Jaap Verweij[3] , Emmanuel Lesaffre [1 4]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics Katholieke Universiteit Leuven & Universiteit Hasselt, David.Dejardin@med.kuleuven.be
[2] Department of Internal Medicine, Sint Franciscus Gasthuis, Rotterdam, The Netherlands
[3] Department of Medical Oncology, Erasmus MC, Rotterdam, The Netherlands
[4] Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands

Abstract: The vast majority of Phase I trials in oncology are using the classical 3+3 design. This design has been criticized for providing rather crude estimates of the Maximum Tolerated Dose (MTD), the estimation of this MTD being the primary objective of phase I trials. Cancer treatments often use combination of agents to provide better activity, and many studies thus involve the combination of a novel agent with an existing one (standard of care). A major challenge is to estimate the MTD in the context of a relatively high incidence of toxicity (Hamberg 2010). When assessing the MTD in such studies, we propose a design that takes advantage a standard of care is involved to obtain 1) a more accurate estimate of the MTD and 2) more accurate information on the toxicity of the combination compared to the control. Our design (BDED) is based on the Bayesian dose escalation design (O'Quigley 1990) and randomizes subjects to a novel combination or a control group. We show that BDED provides a better estimate of the MTD as well as a highest posterior density (HPD) interval for the difference of probability of toxicity.

Keywords: Phase I design, Bayesian dose escalation, adaptive design

1 Introduction

The primary objective of a Phase I dose escalation study in oncology is to find the dose at which the drug (or combination of drugs) will be tested in the subsequent phase II and III trials (Maximum Tolerated Dose or MTD) (see Piantadosi 2005). For cytotoxic anti-cancer drugs, the dose should be chosen as the highest dose for which the toxicity is still acceptable, because it defines the upper boundary of safe dosing. In order to maximize the activity of the treatment, drugs to treat cancer are combined with each other. The combination usually associates drugs that have different mechanisms of action. A different toxicity profile enables to maximize the tolerance of the combination involved.

The vast majority of the phase I dose escalation studies for single agents and combination of agents implement a "3+3" dose escalation scheme (CLD) to find the MTD (Rogatko 2007). This phase I design has been criticized for treating too many subjects at suboptimal doses and providing a rather crude estimated of the MTD estimate (Ratain 1993). Also, this design produces an unreliable estimation of the true rate of toxicity at the optimal dose (Ratain 1993). The 3+3 design is hindered to a great extent by chance (Hamberg 2010) and the alternative 3+3+3 design has been proposed. Although this decreases the impact of chance, the problem basically remains the same. One cause of unreliability is the design itself: the toxicity of the optimal dose is estimated from a small subset of treated subjects, as the majority of subjects in the trial are treated with doses lower or higher than optimal doses.
We propose a randomized Bayesian dose escalation design for combinations of drugs (BDED) that takes advantage of the fact that a standard of care drug is involved in the combination. We aim to obtain, via Bayesian estimation, an improved estimation of the MTD and the toxicity level at the MTD.
Our proposal uses a Bayesian approach and exploits the fact that the regimen is a combination of a new drug with a standard of care, i.e. a drug that is commonly used to treat the disease of the subjects enrolled in the trial. The proposed design implements a randomization between standard of care (also referred to as the control) and the combination regimen for which we want the dose. We estimate the difference between the toxicity of the control and the combination to search for the MTD.

2 Bayesian dose escalation in combinations of drugs with control

The principle of this design it to dynamically adapt the dose at which subjects are treated, based on the excess of probability of toxicity in the combination compared to the control. The excess of probability of toxicity is estimated from the occurrence of toxicity observed in previously treated subjects. If, at a given dose, the probability of toxicity in the combination is too high compared to the control, the subsequent dose will be lowered, and if the probability of toxicity in the combination is close compared to control, the dose will be increased. The randomization and treatment of subjects stops at a given number of subjects (eg. 50).
Let $F(d, \beta)$ be the model of the probability of toxicity in the combination regimen, depending on the dose and parameters β. This model can be a simple logistic model with $F(d, \beta_0, \beta_1) = \frac{1}{1+e^{-(\beta_0+\beta_1 d)}}$ or a more sophisticated model taking into account eg. the severity of the toxicity or the time to get the toxicity.

Let q be the probability of toxicity in the control, which does not depend on the dose as the dose of the control is fixed and let

$$\wp(\theta, d) = \Pr[F(d, \beta) - q < \theta]$$

the probability of the excess of probability of toxicity of the combination compared to the control. For subject i, we compute the posterior distribution of the parameters β and q and estimate the distribution $\wp(\theta, d)$ from the posterior distribution of the parameters:

$$\widehat{\wp}(\theta, d) = \frac{\int_{\beta,q} I[F(d, \beta) - q < \theta] \prod_{j=1}^{i} L_j}{\int_{\beta,q} \prod_{j=1}^{i} L_j}$$

where L_i is the likelihood of the data:

$$L_i = \left(F(d_i, \beta)^{x_i}(1 - F(d_i, \beta))^{1-x_i}\right)^e \left(q^{x_i}(1-q)^{1-x_i}\right)^c$$

where x_i is 1 if a toxicity was observed in subject i and 0 otherwise and $e = 1 - c$ is 1 when the subject has been randomized into the combination and 0 otherwise.

Once $\widehat{\wp}(\theta, d)$ is determined, we propose to choose the next dose to be administered (denoted d^{i+1}) as the highest dose d such that $\widehat{\wp}(\theta, d)$ is above a certain predetermined value (preferably high). The $(i + 1)$ subjects, if randomized to the combination, will be treated at dose d^{i+1}. Subjects are randomized and treated following the algorithm described above, until the maximum number of subjects is reached. The recommended phase II dose will be the dose administered to the last subject randomized to the combination.

3 Simulation Study

To compare the proposed design to the CLD, we simulated data from a phase I study reported by Diaz-Rubio (2002) who tested the combination Oxaliplatin + Capecitabin. The control is Oxaliplatin given at a dose of 130 mg/m^2, while the dose of Capecitabin was to be chosen between 500 and 1250 mg/m^2. Based on the data from Diaz-Rubio (2002), the recommended phase II dose would be 750 mg/m^2. The results given in table 1.

For the BDED, we further obtain an HPD that, at the average recommended dose of 725 mg/m^2, the difference between the probability of toxicity in the combination and the control is included in the 95% HPD interval: [0%, 18%].

4 Conclusions

The number of subjects treated with the BDED is 50 while for the CLD, it is on average 22.5. However, since 33 subjects are treated with the combination in the BDED (on average), the difference between the BDED and

Criteria	CLD	BDED
Average number of subjects treated	22.5	50
Average MTD (True MTD = 750)	703	721
Max MTD	1250	960
Extreme MTD		
% MTD below 700	55%	46%
% MTD above 800	22%	27%
% MTD above 1000	12%	0%

TABLE 1. simulations results

the CLD is only 11 subjects treated with the combination. This increase in the number of subjects allows for a more accurate estimation of the MTD. The recommended phase II doses given by the BDED is closer to the true MTD. Further, extreme doses (too low doses, or too high doses) are much more frequent in the CLD that in the BDED. This may be due to the increase of number of subjects treated. It may also be due to the fact that the toxicity is estimated from a model that takes into account all doses.
We believe that a big advantage of the BDED is the fact that we obtain the distribution of the excess of probability of toxicity, with the advantage that an HPD interval can be derived. Since there is no comparator in CLD, this distribution is not available.

References

O'Quigley, J. et al. (1990) Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer. *Biometrics* **46**, 33-48

Piantadosi, Steven (2005). *Clinical Trials*. John Wiley and Sons.

Rogatko, A. et al. (2007) Translation of Innovative Designs Into Phase I Trials. *Journal of Clinical Oncology* **25**, 4982-4986

Ratain, M. J. et al. (1993) Statistical and Ethical Issues in the Design and Conduct of Phase I and II Clinical Trials of New Anticancer Agents. *Journal of the National Cancer Institute* **85**, 1637-1643

Diaz-Rubio, E. et al. (2002) Capecitabine (xeloda) in Combination with Oxaliplatin: a Phase I, Dose-Escalation Study in Patients with Advanced or Metastatic Solid Tu- mors. *Annals of Oncology* **13**, 558-565

Hamberg P et al. (2010) Dose escalation models for combination phase I trials in oncology. *Eur J Cancer* **46**, 2870-2878.

Using text mining tools to compose structure priors for inferring gene networks.

Johan J. de Rooi[1,2], Paul H. C. Eilers[2]

[1] Department of Bioinformatics, Erasmus Medical Centre, Rotterdam, The Netherlands, email: j.derooi@erasmusmc.nl
[2] Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands

Abstract: We propose a model in which a prior and observed data are combined in order to derive a sparse network of interacting genes. We show how to derive useful information from external sources using text mining tools and translate this into a prior that can be used within the framework of penalized regression. The method is applied to a set of microarray gene expression samples.

Keywords: shrinkage; gene-networks; prior data; text mining

1 Introduction

Many researchers in genetics are concerned with gene interaction networks. Major questions are which genes interact with each other and whether we can distinguish separate groups or clusters of associated genes. The networks derived from the data are often depicted as a graph, in which each gene is represented as a node, a relation between two genes is visualized by an edge between the nodes. Because genetical networks are considered sparse, most of the nodes of the final network should be connected to a single or only a few other nodes.
The task of building a network is often troubled by a low number of observations while the number of variables is large. In an attempt to improve the quality of the network it becomes more common to include secondary data sources in the process of network building. Secondary data used to build a prior can be derived from various sources.
In this paper we introduce a penalized regression model related to the weighted lasso, in which the data and a prior are combined in order to estimate the network. Furthermore we discuss a text mining procedure to translate knowledge from online publications into a usable prior.

2 Constructing a prior

Prior data can be extracted from various sources, examples are online databases like MsigDB, Reactome or KEGG where pathways or networks are curated. Second, additional datasets similar to the primary data under study can be used to build a prior. A third option is to use text mining tools in order to derive a prior based on online publications.
A frequently used method in this context is to represent genes by a set of relevant documents (e.g. Glenisson et al., 2004) and subsequently use the vector space model (see e.g. Liu, 2007) to establish some measure of distance between the genes involved. Here we use a more simple model and derive a co-occurrence prior by querying PubMed using keywords relevant to our data. All retrieved documents are scanned for the presence of gene names. The assumption underlying the method is that genes that are mentioned in the same article have a biological relationship of some type (see e.g. Jenssen et al., 2001).
The co-occurrence of the genes within the documents of the total document set are represented in an term-document incidence matrix Γ. This is a binary matrix, with for every element a one if a certain gene t is cited at least once in document d, and a zero otherwise. To determine the closeness of any pair of genes involved we take the average score for gene j of the document vectors in which gene i appears:

$$w_{ij}^{*} = \frac{1}{N} \sum_{d=1}^{N} \gamma_{di} \gamma_{dj} \frac{N_i}{N}, \tag{1}$$

with N the size of the total document set and a weighting factor N_i/N. Subsequently the weights are normalized and gives the weight matrix:

$$\mathbf{W} = \left\{ \frac{w_{ij}^{*}}{max(w_{ij}^{*})} \right\}. \tag{2}$$

3 Discovering networks using penalized regression

Sparse genetic networks are derived from expression data and can be represented as a graphical Gaussian model. Let $\mathbf{X} = (X_1, X_2, ..., X_p)^T$ be a p-dimensional random vector having a multivariate normal distribution with mean μ and covariance matrix Σ. We have $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ an undirected graphs with $\mathcal{V} = 1, ..., p$ being the set of nodes and $\mathcal{E} = \{e_{ij}\}_{1 \leq i < j \leq p}$ the set of edges. The edge set describes the conditional independence among the genes, in a discrete model this means that e_{ij} is 0 or 1. If not only the presence or absence of a node is relevant but also is weight, e_{ij} could also be a continuous variable with a certain range. The conditional relationships are derived from the inverse of the covariance matrix, $\Phi = \Sigma^{-1}$. A zero

in the inverse covariance matrix corresponds to conditional independence between two nodes:

$$\phi_{ij} = 0 \Leftrightarrow e_{ij} = 0 \Leftrightarrow X_i \perp X_j | X_{-j}. \tag{3}$$

The general aim is to identify the non-zero elements in the precision matrix Φ. Given that $p < n$ and assuming that we have a positive definite covariance matrix Φ we can simply take its inverse and from this calculate the partial correlations. In genetics it is most often the case that $p > n$. As a result the covariance matrix will be singular and its inverse cannot be calculated. The literature provides various solutions to this problem, the one we sue here is the application of penalized regression.
Penalized regression comes in different types, to derive a network of interacting genes the lasso is often used. Here we rely on the implementation proposed by Meinshausen and Bühlmann (2006), where a separate regression model is fit for each variable in the model with all others as predictors. From the regression coefficients we can calculate the partial correlations as:

$$\hat{\rho}_{ij} = \text{sign}(\hat{\beta}_{ij}) \sqrt{\hat{\beta}_{ij} \hat{\beta}_{ji}}. \tag{4}$$

A simple way to incorporate the prior information into the penalty is to weight the tuning parameter according to the prior weight assigned to the particular relation, an approach that is similar to the weighted lasso as introduced by Zou (2006) which looks:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \, ||\mathbf{X}\beta - \mathbf{y}||_2^2 + \kappa \upsilon ||\beta||_1. \tag{5}$$

The first part of the equation is the familiar least squares estimator, the second is the lasso penalty function, with the additional parameter υ being the vector of weights. We propose the following penalized model:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \, ||\mathbf{X}\beta - \mathbf{y}||_2^2 + \kappa_2 ||\beta||_2^2 + \kappa_1 (\mathbf{I} - \mathbb{W}) ||\beta||_1. \tag{6}$$

In order to make the model estimable irrespective of the prior, we add an l_2 penalty with κ_2 as a small constant. The last part of the equation is an l_1 penalty with, $\mathbf{I}$ being an identity matrix, a tuning parameter κ_1 and the set of prior weights in a diagonal matrix $\mathbb{W}$. These weights are one row taken from the matrix $\mathbf{W}$, corresponding with the current $\mathbf{y}$ of the regression model. With κ_1 we can balance between the prior and fidelity to the data and is optimized using cross-validation, or can be tuned by hand. By setting all weights in $\mathbf{W}$ to one and choose a large value for κ_1 the prior will be imposed entirely on the data.

4 Application

The method is applied to 292 microarray gene expression samples (Gravendeel et al., 2009). A network is built using penalized regression in combination with a prior based on a literature search through PubMed. In order to generate a (very) small but insightful example we restricted the number of documents to the first 1000 hits. From the 109 unique genes derived from the document set, 45 were found to be co-cited. Next to the genes present in the prior we included 100 additional genes from the dataset, which are selected based upon their variance. Subsequently we estimated the optimal model using the model explained in the previous paragraph. The posterior network is showed in Figure 1. The green edges are confirmations of relations also present in the prior, the red edges are new relations estimated from the data. The application shows that the procedure is able to verify relations posed in the prior and at the same time determines novel relations from the data.

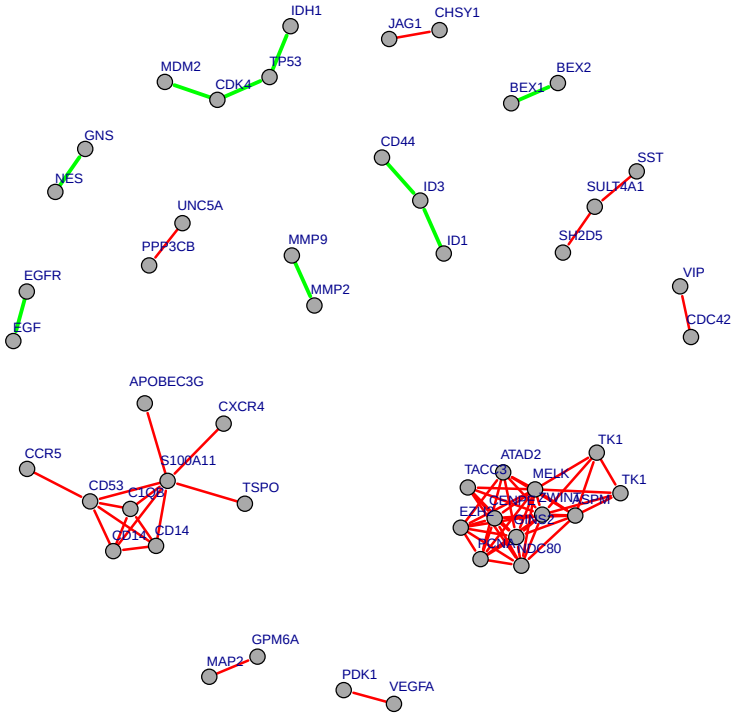


FIGURE 1. The posterior model presented in a graph. Green edges are confirmations of prior edges, red edges are estimates only coming from the data.

5 Discussion

A simple method to build a prior using text mining tools is combined with the framework of penalized regression. The shrinkage estimators seem to fit well with the concept of a weight matrix. Other penalties than the applied l_1 are also considered. In this abstract the prior was built only using text mining tools, however other sources can be used as well. Functional databases can provide binary weight matrices, or auxiliary experiments can be included by e.g. using correlations as weights in the prior matrix. An issue that is not addressed here but should be covered, is the distinction between no evidence and negative evidence with respect to the weighting.

References

Glenisson, P., Coessens, B., Van Vooren, et al. (2004). TXTGate: profiling gene groups with text-based information. *Genome Biology*, **5**:R43.

Gravendeel, L. A. M., Kouwenhoven, M. C. M., Gevaert O. et al. (2009). Intrinsic gene expression profiles of Gliomas Are a better predictor of survival than histology. *Cancer Research* **69(23)**, 9065-9072.

Jenssen, T., Laegreid, A., Komorowski, J. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics* **28**, 21-28.

Liu, B. (2007). *Web Data Mining.* Berlin, Heidelberg, New York: Springer.

Meinshausen, N., Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, **34**, 1436-1462.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101(476)**, 1418-1429.

Markov-Switching Multifractal models within GAMLSS

Abdelmadjid Djennad[31], Robert Rigby[1], Mikis Stasinopoulos[1], Vlasios Voudouris[2]

[1] STORM, London Metropolitan University, Holloway Road, London N7 8DB
[2] Centre for International Business and Sustainability, London Metropolitan Business School
[3] Communicating author: m.djennad@londonmet.ac.uk

Abstract: This paper reports on concepts and methods to incorporate the Markov-Switching Multifractal model for stochastic volatility introduced by Calvet and Fisher (2004) within the GAMLSS model introduced by Rigby and Stasinopoulos (2005), allowing generalization to a non-normal distribution. The software implementation is written in R and the models are fitted and compared using maximum likelihood estimation.

Keywords: Markov-Switching Multifractal models, GAMLSS; returns.

1 Introduction

Generalised additive models for location, scale and shape (GAMLSS) is a general framework for fitting regression type models where the distribution of the response variable does not have to belong to the exponential family and includes highly skew and kurtotic continuous and discrete distribution. GAMLSS allows all the parameters of the distribution of the response variable to be modelled as linear/non-linear or smooth functions of the explanatory variables, (Rigby and Stasinopoulos 2005). In this paper, we describe functions in R for simulating, estimating and forecasting the stochastic volatility, incorporating the Markov-Switching Multifractal (MSM) model within the GAMLSS model, allowing generalisation of the MSM model to a non-normal distribution. The Multifractal processes have recently been proposed as a new formalism for modelling the time series of returns in finance. The major attraction of these processes is their ability to generate various degrees of long memory in different powers of returns. Initial difficulties stemming from non-stationarity and the combinatorial nature of the original model proposed by Mandelbrot et al. (1997), the Multi-Fractal Model of Assets Returns (MMAR), have been overcome by the introduction of an iterative Markov-Switching Multifractal model in Calvet and Fisher (2001) which allows for estimation of its parameters via maximum likelihood (Lux 2006).

Section 2 defines the original GAMLSS model. Section 3 defines the Markov-Switching Multifractal (MSM) Model. Finally in section 4 we use daily returns for oil, to fit and compare the stochastic volatility MSM model with the original GAMLSS (smoothing) model, and with a standard GARCH model. Different distributions were used for the comparison including normal, t and skew t.

2 The GAMLSS Model

A GAMLSS model, assumes independent observations y_i for $i = 1, 2, \ldots, n$ with probability (density) function $f(y_i|\boldsymbol{\theta}^i)$ conditional on $\boldsymbol{\theta}^i$ where $\boldsymbol{\theta}^i = (\mu_i, \sigma_i, \nu_i, \tau_i)$ a vector of four distribution parameters, each of which can be a function to the explanatory variables. The model is defined as follows. Let $\mathbf{y}^T = (y_1, y_2, \ldots, y_n)$ be the n length vector of the response variable. Also for $k = 1, 2, 3, 4$, let $g_k(.)$ be a known monotonic link function relating the k^{th} distribution parameter to explanatory variables by

$$g_k(\boldsymbol{\theta_k}) = \boldsymbol{\eta}_k = X_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(x_{jk}) \tag{1}$$

where $\boldsymbol{\theta}_k$, $\boldsymbol{\eta}_k$ and x_{jk} for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, 4$ are vectors of length n. The GAMLSS model has been implemented in a series of `R` packages and can be obtained in CRAN or at `http://www.gamlss.org/`.

3 The Markov-Switching Multifractal Model

A Markov-Switching Multifractal model with normal errors is defined by the process $y_t = \sigma_t\varepsilon_t$. The innovations ε_t are assumed to drawn from a standard normal distribution $N(0, 1)$, an assumption we will relax. The instantaneous volatility, σ_t, is determined by the product of $\bar{k}$ volatility state components or multipliers $M_t^{(1)}, M_t^{(2)}, \ldots, M_t^{(\bar{k})}$ and a constant scale factor $\bar{\sigma}$:

$$\sigma_t = \bar{\sigma}(\prod_{i=1}^{\bar{k}} M_t^{(i)})^{1/2}. \tag{2}$$

It is assumed that each volatility state component can take one of two values m_0 and $m_1 = 2 - m_0$. Each volatility state component is renewed at time t with probability γ_i depending on its rank within the hierarchy of multipliers or remains unchanged with probability $1 - \gamma_i$, (Lui and Lux, 2006).
Hence for $i = 1, 2, \ldots, \bar{k}$,

$$M_{t+1}^{(i)} = \begin{cases} M_t^{(i)}, & \text{with probability } 1 - \gamma_i/2, \\ 2 - M_t^{(i)}, & \text{with probability } \gamma_i/2. \end{cases} \tag{3}$$

Let $\mathbf{M}_t = \left[M_t^{(1)}, M_t^{(2)}, \ldots, M_t^{(\bar{k})}\right]$ be the vector of state components at time t. Since each state component can take one of two possible values m_0 and $m_1 = 2 - m_0$, there are $2^{\bar{k}}$ possible states for the vector $\mathbf{M}_t$. Hence the transition matrix $\mathbf{A}$ from $\mathbf{M}_t$ to $\mathbf{M}_{t+1}$ is a $2^{\bar{k}}$ x $2^{\bar{k}}$ matrix. For example for $\bar{k} = 3$ the transition matrix A is 8 x 8. The transition probabilities can be obtained using (3). For example the transition from $\mathbf{M}_t = (m_0, m_1, m_0)$ to $\mathbf{M}_{t+1} = (m_0, m_1, m_1)$ is given by $\left(1 - \frac{\gamma_1}{2}\right)\left(1 - \frac{\gamma_2}{2}\right)\left(\frac{\gamma_3}{2}\right)$.
The relationship between the $\gamma's$ is specified by

$$\gamma_k = 1 - (1 - \gamma_1)^{b^{(\bar{k}-1)}} \tag{4}$$

for $i = 1, 2, \ldots, \bar{k}$. Estimation of MSM model involves estimating the parameters m_0, $\gamma_{\bar{k}}$, b and $\bar{\sigma}$. Lux (2006) suggested the Generalised Method of Moments (GMM) approach to speed up the computational limitation choice of $\bar{k}$ when the $\bar{k}$ is higher than 10 because of the implied evaluation of the transition matrix in each iteration. For fitting the models in `R`, we have built a function which, in order to avoid loops, builds a transition matrix index as a way to calculate the probability transition matrix. Our function also has the facility of including different *gamlss.family* distributions, (Stasinopoulos *et al.*, 2008), so the normality of the innovation in the original assumption of MSM can be relaxed by using a kurtotic or skew distribution, providing a new framework for modelling stochastic volatility. The likelihood function for MSM model is defined as:

$$\text{Ł} = \prod_{t=1}^{n} f(y_t|\mathbf{H}_t), \tag{5}$$

where $\mathbf{H}_t = [\,\mathbf{M}_0, \mathbf{M}_1, \ldots, \mathbf{M}_t]$ is the history up to time t. The individual contribution to the likelihood is given by:

$$f(y_t|\mathbf{H}_t) = \sum_{j=1}^{2^{\bar{k}}} f(y_t, \mathbf{M}_t = S_j|\mathbf{H}_t) = \sum_{j=1}^{2^{\bar{k}}} f(y_t|\mathbf{M}_t = S_j)\xi_{tj}, \tag{6}$$

where, S_j for $j = 1, 2, \ldots, 2^{\bar{k}}$ represent the $2^{\bar{k}}$ possible state vector that $\mathbf{M}_t$ can take, and $\xi_{tj} = P(\mathbf{M}_t = S_j|\mathbf{H}_t)$. Let $\boldsymbol{\xi}_t = (\xi_{t1}, \xi_{t2}, \ldots, \xi_{t2^k})$ then $\boldsymbol{\xi}_t = \mathbf{A}\boldsymbol{\xi}_{t-1}$.

4 The Data

In this section shows some of our preliminary results in comparing the MSM(7) model to both the standard GARCH(1,1) and original GAMLSS models. For the comparison we analysed oil daily returns for the period of 13-6-2003 to 30-11-2010. Here we model the oil daily returns. The Figure 1 shows the 2.5% and 97.5% centile estimates for the MSM(7), GARCH(1,1),

and a GAMLSS model with a smoothing term for time for each of the four parameters of the distribution, using a Skew Student-t distribution.
The table 1 shows the Akaike Information Criterion in-sample goodness of fit for the three models with different distributions.

TABLE 1. In-sample model comparison

AIC	Normal	Student-t	Skew Student-t
GAMLSS	-12614.3	-12694.7	-12696.9
MSM	-12647.2	-12662.8	-12666.4
GARCH	-12582.9	-12692.5	-12695.9

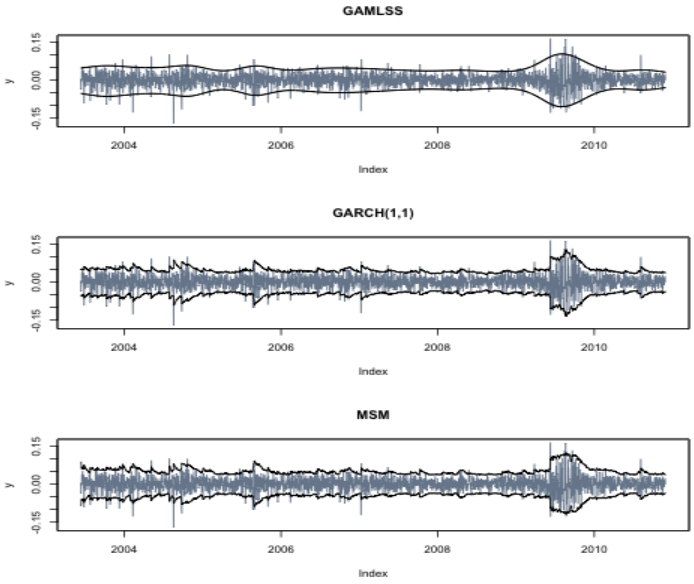


FIGURE 1. Retuns and fitted σ model for GAMLSS, GARCH(1,1) and MSM(7) models

References

Calvet, L., Fisher, A. (2004) How to Forecast Long-Run Volatility: Regime Switching and the Estimation of Multifractal Processes, *Journal of Financial Econometrics.*, **Vol. 2, No. 1**, 49-83

Calvet, L., Fisher, A. (2008) Multifractal Volatility: Theory, Forecasting and Pricing. USA: Elsevier Academic Press.

Liu, R., Lux, T. (2006) Bivariate Multi-Fractal Model: Estimation of parameters and Application to Risk Management, *Economics Working paper 2006*

Lux, T. (2008) The Markov-Switching Multifractal Model of Asset Returns: Estimation via GMM and Linear Forecasting of Volatility, *Journal of Business & Economics Statistics*, **26**, 194-210

Mandelbrot, B. (1997) *Fractals and scaling in finance: Discontinuity, concentration, risk*. New York: Springer-Verlag.

Mandelbrot, B., Fisher, A., Calvet, L. (1997) A multifractal model of asset returns. *Cowles Foundation Discussion.*, **1164**,

Rigby, R.A. and Stasinopoulos, D.M (2005) Generalized Additive Models for Location, Scale and Shape., *App. Statist.*, 54, pp 507-554.

Stasinopoulos, D.M., Rigby, R.A. (2007) Generalized Additive Models for Location Scale and Shape (GAMLSS) in R, *Journal of Statistical Software*, Vol. 23, Issue 7.

Smooth mixed models for nested curves

Viani A. B. Djeundje[1] , Iain D. Currie[1]

[1] Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS.

Abstract: We consider grouped longitudinal data where the functional form of the effect of time varies across groups. One approach to capturing these functional forms uses penalized splines with truncated polynomials as bases for the smooth functions. This, together with a standard assumption on the covariance structure, allows the model to be expressed as a mixed model. We show that this approach can be seriously biased. We propose an alternative approach where the covariance structure is derived via a penalty argument. We illustrate our methods with some Canadian weather data.

Keywords: Longitudinal data, mixed models, nested curves, penalties.

1 Introduction

Repeated observations on the same subjects over time are common in many areas such as medicine, psychology, environmental science, etc. Such data are referred to as longitudinal data and often have a grouped or nested structure. A popular inferential approach is through mixed models and, when the number of observations permits, the modelling process can additionally incorporate smoothing. In this context, one approach to modelling the time varying trends at both group and subject levels uses truncated polynomials as bases with a standard covariance structure for the "random" effects. To our knowledge, the impact of this covariance structure on the estimation of model terms has received very little attention. Here, we first illustrate some of its unfortunate effects, and then derive a more appropriate covariance structure via a penalty argument.

2 Standard approach for nested curves

Data are collected on n subjects, partitioned into k groups of sizes $r_1, \ldots, r_k$, with data on subject i represented by $(g(i), t_{i,j}, Y_{i,j})$, $j = 1, ..., n_i$; here $Y_{i,j}$ is the value of the response collected on subject i at time $t_{i,j}$, and $g(i)$ is the group to which subject i belongs. For notational convenience, we assume that the data are entered in group order. We will denote by $\boldsymbol{Y}_i$ the response vector on subject i and by $\boldsymbol{t}_i$ the corresponding time vector. A typical example, illustrated by panel ($\boldsymbol{a}$) in Figure 1, shows the daily

average temperature in 35 Canadian cities; these cities have been grouped into four regions. Such data can be modelled as

$$Y_{i,j} = S_{g(i)}(t_{ij}) + S_i(t_{i,j}) + \varepsilon_{i,j}, \quad \varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $S_{g(i)}$ measures the group/region effect to which subject i belongs, and S_i captures the ith subject/city effect relative to its group effect. These functions are designed to capture the underlying patterns in the data, and depending on the structure of the data, a simple approach may be to treat them either as straight lines or as some low degree polynomials. But for general modelling purposes, a flexible structure may be required to reflect the true dynamism driving the data. One popular approach is to account for this flexibility with truncated polynomials. In this setting, truncated lines are often used, since truncated polynomials of high order tend to be more unstable. With truncated lines, these functions can be expressed as

$$\begin{aligned} S_{g(i)}(t) &= \delta_{g(i),0} + \delta_{g(i),1}t + \sum_{k=1}^{q} \xi_{g(i),k}(t - \tau_k)_+ \\ S_i(t) &= \breve{\delta}_{i,0} + \breve{\delta}_{i,1}t + \sum_{l=1}^{\breve{q}} \breve{\xi}_{i,l}(t - \breve{\tau}_l)_+ \end{aligned} \tag{2}$$

where $x_+ = max\{x, 0\}$, and $\tau = \{\tau_1, ...\tau_q\}$ and $\breve{\tau} = \{\breve{\tau}_1, ...\breve{\tau}_{\breve{q}}\}$ are sets of internal knots at the group and subject levels respectively. Model (1) can now be expressed in matrix form as

$$\boldsymbol{Y}_i = \{\boldsymbol{X}_{g(i)}\boldsymbol{\delta}_{g(i)} + \boldsymbol{T}_{g(i)}\boldsymbol{\xi}_{g(i)}\} \; + \; \{\breve{\boldsymbol{X}}_i\breve{\boldsymbol{\delta}}_i + \breve{\boldsymbol{T}}_i\breve{\boldsymbol{\xi}}_i\} \; + \; \boldsymbol{\varepsilon}_i, \quad i = 1, ..., n. \tag{3}$$

With these components in place, one approach (Coull *et al.* 2001, Ruppert *et al.* 2003, Durban *et al.* 2005) achieves smoothness and identifiability of the model by expressing it as a mixed model with the following covariance structure:

$$\boldsymbol{\xi}_{g(i)} \sim \mathcal{N}(\boldsymbol{0}, \sigma_g^2\boldsymbol{I}_q), \;\; \breve{\boldsymbol{\delta}}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \;\; \breve{\boldsymbol{\xi}}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_s^2\boldsymbol{I}_{\breve{q}}), \tag{4}$$

where σ_g^2 is the variance parameter driving the smoothness at the group level, $\boldsymbol{\Sigma}$ is some 2×2 symmetric, positive definite matrix, σ_s^2 is the variance parameter driving the smoothness at the subject level, and $\boldsymbol{I}_r$ is the $r \times r$ identity matrix. We refer to (4) as the *standard covariance structure*.

Under these assumptions, model (1) can be fitted with the function `lme` available in the package `nlme` in `R` (R Development Core Team 2010). Panel (***b***) in Figure 1 illustrates this model fitted to the Canadian weather data with $q = 39$ and $\breve{q} = 19$ equi-spaced internal knots at the regional and city levels respectively. As we can see, the fitted subject means (obtained by adding the subject effects, $S_i(\boldsymbol{t})$, to their group effect, $S_{g(i)}(\boldsymbol{t})$), capture the data well; here, and below, $S_i(\boldsymbol{t}_i)$ represents the element-wise action of S_i on the components of $\boldsymbol{t}_i$ with a similar definition for $S_{g(i)}(\boldsymbol{t})$. Hence, one may be tempted to conclude that the fitted group means will also follow

the data. However, panel (***c***) of the same figure shows the fitted regional effects. By giving different values to $\breve{q}$, we make two observations: (i) the fitted group effects are sensitive to the knot locations at the subject level, and (ii) the confidence intervals exhibit an unexpected widening fan effect. This behaviour of the fitted group effects is balanced by a similar behaviour of the fitted subject effects in such a way that the fitted subject means are appropriately recovered as illustrated in panel (***b***).
This unexpected behaviour occurs as the result of the mis-specification of the covariance structure (4). There are three major reasons for the choice of (4): (a) a ridge penalty on a truncated line basis works well when one deals with smoothing at a single level, (b) its simplicity is attractive and (c), it appears to offer sufficient flexibility so that proper identification of the components of the model is possible. However, the covariance structure (4) does not perform satisfactorily and we are faced with a common challenge in mixed models, namely the appropriate specification of the covariance structure of the random effects. In the next section, we use penalization to provide a solution to this issue.

3 Penalty approach

We consider model (1) with its components given by (2) or, equivalently, (3). Instead of focusing directly on the covariance structure of the parameters, we specify the modelling effects we wish to achieve, and from them derive an appropriate covariance structure. In terms of modelling effects, two issues need to be addressed: smoothness and identifiability.

- *Smoothness*: We control the jumps in the derivative at the knots of $S_{g(i)}$ and S_i, ie, we impose the constraints $\|\boldsymbol{\xi}_{g(i)}\|^2 < \rho$ and $\|\breve{\boldsymbol{\xi}}_i\|^2 < \breve{\rho}$, for some well chosen constants ρ and $\breve{\rho}$.
- *Identifiability*: We shrink the subject effects towards 0, $\|S_i(\boldsymbol{t}_i)\|^2 < \breve{\delta}$.

Using Lagrange arguments, we find that the penalized residual sum of squares, PRSS, of (3) under the above three inequality constraints can be expressed as

$$\texttt{PRSS} = \sum_{i=1}^{n} \|\boldsymbol{Y}_i - S_{g(i)}(\boldsymbol{t}_i) - S_i(\boldsymbol{t}_i)\|^2 + \lambda \sum_{l=1}^{k} \|\boldsymbol{\xi}_l\|^2 + \sum_{i=1}^{n} (\breve{\lambda}\|\breve{\boldsymbol{\xi}}_i\|^2 + \breve{\gamma}\|S_i(\boldsymbol{t}_i)\|^2). \quad (5)$$

Setting $\boldsymbol{\xi} = vec(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k)$ and $\breve{\boldsymbol{b}} = vec(\breve{\boldsymbol{b}}_1, \ldots, \breve{\boldsymbol{b}}_n)$ with $\breve{\boldsymbol{b}}_i = vec(\breve{\boldsymbol{\delta}}_i, \breve{\boldsymbol{\xi}}_i)$ reduces (5) to

$$\texttt{PRSS} \quad = \quad \sum_{i=1}^{n} \|\boldsymbol{Y}_i \; - \; S_{g(i)}(\boldsymbol{t}_i) \; - \; S_i(\boldsymbol{t}_i)\|^2 \; + \; \lambda \boldsymbol{\xi}'\boldsymbol{\xi} \; + \; \breve{\boldsymbol{b}}'\boldsymbol{P}\breve{\boldsymbol{b}} \quad (6)$$

where

$$\boldsymbol{P} = blockdiag(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_n) \text{ with } \boldsymbol{P}_i = \breve{\lambda}\boldsymbol{J} + \breve{\gamma}\breve{\boldsymbol{L}}_i'\breve{\boldsymbol{L}}_i$$

is the penalty matrix on the subject coefficients; here $\breve{\boldsymbol{L}}_i = [\breve{\boldsymbol{X}}_i : \breve{\boldsymbol{T}}_i]$, and $\boldsymbol{J}$ represents the identity matrix of appropriate size, but with its two upper diagonal elements replaced by zeros.

4 Inference and application

It is well known that REML estimates of the variance parameters tend to behave well. With this motivation, we find after some algebra that expression (6) corresponds (up to additive and multiplicative constants) to the log likelihood of $(\boldsymbol{y}, \boldsymbol{b})$ in the mixed model representation

$$\boldsymbol{y}|\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \sigma^2\boldsymbol{I}), \;\; \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi}), \quad (7)$$

where $\boldsymbol{y} = vec(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n)$ is the data vector, $\boldsymbol{\beta} = vec(\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_k)$ is the fixed effect, $\boldsymbol{b} = vec(\boldsymbol{\xi}, \breve{\boldsymbol{b}})$ is the random effect and $\boldsymbol{\Phi} = \sigma^2\, blockdiag(\lambda^{-1}\boldsymbol{I}_q, \boldsymbol{P}^{-1})$ is its covariance matrix. The regression matrix $\boldsymbol{X}$ for the fixed effect is defined as follows: let $\boldsymbol{G}_1 = stack(\boldsymbol{X}_{g(1)}, \ldots, \boldsymbol{X}_{g(r_1)})$ with similar definitions for $\boldsymbol{G}_2, \ldots, \boldsymbol{G}_k$, then

$$\boldsymbol{X} = blockdiag(\boldsymbol{G}_1, \ldots, \boldsymbol{G}_k); \quad (8)$$

here, and below, $stack(\boldsymbol{A}, \boldsymbol{B})$, indicates that the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with the same number of columns are stacked on top of each other. The regression matrix $\boldsymbol{Z} = [\boldsymbol{Z}_{\text{gp}} : \boldsymbol{Z}_{\text{subj}}]$ is partitioned into the regression matrix for the group "random effects", $\boldsymbol{Z}_{\text{gp}}$, and the regression matrix for the subject random effects, $\boldsymbol{Z}_{\text{subj}}$. We define $\boldsymbol{Z}_{\text{gp}}$ and $\boldsymbol{Z}_{\text{subj}}$ as follows: let $\boldsymbol{Z}_1 = stack(\boldsymbol{T}_{g(1)}, \ldots, \boldsymbol{T}_{g(r_1)})$ with similar definitions for $\boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_k$, then

$$\boldsymbol{Z}_{\text{gp}} \quad = \quad blockdiag(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k); \quad (9)$$

$$\boldsymbol{Z}_{\text{subj}} \quad = \quad blockdiag(\breve{\boldsymbol{L}}_1, \ldots, \breve{\boldsymbol{L}}_n). \quad (10)$$

This mixed model representation allows us to estimate the fixed effect $\boldsymbol{\beta}$ and the random effect $\boldsymbol{b}$ by their best linear unbiased estimator/predictor, with the REML estimates of variance parameters plugged in. Although λ and $\breve{\lambda}$ are treated as variance parameters in the fitting process, they are purely smoothing parameters, in the sense that they act on the shape of the corresponding effects only. In contrast, $\breve{\gamma}$ is a variance parameter for the subject effects in the original sense of a mixed model; ie, $\breve{\gamma}$ controls the overall size of the subject effects, in the same way that the departures of the response data $Y_{i,j}$ from the mean are modulated by σ^2.

We now apply this approach to the Canadian weather data described earlier. The Canadian data are balanced in the sense that observations on each subject are made on a common time vector; in this case, the above formulae are considerably simplified. An illustration of the fitted region effects is shown on panel (***d***) in Figure 1. Through this example we see the difference between our penalty approach and the standard approach (4). Specifically, the penalty approach allows the appropriate identification of the underlying effects as well as a correction to the confidence intervals.

5 Conclusion

In this paper, we have first illustrated some problems that occur by fitting flexible nested curves with the standard covariance structure. We have then proposed an alternative approach to deal with this issue, and illustrated that it leads to satisfactory results. This new approach is presented here with truncated polynomials so that we can illustrate the fundamental problem with the covariance structure (4). However, our approach is easily adapted to other bases such as B-splines bases. With B-splines, smoothness is obtained via a roughness penalty, and identifiability is obtained via direct shrinkage of the B-splines coefficients. More detail and a fuller discussion of using penalties to define appropriate covariance structures in mixed models can be found in Djeundje & Currie (2010). There we estimated the smoothing/shrinkage parameters by minimizing the BIC, whereas here we use the mixed model perspective and optimize the restricted likelihood.

References

Coull, B. A., Ruppert, D. and Wand, M. P. (2001). Simple incorporation of interactions into additive models. *Biometrics*, **57**, 539-545.

Djeundje, V. A. B. and Currie, I. D. (2010). Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. *Electronic Journal of Statistics*, **4**, 1202-1224.

Durban, M., Harezlak, J., Wand, M. P. and Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**, 1153-1167.

R Development Core Team (2010).*R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ruppert, D., Wand, M.P. and Carroll, R. J. (2003). Semi-parametric regression. Cambridge: Cambridge University Press.

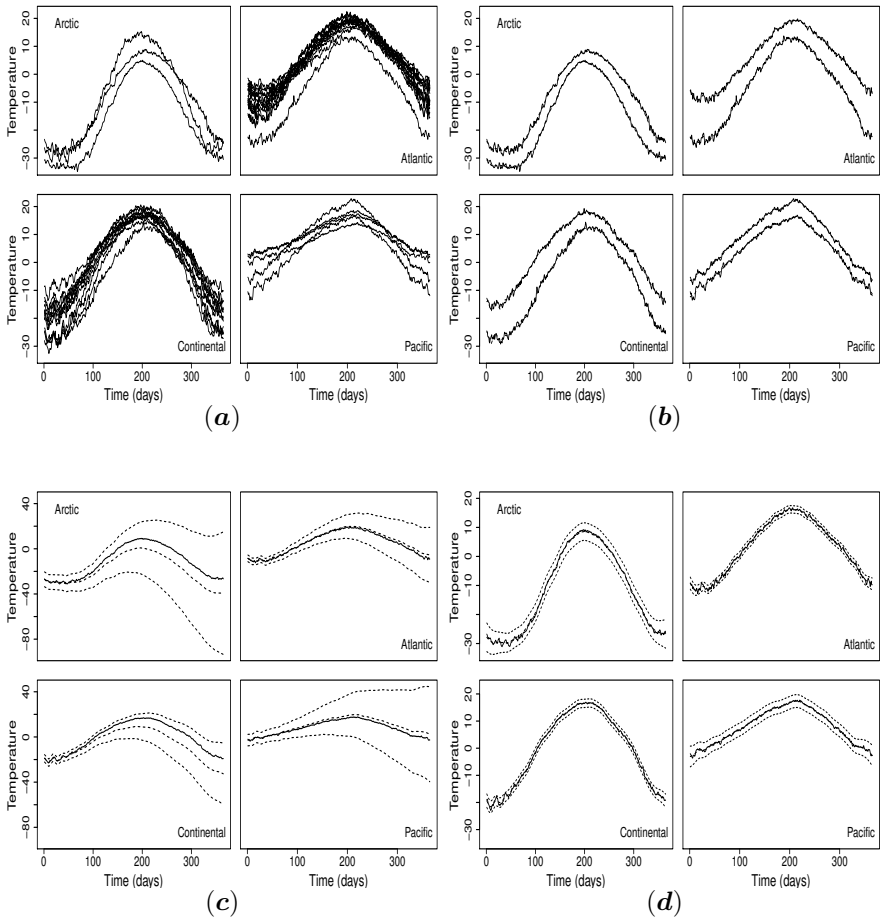


FIGURE 1. (***a***): Canadian weather data. (***b***): fitted cities (dashed) using the standard approach, together with the observed data (solid). (***c***): fitted region effects with 95% CI (dashed) using the standard approach, together with the observed average (solid). (***d***): fitted region effect with 95% CI (dashed) using the penalty approach, together with the observed average (solid).

A Bayesian regression and multiple changepoint model for systems biology

Frank Dondelinger[1,2,6], Andrej Aderhold[1], Sophie Lèbre[3], Marco Grzegorczyk[4,5], Dirk Husmeier[1]

[1] Biomathematics and Statistics Scotland, JCMB, Edinburgh, EH9 3JZ, UK.
[2] Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, UK
[3] LSIIT - UMR 7005, Université de Strasbourg, 67412 Illkirch, France.
[4] Department of Statistics, TU Dortmund, 44221 Dortmund, Germany
[5] Department of Mathematics, Carl von Ossietzky University Oldenburg, Germany
[6] Communicating Author. Email: frankd@bioss.ac.uk

Abstract: We propose a Bayesian regression and multiple changepoint model for reverse engineering gene regulatory networks from high-throughput gene expression profiles. We report results from a recently held international gene network reconstruction competition, in which our method was objectively assessed in a blind study. While we did not win the competition, the scores indicate that the proposed method favourably compares with the majority of competing approaches and clearly belongs to the group of highest-ranked performers.

Keywords: Systems biology; gene regulatory network inference; Bayesian multiple changepoint model; RJMCMC; DREAM

1 Introduction

The objective of the highly topical field of systems biology is the reverse engineering of molecular regulatory networks and signalling pathways from high-throughput post-genomic data, and a flurry of activities in the statistics and machine learning communities are currently aimed at solving this problem. A variety of methods from statistics and machine learning have been applied to this end. See e.g. Grzegorczyk et al. (2008) and Cantone et al. (2009) for brief reviews. In the present paper, we propose a Bayesian regression and multiple changepoint model, with Bayesian inference based on reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). We participated in a recently held gene regulatory network prediction competition (DREAM 5), which assures that the comparative evaluation with other methods was done objectively.

2 Model

Multiple changepoints: Let p be the number of target genes, whose expression values $y = \{y_i(t)\}_{1\leq i\leq p, 1\leq t\leq N}$ are measured on N separate chips. $\mathcal{M}_i$ is the set of parents (regulators) associated with target gene i in the gene regulatory network. We model the differences in the regulatory relationships measured by different chips (assumed to be in some natural order, e.g. a time series) with a multiple changepoint process. For each target gene i, an unknown number k_i of changepoints define $k_i + 1$ non-overlapping segments. Segment $h \in \{1, .., k_i+1\}$ starts at changepoint ξ_i^{h-1} and stops before ξ_i^h, so that $\xi_i = (\xi_i^0, ..., \xi_i^{h-1}, \xi_i^h, ..., \xi_i^{k_i+1})$ with $\xi_i^{h-1} < \xi_i^h$. This changepoint process induces a partition of the chip ordering, $y_i^h = (y_i(t))_{\xi_i^{h-1}\leq t<\xi_i^h}$. The network structure $\mathcal{M}_i$ remains the same for each segment h, but the other parameters of the model can vary.

Regression model: For all genes i, the random variable $Y_i(t)$ refers to the expression of gene i on chip t. Within any segment h, the expression of gene i at chip t depends on the gene expression values on chip t of a set R_i of m potential regulator genes (parents), with $i \notin R_i$. We define a regression model by (a) the set of s_i parents denoted by $\mathcal{M}_i = \{j_1, ..., j_{s_i}\} \subseteq R_i$, and (b) a set of parameters $((a_{ij}^h)_{j\in R_i},\ \sigma_i^h)$; $a_{ij}^h \in \mathbb{R}$, $\sigma_i^h > 0$. For all $j \neq 0$, $a_{ij}^h = 0$ if $j \notin \mathcal{M}_i$. For all genes i, for all chips t in segment h ($\xi_i^{h-1} \leq t < \xi_i^h$), the random variable $Y_i(t)$ depends on the m variables $\{Y_j(t)\}_{j\in R_i}$ according to

$$Y_i(t) \ = \ a_{i0}^h + \sum\nolimits_{j\in\mathcal{M}_i} a_{ij}^h\ Y_j(t) \ + \varepsilon_i(t) \tag{1}$$

where the noise $\varepsilon_i(t)$ is assumed to be Gaussian with mean 0 and variance $(\sigma_i^h)^2$, $\varepsilon_i(t) \sim N(0, (\sigma_i^h)^2)$. We define $a_i^h = (a_{ij}^h)_{j\in R_i}$.

Prior: The k_i+1 segments are delimited by k_i changepoints, where k_i is distributed a priori as a truncated Poisson random variable with mean λ and maximum $\overline{k} = N-2$: $P(k_i|\lambda) \propto \frac{\lambda^{k_i}}{k_i!}\mathbb{1}_{\{k_i\leq\overline{k}\}}$. Conditional on k_i changepoints, the changepoint positions vector $\xi_i = (\xi_i^0, \xi_i^1, ..., \xi_i^{k_i+1})$ takes non-overlapping integer values, which we take to be uniformly distributed a priori. For all genes i, the number s_i of parents for node i follows a truncated Poisson distribution with mean Λ and maximum $\overline{s} = 5$: $P(s_i|\Lambda) \propto \frac{\Lambda^{s_i}}{s_i!}\mathbb{1}_{\{s_i\leq\overline{s}\}}$. Conditional on s_i, the prior for the parent set $\mathcal{M}_i$ is a uniform distribution over all parent sets with cardinality s_i: $P(\mathcal{M}_i\,||\mathcal{M}_i| = s_i) = 1/(\,{}_{s_i}^{p})$. The overall prior on the network structures is given by marginalization:

$$P(\mathcal{M}_i|\Lambda) = \sum\nolimits_{s_i=1}^{\overline{s}} P(\mathcal{M}_i|s_i)P(s_i|\Lambda) \tag{2}$$

Conditional on the parent set $\mathcal{M}_i$ of size s_i, we assume for the prior distribution $P(a_i^h|\mathcal{M}_i, \sigma_i^h)$ of the s_i+1 regression coefficients for each segment h a zero-mean multivariate Gaussian with covariance matrix $(\sigma_i^h)^2\Sigma_{a_i^h}$, where following Andrieu and Doucet (1999) we set $\Sigma_{a_i^h} = \delta^{-2}D_{a_i^h}^{\dagger}(y)D_{a_i^h}(y)$, and

$D_{a_i^h}(y)$ is the $(\xi_i^h - \xi_i^{h-1}) \times (s_i + 1)$ matrix whose first column is a vector of 1 (for the constant in model (1)) and each $(j+1)^{th}$ column contains the observed values $(y_j(t))_{\xi_i^{h-1}-1 \leq t < \xi_i^h - 1}$ for all regulatory genes j in $\mathcal{M}_i$. Finally, the conjugate prior for the variance $(\sigma_i^h)^2$ is the inverse gamma distribution, $P((\sigma_i^h)^2) = \mathcal{IG}(\upsilon_0, \gamma_0)$. Following Lèbre et al. (2010), we set the hyperparameters for shape, $\upsilon_0 = 0.5$, and scale, $\gamma_0 = 0.05$, to fixed values that give a vague distribution. The terms λ and Λ can be interpreted as the expected number of changepoints and parents, respectively, and δ^2 is the expected signal-to-noise ratio. These hyperparameters are drawn from vague conjugate hyperpriors, which are in the (inverse) gamma distribution family: $P(\Lambda) = P(\lambda) = \mathcal{G}a(0.5, 1)$ and $P(\delta^2) = \mathcal{IG}(2, 0.2)$.

Posterior: Equation (1) implies that

$$P(y_i^h | \xi_i^{h-1}, \xi_i^h, \mathcal{M}_i, a_i^h, \sigma_i^h) \quad \propto \quad \exp\left(-\frac{(y_i^h - D_{a_i^h}(y)a_i^h)^\dagger \ (y_i^h - D_{a_i^h}(y)a_i^h)}{2(\sigma_i^h)^2}\right) \tag{3}$$

From Bayes theorem, the posterior is given by the following equation:

$$P(k, \xi, \mathcal{M}, a, \sigma, \lambda, \Lambda, \delta^2 | y) \ \propto \ P(\delta^2)P(\lambda)P(\Lambda) \prod_{i=1}^{p} P(k_i|\lambda)P(\xi_i|k_i)P(\mathcal{M}_i|\Lambda) \tag{4}$$
$$\prod_{h=1}^{k_i} P([\sigma_i^h]^2)P(a_i^h|\mathcal{M}_i, [\sigma_i^h]^2, \delta^2)P(y_i^h|\xi_i^{h-1}, \xi_i^h, \mathcal{M}_i, a_i^h, [\sigma_i^h]^2)$$

Inference: An attractive feature of the chosen model is that the marginalization over the parameters a and σ in the posterior distribution of (4) is analytically tractable: $P(k,\xi,\mathcal{M},\lambda,\Lambda,\delta^2|y) = \int P(k,\xi,\mathcal{M},a,\sigma,\lambda,\Lambda,\delta^2|y)dad\sigma$ See Andrieu and Doucet (1999), Lèbre et al. (2010) for details and an explicit expression. The number of changepoints and their location, k, ξ, the network structure $\mathcal{M}$ and the hyperparameters λ, Λ and δ^2 can be sampled from the posterior $P(k, \xi, \mathcal{M}, \lambda, \Lambda, \delta^2|y)$ with RJMCMC. A detailed description can be found in Lèbre et al. (2010). The posterior probabilities of the gene interactions submitted to DREAM are obtained from the posterior sample of network structures $\mathcal{M}$ by marginalization.

3 Simulations and Results

To assess the performance of the proposed method we participated in a competition organised by the DREAM (Dialogue for Reverse Engineering Assessments and Methods) consortium in autumn of 2010. The goal was to reverse engineer gene regulatory networks from gene expression data sets. Participants were given four microarray compendia and were challenged to infer the structure of the underlying transcriptional regulatory networks. The first compendium was based on an in-silico (i.e. simulated) network,

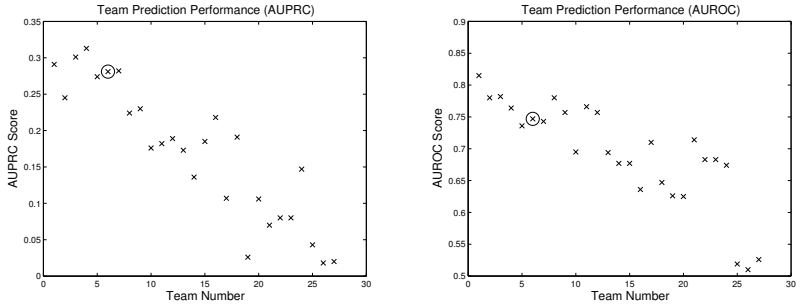


FIGURE 1. Areas under the precision recall (left) and ROC (right) curves obtained on an in silico data set by all teams participating in the DREAM 5 competition. The circles indicate the performance of our proposed method.

TABLE 1. This table summarises the information about the DREAM 5 Network Inference Challenge data sets. For each data set, we show which organism it came from, how many genes were measured, how many of those genes were identified as transcription factors (possibly regulatory genes) and how many chips (datapoints) were included.

Data Set	Organism	Genes	Transcription Factors	Chips
1	Synthetic	1643	195	806
2	*S. Aureus*	2810	99	160
3	*E. Coli*	4511	334	805
4	*S. Cerevisiae*	5950	333	536

the other three compendia were obtained from microorganisms. Each compendium consisted of hundreds of microarray experiments, which included a wide range of genetic, drug, and environmental perturbations. More information is available in Table 1 and at `http://wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project`. Network predictions were evaluated by the organisers on a subset of known interactions for each organism, or on the known network for the in-silico case (which is more objective).
Our method assumes an ordering of the microarray chips. While this condition is naturally met for time course experiments, it does not hold for the varying experimental conditions of the DREAM data. We therefore resorted to the heuristic pre-processing step of mapping the high-dimensional gene expression profiles onto a one-dimensional self-organising map (SOM) initialized by the first principal component. We applied the software package *som* in R with default parameter settings. To reduce the computational complexity of the RJMCMC simulations we applied a pre-filtering

step based on TESLA (Ahmed and Xing, 2009), a time-varying network inference method based on L1-regularised linear regression. For each gene we identified a set of 20 potential candidate regulators, based on the 20 regression coefficients with the largest modulus.
We assessed the convergence of our simulations with standard diagnostics based on Gelman-Rubin potential scale reduction factors (PSRF). Owing to unexpected downtime of the computer cluster we were using, only the simulations on the first two data sets showed a sufficient degree of convergence (PSRF≤ 1.2); for the latter data sets we submitted the results from TESLA. The second data set was later removed from the evaluation by the organisers. Figure 1 shows the results for the in silico data set obtained from the rankings of interactions submitted by all participating teams, using two criteria: the area under the precision-recall curve (AUPRC), and the area under the receiver-operator characteristic (AUROC) curve. As discussed in Davis and Goadrich (2006), AUPRC gives a more faithful indication of the network reconstruction accuracy than AUROC, and it is thus seen that our method clearly lies in the group of the 5 top-ranked models. This suggests that it compares favourably with the majority of existing schemes and provides a useful tool for contemporary research in systems biology.

Acknowledgments: Marco Grzegorczyk is supported by the Graduate School Statistische Modellbildung of the Department of Statistics, TU Dortmund University. Dirk Husmeier is supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD) and under the EU FP7 project TiMet. Andrej Aderhold's involvement in this project was funded by RERAD. Frank Dondelinger's research is funded by RERAD and the UK Engineering and Physical Sciences Research Council (EPSRC). We are grateful to Tony Travis for granting us user time on the Beowulf cluster at the Rowett Institute in Aberdeen and for providing support with respect to a parallelisation of the processes.

References

Ahmed, A., and Xing, E.P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, **106**:29, 11878-11883.

Andrieu, C., and Doucet, A. (1999). Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Transactions on Signal Processing*, **47**, 2667-2676.

Cantone, I., Marucci, L., Iorio, F., Ricci, M. A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., and Cosma, M. P. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, **4**:130.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proc. of the 23rd Int. Conf. on Machine Learning*

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**:4, 711-732.

Grzegorczyk, M., Husmeier, D. and Werhli, A.V. (2008). Reverse engineering gene regulatory networks with various machine learning methods. In: *Analysis of Microarray Data: A Network-Based Approach*, Wiley Online Library.

Lèbre, S., Becq, J., Devaux, F., Stumpf, M.P.H. and Lelandais, G. (2010). Statistical inference of the time-varying structure of gene regulation networks. *BMC Systems Biology*, **137**, 172-181.

Analysis of an Observational Study

Cara Dooley[1], John Hinde[1], Harry Comber[3], Larry Egan[2], John Newell[2]

[1] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, University Rd, Galway, Ireland. c.dooley6@nuigalway.ie
[2] HRB Clinical Research Facility, School of Medicine, National University of Ireland, Galway, Ireland.
[3] National Cancer Registry, Cork, Ireland.

Abstract: The study presented below aimed to compare survival of colorectal cancer patients against survival of a sub-population with a secondary disease, inflammatory bowel disease (IBD).
The data were taken from a observational study, that is there was no explicit design. The study had many complications, but the most significant aspect was that the number of controls was much greater than the number of cases of interest. Some techniques are used to overcome these obstacles, including: matching of the dataset, to make the controls and cases as similar as possible at time of diagnosis, effectively retrospectively fitting a design; weighting of the data, using both the propensity score and the number of similar patients found in matching.

Keywords: Observational Study; Propensity Score; Matching; Kaplan-Meier; Cox Model.

1 Introduction

The aim of the study was to compare survival of colorectal cancer patients in the whole population against the survival of patients in a sub-population who also had inflammatory bowel disease (IBD). All individuals who suffered from colorectal cancer were drawn from the entire Irish population using data from January 1994 to December 2005 provided by the National Cancer Registry of Ireland (NCRI). The control group contained many more observations ($n > 20000$) when compared to the IBD group ($n = 170$). Given the number of control patients, there was large diversity in this group. In a conventional designed experiment or trial, patients entering the trial would be randomised across arms of the study, with similar numbers in each group. Usually patients would be similar in age, health, etc. As this was an observational study, there was no design prior to collecting the data and so no benefit, in terms of bias protection, from randomization in terms of the balance of the distribution of unobserved explanatory variables.

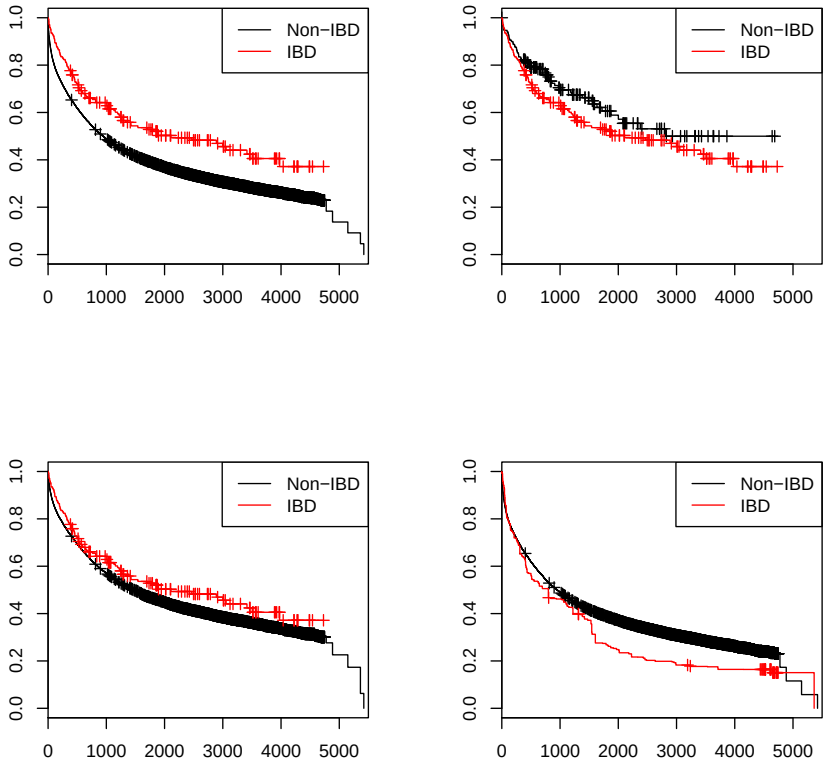


FIGURE 1. Comparison of the four methods used to produced Survival Curves: (a) The whole data set using the conventional KM estimates; (b) The matched dataset using conventional KM estimates; (c) The whole dataset using the number matched as weights; (d) the whole dataset using Adjusted KM estimates.

1.1 Analysis of the full dataset

Initially, the whole data set was analysed. Kaplan-Meier estimates were examined, as seen in Figure 1(a). A Cox proportional hazards model was fitted and all factors except for IBD were found to be significant ($p = 0.4121$). These factors included age, gender and various descriptors of the disease, including tumour type, location and stage of illness, i.e., the effect of IBD as seen in Figure 1(a), was eliminated by covariate adjustment.

2 Matching

One approach to implement a design in a observational study is to use matching. In this example, we match the IBD patients to the nearest control by minimizing the Mahalanobis distance between them using the `optmatch` package in R. The Mahalanobis distance has an added calliper (or penalty) calculated using propensity scores, as suggested by Rosenbaum (2010). Following the matching, Kaplan-Meier estimates were again calculated, as shown in Figure 1(b). As there may still be heterogeneity between the members of a pair that is unexplained by the matching variables, a Cox proportional hazards model with a frailty term was also fitted to compare the risk of death for IBD and non-IBD patients while adjusting for the matching variables. Again, IBD was found to be non-significant ($p = 0.29$), the frailty term was also non-significant ($p = 0.92$), the two variables describing the severity of the illness were still found be significant, all other terms were non-significant.

2.1 Propensity Score

The conditional probability of being in the treated group ($Z = 1$) given the observed covariates x, is called the propensity score,

$$e(X) = P(Z = 1|x)$$

The propensity score $e(x)$ balances on observed bias, but not on the unobserved bias. In practice the estimated propensity score $\hat{e}(x)$ is used. To obtain the estimated propensity score, we fit a logistic regression model and use the estimated fit as the estimated propensity score, $\hat{e}(x)$, however other models may be used. The model can be over-fitted, including all variables available at time of diagnosis. The estimated propensity score, $\hat{e}(x)$ will not only balance on observed bias, but also on some of the unobserved bias (Rosenbaum and Rubin, 1983).

3 Analysis

While matching is a useful technique, in a simple 1:1 match much of the data remains unused. Some alternatives which are useful in this situation include the Weighted Kaplan-Meier (Winnett and Sasieni, 2002), the Adjusted Kaplan-Meier (Xie and Liu, 2005) and the adjusted Cox proportional hazard model (Sugihara, 2010).

3.1 Weighted Kaplan-Meier

Winnett and Sasieni (2002) suggest full matching, that is matching all available controls to cases and then weighting the Kaplan-Meier estimates

by the number of controls matched to each case.

$$\hat{S}^w(t) = \prod_{u:u\leq t}\left[1 - \frac{\sum_{j=1}^{k} w_j d_j(u)}{\sum_{j=1}^{k} w_j r_j(u)}\right]$$

where, $d_j(u)$ = number of events at time u in stratum j, $r_j(u)$= number at risk at u in stratum j and $w_j = 1/mj$ is the reciprocal of the stratum size. When the same number of controls are matched to each case this reduces to the usual KM estimates. The results of this can be seen in Figure 1(c).

3.2 Adjusted Kaplan-Meier Estimator - AKME

Xie and Liu (2005) suggest using the inverse of the propensity score to weight the Kaplan-Meier, assigning a weight $w_{ik} = 1/p_{ik}$ to each individual, where p_{ik} is the propensity score for individual i in group k.
So the AKME for the kth group is

$$\hat{S}^k(t) = 1 \quad \text{if } t < t_i$$

or

$$\hat{S}^k(t) = \prod_{t_j\leq t}\left[1 - \frac{d_{jk}^w}{Y_{jk}^w}\right] \quad \text{if } t_i \leq t$$

where, d_{jk}^w is the weighted number of events and Y_{jk}^w is the weighted number at risk.
The results of this are shown in Figure 1(d).

3.3 Adjusted Cox Proportional Hazards Model

In the same way that Kaplan-Meier estimates were adjusted using the inverse propensity score as weights, the Cox proportional hazards model may be modified as proposed by Sugihara (2010). After fitting the adjusted Cox proportional hazards model, except for gender, all factors, including IBD, were significant ($p < 0.0001$).

4 Results

As mentioned, matching is a useful technique, however, when using 1 : 1 matching, much of the data remains unused. The three methods mentioned is Section 3, all use the whole dataset adjusting for the disparity in numbers between the two groups. The adjusted Cox proportional hazards model is the only model which finds a significant difference between the IBD group and the control. Further work is required to see if this an artifact of the weighting or a true difference.

The propensity score is known to be unstable when the data set is large or contains a great disparity between the number of cases and controls. There are stabilization techniques in the literature that attempt to address this issue, however one such method was applied to this data which showed little effect.

Acknowledgments: This work was supported by Science Foundation Ireland grant 07/MI/012

References

Hansen, B. (2007). Optmatch: Flexible, Optimal Matching for Observational Studies. *R News*, **7**, 18-24.

Rosenbaum, P.R. (2010). *Design of observational studies*. London: Springer Series in Statistics.

Rosenbaum, P. R. and Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41-55.

Sugihari, M. (2010). Survival analysis using inverse probability of treatment weighted methods based on the generalized propensity score. *Pharmaceutical Statistics*, **9**, 21-24.

Winnett, A. and Sasieni, P. (2002). Adjusted Nelson-Aalen estimates with retrospective matching. *Journal of American Statistical Association*, **97**, 245-256.

Xie, J. and Lui, C. (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine*, **24**, 3089-3110.

Sea Level Trend Estimation by Seemingly Unrelated Penalized Regressions

Paul H. C. Eilers[1], Richard Duin[2] , Douwe Dillingh[3]

[1] Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands
[2] Dutch Ministry of Public Works, The Hague, The Netherlands
[3] Deltares, Delft, The Netherlands

Abstract: Fluctuations around long-term sealevel trends at different monitoring sites are very similar. Explicit modelling of the fluctuations, similar to Seemingly Unrelated Regressions, strongly reduces the uncertainty in trend estimates.

Keywords: P-splines, sea levels, trend estimation

1 Introduction

One of the potential effects of global warming is a rise in sea levels. Sea water expands if its temperature rises, and the melting of gletschers and the ice caps at the poles adds more volume. Some predictions amount to a rise of one meter or more in the next century. Over 50 per cent of the area of the Netherlands lies below the present average sea level, protected by dunes and dikes. It is clear that global warming can have serious consequences for our safety and economy.

The Dutch government operates a network of monitoring stations along the coast. Some of the stations are already in operation for almost 200 years. Data are collected continuously, and summarized to different levels of detail. Here we will be concerned with yearly averages.

From a series of yearly levels one can compute trends, using statistical models, and use these to forecast future levels. This is being done on a regular basis. From the data of each station a separate trend is computed, assuming independent errors around the trend. When looking at each station in isolation, this looks like a reasonable choice. However, when plotting the residuals of the individual stations, a striking similarity is seen. Apparently large-scale processes in the weather, or in the North Sea as a whole, strongly influence the yearly levels at all stations in the same way.

We can exploit the similarity by adapting the model known as Seemingly Unrelated Regressions (SUR), familiar to econometricians (Mittelhammer et al., 2000). We combine it here with trend estimation by penalized regression and use a simplified correlation structure. We propose to call it Seemingly Unrelated Penalized Regressions (SUPR).

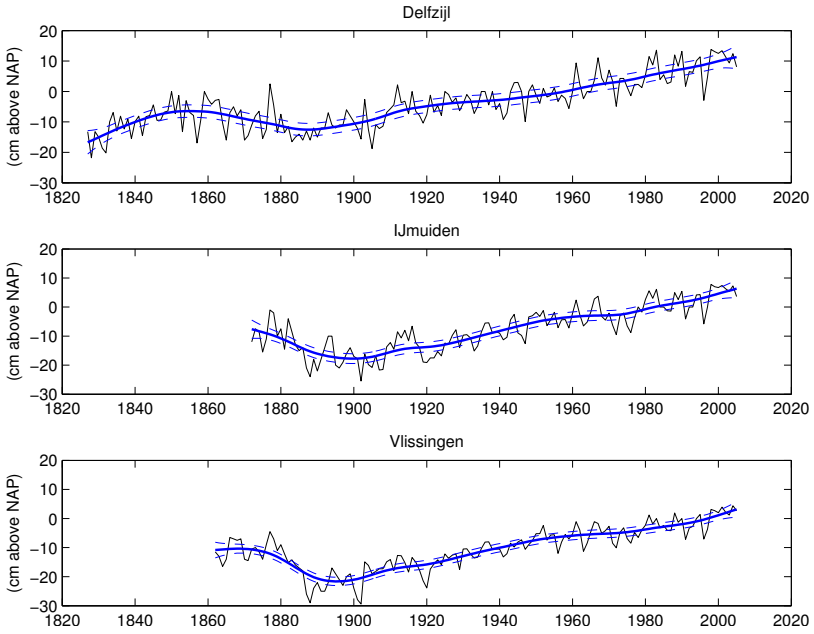


FIGURE 1. Data for three sea level monitoring stations, with individually estimated trends (full line) plus and minus point-wise standard errors (broken lines).

2 The model

We have data from n stations for m years, in a data matrix Y. To simplify the presentation, we assume that Y is complete. This is not the case in practice, because not all stations have been operating during all m years. But this introduces no problems if 0/1 weights are used in the implementation of the calculations.
The model is

$$y_{ij} = \mu_{ij} + e_{ij} = f_{ij} + u_i + e_{ij}, \tag{1}$$

where f_{ij} gives the smooth trend for station j in year i, u_i is a common disturbance and e_{ij} a random error, assumed to be independent. The smooth trend is modeled by P-splines, a combination of a B-spline basis $B = [b_{ik}]$ and a difference penalty on the B-spline coefficients α (Eilers and Marx, 1996). For just one time series, the P-spline objective function to be minimized is

$$S = ||y - B\alpha||^2 + \lambda||D\alpha||^2, \tag{2}$$

where D is a matrix that forms second order differences. The number of B-splines in the basis is chosen large enough, so that the potential flexibility

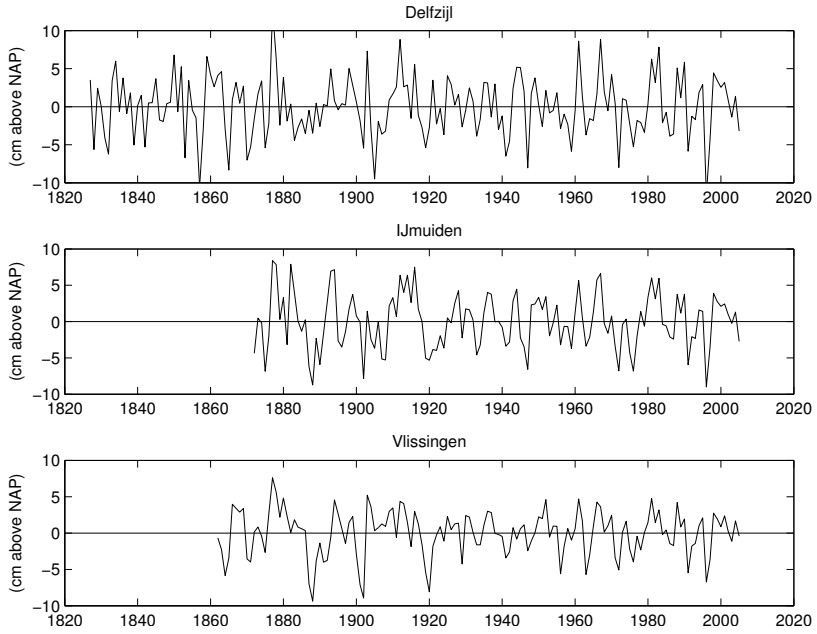


FIGURE 2. Differences between the data and the estimated trends, that are shown in Figure 1.

of the fit is larger than needed. The parameter λ is used to tune smoothness to the desired level.

In the case of n time series we have n vectors of coefficients α_{kj}, forming the columns of a matrix A, and $f_{ij} = \sum_k b_{ik}\alpha_{kj}$. The objective function is

$$S = \sum_i \sum_j (y_{ij} - \sum_k b_{kj}\alpha_{kj} - u_i)^2 + \lambda \sum_j \sum_k (\Delta^2 \alpha_{kj})^2, \qquad (3)$$

where Δ^2 is the operator for second order differences. Using Kronecker products, this can be written as a large penalized regression problem. The system contains mainly zeros and can be solved quickly with sparse matrix functions (we use Matlab).

In principle the parameter λ can be optimized using cross-validation or AIC. But this only works if the errors are really independent. Although this assumption was made when introducing the model, it turns out not to be true in practice: there is serial correlation, indicating light smoothing for optimal prediction. But we are interested in long-term trends, hence we have used our carpenter's eye to set λ.

The serial correlation of the errors also has consequences for standard error estimates, because effective degrees of freedom are smaller. Presently our model does not yet account for this.

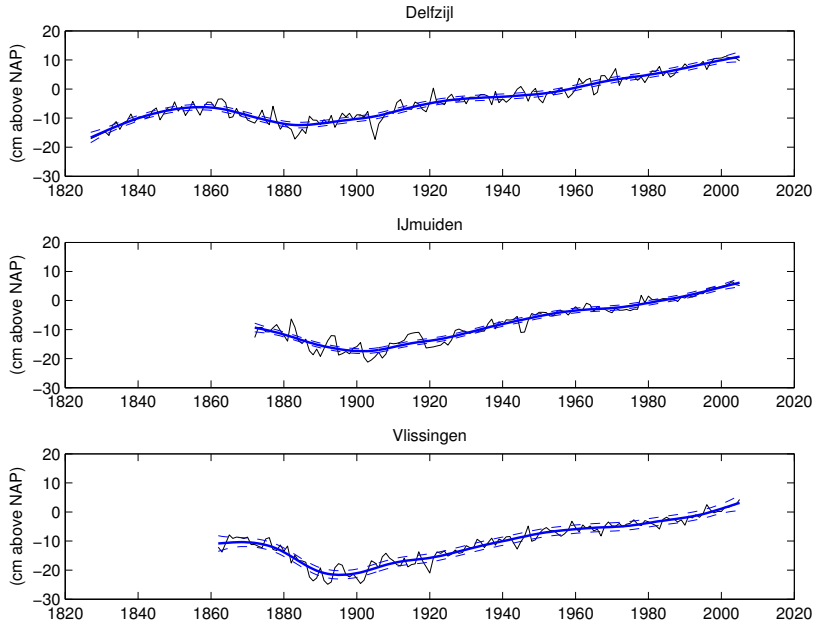


FIGURE 3. Data for three sea level monitoring stations, after subtraction of the shared disturbances, with individually estimated trends (full line) plus and minus point-wise standard errors (broken lines).

The model as described gives a good fit to data from monitoring stations along the Dutch coast. A careful study of the residuals showed that it can be improved. The shared disturbances have similar shapes, but they gradually decrease in strength, going form North to South. To account for this, the model is modified as follows:

$$y_{ij} = \mu_{ij} + e_{ij} = f_{ij} + c_j v_i + e_{ij}, \tag{4}$$

where v_i represents the common pattern in year i and c_j the local strength. For identifiability, the condition $\sum_j c_j^2 = n$ is imposed. We now have a bilinear structure for the shared disturbances, and estimation becomes a bit more complex. We iterate between smoothing of each individual series $y_{ij} - \tilde{c}_j \tilde{v}_{ij}$ and the singular value decomposition of $y_{ij} - \tilde{f}_{ij}$ (where a tilde indicates the current approximation).

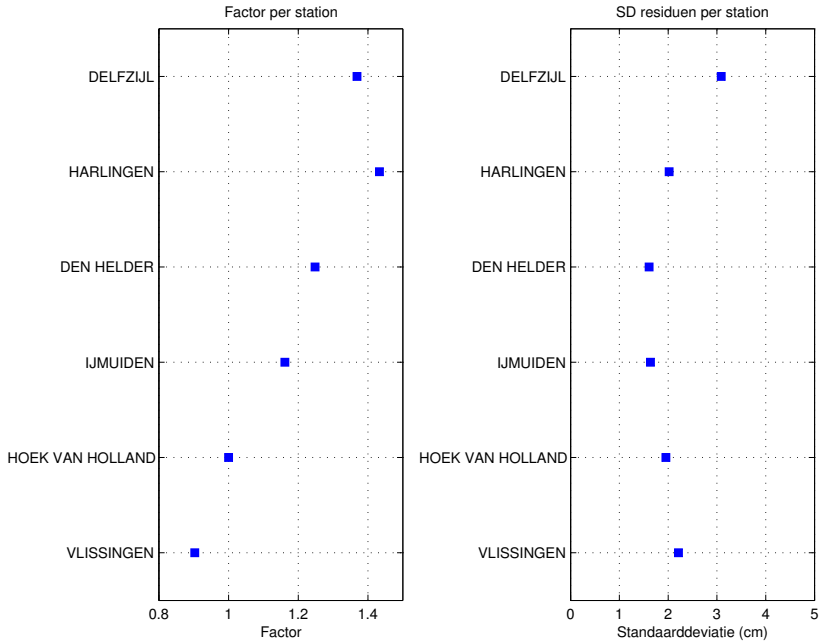


FIGURE 4. Left panel: the strength parameter (c) for the shared disturbances in the extended model, for each monitoring stations. Right panel: standard deviations of residuals.

3 Application to North Sea levels

Figure 1 shows time series for three monitoring stations along the North Sea coast, one in the North (Delfzijl), one in the South (Vlissingen), and one approximately half-way in between (IJmuiden). The trends have been computed with simple smoothing. The differences between data and trends are shown in Figure 2; their similarity is quite clear. A simple way to show the effectiveness of the model in (1) is to present $y_{ij} - \hat{u}_i$, as is done in Figure 3. The error bands around the trends are much smaller there. Note that the shared disturbances have been estimated from a set of eight stations, not only from the three stations shown here.
Figure 4 summarizes results form the extended model in (4). It shows how c decreases from North (Delfzijl) to South (Vlissingen).

4 Discussion

We have presented a modification of the SUR model and applied it to the simultaneous estimation of sea level trends at multiple monitoring stations, dramatically improving precision. Compared to usual SUR, the regressions are more complicated, due to the penalties. The error structure is simplified however. In SUR a general covariance matrix between the equations is used, and has to be estimated and inverted. In our model all equations share the same fluctuations, and in addition there is noise, assumed to be independent (but it does not seem to hard to introduce an AR process).
We certainly are not the first to combine SUR ideas with smoothing. Lang et al (2003) introduced a model for additive spatial modelling using Bayesian P-splines.
AS far as we know, the bilinear error structure in (4) has not been proposed before.
There is no room for describing the details of a useful structural extension that has been studied, individual step functions per station, to model the effects of sudden changes of levels, caused by large artificial waterworks. An example is the Afsluitdijk, a dike that closed off the Zuiderzee in 1937. For the size of each step a parameter is added to the model.
It was found that on a monthly scale the model gives an even better fit. Strong seasonal patterns were found in the shared disturbances.
The data were obtained form the public database PSMSL (Permanent Service for Mean Sea Level at www.psmsl.org). It contains sea level records of over one thousand monitoring stations worldwide. Such a rich collection of data offers many opportunities for additional research.

References

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

Lang, S., Adebayo, S., Fahrmeir, L. and Steiner, W. (2003) Bayesian Geoadditive Seemingly Unrelated Regression. *Computational Statistics* **18**, 163–192.

Mittelhammer, R.C., Judge G.G., and Miller, D.J. (2000) *Econometric Foundations*. Cambridge University Press.

Generalized random intercept log-gamma exponential family models

Lizandra C. Fabio[1], Gilberto A. Paula[1] , Mário de Castro[2]

[1] Dept of Statistics, University of São Paulo, Brazil, e-mail:lcfabio@ime.usp.br, Dept of Statistics, University of São Paulo, Brazil, e-mail:giapaula@ime.usp.br
[2] Institute of Mathematical Sciences and Computation, University of São Paulo, Brazil, e-mail:mcastro@icmc.usp.br

Abstract: We propose in this work the exponential family (EF) models in which the random effect distribution is assumed to follow a generalized log-gamma (GLG) distribution. An application in which the outcome follows a gamma distribution is presented.

Keywords: Exponential family; Random effect model; Generalized log-gamma distribution.

1 Introduction

The generalized linear mixed model (GLMM) class (Breslow and Clayton, 1993; MucCulloch and Searle, 2001) is a extension of the random effect model proposed by Laird and Ware (1992) to model correlated data structures and accommodate the overdispersion often observed in counting data. Lee and Nelder (1996) have investigated the flexibilization of the random effect distribution in this model class under a hierarchical framework. Recently, Molenberghs et al. (2007) have suggested a combination between gamma and normal random effects in Poisson mixed models deriving the marginal distribution of the response variable.
The aim of this paper is to present an alternative distribution for the random effect in random intercept exponential family (EF) models, that is characterized by assuming a generalized log-gamma (GLG) distribution for the random effect component. This distribution introduced by Prentice (1974) has as particular cases the normal and extreme value distributions and it assumes skew forms to right and left. The generalized log-gamma distribution has been widely applied by Lawless, 2002 and Ortega et al., 2009). In general, numerical integration methods are required to a previously analyzed data set (Hadgu and Koch, 1999) is presented.

2 Generalized random intercept log-gamma EF models

Let y_{ij} denote the jth outcome measured for the ith cluster (subject), $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. We will assume the following random intercept Poisson model:

(i) $y_{ij}|b_i \overset{\text{ind.}}{\sim} \text{P}(u_{ij})$,

(ii) $u_{ij} = \exp(\mathbf{x}_{ij}\boldsymbol{\beta} + b_i)$ and

(iii) $b_i \overset{\text{i.i.d}}{\sim} \text{GLG}(0, \sigma, \lambda)$,

where $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$ contains values of explanatory variables and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. When $\lambda = 0$, model (i)-(iii) reduces the generalized mixed model proposed by Breslow and Clayton (1993). Let $f_{Y|b}(y_{ij}|b_i, \boldsymbol{\beta})$ and $f_b(b_i; \sigma, \lambda)$ be the pdf of $y_{ij}|b_i$ and the pdf of u_i, respectively. Then, the marginal pdf of $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^\top$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^\top$, is given by

$$f_Y(\mathbf{y}_i; \boldsymbol{\beta}, \sigma, \lambda) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{m_i} f_{Y|b}(y_{ij}|b_i, \boldsymbol{\beta}) \right\} f_b(b_i; \sigma, \lambda) db_i, \qquad (1)$$

which in general does not have random closed-form. The marginal likelihood function presented in (1) from the intercept GLG-EF model is given by

$$L(\boldsymbol{\theta}) = \log \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{m_i} f_{Y|b}(y_{ij}|b_i, \boldsymbol{\beta}) \right\} f_b(b_i; \sigma, \lambda) db_i. \qquad (2)$$

We use the NLMIXED procedure in SAS to maximize (2) returning the parameter estimates. Considering the Poisson distribution in (i), in the proposal model, we have shown for some particular parameter setting that the marginal distribution assumes a closed-form expression, such as, the multivariate negative binomial (MNB) distribution (see, for instance, Johnson et al. 1997). Person analysis for the model proposed and local influence for the MNB model have been made.

3 Application

We present an example with dental plaque data set described by Hadgu and Koch (1999) who discussed the results of a clinical trial with 109 adult volunteers with pre-existing dental plaque. In the study, subjects were randomly distributed to receive a liquid type A (34 subjects), a liquid type B (36 subjects) e um liquid control (39 subjects). The dental plaque score of

each individual was assessed and classified in the early of treatment, after 3 moths and 6 months. The aim of the study was to verify if at least one of the new liquids reduce the average score of dental plaque. The score measures the influence of the liquids on the dental plaque and lower is the score, greater is the effect of liquids in reducing dental plaque. Let y_{ijk} denote the score of the kth subject in the ith group and jth period for $i, j = 1, 2, 3$ and $k = 1, \ldots, n_{ij}$, with $n_{1j} = 39$, $n_{2j} = 34$ and $n_{3j} = 36$. We assume the model given by (i)-(iii), in that $y_{ijk}|b_k \sim$Gamma(u_{ijk}, ϕ) and $\eta_{ijk} = \log(u_{ijk}) = \alpha + \beta_i + \gamma_j + \delta_{ij} + b_k$, in which β_i and γ_j are the main effects and δ_{ij} are the interactions between treatment and period. The restriction $\beta 1 = 0$, $\gamma_1 = 0$, $\delta_{1j} = 0$ and $\delta_{i1} = 0$ were considered, for $i = 1, 2, 3$ and $j = 1, 2, 3$.

TABLE 1. Parameters estimates with the respective approximate standard errors for the random intercept Gamma-Normal and Gamma-GLG models

Effect	Estimate	Std. error	Estimate	Std. error
α	2.3035	0.0906	1.1111	0.1017
β_2	0.0021	0.0722	0.0828	0.1192
β_3	-0.0297	0.0717	-0.0380	0.1164
γ_2	-0.4135	0.0776	-0.3966	0.0944
γ_3	-0.4299	0.0799	-0.4183	0.0940
δ_{22}	-0.4984	0.1235	-0.3852	0.1368
δ_{23}	-0.4164	0.1243	-0.4027	0.1377
δ_{33}	-0.3182	0.1186	-0.3201	0.1339
ϕ	0.2671	0.0213	6.1849	0.5540
σ	1E-8	-	0.0813	0.0172
λ			2.7687	0.7147
AIC	638.36			634

The parameter estimates presented in the Table 1 show that the liquids A e B decrease in average the amount of dental plaque and a marked reduction of the liquid B from 3 to 6 months of brushing in both models. However, the AIC criterion suggests evidence that the GLG-Gamma model yields the better fit. We must also observe that the zero value does not belong to the 95% confidence interval of the λ parameter given by [1.368, 4.17]. Moreover, the estimate $\hat{\lambda} = 2.7687$ indicates that the random effect is skewed to the left.

Acknowledgments: The authors are grateful to CNPq, CAPES and FAPESP, Brazil.

References

Breslow and Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

Hadgu, A., and Koch, G. (1999). Application of generalized estimating equations to a dental randomized clinical trial. *Journal of Biopharmaceutical Statistical*, **9**, 161-178.

Johnson, N.L., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York.

Laird and Ware (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Lee, Y., and Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58**, 619-678.

McCulloch, C.E., and Searle, S.R. (2001). *Generalized Linear and Mixed Models*. Wiley, New York.

Molenbergs, G., Verbeke, G., and Demétrio, G.G.B. (2007). An extended random-effects approach to modeling repeated, overdispersed and count data. *Lifetime Data Analysis*, **13**, 513-531.

Ortega, E.M.M., Cancho, V.G., and Paula, G.A. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, **15**, 79-106.

Prentice, R.L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika*, **61**, 539-544.

Modelling Financial Data using Poisson Mixture Approach

S. Faria[1], F. Gonçalves[2]

[1] Department of Mathematics and Applications, Mathematical Research Centre, University of Minho, 4800-058 Guimarães, Portugal *sfaria@math.uminho.pt*
[2] University of Minho, 4800-058 Guimarães, Portugal *fat.rod.goncalves@sapo.pt*

Abstract: Poisson mixture regression models are commonly used in financial applications to analyze heterogeneous count data. In these models, the observed counts are assumed to come from two or more subpopulations and parameter estimation is typically performed by means of maximum likelihood via the EM algorithm. In this study, we discuss briefly the fitting of Poisson mixture regression models using maximum likelihood methods. These models' methodology is applied to a real data set for credit-scoring purposes. We model the number of defaulted payments of clients for a bank, who had obtained loans for consumption.

Keywords: Poisson mixture regression models; EM algorithm; count data; heterogeneity

1 Introduction

Finite mixture models are a well-known method for modelling unobserved heterogeneity (see e.g. McLachlan and Peel (2000) and Frühwirth-Schnatter (2006) for a review). In particular, Poisson mixture regression models (PMR) are commonly used to analyze heterogeneous count data.
Let the random variable Y_i denote the ith response variable, and let $(\mathbf{x}_i, y_i), i = 1, \ldots, n$ denote observations where y_i is the observed value of Y_i and $\mathbf{x}_i$ a $(p+1)$-dimensional covariate vector. It is assumed that the marginal distribution of Y_i follows a mixture of Poisson distributions,

$$Y_i \sim \sum_{k=1}^{K} \pi_k f_k(y_i|\mathbf{x}_i, \lambda_{i|k}) \tag{1}$$

where

$$f_k(y_i|\lambda_{i|k}) = \frac{\exp(-\lambda_{i|k})(\lambda_{i|k})^{y_i}}{y_i!}, \quad i = 1, \ldots, n, k = 1, \ldots, K \tag{2}$$

and $\lambda_{i|k} = \exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)$, with $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})^T$ denoting the $(p+1)$-dimensional vector of regression coefficients for kth component. The proportions π_k are the mixing probabilities ($0 < \pi_k < 1$, for all $k = 1, \ldots, K$

and $\sum_k \pi_k = 1$) and can be interpreted as the unconditional probability that an observation arises from component k of the mixture.
In this work, the methodology of Poisson mixture regression models is applied to a real data set to predict a client's number of defaulted payments. Using covariates in all components we aim to reveal the impact of demographic and financial variables in creating different groups of clients and to predict the group to which each client belongs, as well as, his expected number of defaulted payments.

2 Data

The data consist of a random sample clients who had been granted credits for consumption from a Portuguese bank.
The credits focused in these data are credits taken on by individual consumers for personal, family or household purposes.
A description of the data is presented in Table 1. The sample was taken on 31st December, 2008. All records correspond to clients who were granted credit and whose contract is not completed yet. For each client, there is available information on his characteristics at the beginning of the contract and there is also recorded the total number of defaulted payments, i.e. the number of consecutive monthly payments that were not paid and should have been paid, by the sampling date (variable *Nnonpay*). This variable is taken as the dependent variable. The sample mean and variance of the number of defaulted payments are 0.524 and 1.320, respectively, suggesting the data are overdispersed and indicating the inadequacy of the standard Poisson regression model.

3 Model

To estimate the model we used the methods available in the R package `flexmix` (see Leisch (2004) and Grün and Leisch (2008)).
The number of components of Poisson regression models to be fitted was unknown needing, therefore, to be estimated from the data. To determine it we employed information criteria: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). It was interesting to notice the two criteria resulted in different number of components with BIC selecting a model with fewer parameters. Following Wang *et al.* (1996) recommendation, we relied on the BIC criterion selecting 3 components.
Attempting to avoid convergence to a local maximum, the EM algorithm was run 15 times, using different starting values and the model with maximum likelihood was chosen. For each trial the algorithm was stopped when the relative change in the log likelihood between two successive iterations was smaller than 10^{-12}.

TABLE 1. Description of the variables used in the study

Variable	Description
NMDec08.c1	1 if the age of the contract at the sampling date is less than 7.5 months, 0 otherwise.
NMDec08.c2	1 if the age of the contract at the sampling date is between 7.5 and 11.5 months, 0 otherwise.
NMDec08.c3	1 if the age of the contract at the sampling date is more than 11.5 months, 0 otherwise.
Install.c1	1 if the monthly installment is less than 140.7 euros, 0 otherwise.
Install.c2	1 if the monthly installment is between 140.7 and 506.7 euros, 0 otherwise.
Install.c3	1 if the monthly installment is more than 506.7 euros, 0 otherwise.
Age.c1	1 if the age group is 18 − 42 years, 0 otherwise.
Age.c2	1 if the age group is 42 years or more, 0 otherwise.
AvgSBl.c1	1 if the semesterly average account balance of the client is less than 11.5 euros, 0 otherwise.
AvgSBl.c2	1 if the semesterly average account balance of the client is between 11.5 and 136.5 euros, 0 otherwise.
AvgSBl.c3	1 if the semesterly average account balance of the client is between 136.5 and 325 euros, 0 otherwise.
AvgSBl.c4	1 if the semesterly average account balance of the client is more than 325 euros, 0 otherwise.
Gender	Gender of the client.
NYClient.c1	1 if the number of years as client of the bank is less than 14.5, 0 otherwise.
NYClient.c2	1 if the number of years as client of the bank is more than 14.5, 0 otherwise.
Education	Level of Education: 0− Unknown, 1− Primary Education, 2− High School, 3− Professional education, 4− University Degree.
Occupation	Professional occupation: 0− Unknown, 1− Student or housewife 2− Sales, Service or Technical, 3− Small or medium enterprises, 4− Professional, 5− Other.
RecSalary	Indicator of wether the client receives the salary through the bank.
Region	Region of residence in Portugal: 1− North, 2− Center, 3− Lisbon, 4− Alentejo/Algarve, 5− Madeira and 6− Azores.
Nnonpay	Number of consecutive monthly defaulted payments.

Table 2 reports the estimated coefficients for the Poisson regression model with 3 components. We can see how the effect of covariates differs between components. There are variables with large coefficients (in absolute value) for some components and small for others. There are also variables with a different sign between the components. This shows the regression part of the model captures the characteristics of each group which differ from one another.

4 Discussion

This study shows the application of poisson mixture regression modelling financial data. The results are very interesting, revealing that the population consists of three groups, contrasting with the typical good versus bad categorization approach of the credit-scoring systems.

Acknowledgments: S. Faria wants to acknowledge the financial support provided by the Research Centre of Mathematics of the University of Minho through the FCT Pluriannual Funding Program.

TABLE 2. Estimated coefficients for the Poisson regression model with 3 components.

Variable	Component					
	1		2		3	
	Coefficient	Std Error	Coefficient	Std Error	Coefficient	Std Error
Intercept	-0.1979	0.2662	-0.4785	0.1899	-0.4472	0.6146
Gender1	0.3138	0.1346	0.1532	0.1113	-0.0609	0.0929
RecSalary1	-0.8315	0.1506	-0.3645	0.1318	-3.5877	0.4959
Education1	-0.0820	0.1693	-0.0968	0.1351	0.1520	0.1363
Education2	-0.0869	0.1672	-0.0635	0.1241	0.2345	0.1360
Education3	-0.0745	0.3259	-0.1798	0.2668	-0.8166	0.3491
Education4	-0.5107	0.4295	-0.1900	0.3025	0.4302	0.2970
Occupation1	-0.1268	0.3318	-0.1448	0.2979	-0.1411	0.1788
Occupation2	0.0502	0.2042	0.2633	0.1372	-0.2179	0.1187
Occupation3	1.5712	0.1746	0.9342	0.1463	-1.4841	0.2946
Occupation4	-15.0525	0.6430	0.9554	0.4470	0.0258	0.3655
Occupation5	1.6131	0.1868	1.5122	0.1404	-0.9756	0.2139
Region2	-0.0776	0.1763	0.0392	0.1273	0.0330	0.1572
Region3	-0.3032	0.2401	-0.0480	0.1754	0.3128	0.1809
Region4	-0.0554	0.2028	-0.0879	0.1523	0.5573	0.1832
Region5	-0.0570	0.2355	0.1641	0.1496	0.4021	0.1557
Region6	0.2650	0.1694	-19.1952	1.9270	0.5703	0.1466
NYClient.c2	-0.2600	0.1688	-0.3485	0.2024	-0.0890	0.1962
Age.c2	-0.1749	0.1090	0.1715	0.0961	-0.2823	0.1024
Install.c2	-0.1352	0.0958	-0.0268	0.0974	-0.1477	0.0915
Install.c3	0.3177	0.1696	-0.1456	0.2495	0.5311	0.3442
AvgSBl.c2	-0.6440	0.1527	-0.1308	0.1099	-1.8328	0.1717
AvgSBl.c3	-0.9562	0.2673	-0.2491	0.1456	-3.7184	0.6384
AvgSBl.c4	-0.7719	0.1472	-18.4007	1.3086	-3.5018	0.3424
NMDec08.c2	-0.0441	0.2274	-0.0638	0.2121	1.7627	0.6360
NMDec08.c3	0.3235	0.1568	0.0931	0.1444	2.1559	0.6221
proportions	0.210		0.499		0.291	

References

Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models*, Springer, Heidelberg.

Grün, B. and Leisch, F. (2008) Flexmix version2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, **28**, 1-35.

Leisch, F. (2004) Flexmix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, **11**, 1-18.

Mclachlan, G. J., and Peel,P. (2000) *Finite Mixture Models*, Wiley, New York.

Wang, P., Puterman, M. Cockburn, I. and Le, N. (1996) Mixed Poisson Regression Models with Covariate Dependent Rates. *Biometrics*, **52**, 381-400.

A multivariate space-time model for heterogeneous air quality networks

Francesco Finazzi[1], E. Marian Scott[2], Alessandro Fassò[1]

[1] Dept. of IT and Mathematical Methods, University of Bergamo, Via Marconi 5, 24044 Dalmine BG, Italy. francesco.finazzi@unibg.it
[2] Dept. of Statistics, University of Glasgow, 15 University Gardens, Glasgow, G12 8QQ, UK.

Abstract: Multivariate space-time models are useful tools for mapping pollutant concentrations over a region of interest. The mapped concentrations are then used to evaluate both aggregate statistics and air quality indicators. In this study, the spatio-temporal cross-correlation between pollutants is exploited in order to obtain more accurate concentration estimates even for those pollutants which are observed at a limited number of sites. The advantage of considering a multivariate model is verified by means of the leave-one-out cross-validation technique. Scottish air quality data for the year 2009 are considered.

Keywords: Multivariate space-time models; cross-validation; EM.

1 Introduction

Aggregate statistics produced as output from space-time air quality models, including air quality indicators, are useful measures summarizing the concentration of airborne pollutants over a region. The role of these statistics is twofold: to provide an air quality measure that can be easily understood by the public and to verify if the air quality targets imposed by both the local and the European legislation are met.
Although aggregate statistics and air quality indicators are used by the environmental agencies of many countries, a common and shared methodology for their definition is far from being developed. See, for example, Bodnar et al. (2008) on the comparison of air quality across states and Lee et al. (2011) for a Bayesian approach applied to a UK case.
An issue related with the definition and the estimation of the aggregate statistics is the fact that not every pollutant is measured at all the monitoring stations. In many cases, either the number of stations measuring a particular pollutant is too small or the stations do not cover the region properly; this results in poor pollutant concentration estimates (see Bodnar et al, 2008).
The aim of this study is to exploit the spatio-temporal correlation and cross-correlation between pollutants in order to provide, for each pollutant, daily

high resolution concentration maps characterized by lower uncertainty and to gain a better insight into the dynamics and the interactions between pollutants. This is done by considering a multivariate spatio-temporal model able to deal with heterogeneous monitoring networks (with respect to the number of pollutants measured at each station) and missing data. The concentration maps are eventually used to evaluate, for each pollutant, a simple aggregate statistic defined as the daily map average concentration. The aggregate statistic derived from the multivariate model is compared with the same statistic obtained as output of a univariate model concerning the single pollutant. The model is applied to Scottish air quality data for the year 2009 comprising 6 different pollutants, namely Ozone (O_3), Carbon monoxide (CO), Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2) and Particulate Matter (PM_{10} and $PM_{2.5}$).

2 Data description

2.1 Ground level data

The ground level air quality monitoring network of Scotland composes 81 stations each providing hourly data on the concentration of the above mentioned pollutants. The network is heterogeneous in the sense that each station measures only a subset of the pollutants, from a minimum of 6 to a maximum of 67 stations measuring the same pollutant. Missing data are also common but, for the time period considered, they do not exceed 21% with respect to each pollutant. As far as the geographical location of the stations is concered, these are unevenly distributed over Scotland, with a higher concentration of monitoring stations within the Grampian, Tayside, Lothian, and Greater Glasgow regions.

2.2 Covariates

In order to better understand the spatio-temporal dynamics of each pollutant and to improve mapping capability, a set of 7 meteorological and morphological covariates is considered. The two morphological covariates are time-invariant and are the land elevation (ele) and the percentage of urban area (urb). The meteorological covariates considered come from the NASA MERRA database and are sea level pressure (slp), wind speed at 2 meter (ws2), temperature at 2 meter (t2), specific humidity at 2 meter (sh2) and the planetary boundary layer height (blh).

3 The multivariate model

Air pollutants are characterized by both temporal and spatial dynamics. Moreover, different pollutants may exhibit cross-correlation in space and

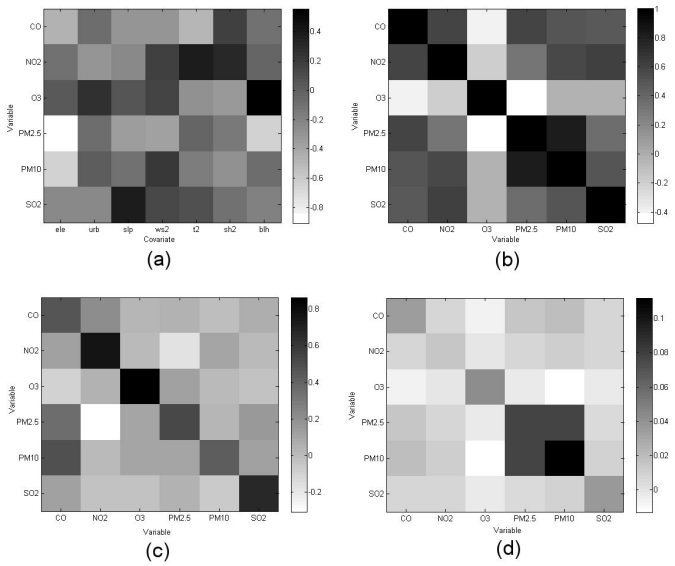


FIGURE 1. EM estimation results. (a) β coefficients; (b) coregionalization matrix V; (c) autoregressive transition matrix G; (d) error variance matrix Σ_η .

time in a non-trivial way. For these reasons, the multivariate dynamic coregionalization model introduced by Fassò and Finazzi (2011) is considered here, which is suitable for dealing with latent temporal and spatial variables in the presence of missing data.

Let $Y(s,t) = (Y_1(s,t), ..., Y_i(s,t), ..., Y_q(s,t))$ be a $q-$dimensional vector of pollutant concentrations at time $t = 1, ..., T$ and site $s \in D \subset \Re^2$, with i running through the set of pollutants {O_3, CO, SO_2 ,NO_2, PM_{10}, $PM_{2.5}$}, the model equation is given by

$$Y(s,t) = X(s,t)\beta + KZ(t) + W(s,t) + \varepsilon(s,t)$$

where $X(s,t)$ is the covariate matrix for time t and site s, $Z(t)$ is the $p-$dimensional latent temporal variable at time t, $W(s,t)$ is the $q-$dimensional latent spatial variable and $\varepsilon(s,t)$ is the measurement error, which is white noise in space and time with $q \times q$ variance covariance matrix Σ_ε. The latent temporal variable $Z(t)$ is characterized by the Markovian dynamic $Z(t) = GZ(t-1) + \eta(t)$, with $\eta(t) \sim N_p(0, \Sigma_\eta)$. In the simplest case, either $p = 1$ or $p = q$, namely all pollutants share the same temporal dynamic or each pollutant has its own. The loading matrix K is a $q \times p$ matrix of known coefficients and is fixed in space and time. Finally, $W(s,t)$ is modelled as a $q-$dimensional linear coregionalization

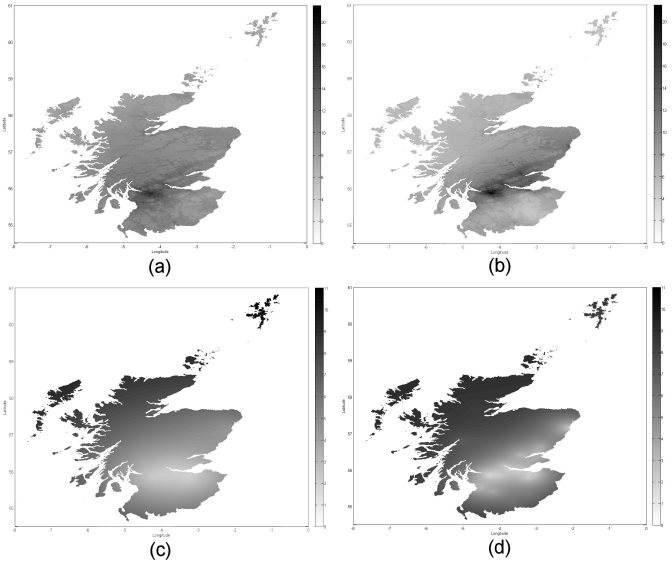


FIGURE 2. 1 km resolution kriging results ($\mu g/m^3$). (a) yearly average $PM_{2.5}$ concentration for model M_1; (b) yearly average $PM_{2.5}$ concentration for model M_2; (c) estimation variance for model M_1; (d) estimation variance for model M_2.

model, namely $W(s,t) = (W_1(s,t), ..., W_q(s,t))$ is white noise in time but correlated over space with a $q \times q$ covariance and cross-covariance matrix function given by $\Gamma(h,\theta) = (\text{cov}(W_i(s), W_j(s')))_{i,j=1,...,q} = V \cdot \rho(h,\theta)$ where $h = \|s - s'\|$ is the Euclidean distance between two sites $s, s' \in D$, V is a positive semi-definite $q \times q$ matrix and $\rho(h,\theta) = \exp(-h/\theta)$ is the exponential correlation function of parameter θ.
The set of model parameters is $\Psi = \{\beta, \Sigma_\varepsilon, G, \Sigma_\eta, V, \theta\}$. The estimation of Ψ and the evaluation of confidence intervals for every parameters are carried out using the EM algorithm. Following a plug-in approach, the daily concentration of each pollutant is kriged over the area of interest as detailed in Fassò et al. (2009).

4 Model estimation and mapping

Considering the data set discussed in Section 2, two models are estimated. The first model, M_1, is a univariate model for the $PM_{2.5}$ pollutant concentration only, which is measured at 6 monitoring stations mainly located in the Lothian and Greater Glasgow regions. The second model, M_2, is a multivariate model which considers the data for all pollutants at all the

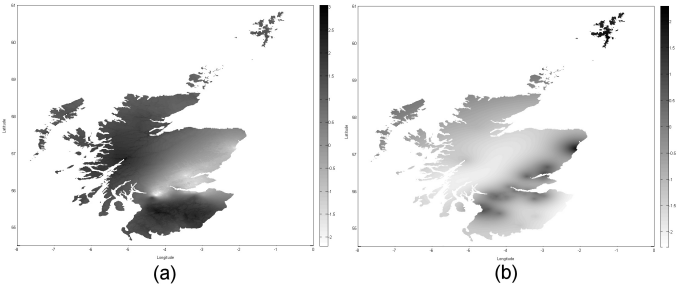


FIGURE 3. Model output comparison (a) Difference in the yearly average $PM_{2.5}$ concentration between model M_1 and M_2; (b) Difference in the estimation variance between model M_1 and M_2.

81 monitoring stations. Moreover, M_2 is characterized by $p = q$, namely each pollutant has its own temporal dynamic. The models are compared by applying the leave-one-out cross-validation technique. The daily concentration maps are used to evaluate the daily map average aggregate statistic.

5 Result analysis and conclusion

The EM estimation results are partially reported in Figure 1 only for the more interesting model M_2. The images in the figure graphically represent the estimated β coefficients, the coregionalization matrix V, the autoregressive transition matrix G and the autoregressive error variance-covariance matrix Σ_η.
Each variable and each covariate being standardized, the beta coefficients can be directly compared within and across variables. By analyzing the image (a) of Figure 1, it can be noted that the most significant covariates are the temperature for NO_2, the planetary boundary layer height for O_3, the land elevation for PM_{10} and $PM_{2.5}$ and the sea level pressure for SO_2. The estimated matrix V clearly shows the negative correlation between O_3 and all the other pollutants and the strong positive correlation between PM_{10} and $PM_{2.5}$. The matrix G is a stable transition matrix and the positive value of its diagonal elements reflects the temporal persistence of the pollutants, which is stronger for O_3 and NO_2. The cross-validation RMSE based on the standardized data goes from 0.4995 for model M_1 to 0.3746 for model M_2, which corresponds to a 25% reduction. This clearly demonstrates the advantage of considering the multivariate model. The images of Figure 2 show the yearly average $PM_{2.5}$ concentration and its estimation variance for both the model M_1 and M_2. If M_2 is considered as the best model (due to its lower cross-validation RMSE), then it can be said that

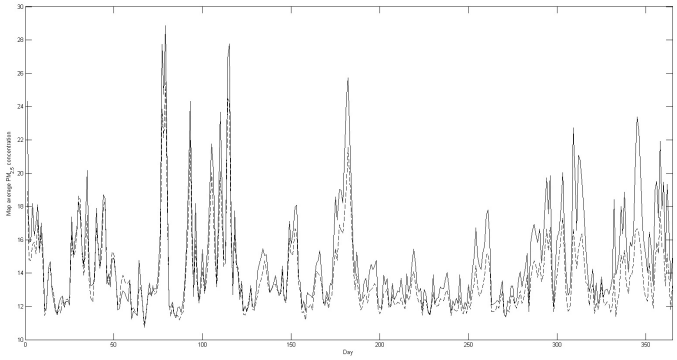


FIGURE 4. Daily map average aggregate statistics derived from model M_1 (solid line) and model M_2 (dashed line).

model M_1 underestimates the $PM_{2.5}$ concentration over the highly populated central-east part of Scotland while it overestimates the concentration over the remaining part of Scotland. The output difference is displayed in Figure 3. The daily map average aggregate statistics derived from the output of model M_1 and M_2 are depicted in Figure 4. The statistic related with M_1 reflects the fact that model M_1 overestimates the $PM_{2.5}$ concentration.

Acknowledgments: This research is part of Project EN17, 'Methods for the integration of different renewable energy sources and impact monitoring with satellite data', Lombardy Region under 'Frame Agreement 2009'

References

Bodnar, O., Cameletti, M., Fassò, A. and Schmid, W. (2008). Comparing air quality in Italy, Germany and Poland using BC indexes. *Atmospheric Environment*, **42**, 8412-8421.

Fassò, A., Finazzi, F. and D'Ariano, C. (2009). Integrating satellite and ground level data for air quality monitoring and dynamical mapping. *GRASPA Working Paper n.34*, http://www.graspa.org/.

Fassò, A., Finazzi, F. (2011). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics. Accepted for publication.*

Lee, D., Ferguson, C. and Scott, E.M. (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society series A*, **174**, 109-126.

Predictive distributions for non-regular parametric models

Giovanni Fonseca[1], Federica Giummolè[2] , Paolo Vidoni[1]

[1] University of Udine, Department of Economics and Statistics, via Treppo 18, I-33100 Udine, ITALY. e-mail: giovanni.fonseca@uniud.it, paolo.vidoni@uniud.it

[2] Ca' Foscari University - Venice, Department of Environmental Sciences, Informatics and Statistics, San Giobbe, Cannaregio 783, I-30121 Venice, ITALY. e-mail: giummole@unive.it

Abstract: Improved prediction distributions based on asymptotic methods are a well known tool for prediction in the context of regular parametric models. On the contrary, for non-regular cases, prediction is mainly based on the estimative or plug-in distribution. The aim of this work is to define calibrated predictive distributions which quantiles have coverage probability equal or close to the target nominal value. Whenever the computation is not feasible, a suitable bootstrap procedure easily provides a good estimate for the proposed distribution. A simulation example is provided for a particular non regular model, the generalized extreme value distribution, which support depends on unknown parameters.

Keywords: Coverage probability; Extreme value distributions; Non-regular models; Parametric bootstrap; Prediction limits; Predictive distributions.

1 Introduction

In this work, we consider the problem of prediction of a future, or unobservable, unidimensional absolutely continuous random variable Z, on the basis of an observed sample $y = (y_1, \ldots, y_n)$ from a random vector $Y = (Y_1, \ldots, Y_n)$. We assume that the joint distribution of (Y, Z) is known, up to a k-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^k$. In this case, a possible solution can be given in terms of prediction limits, i.e. functions $\tilde{z}_\alpha(\hat{\theta})$ such that, for all $\alpha \in (0, 1)$, the coverage probability

$$P_{Y,Z}\left[Z \leq \tilde{z}_\alpha(\hat{\theta}(Y))\right] = \alpha, \tag{1}$$

at least to a high order of approximation. Here $\hat{\theta} = \hat{\theta}(Y)$ is an asymptotically efficient estimator for θ, usually the maximum likelihood estimator. When exact results are not available, an easy solution is given by considering the estimative prediction limits, obtained by substituting the unknown parameter θ by $\hat{\theta}$ in the α-quantiles of the conditional distribution of Z given $Y = y$. Unfortunately the associated coverage error has order $O(n^{-1})$, which is often considerable. Improved prediction limits with coverage error of order $o(n^{-1})$ have been proposed by Barndorff-Nielsen and

Cox(1996) and Vidoni (1998), as modifications of the estimative prediction limits. Their results rely on asymptotic expansions and only hold under regularity assumptions on the model. Calibrated prediction limits can be obtained by means of a bootstrap based procedure, as proposed by Hall et al. (1999). Though very interesting, this approach provides solutions for specific fixed values of the target coverage α.
In this work, following Fonseca et al. (2010), we define a predictive distribution which α-quantiles provide exact prediction limits for every $\alpha \in (0,1)$. When this predictive distribution is not explicitly available, it can be approximated using a suitable bootstrap technique. The coverage error associated to the resulting approximated quantiles is of order $o(n^{-1})$, improving on the estimative solution. The proposed method for prediction is general, easy to compute and does not require regularity assumptions on the underlying model. Thus, it also applies to non-regular cases when the support of the model depends on an unknown parameter. This extension is very useful, for instance, in the applications to survival analysis and in the studies of extreme events.

2 Calibrated predictive distributions

Let us assume, for simplicity, that $Y_1, \ldots, Y_n, Z$ are independent continuous random variables with the same distribution. Denote by $G(z;\theta)$ the distribution function of Z.
Consider the estimative prediction limit $z_\alpha(\hat{\theta}) = G^{-1}(\alpha;\hat{\theta})$, where $G^{-1}(\cdot;\hat{\theta})$ is the inverse of function $G(\cdot;\hat{\theta})$. The associated coverage probability is

$$P_{Y,Z}\{Z \leq z_\alpha(\hat{\theta});\theta\} = E_Y[G\{z_\alpha(\hat{\theta});\theta\};\theta] = C(\alpha,\theta).$$

Function $C(\alpha,\theta)$ depends on the true parameter value θ and on the nominal coverage probability α. However, its explicit expression is rarely available. It is well known that it does not match the target value α, although asymptotically $C(\alpha,\theta) = \alpha + O(n^{-1})$, as $n \to +\infty$.
As suggested by Fonseca et al. (2010), a predictive distribution function can be defined by substituting α with $G(z;\hat{\theta})$ in $C(\alpha,\theta)$:

$$G_c(z;\hat{\theta},\theta) = C\{G(z;\hat{\theta}),\theta\}. \tag{2}$$

$G_c(\cdot;\hat{\theta},\theta)$ is a proper predictive distribution function in regular parametric models. When the support of Z depends on θ, $G_c(z;\hat{\theta},\theta)$ may not satisfy one or both the limit conditions as $z \to \infty$. Nevertheless, it can still be fruitfully employed for obtaining good prediction limits, far from the boundary of the support of Z.
The predictive distribution (2) gives, as quantiles, prediction limits $z^c_\alpha(\hat{\theta},\theta)$ which coverage probability equals the target nominal value α, for all $\alpha \in (0,1)$.

Though interesting from a theoretical perspective, the calibrated predictive distribution $G_c(z;\hat{\theta},\theta)$ is in fact inapplicable since it usually depends on the unknown parameter θ. A useful surrogate is the corresponding plug-in estimator

$$\hat{G}_c(z;\hat{\theta}) = G_c(z;\hat{\theta},\hat{\theta}) = C\{G(z;\hat{\theta}),\hat{\theta}\}.$$

The associated α-prediction limit is defined as $\hat{z}^c_\alpha(\hat{\theta}) = z^c_\alpha(\hat{\theta},\hat{\theta}) = z_{\hat{\alpha}^c}(\hat{\theta})$, with $\hat{\alpha}_c = C^{-1}(\alpha,\hat{\theta})$, and it satisfies (1) to a closer approximation than the estimative prediction limit $z_\alpha(\hat{\theta})$, that is with an error term of order $o(n^{-1})$.

A closed form expression for the coverage probability $C(\alpha,\theta)$ is rarely available so that even the predictive distribution function $\hat{G}_c(z;\hat{\theta})$ is not very useful in practice. Anyway, there is a suitable parametric bootstrap estimator for $G_c(z;\hat{\theta},\theta)$, to be considered when $C(\alpha,\theta)$ is not available. Let $y^*(j)$, $j = 1,\ldots,B$, be parametric bootstrap samples generated from the estimative distribution of the data and let $\hat{\theta}^*(j)$, $j = 1,\ldots,B$, be the corresponding maximum likelihood estimates. Since $C(\alpha,\theta) = E_Y[G\{z_\alpha(\hat{\theta});\theta\};\theta]$, we define the bootstrap-calibrated predictive distribution as

$$G^b_c(z;\hat{\theta}) = \frac{1}{B}\sum_{j=1}^{B} G\{z_\alpha(\hat{\theta}^*_j);\hat{\theta}\}|_{\alpha=G(z;\hat{\theta})}. \tag{3}$$

The corresponding α-quantile defines, for each $\alpha \in (0,1)$, a prediction limit having coverage probability equal to the target α, with an error term which depends on the efficiency of the bootstrap simulation procedure. It is important noticing that the computation of (3) does not require any assumption on the regularity of the parametric models involved, as long as the bootstrap applies.

3 Generalized extreme value distribution

Let $Y_1,\ldots,Y_n$ be independent random variables with common generalized extreme value distribution, that is

$$G(y;\mu,\sigma,\xi) = \exp\left\{-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}\right\},$$

where $1+\xi(y-\mu)/\sigma > 0$ and $\theta = (\mu,\sigma,\xi)$ is an unknown parameter with $\sigma > 0$ a scale parameter, $\mu \in \mathbb{R}$ a location parameter and $\xi \in \mathbb{R}$ a shape parameter. The generalized extreme value distribution includes the Frechet, the Gumbel and the Weibull distributions as particular cases and is usually used for the study of extreme events, such as extreme flood of a river or maximum sea level. In this context it can be useful to consider the problem of prediction of a future value $Z = Y_{n+1}$, independent of $Y_1,\ldots,Y_n$ and with the same distribution.

TABLE 1. Generalized extreme value distribution. Coverage probabilities for estimative and bootstrap calibrated prediction limits of level α=0.9, 0.95, 0.99.

α	n	Estimative	Bootstrap
0.9	10	0.880	0.899
	20	0.893	0.905
0.95	10	0.933	0.954
	20	0.942	0.951
0.99	10	0.976	0.987
	20	0.982	0.986

In this case, an explicit expression for the coverage probability $C(\alpha, \mu, \sigma, \xi)$, associated to the estimative α-prediction limit, is not available. As explained in Section 2, we can estimate (2) using the bootstrap estimator (3) and calculate calibrated prediction limits as quantiles of this approximated predictive distribution.
Table 1 shows the results of a simulation study for comparing the performance of estimative (Estimative) and bootstrap calibrated (Bootstrap) prediction limits, with respect to the corresponding coverage probabilities. Estimation is based on 5,000 Monte Carlo replications. Bootstrap procedure is based on 1,000 bootstrap samples. Estimated standard errors are always smaller than 0.005. Different values of the target level, α=0.9, 0.95, 0.99, and of the sample size, $n = 10, 20$, are considered. The parameters of the generalized extreme value model are fixed to $\mu = 5$, $\sigma = 2$ and $\xi = 0.4$. It can be seen that the bootstrap solution remarkably improves on the estimative one.

References

Barndorff-Nielsen, O.E., and Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli*, **2**, pp. 319-340.

Fonseca, G., Giummolè, F., and Vidoni, P. (2010). Calibrating predictive distributions. *Redazioni Provvisorie*, **2/2010**, Department of Statistics, Ca' Foscari University, Venice.

Hall, P., Peng, L. and Tajvidi, N. (1999). On prediction intervals based on predictive likelihood or bootstrap methods. *Biometrika*, **86**, pp. 871-880.

Vidoni, P. (1998). A note on modified estimative prediction limits and distributions. *Biometrika*, **85**, pp. 949-953.

Objective Bayes Criteria for Variable Selection.

A. Forte[1], M. J. Bayarri[1], J. O. Berger[2], G. García-Donato[3]

[1] Universitat de València
[2] Duke University and SAMSI
[3] Universidad de Castilla La-Mancha

Abstract: Elicitation of objective priors that are suitable for model selection is a rather difficult problem and not yet entirely understood. In the variable selection scenario, many priors have been proposed. Most of them follow the guidelines given in Jeffreys-Zellner-Siow, namely 1) orthogonalize common and non-common parameters 2) give the common parameters a common prior, taken to be the usual improper estimation prior and 3) give a flat-tailed prior to the model specific parameters. This extended practice, although intuitively sound, seems to be basically ad-hoc: no formal arguments are usually given. In this talk, we propose a general class of priors , generalizing several earlier proposals, and show that 1) use of the right Haar prior for the common parameters is the right prior for this problem; no orthogonalization is needed, and justification is in terms of invariance, 2) flat-tailed distributions generalizing priors proposed previously for minimax and robust Bayes estimations are very well suited for this problem, resulting in consistent and information consistent Bayes factors, and 3) the recommended prior exhibits a novel form of predictive matching, which we believe is the appropriate one when selecting models of differing dimensions. Finally, it even produces close-form Bayes factors.

Keywords: Variable Selection; Objective Bayes; Bayes Factors.

1 The problem

Many of today's scientific problems require identifying which variables from an entertained set are involved in a specific phenomenon. For instance, many public health studies require the identification of the causes of a certain disease.

This problem is referred to as variable selection and can be seen as a particular case of model selection. In this specific model selection problem each model contains a certain subset of the entertained covariates. This means a total of 2^p possible models for a problem with p potential covariates. The variable selection problem is difficult to address both from a theoretical and from a computational point of view.

In particular, in this work the problem of variable selection is addressed in the framework of linear regression, but it also appears in many other

scenarios such as generalized linear models and non-parametric function estimation.

Our preferred Bayesian way for solving model selection, and, in particular variable selection, is to base the choice on the posterior probabilities of the competing models. These posterior probabilities can be expressed in terms of the prior probabilities of the models and the 2^p Bayes factors.

For the assignment of prior probabilities over the model space we entertain and compare some approaches, and state our preferred choice. However, this is not the main topic of this work.

Posterior probabilities require the computation of 2^p Bayes factors in favor of each model M_i and against a base model M_d for $i = 0, \ldots 2^p - 1$. Our choice for M_d is the simplest model explaining the data, which as usual we denote M_0; M_0 is nested in every model M_i. The computation of those 2^p Bayes factors require the elicitation of priors for the corresponding parameters under each model. Subjective elicitation of priors assessed by experts knowledge in this scenario is practically impossible due to the very large number of models, and model-specific parameters. The idea is hence, to adopt an objective point of view (see Berger, 2006, and references therein) but the objective elicitation of priors in model selection has to be done carefully due to the high sensitivity of Bayes factors to the choice of objective priors. In fact, the usual non-informative (usually improper) priors, which work well in estimation problems do not always produce sensible results in model selection (see Berger and Pericchi, 2001, and references therein) often resulting in indeterminate Bayes factors.

The large number of models also posses a computational challenge since the numerical computation of the 2^p Bayes factors is required. When p is so large that the models space can not even be enumerated (for all practical purposes), many authors (see, for example, George and McCulloch, 1993; Carlin and Chib, 1995; Berger and Molina, 2005, and references therein) propose methods for searching over the model space trying to find models with high posterior probabilities. But usually Bayes factors are hard to compute, so that even this solution can be computationally very demanding. This difficulty can be largely alleviated if simple expressions for Bayes factors are available.

2 Our Solution

The aim of this work is to propose a novel, suitable and rigorously justified prior distribution for the variable selection problem. In particular, we look for a prior distribution which achieves many desirable properties and provides simple expressions for the Bayes factors.

We follow the Conventional approach of Jeffreys (1961), who outlined a number of desiderata for a good objective prior distribution to have in the variable selection problem.

Following Jeffrey's Conventional scheme, the prior distribution under each model M_i is assessed in two steps. The first one consists in assigning a proper prior distribution for those parameters in M_i that are not in M_0 conditionally on those parameters in both models (in particular, as M_0 is nested in M_i this means conditionally on the parameters in M_0). The second step consists in assigning a non-informative prior for the parameters in M_0.

For assessing the conditional prior in the first step we found some interesting ideas in the work of Strawderman (1971) and Berger (1976, 1980, 1985). Their work, originally developed in a context of robust and minimax normal mean estimation, is extended and adapted here to solve the variable selection problem.

For the prior distribution of the parameters in M_0 (occurring in all models) we consider a prior which makes the problem invariant. In this case, it happens to coincide with the reference prior or independent Jeffreys' prior which is the usual choice in the literature. Hence, the usual choice gets fully justified.

The result is a joint prior distribution in the parametric space which, following Berger (1985) we call Conventional Robust prior. This prior distribution is defined up to some parameters that can be tuned to achieve a number properties. Our specific proposal for these parameters is based in certain optimality properties of the resulting procedure.

The theoretical highlights of this distribution for variable selection are

- *The choice of the prior is justified from a theoretical point of view.* Jeffreys (1961)'s Conventional approach scheme for the elicitation of prior distributions was based on the orthogonal parameterization of the model. Our choice is instead completely justified by a sensible choice of the scale matrix and the use of invariance ideas in Berger et al. (1998). This fully theoretical justification makes the orthogonal parameterization no longer required.

- *It produces well defined Bayes factors with good consistency properties from many points of view.* The resulting Bayes factors are well defined in the sense that they are not indeterminate as is usually the case when using objective (improper) priors. This indeterminacy is avoided here through invariance arguments. On the other hand, the consistency properties of the resulting Bayes factor, closely related to the shape of the prior's tails, makes this choice a suitable prior for variable selection.

- *It agrees with the predictive matching idea.* In particular, our prior distribution accords with our preferred and weaker interpretation of predictive matching for this problem. Specifically, we require that, if the information in the sample is barely enough for estimating the specific parameters of *any* model entertaining k extra covariates (i.e.

$n = k_0 + k$), then this information should not be enough to discriminate among those models.

In addition, our approach produces simple, tractable, closed-form expressions for Bayes factors considerably simplifying computation.

Acknowledgments: This research has been partially suported by the Ministerio de Ciéncia e Innovación grant MTM2010-19528.

References

Berger, J.O. (1976). Admissible Minimax Estimation of a Multivariate Normal Mean with Arbitrary Quadratic Loss, *The Annals of Statistics*, **4**, 1.

Berger, J.O. (1980). A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean, *The Annals of Statistics*, **8**, 4.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.*, Springer, 2nd edition.

Berger, J.O. (2006). The Case for Objective Bayesian Analysis. *Bayesian Analysis*, **1**, 3.

Berger, J.O. and Molina, G. (2005). Posterior Model Probabilities Via Path-Based Pairwise Priors. *Statistica Neerlandica*, **59,** 1.

Berger, J.O. and Pericchi, L.R. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison. *Lecture Notes- Monograph Series*, **38**, 3.

Berger, J.O. et al. (1998). Bayes Factors and Marginal Distributions in Invariant Situations, *Sankhya: The Indian Journal of Statistics, Series A*, **60**, 3.

Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 3.

George, E.I. and McCulloch, R.E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 423.

Jeffreys, H. (1961). *Theory of Probability.* Oxford University Press, 3rd edition.

Strawderman, W.E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *The Annals of Mathematical Statistics*, **42**, 1.

Conditional Probability of Flood Risk in Scotland

Maria Franco-Villoria[1], Marian Scott[1], Trevor Hoey[2], Denis Fischbacher-Smith[3]

[1] School of Mathematics and Statistics, 15 University Gardens, University of Glasgow, G12 8QQ. Contact: mvilloria@stats.gla.ac.uk
[2] School of Geographical and Earth Sciences, University of Glasgow
[3] Business School, University of Glasgow

Abstract: Climate change impacts are expected to vary spatially and to produce changes in precipitation patterns that control river flows, the extremes of which are critical for flood risk estimation. A multivariate conditional model, as proposed by Keef et al.(2009), is applied here to a set of Scottish rivers to estimate the spatial dependence in extreme river flows. The results reveal relationships between extreme flows that agree with what would be expected based on catchment properties and will potentially prove useful for planning purposes.

Keywords: River Flow Series; Flood; Conditional Probability; Semi-parametric

1 Introduction

Understanding temporal patterns in river flows and their relationship to flooding is critical to flood planning and risk management. Delivering effective and efficient flood risk management in future requires new approaches: in Scotland, the Flood Risk Management Act (2009) was passed with the aim of introducing "a more sustainable and modern approach to flood risk management" [Scottish Government (2010)]. To do so, new and improved estimates of flood risk which take into account the impact of climate change and possible spatial heterogeneity are needed. Much flood-risk management is based on the concept of a return period for an event of given magnitude and, despite ongoing debate about the utility of the return period approach [White (2001),Young and Davies (1989)], considerable effort continues to be made to refine predictions of the 1 in 100 year event. The weather systems that generate extreme flood events operate at regional scales, however, extreme events require combinations of conditions that are rarely coincident between catchments. Thus, while a period of wet weather may generate high flows simultaneously in all rivers in a region the particular conditions required to produce an extreme event may only be found in a small subset of these rivers. Isolating the factors that produce extreme conditions is difficult, as catchment-scale meteorology and hydrology are controlled by

many parameters all of which are spatially and temporally variable. Pooling data from several catchments is an efficient way of optimising prediction of extremes. Here we use a conditional probability approach to investigate spatial interdependence in extreme flows.

1.1 Data

Daily river flow data from 3 rivers across Scotland over a period of 20 years (1985-2005) were selected based on data quality and spatial location (Figure 1). Data were provided by the National River Flow Archive(NRFA) and the Scottish Environment Protection Agency (SEPA).

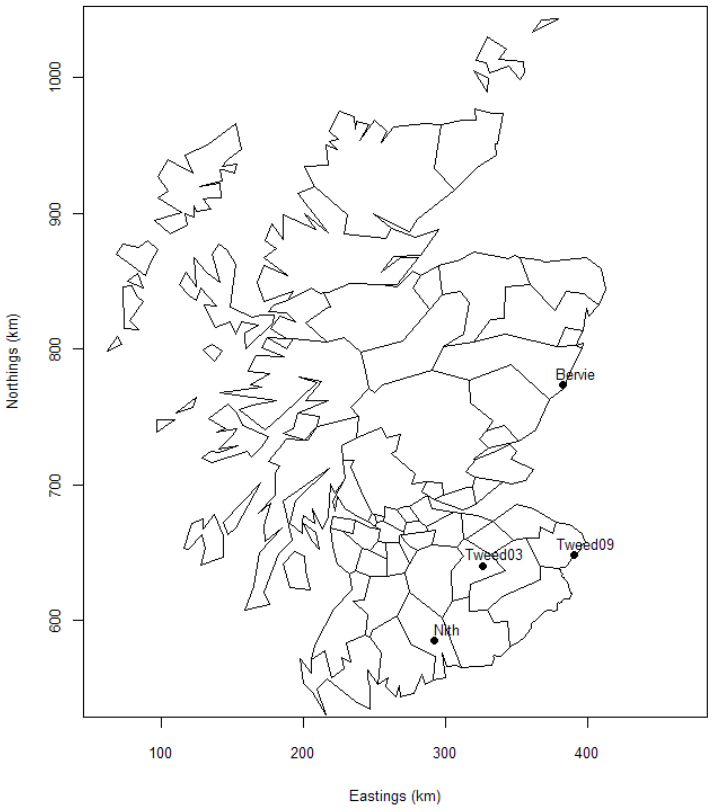


FIGURE 1. Rivers location

2 Methods

Following Keef et al.(2009), two conditional measures of spatial flood risk, firstly the probability ($P_C(p)$) that a set of rivers $Y = (Y_1, ..., Y_d)$ within a region C of interest flood at time t (or any lag of t), given that another river X has already flooded, and secondly the expected number of rivers (N(p)) that will flood on average within that region (given that X has flooded) were estimated. These estimates are calculated based on the conditional distribution of $Y|X = x$ (for large x), which can be modeled using a semi-parametric approach [Heffernan and Tawn (2004)]. Assuming the random variables $(Y_1, \ldots, Y_d, X)$ marginally follow a standard Gumbel distribution, there exist normalizing functions $a(x)$, $b(x)$ such that $\forall$ fixed z, $lim_{x\to\infty}[Y \leq a(x)+b(x)z|X = x] = G(z)$, for $x > u_X$ (u_X being a suitable high threshold). Note that this is a POT approach, as only values over the chosen threshold are used to fit the model. $a(x)$, $b(x)$ are estimated parametrically and take the form $a(x) = \alpha x$ and $b(x) = x^\beta$, where $0 \leq \alpha \leq 1$, $-\infty < \beta < 1$. $a(x)$ describes the overall strength of the dependence structure, while $b(x)$ relates to the variability (Keef et al. (2009)). The limiting distribution $G(z)$ is estimated non-parametrically using kernel smoothing. Once the model is fitted, pseudo-samples can be generated in order to estimate functionals of the joint tails of (Y,X). Confidence intervals can be calculated using block bootstrapping methods.

3 Results

To illustrate how the method works, results (based on mean daily flow) for 3 Scottish rivers are presented here, the River Tweed, for which data on two different gauging stations are available (Tweed09(4390km^2)* and Tweed03(694km^2)), the River Nith(799km^2) and the River Bervie(123km^2). A unique threshold u_X corresponding to p=0.9 (where 1-p is the probability of the flow exceeding the chosen threshold) is used to condition on the four records. Before fitting the model, data were transformed to follow a standard Gumbel distribution. The estimated parameters $\hat{a}$ and $\hat{b}$ are presented on Table 1. The results point towards relationships that agree with what would be expected; the two stations on the Tweed are expected to be similar as both have large catchment areas and are influenced by the same weather patterns, and they are closely related (Table 1) with the influence of each station on the other being of similar order. The River Bervie drains a small catchment in a different hydrological region from the other rivers, and so is influenced by localized weather events that are often not experienced at the other sites. The influence of the Rivers Tweed and Nith on the River Bervie appears to be stronger than that in the opposite direction, which seems reasonable given that the River Bervie has a very small catchment. The dependence between the Rivers Nith and Tweed is

stronger for station03, which is geographically closer to the former than station09.*(catchment area)

TABLE 1. Parameter estimates $\hat{a}(x)$ and $\hat{b}(x)$ of the dependence model

Conditioning on	Tweed09	Tweed03	Nith	Bervie
Tweed09		$\hat{a}$=0.749	$\hat{a}$=0.094	$\hat{a}$=0.229
		$\hat{b}$=0.677	$\hat{b}$=0.679	$\hat{b}$=0.340
Tweed03	$\hat{a}$=0.646		$\hat{a}$=0.513	$\hat{a}$=0.098
	$\hat{b}$=0.763		$\hat{b}$=0.676	$\hat{b}$=0.350
Nith	$\hat{a}$=0.152	$\hat{a}$=0.455		$\hat{a}$=0.164
	$\hat{b}$=0.544	$\hat{b}$=0.684		$\hat{b}$=0.189
Bervie	$\hat{a}$=0.000	$\hat{a}$=0.074	$\hat{a}$=0.051	
	$\hat{b}$=0.422	$\hat{b}$=0.211	$\hat{b}$=0.202	

Estimates of Pc(p) for the rivers (b)Tweed(station03), (c)Nith, (d)Bervie and (e)N(p) conditioning on the River Tweed (station 09) are shown in Figure 2, along with 95%confidence intervals. The conditional probability of each river flooding decreases as the event in Tweed09 (conditioning river) becomes more and more extreme (ie, as p$\rightarrow$1), and is higher for stations that are close by. Flood risk estimates tend to be expressed in terms of return periods rather than probabilities. Figure 3 shows the expected number of rivers in the set that would flood for events with different return period occurring in the conditioning river. For example, if an event with associated return period T=0.5years happens in the river Tweed(station09) (conditioning river), we would expect, on average, one river out of the three considered here to have an event of (at least) similar magnitude.

4 Summary and Future Work

A conditional multivariate model to estimate flood risk was fitted to a set of Scottish rivers. So far, the results reveal relationships between extreme flows that agree with what would be expected given the spatial location of the rivers. The analysis is currently being extended to a further 30 rivers and will potentially prove useful for planning purposes. In addition to the two measures of spatial risk estimated here, the joint conditional distribution can be used to investigate the dependence structure in the upper tail of the distribution and estimate quantities such as (conditional) return levels. Alternative methods of spatial analysis, in particular, spatially varying coefficients models, will be explored next.

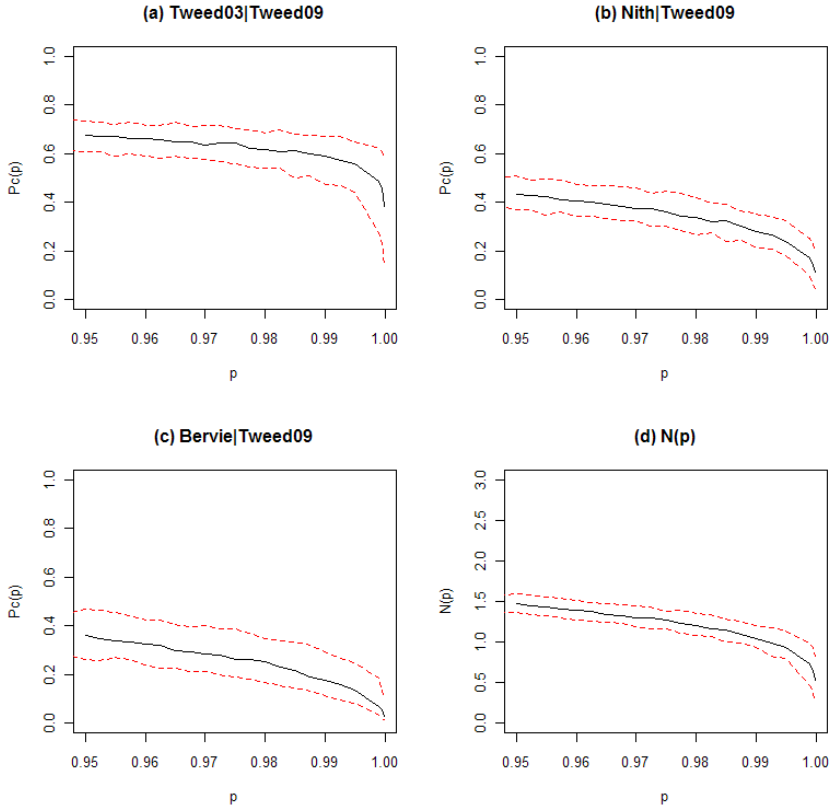


FIGURE 2. Model based estimates of the conditional probability of flooding for the rivers (a)Tweed, (b)Nith and (c)Bervie, and (d)expected number of sites to flood in the region.

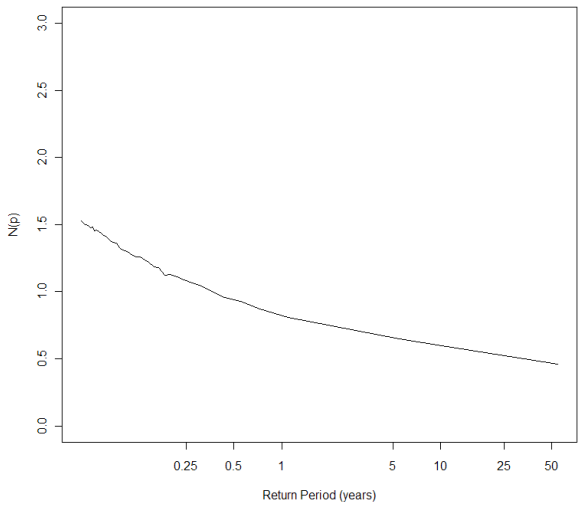


FIGURE 3. Expected number of sites (N(p)) that will flood given a T-year event in the river Tweed(station03).

References

Heffernan, J.E. and Tawn, J.A. , (2004). A conditional approach for multivariate extreme values *Journal of the Royal Statistical Society - Series B*, **66(3)**,n 497-546.

Keef, C. and Svensson, C. and Tawn, J.A. , (2009). Spatial dependence in extreme rive flows and precipitation for Great Britain *Journal of Hydrology*, **378**, 240-252.

Keef, C. and Tawn, J. and Svensson, C. (2009). Spatial risk assessment for extreme river flows *Applied Statistics*, **58(5)**, 601-618.

The Scottish Government (2010). URL: http://www.scotland.gov.uk/Topics/Environment/Water/Flooding/FRMAct. Electronic Resource [Accessed 13/09/2010]

White, W.R. (2001). Water in rivers: flooding *Proceedings of the Institution of civil engineers - Water, Maritime and Energy*, **148(2)**, 107-118.

Young, J.R. and Davies,T.R.H. (1989). The realistic criteria for flood-control design *Hydrology and water resources symposium 1989: comparisons in austral hydrology, institution of engineers, Australia, national conference publications*, **89**, 227-231.

Outliers and interventions in INGARCH time series

Roland Fried[1], Hanan El-Saied[1], Konstantinos Fokianos[2]

[1] Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany. Email: fried@statistik.tu-dortmund.de
[2] University of Cyprus, Nicosia, Cyprus

Abstract: We consider outliers and intervention effects in INGARCH-models for time series of counts. An iterative procedure based on conditional maximum likelihood estimation and maximum score test statistics is constructed for stepwise detection and elimination of intervention effects which enter the dynamics of the process. Purely additive outliers representing e.g. measurement artifacts cannot be treated in this way. We propose robust M-estimation of the model parameters in the possible presence of this type of outliers.

Keywords: Bootstrap test; Robustness; Huber M-estimator; Tukey M-estimator.

1 Introduction

Time series of counts are observed e.g. in epidemiology, where we measure the number of new infections within a certain time period. Ferland et al. (2006) propose integer-valued GARCH (INGARCH) models for such data, which have been studied by Fokianos et al. (2009) later on. An INGARCH$(1,1)$ process $(Y_t : t \geq 1)$ is defined through the relationships

$$\begin{aligned} Y_t|\mathcal{F}_{t-1} &\sim \text{Poisson}(\lambda_t), \\ \lambda_t &= \beta_0 + \beta_1 Y_{t-1} + \alpha_1 \lambda_{t-1}, \end{aligned} \tag{1}$$

for $t \geq 1$. The dynamics of the process is modeled via the conditional mean $\lambda_t = \mathrm{E}(Y_t|\mathcal{F}_{t-1})$ of Y_t, where $\mathcal{F}_t$ stands for the σ–field generated by $\{Y_0, \ldots, Y_t, \lambda_0\}$ representing the whole information up to time t, $\beta_1 \geq 0$ and $\alpha_1 \geq 0$ are regression parameters and $\beta_0 > 0$ is an intercept. A stationary process fulfilling model (1) with marginal mean $\lambda = \beta_0/(1-\alpha_1-\beta_1)$ exists if $\alpha_1+\beta_1 < 1$. Model (1) closely resembles the popular GARCH$(1,1)$-model since the mean of the Poisson distribution equals its variance.
An interesting question is whether a certain model fits some given data well, or whether special effects need to be included to achieve a good model fit. Such a need indicates the existence of particular events or model deficiencies. Fokianos and Fried (2010) model different types of outliers and interventions in INGARCH-processes and propose an iterative procedure

for the detection of such effects. All outliers considered in their work cause a spill-over effect and influence the future of the process via its dynamic. We extend this approach by including additive outliers affecting single observations. Such outliers are particularly harmful for classical maximum likelihood estimators. We construct robust M-estimators for this reason.

2 Outliers and interventions in INGARCH-models

Fokianos and Fried (2010) introduce intervention effects into INGARCH-models assuming that instead of the "clean" INGARCH process $(Y_t : t \geq 1)$ we observe a contaminated process $(Z_t : t \geq 1)$, which includes the effect of an intervention at time τ,

$$\begin{aligned} Z_t|\mathcal{F}_{t-1}^Z &\sim \text{Poisson}(\mu_t), \\ \mu_t &= \beta_0 + \beta_1 Z_{t-1} + \alpha_1 \mu_{t-1} + \nu\delta^{t-\tau} I(t \geq \tau), \end{aligned} \tag{2}$$

for $t \geq 1$, where $\mathcal{F}_t^Z$ is the σ–field generated by $\{Z_0, \ldots, Z_t, \mu_0\}$, ν is the size of the intervention effect, $I(t \geq \tau)$ is the indicator function for $t \geq \tau$, and $\delta \in [0, 1]$ regulates whether the effect is concentrated on the observation at time τ ($\delta = 0$), is spread out over all observations from time τ on ($\delta = 1$), or it is something in between ($\delta \in (0, 1)$).

Conditional maximum likelihood (CML) estimation can be applied for joint estimation of the model parameters β_0, β_1, α_1 and ν, if the type and the time τ of the intervention effect is assumed to be known. Score tests allow simultaneous testing for all types and times of outliers, fitting the model only once under the common null hypothesis H_0 of observing a clean INGARCH series without interventions. The score test statistic for a given type and time of intervention is asymptotically χ_1^2-distributed under H_0 with one degree of freedom.

In practice we often do not know neither the type nor the time of an outlier. Maximizing the score test statistics for the same type of intervention over all time points is intuitive, but bears the problem that the resulting maximum score test statistics have different asymptotic distributions for the different types of interventions and are not comparable. Accordingly, we propose a parametric bootstrap procedure: the model is fitted under H_0 and a large number b, say $b = 200$, of bootstrap replicate time series is generated from the fitted model. The maximum score test statistics are calculated for all types of outliers and each of the $b + 1$ time series. If an INGARCH model fits the real data well, we expect the maximum score test statistics for these data to be comparable to those for the bootstrap replicates and calculate an approximate p-value for each type of intervention as the fraction of time series for which the corresponding maximum score test statistic is at least as large as for the real data. If different types of interventions with different values of δ are significant, preference should be given to level shifts, since these are rarely detected in the absence of shifts.

For dealing with data scenarios with several intervention effects, Fokianos and Fried (2010) design a stepwise outlier detection and elimination procedure. They make use of an equivalent formulation of model (2), namely

$$\begin{aligned} Z_t &= Y_t + C_t \\ C_t|\mathcal{F}_{t-1}^C &\sim \text{Poisson}(\kappa_t) \\ \kappa_t &= \beta_0 + \beta_1 C_{t-1} + \alpha_1 \kappa_{t-1} + \nu\delta^{t-\tau} I(t \geq \tau) \end{aligned} \tag{3}$$

for $t \geq 1$, where $(Y_t : t \geq 1)$ follows the INGARCH(1,1) model (1), $(C_t : t \geq 1)$ is an additive contamination and $\mathcal{F}_t^C = \{C_0, \ldots, C_t, \kappa_0\}$. Predicting C_t by its conditional expectation given $\mathcal{F}_t^Z$, with the estimated parameters plugged in, and subtracting this prediction from Z_t allows us to clean the observed time series from a detected intervention. Then we can continue testing for further interventions. Applications to real and simulated data indicate the reliability of the resulting stepwise procedure.
A drawback of the above approach is that only interventions which influence the future of the process via its dynamics are considered. Another type of outliers are additive outliers (AOs), representing e.g. simple measurement errors. AOs can be modeled by modifying (3) as follows:

$$\begin{aligned} Z_t &= Y_t + I(t = \tau)C_t \\ C_t &\sim \text{Poisson}(\nu), \end{aligned} \tag{4}$$

for $t \geq 1$, with $C_1, \ldots, C_n$ being independent identically distributed and C_t being independent of $\mathcal{F}_{t-1}$. Extension of model (4) for inclusion of more than one AO is straightforward. Joint estimation of model parameters and outlier effects by maximum likelihood methods is difficult since $Z_{\tau+1}$ needs to be conditioned on the unobserved Y_τ instead of Z_τ, when constructing the likelihood. Therefore we propose M-estimators for robust estimation of the model parameters in the presence of AOs in the next section.

3 M-estimation in the INARCH model

Assume $y_1, \ldots, y_n$ is an observed time series of counts and we want to fit an INARCH(1) model to these data, applying model (1) with $\alpha_1 = 0$. Application of conditional likelihood, conditioning on the first observation, gives the following set of estimation equations:

$$\sum_{t=2}^{n} \left(\frac{y_t - \lambda_t}{\sqrt{\lambda_t}} \right) \frac{1}{\sqrt{\lambda_t}} \begin{pmatrix} 1 \\ y_{t-1} \end{pmatrix} = 0 \ . \tag{5}$$

Downweighting the influence of unusual observations in these equations leads to a straightforward robustification of the conditional likelihood estimators. For this, we truncate observations with large standardized residuals $(y_t - \lambda_t)/\sqrt{\lambda_t}$ using Huber's or Tukey's ψ function ψ_H and ψ_T,

$$\psi_H(x) = xI(|x| < k) + k\text{sign}(x)I(|x| \geq k),$$

$$\psi_T(x) \quad = \quad x[1-(x/k)^2]^2 I(|x| \leq k),$$

where k is a tuning constant regulating the robustness and the efficiency of the estimator. In ordinary location estimation, this constant is usually chosen in the range between 1 and 2 for the Huber and in the range between 3 and 5 for the Tukey function, see e.g. Maronna, Martin and Yohai (2006, p. 27 and p. 30). The Huber function ψ_H is monotone and usually leads to unique solutions, which can be obtained by straightforward iterations. As opposed to this, the Tukey function ψ_T is redescending to zero and can suppress the influence of very large outliers completely. A drawback is the possible existence of multiple roots, so that good initial values are needed for the iterations to get to the right solution.
Regressors y_{t-1} which are outlying w.r.t. the marginal distribution should also be downweighted. This leads us to the following set of generalized M-estimation equations

$$\sum_{t=2}^{n} \psi\left(\frac{y_t-\lambda_t}{\sqrt{\lambda_t}}\right) \frac{1}{\sqrt{\lambda_t}} \begin{pmatrix} 1 \\ \sigma\psi\left(\frac{y_{t-1}-\lambda}{\sigma}\right)+\lambda \end{pmatrix} - (n-1) \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{6}$$

where $\sigma^2 = \lambda/(1-\beta_1^2)$ is the marginal variance of the process. Recall that $\lambda = E(Y_t)$. The term $(a_0, a_1)'$ is an asymptotic bias correction such that the expectation of the left hand side equals $(0,0)'$. It can be approximated by simulation.
Since the autocorrelation function ρ of an INARCH(1)-model resembles that of an autoregressive model of first order, $\rho(h) = \beta_1^h$ for all $h = 0, 1, 2, \ldots$, we can use the median of all data points as a robust estimator of λ in combination with any robust autocorrelation estimate for initialization of $\beta_1 = \rho(1)$ in the iterative calculations. We apply the highly robust estimate of Ma and Genton (2000) in the following for this,

$$\hat{\rho}(1) = \frac{Q_{n-1}^2(y_2+y_1, \ldots, y_n+y_{n-1}) - Q_{n-1}^2(y_2-y_1, \ldots, y_n-y_{n-1})}{Q_{n-1}^2(y_2+y_1, \ldots, y_n+y_{n-1}) + Q_{n-1}^2(y_2-y_1, \ldots, y_n-y_{n-1})},$$

using Rousseeuw and Croux' (1993) Q_n for estimation of the unknown variances $Var(Y_t + Y_{t-1})$ and $Var(Y_t - Y_{t-1})$ because of its high robustness and considerable efficiency. The Q_n scale estimator of a sample $x_1, \ldots, x_m$ roughly corresponds to the 25% percentile of the sample of pairwise differences and is defined by

$$Q_m(x_1, \ldots, x_m) = c_m\{|x_i - x_j| : 1 \leq i < j \leq m\}_{(l)}, \quad l = \binom{\lfloor m/2 \rfloor + 1}{2}.$$

c_m is a finite sample correction factor to achieve unbiasedness at a sample of size m. It can be omitted in our context since it cancels out.

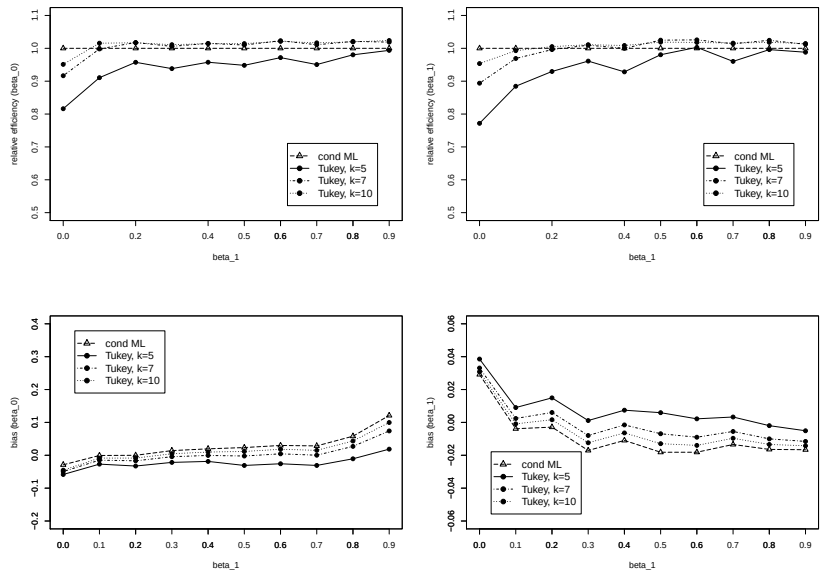


FIGURE 1. Simulated relative efficiencies of the Tukey M estimator with different tuning constants k relative to the CML estimator (top) and empirical biases (bottom) for β_0 (left) and β_1 (right) as a function of the true β_1, $n = 200$.

4 Simulations

We perform some simulation experiments to compare the performance of the CML and the generalized M-estimator in finite samples from an INARCH(1) model. We concentrate on the Tukey M-estimator because of its ability to reject observations which strongly deviate from the model for the bulk of the data.

First we consider situations with clean data, generating 500 time series of length $n = 200$ from each of different INARCH(1) models with fixed $\beta_0 = 1$ and β_1 varying in $\{0, 0.1, \ldots, 0.9\}$. Figure 1 illustrates the resulting finite-sample biases and relative efficiencies of the generalized M-estimator with different values of the tuning constant $k \in \{5, 7, 10\}$ relative to the CML estimator as a function of β_1. All estimators show a similar bias behavior, and the performance of the Tukey M estimator approaches that of the conditional ML estimator as k increases. Choosing $k \geq 7$ leads to relative efficiencies of 90% or larger for all parameter combinations considered here.

Figure 2 depicts the bias caused by an AO of increasing size ν at time $\tau = 50$ in a time series of length $n = 200$, generated from model (4) with $\beta_0 = 1$ and $\beta_1 = 0.4$. 200 data sets have been simulated for each value of ν.

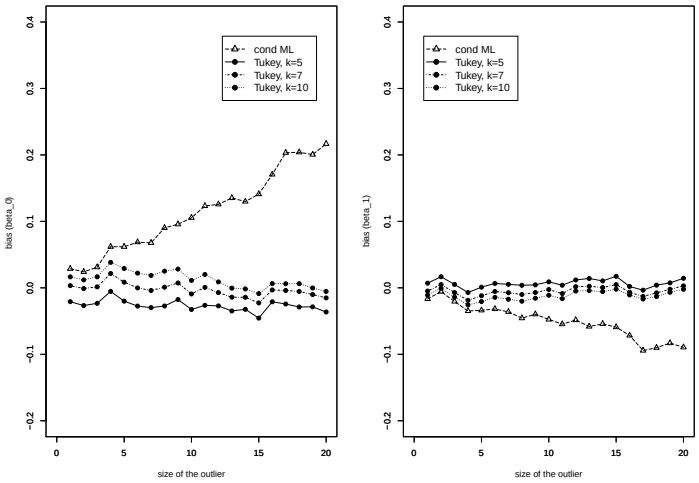


FIGURE 2. Simulated bias of the conditional ML estimator and of the Tukey M estimator with different tuning constants k for β_0 (left) and β_1 (right) in case of an additive outlier of increasing size.

While the CML estimator overestimates β_0 and underestimates β_1 because of an AO, the M estimator is little affected, even for a large tuning constant $k = 10$. The MSE is dominated by the bias and not shown here.

References

Ferland, R., Latour, A., and Oraichi, D. (2006). Integer–valued GARCH processes. *Journal of Time Series Analysis*, **27**, 923-942.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, **104**, 1430-1439.

Fokianos, K., and Fried, R. (2010). Interventions in INGARCH processes. *Journal of Time Series Analysis*, **31**, 210-225.

Ma, Y., and Genton, M.G. (2000). Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis*, **21**, 663-684.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006). *Robust Statistics*. Wiley: New York.

Rousseeuw, P.J., and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**, 1273-1283.

Bivariate Ordinal Regression Models for the Analysis of Neural Data

Julia Furche[1], Jutta Kretzberg [2], Thomas Kneib[1]

[1] Institute of Mathematics, Carl von Ossietzky University Oldenburg, D-26111 Oldenburg, {julia.furche,thomas.kneib}@uni-oldenburg.de
[2] Institute of Biology and Environmental Sciences, Carl von Ossietzky University Oldenburg, D-26111 Oldenburg, jutta.kretzberg@uni-oldenburg.de

Abstract: A bivariate cumulative probit model with linear effects and penalized splines is proposed to reconstruct stimulus properties from spike trains of retinal ganglion cells. The use of the ordinal regression model aims at evaluating the suitability of this model for the analysis of neural data, in particular with respect to gaining insights about which spike train features encode which stimulus properties (light intensity and velocity of a moving dot pattern).

Keywords: Ordinal Regression; MCMC; Penalized Splines; Stimulus Reconstruction; Spike Train

1 Introduction:

The surroundings of an organism can be described in terms of multidimensional stimuli accessible to the organism's sensory organs. Information about these stimuli is transfered to the central neural system via electrical impulses of nerve cells or *neurons*. The shape and especially the size of these impulses called *action potentials* or *spikes* are independent of the strength of the stimulus that evoked the potential. Hence, information about the environment must be transmitted by using temporal sequences of spikes, so called *spike trains*. Exploring the neural code includes especially finding out which features of a spike train (e.g. the number of evoked spikes, the relative timing of the first spike etc) encode which stimulus properties (Rieke et al (1999), see also Figure 1).

A common approach to gain insights about the encoding strategies is to reconstruct the stimulus properties from the observed neuronal responses in an experiment. Using statistical classification methods, the recorded spike trains are assigned to stimulus property classes according to their features. This idea yields a measure for the relative importance of certain spike train features by comparing the predictive performance. Most of the common methods for this classification task yield good classification results but do not take into account the possible ordinal structure of the properties even

though in fact most stimuli can be ordered considering for example properties like light intensity or velocity of visual objects. Even more important is the problem that most methods are not easily to a stimulus of more than one dimension whereas e.g. each visual object can be described by different properties like its colour, light intensity, velocity and so on.

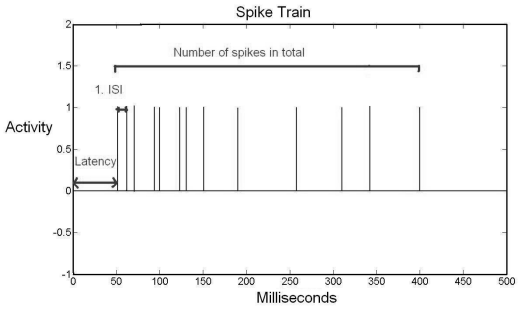


FIGURE 1. Spike train - each line represents the occurrence of an action potential

In order to fill this gap, we examine the suitability of a cumulative probit regression model for stimulus reconstruction. Based on the work of McCullagh in 1980 the idea is to introduce a latent variable in order to relate the problem to a linear regression approach (McCullagh (1980)). A multivariate extension of this model was presented in 1995 by Kim (1995) for the analysis of ophthalmological data where the response variable - in this case severity of diabetic retinopathy for both eyes - is bivariate due to paired organs. Similar studies followed (Zayeri et al (2006)) but to our knowledge no research has been made concerning the applicability of the ordinal regression model to neurobiological questions where correlation between both univariate responses is not as obvious as in measurements at paired organs. Whereas in these articles only linear effects were investigated, we also extended the approach for non-linear effects by modeling the latent variable with penalized splines.

2 Bivariate Cumulative Probit Model

Analogue to the univariate model, the idea of the bivariate cumulative probit model is to relate both components of the ordinal response variable $Y = (Y_a, Y_b) \in \{1, \ldots, K\} \times \{1, \ldots, L\}$ to latent variables Z_a, Z_b that can not be observed directly. Those real-valued variables span the whole R^2 and are cut off by $-\infty = \gamma_0^a < \gamma_1^a \ldots < \gamma_{K-1}^a < \gamma_K^a = \infty$ and $-\infty = \gamma_0^b < \gamma_1^b \ldots < \gamma_{L-1}^b < \gamma_L^b = \infty$ respectively, such that one gets for the i-th observation $Y_i = (k, l)$ iff $\gamma_{k-1} < Z_{a,i} \leq \gamma_k$ and $\gamma_{l-1} < Z_{b,i} \leq \gamma_l$. The

difference of the bivariate model to two univariate ones originates in the definition of the residuals of the regression

$$Z_{a,i} = \eta_{a,i} + \varepsilon_{a,i} \qquad Z_{b,i} = \eta_{b,i} + \varepsilon_{b,i},$$

where $\varepsilon_{a,1}, \ldots, \varepsilon_{a,n}$ (as well as $\varepsilon_{b,1}, \ldots, \varepsilon_{b,n}$) are assumed to be independent, but for each observation i there exists a constant correlation $\mathrm{Cor}(\varepsilon_{a,i}, \varepsilon_{b,i}) = \rho$. The predictor $\eta_{a,i}$ can either be linearly defined as $\eta_{a,i} = x'_{a,i}\beta_a$ or include non-linear effects for the explanatory variable $x_{a,m}$ in terms of penalized spline functions. In this case

$$\eta_{a,i} = \ldots + \sum_{j=1}^{M+d} B_j(x_{a,m,i})\alpha_{a,j} + \ldots$$

with $x_{a,m,i}$ denoting the realization of the m-th explanatory variable for the i-th observation and $B_j(x_i)$ the j-th basis function of the B-Spline basis of degree d for M knots. The penalization term can either be included into the least squares criterion or one equivalently assumes the following normal prior distribution of α:

$$p(\alpha) \propto \exp\left(-\lambda\alpha' K\alpha\right).$$

with a penalty matrix K.
Bayes-optimal parameters can be obtained numerically using a Gibbs-sampler of the full conditionals (Albert and Chib (1993)). Combination of linear effects and penalized splines into one model is possible as well, resulting in slightly different full conditionals for the individual regression parameters and the spline weights.

3 Stimulus Reconstruction

After building up the model and deriving the full conditionals, an examination of the suitability of the cumulative probit model for the analysis of neural data was performed. The question which spike train features encode which stimulus properties was considered in the light of the visual system. In the corresponding experiment, an extracted carp retina was stimulated with a moving dot pattern of different light intensity and velocity. Patterns of three velocities with movement to the right and four different light intensities were presented, but we considered only the intensity changes from the lowest to the remaining three for the classification task since former studies have detected a stronger response to changes than to constant intensities. Each stimulus combination has been applied 128 times resulting in 1152 trials in total. Spike trains were recorded from 114 ganglion cells.

We examined the predictive performance of the model first using features of a spike train recorded from a single cell and second combining several

cells to construct further possible explanatory variables. For the analysis based on single cells, we chose one neuron manually that showed clear responses to velocity as well as intensity changes, with increasing firing rate for larger realisations of both stimulus variables. Besides the number of evoked spikes we used the latency of the first spike, the length of the first interspike interval and a binary variable concerning the state of activity (at least one spike or none) for the analysis. Additionally to the linear model containing all four variables we built up models with P-splines for the most promising covariates spike count and first spike latency.
Making up a population of the total 114 neurons, we considered the number of active cells, the total spike count as well as the number of spikes observed for each single cell as potential explanatory variables.

For univariate response variables, fixing either the light intensity or the velocity, the latency as well as the spike count yielded good classification results. Consistent with former studies, classification with latency was superior for intensity reconstruction whereas the total number of spikes was superior for reconstruction of velocity. In this application, a linear model was not sufficient to cover the effect of first spike latency such that the use of penalized splines was preferred. For bivariate classification, linear models of at least two covariates outperformed one-covariate-models. Here, the best results were obtained using the total number of spikes of each neuron, resulting in a model with 114 explanatory variables. In general, the classification rates assigning the spike trains with the cumulative probit model were comparable to those obtained with other methods, especially when P-splines accounted for non-linear effects of certain variables.
sectionSection 3.1

References

Albert, J.H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Data. In: *Journal of the American Statistical Association*, **88**-422, 669-679.

Kim, K. (1995). A Bivariate Cumulative Probit Regression Model for Ordered Categorical Data. In: *Statistics in Medicine*, **14**, 1341-1352.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). In: *Journal of the Royal Statistical Society, Series B*, **42**, 109-142.

Rieke, F., Warland, D., de Ruyter van Steveninck, R. and Bialeck, W. (1999). *Spikes - Exploring the neural code*. The MIT Press.

Zayeri, F. and Kazemnejad, A. (2006). A Latent Variable Regression Model for Asymmetric Bivariate Ordered Categorical Data. In: *Journal of Applied Statistics*, **33**-7, 743-753.

Modelling endocytosis by means of non-homogeneous temporal Boolean models.

M. Ángeles Gallego[1], M. Victoria Ibáñez[1], Amelia Simó[1]

[1] Department of Mathematics. University Jaume I. 12071 Castellón. Spain.

Abstract: Many medical and biological problems require the analysis of large sequences of microscope images. These images capture phenomena of interest and it is essential to characterize their spatial and temporal properties. The purpose of this paper is to show the application of the Non-Homogeneous Temporal boolean model, and of a statistical methodology to estimate these parameters of interest in image sequences obtained in the observation of endocytosis. Endocytosis is a process by which cells traffic molecules from the extracellular space into different intracellular compartments.
In this paper, we introduce the concept of Non-Homogeneous Temporal Boolean Model; a hypothesis testing procedure to check the spatial homogeneity assumption; and a reformulation of the existing methodology to work with underlying non-homogeneous point processes. Finally we apply it, to three sequences of endocytic images.

Keywords: Temporal Boolean model, Endocytosis, Spatial non-homogeneity, parameter estimation.

1 Introduction

Endocytosis is a cellular process whereby some materials (e.g. nutrients) are drawn into the cell by means of invagination of the plasma membrane. This process happens in discrete events and it is required for a vast number of vital functions for the well-being of a cell.
A microscopical technique called Total Internal Reflection Fluorescence Microscopy (TIRFM), allows real-time imaging for endocytosis with a high degree of accuracy. Using TIRFM, the assembly of fluorescently labelled clathrin where endocytosis is taking place, results in the appearance of a diffraction-limited spot. The areas of fluorescence generated by different endocytic spots overlap and form random clumps which have different size, shape and duration. The time which elapses between the appearance and the disappearance of a fluorescent clathrin spot is defined as the duration, or lifetime, of a discrete endocytic event.
The spatial and temporal distribution of these clumps is influenced by many biological factors and there is no precise biological knowledge about

their spatial distribution in the plasma membrane. In fact, this is one of the unsolved questions in the biological understanding of the endocytic process. Therefore, to characterize endocytic events it is crucial to estimate the mean number of endocytic events per unit area and per unit time at different spatial sites and their lifetime. Due to endocytic spots overlapping and clump formation, it is not possible to carry out these tasks in a trivial way. In [3] and [1], Sebastián et al. used the homogeneous temporal Boolean model (HTBM) to estimate these parameters of interest.
The novelties introduced by our work are: the relaxation of the spatial homogeneity hypothesis by introducing the concept of Non-Homogeneous Temporal Boolean Model (NHTBM); the introduction of a hypothesis testing procedure to check the non-homogeneity hypothesis; and a generalization of the methodology to estimate the new parameters of interest. We apply it to analyze the behavior of the clathrin-dependent endocytic machinery. The use of a model that is more closely adjusted to the physiological characteristics of the real problem leads to more accurate estimators, and it solves one of the open biological questions regarding which parts of the membrane present a greater accumulation of events.

2 Models and methods

2.1 Non-homogeneous temporal Boolean model

Let $\Psi = \{(x_i, t_i)\}_{i\geq 1}$ be a Poisson point process in $\mathbf{R}^2 \times \mathbf{R}_+$, homogeneous in time but non-homogeneous in space, with intensity function $\Lambda(x)$, $x \in \mathbf{R}^2$. Let $\{A_i\}_{i\geq 1}$ be a sequence of independent and identically distributed random compact sets in $\mathbf{R}^2$, and let $\{d_i\}_{i\geq 1}$ be a sequence of independent and identically distributed (as D) positive random variables and that $E\nu_3(A_0 \times [0, D] \bigoplus \check{K}) < +\infty$ for any compact subset K of $\mathbf{R}^3$. Then, the non-homogeneous temporal Boolean model is defined as: $\Phi = \bigcup_{i\geq 1}(A_i + x_i) \times [t_i, t_i + d_i]$, where E denotes the expectation; for any sets A and B in $\mathbf{R}^3$ $\nu_3(A)$ denotes the volume of A, and $A \bigoplus B$ denotes their Minkowsky addition.
In the applications, we will work with binary images sequences that will be considered as samples of a spatiotemporal infinite process, as the defined below. The spatiotemporal sampling window will be denoted by $W \times [0, T]$ and the sampling times will be denoted by $s_1 < s_2 < \cdots < s_m$, with $0 \leq s_1$; $s_m \leq T$. Then, the observed data set will be: $\{\Phi_{s_i}\}_{i=1,\cdots,m}$ with $\Phi_{s_i} = \Phi \bigcap (W \times \{s_i\})$ $\forall i = 1, \cdots, m$.

2.2 Parameter estimation

Two different approaches are found in [1] and [3] to manage parameter estimation in a HTBM, but only one of them can be generalized to the non-homogeneous case. This approach uses several cross-section aggregations, $\widetilde{\Phi}_{s_i} = \bigcup_{j=i}^{i+k} \Phi_{s_j}$, $i = 1, \cdots, m-k$, to analyze the increase in intensity.

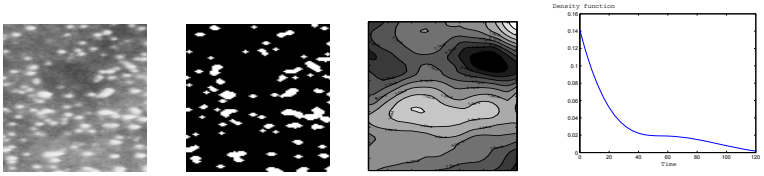
In these sequences, the grains size will keep its original distribution, although the spatial intensity for the germs process will be higher. This rise in intensity will only depend on the number of frames aggregated and their time lags. Each $\widetilde{\Phi}_{s_i}$ is a realization of a spatial non-homogeneous Boolean model. Algorithms to estimate the intensity function, $\lambda_s(k, \delta, x)$, are scarce in the non-homogeneous literature. We will use the proposed by Molchanov and Chiu [2] , that will be repeated for different values of k and δ.
Once $\lambda_s(k, \delta, x)$ has been estimated we follow the ideas (and notation) stated in [3], getting an estimate of the intensity function $\Lambda(x)$: $\hat{\Lambda}(x) = \hat{\alpha}'(0, x)$; an estimation of the probability density of D for each site, $\hat{f}_D(\delta) = -\frac{1}{\hat{\Lambda}(x)}\hat{\alpha}''(\delta, x)$ (we will use their mean as the final estimate of the probability density function); and an estimate of ED, $\hat{ED} = \frac{1}{\sharp W}\sum_{x \in W}\left[\frac{\frac{1}{m}\sum_{j=1}^{m}\hat{\Lambda}_{s_j}(x)}{\hat{\Lambda}(x)}\right]$

2.3 A simple test for spatial homogeneity

There are no formal homogeneity tests in spatial random sets and point processes literature. Usually, a single observation in a sample window is available and homogeneous spatial patterns could look like non-homogeneous depending on the size of the window. Nevertheless, we propose the following scheme to check homogeneity in spatio temporal problems: **Step 1.** Use the Molchanov method [2] to estimate the intensity function from each frame. **Step 2.** Use a batch-means type method to obtain independent replications based on the temporal observations. **Step 3.** Under the null hypothesis of homogeneity, the estimated value does not depend on the coordinate. Apply the Friedman non-parametric ANOVA test to the sample obtained in the previous step, to compare the estimated values at each position of each frame.
A simulation study has been carried out order to check the performance of the parameter estimation and the homogeneity testing procedures.

3 Application

Lets analyze clathrin-mediated endocytosis dynamics, from three sequences of images of COS-7 monkey fibroblast cells. Each sequence consists of 300 frames acquired at one frame every four seconds. One of the original frames is shown in fig 1 (a). Frames are preprocessed and transformed to binary images (fig 1 b). The homogeneity test exposed below allows us to reject the homogeneity hypothesis and to consider our data set as a realization of a NHTBM. Fig. 1(c) shows the spatial intensity function estimated for one of the analyzed cells (Cell 2). We can clearly observe a greater density of endocytic spots in the image centre. Fig. 1 (d) shows the estimation of the density function of event durations for Cell 2.



(a) (b) (c) (d)

FIGURE 1. (a) A frame of a sequence of 300 TIRFM images of a cell; (b) segmented endocytic spots of the frame showed in (a); (c) Estimated spatial intensity function for the Cell 2; (d) Estimated density functions of durations for Cell 2.

4 Conclusions

In this paper we have proposed both a probabilistic model and a statistical methodology that generalize the methodology proposed in [3] to study the kinetics of endocytosis in living cells. The novelty spatial homogeneity hypothesis has been relaxed by introducing the concept of a non-homogeneous temporal Boolean model. Regarding the endocytosis, we have detected parts of the cellular membrane with a higher accumulation of endocytic spots and slightly lower estimates for the durations of the endocytic events than the obtained with the methods nowadays in use.

Acknowledgments: We would like to thank Dr. M.E. Díaz, Dr. G. Ayala, Dr. D. Toomre and R. Zoncu, for introducing us in this problem, obtaining the images and allowing us to use them. This work has been supported by projects TIN2007-67587, TIN2009-14392-C02-01, and P11A2009-02.

References

[1] G. Ayala, R. Sebastián, M.E. Díaz, E. Díaz, R. Zoncu, and D. Toomre. Analysis of spatially and temporally overlapping events with application to image sequences. *IEEE Transactions on Pattern Analysis and machine intelligence*, 28(10):1707–1712, 2006.

[2] I.S. Molchanov and S.N. Chiu. Smoothing techniques and estimation methods for nonstationary boolean models with applications to coverage processes. *Biometrika*, 87(2):265–283, 2000.

[3] R. Sebastián, E. Díaz, G. Ayala, M.E. Díaz, R. Zoncu, and D. Toomre. Studying endocytosis in space and time by means of temporal boolean models. *Pattern Recognition*, 39(11):2775–85, 2006.

A Prior for multiplicity control and closed-form Bayes factors in variable selection

Gonzalo García-Donato[1], M. Jesús Bayarri[2], James O. Berger[3], Anabel Forte[2]

[1] Universidad de Castilla La Mancha
[2] Universitat de Vàlencia
[3] Duke University and SAMSI

Abstract: In model selection problems posterior probabilities of entertained models are simple expressions of the Bayes factors and the prior distribution over the model space. For the variable selection problem in normal regression models, in this work we consider the posterior probabilities that arise of combining the Bayes factors in (Bayarri, Berger, García-Donato, and Forte, 2011) and the proposal in (Scott and Berger, 2010) for the model prior probabilities. The result is a default Bayesian approach, based on theoretical arguments, where posterior probabilities i) are closed-form (in terms of hypergeometric functions), and ii) automatically control for multiplicity. Notice that both properties are very appealing specially when the number of potential explanatory variables initially considered is very large. We compare this approach with other existing methodologies like (Zellner and Siow, 1980) and (Liang et al., 2008) for the Bayes factors and the constant prior for the prior distribution over the model space. A number of different scenarios are considered, including real data sets with large model spaces.

Keywords: Variable Selection; Bayes Factors; Objective Priors.

References

Bayarri, M.J., Berger, J.O., Garía-Donato, G. and Forte, A. (2011). Objective Bayes Criteria for the Variable Selection Problem.

Scott, J. and Berger, J.O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *The Annals of Statistics*, **38**(5) pp: 2587-2619

Zellner, A. and Siow, A. (1980). Posterior Odds Ratio for Selected Regression Hypotheses. In *Bayesian Statistics 1.* 585-603, Valencia University Press.

Liang, F., Paulo, R., Molina, G., Clyde, M.A. and Berger, J.O. (2008). Mixtures of g-Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, **103** pp: 410-423.

Approximated Survival function in the Sum of Two Independent Homogeneous Markov Processes: Application to Bladder Carcinoma.

B. García-Mora[1], C. Santamaría[1], E. Navarro[1], G. Rubio[1]

[1] Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, Edificio 8G, piso 2, Camino de vera s/n, 46022, Valencia, España. E-mail address: {magarmo5, crisanna, entorres, grubio}@imm.upv.es

Abstract: The study of the sum of two independent phase–type (PH) distributed variables is considered, each of them being associated with a Markovian process with one absorbing state. The distribution function of the variable sum, PH–distributed, is computed. The exponential function of a block upper triangular matrix is calculated in terms of its respective blocks to reduce the dimension from the original processes. In a second step an approximated solution to this previous method is modelled. An application in bladder carcinoma is shown taking into account two absorbing states.

Keywords: Markov process; Phase–type distribution; Bladder carcinoma.

1 Introduction and Motivation.

Bladder carcinoma is the fourth most frequent solid tumor among men and the seventh most frequent among women, with more than 350.000 new cases diagnosed annually worldwide. 80% of patients present *superficial* transitional cell carcinoma (TCC), which can be managed with transurethral resection (*TUR*), a surgical endoscopic technique. However, more than 50% of the patients will have *recurrences* (reappearance of a new superficial tumor) and 10–30% of patients will have *progression* to muscle invasive disease which leads to a more aggressive treatment including the *bladder extirpation*.

In this regard, *Markov models* have proven to be useful in the analysis of the course of chronic diseases with relapse times. The Markov model is also complemented by the use of phase–type distributions (Aalen, 1995). A *Phase–Type (PH) distribution* is the absorption distribution time in a homogeneous Markov process in a finite state space with one absorbing state. In our modelling we consider two random continuous variables representing two independent absorption times, each one of them *PH*-distributed. In order to study the risk of the *bladder extirpation*, we are interested in

obtaining the distribution function of the sum of these two variables. In a second step we will obtain an approximation of this function.

2 Approximated Survival Function of the Sum of Two Independent Markov Processes.

Let X_1 and X_2 be nonnegative random independent variables representing the absorption times in two homogeneous Markov processes with m and n transient states respectively and $m+1$ and $n+1$ the absorbing ones. Both variables are PH–distributed with representation $(\alpha,\, T)$ and $(\beta,\, S)$ respectively and distribution function $F_1(\cdot)$ and $F_2(\cdot)$. The distribution of $X = X_1 + X_2$ is the convolution of the distributions of X_1 and X_2 (Neuts, 1998) and $F(\cdot) = (F_1 * F_2)(\cdot)$ is a PH-distribution with (γ, L), given by

$$\gamma = (\alpha, \alpha_{n+1}\beta) \tag{1}$$

$$L = \left[\begin{array}{c|c} T & T^0\beta \\ \hline 0 & S \end{array}\right] \tag{2}$$

The transition rates of X_1 and X_2 within the set of transient states are given by the matrix T and S respectively. X is the total time duration of the whole process: from initial state until the second absorbing *n+1* state, passing through the first absorbing *m+1* state. Therefore we consider a new Markov process with state space $\{1,\, 2,\, \ldots, m,\, m+1,\, m+2, \ldots, m+n,\, m+n+1\}$, where $m+n+1$ is the only absorbing one. In this process the first states $\{1,\, 2,\, \ldots, m\}$ are the transient ones of the first process and the state $m+j$, $1 \leq j \leq n+1$ represents the transient jth state of the second process when the first process has already arrived at the absorbing state $m+1$. After having reached the state $m+1$, the chain immediately proceeds to the second stage. The initial probability vector and the infinitesimal generator of this new Markov process are $(\gamma, \alpha_{m+1}\beta_{n+1})$ with $\gamma = (\alpha, \alpha_{m+1}\beta)$, and

$$Q = \left[\begin{array}{c|c} L & L^0 \\ \hline 0 & 0 \end{array}\right] \text{ with } L = \left[\begin{array}{c|c} T & T^0\beta \\ \hline 0 & S \end{array}\right]$$

The distribution function $F(x)$ for the variable sum X is PH–distributed with representation $(\gamma,\, L)$ given by

$$F(x) = 1 - \gamma \exp(Lx)\mathbf{e_{m+n}} \tag{3}$$

Notice that in (3) the dimension of the problem increases given that matrix L is greater than the matrices T and S. In order to reduce the dimension we apply the Fréchet derivative (Kenny and Laub, 1998) to the term $\exp(Lx)$

$$F(x) = 1-(\alpha\ \ \alpha_{m+1}\beta)\left[\exp\begin{pmatrix} Tx & 0 \\ 0 & Sx \end{pmatrix} + \int_0^1 \exp\left[(1-s)\begin{pmatrix} Tx & 0 \\ 0 & Sx \end{pmatrix}\right]\begin{pmatrix} 0 & T^0\beta x \\ 0 & 0 \end{pmatrix} exp\left[s\begin{pmatrix} Tx & 0 \\ 0 & Sx \end{pmatrix}\right] ds\right]\mathbf{e_{m+n}} \tag{4}$$

and operating we arrive at the expression for the distribution function

$$F(x) \quad = \quad 1-\alpha\,\exp(Tx)\mathbf{e_m}-\left(\alpha_{m+1}\beta\exp(Sx)+\alpha\int_0^1\exp(1-s)TxT^0\beta x\exp(sSx)\,ds\right)\mathbf{e_n}$$

Applying the *Kronecker matrix form* on the integral of this last expression,

$$F(x)=1-\alpha\,\exp(Tx)\mathbf{e_m}-\alpha_{m+1}\beta\exp(Sx)\mathbf{e_n}-(\mathbf{e_n}'\otimes\alpha)[S'x\oplus(-Tx)]^{-1}(\exp(S'x)\otimes\mathrm{I_m}-\mathrm{I_n}\otimes\exp(Tx))vec(T^0\beta x) \tag{5}$$

Notice that the calculation of the inverse of the matrix $sS'x\oplus(-Tx)$ in (5) can present serious difficulties if this matrix is bad conditioned. With the aim of avoiding a possible bad conditioning we propose an approach for calculating the integral $\int_0^1\exp\left(sS'x\oplus(1-s)Tx\right)ds$ in $F(x)$. For this we use the Taylor series expansion of the *exponential function* in the integrate. We apply the Weierstrass' criterion of uniform convergence and the Newton Binomio and finally we arrive to the approximated function of $F(x)$

$$\begin{aligned}\hat{F}_1(x) \quad &= \quad 1-\alpha\,\exp(Tx)\mathbf{e_m}-\alpha_{m+1}\beta\exp(Sx)\mathbf{e_n}- \\ &\quad (\mathbf{e_n}'\otimes\alpha)\left(\mathrm{I}+\sum_{k=1}^{p}\frac{x^k}{(k+1)!}\sum_{j=0}^{k}(S')^{k-j}\otimes T^j\right)vec(T^0\beta x)\end{aligned} \tag{6}$$

In a step more with the aim to get a greater convergence and accuracy for $F(x)$, improving $\hat{F}_1(x)$, we construct a second approximation. Applying again the Frechet derivative and recurrently the Kronecker matrix properties we arrive to a second approximated distribution function.

$$\begin{aligned}\hat{F}_2(x) \quad &= \quad 1-\alpha\,\exp(Tx)\mathbf{e_m}-\alpha_{m+1}\beta\exp(Sx)\mathbf{e_n}- \\ &\quad (\mathbf{e_n}'\otimes\alpha)\,[(e^{s\frac{S'x}{2^k}}\otimes I+I\otimes e^{\frac{Tx}{2^k}})\,(\int_0^1 e^{s\frac{S'x}{2^k}}\otimes I+I\otimes e^{(1-s)\frac{Tx}{2^k}}\,ds)]\,vec\frac{T^0\beta x}{2^k}\end{aligned} \tag{7}$$

3 Application to Bladder Carcinoma.

3.1 Initial state assumptions

Three states are distinguished (see Diagram) in bladder carcinoma with the first *moderate progression* and the *bladder extirpation* as absorbing states.

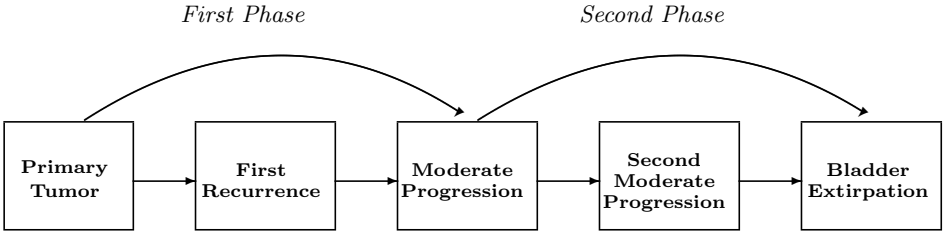


Diagram: Markov processes with two absorbing states

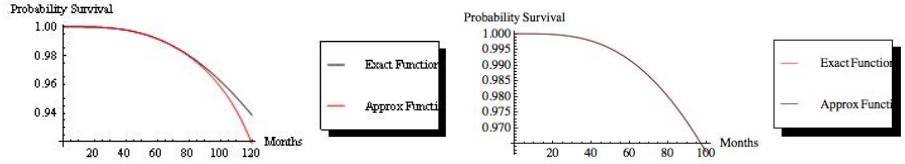


FIGURE 1. Survival function $S(x)$ and the first and second approximation, $\hat{S}_1(x)$ and $\hat{S}_2(x)$ of the distribution function in the Markov process.

Two well differentiated follow–up protocols are distinguished according to the *two phases* of the study: one treatment for superficial tumors and a more specific and different for invasive for invasive tumors. Two independent databases have been considered from *La Fe University Hospital*.

3.2 Computing the approximated survival function

In the Figure 1 the first (6) and second (7) approximated survival functions has been compared with the theoretical model (5). We can observe a light mismatch between both survival functions with the first approximation while the second is more secure convergence.

Acknowledgments: This study has been funded by *First Research Projects of Universitat Politècnica de València. Code 20100975. Call 2010.*

References

Aalen, O.O. (1995). On phase type distributions in survival analysis. *Scand. J. Stat.*, **22** 447–463.

Kenney, C.S. and Laub, A.J. (1998). Condition estimates for matrix functions. *SIAM J. Matrix Anal. Appl.*, **10 (3)**, 191–209.

Neuts, M.F. (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. John Hopkins University Press.

On using the Hellinger distance in checking the validity of approximations based on dynamic generalized linear models

Ali S. Gargoum

[1] Department of Statistics, UAE University, United Arab Emirates, P. O. Box 17555, alig@uaeu.ac.ae

Abstract: The purpose of this paper is to provide validation for the approximate algebraic propagation algorithms to accommodate non-Gaussian dynamic processes. These algorithms have been developed to carry out Bayesian analysis based on conjugate forms. The validity of the approximation algorithms can be checked by introducing a metric (Hellinger divergence measure) over the distribution of the states (parameters) and use it to judge the approximation. Theoretical bounds for the efficacy of such procedure are discussed.

Keywords: Dynamic generalized linear models; Hellinger distance; Probability propagation.

1 Introduction

Over the past two decades non-Gaussian time series have been addressed by many authors, see, e.g., Kitagawa (1987) and Durbin and Koopman (2000). The dynamic generalized linear models (DGLM) provide a general framework for dealing with time series data which considers generalized linear models with time-varying parameters. DGLM have been widely used for non-Gaussian time series data, see, e.g. West and Harrison. (1997). In dynamic processes which are known in sufficient detail to be described in terms of parametric models, the model parameters or states can be regarded as the means summarizing the information necessary to forecast the future system behavior. The learning process sequentially revises the uncertainty about the parameters, by adjusting the probability distribution attached to its state variables. An important example of such a process is the environmental problem of forecasting the geographical spread of a release of toxic gases in the event of an accident at a chemical or nuclear plant Smith and French (1993). Puffs of contaminated masses are emitted from a release source, dispersed by a wind field and fragment into other puffs over time. The wind field, mass release and fragmentation process follows a complicated physical model. The problem here is to produce realistic probability estimates of contamination concentration over space and time. In

this high-dimensional dynamic system, where the states (parameters) are allowed to change over time, computational efficiency is essential. To model such scenarios Bayesian networks were defined over state spaces. When the system is Gaussian i.e. the states are normally distributed and the observations have Gaussian density, quick exact propagation algorithms that calculate the posterior distribution in closed form in the light of incoming data are well known. Gargoum (2006) and Settimi and Smith (2000) described approximate algorithms of propagation and probability updating for non-Gaussian dynamic systems incoming data. These algorithms which are based on dynamic generalized linear models. They are extremely efficient and provide a fast method compared to numerical methods based on MCMC algorithms to update the probabilities in dynamic systems. The validity of the these updating algorithms depends critically on how well the posterior density is approximated. Checking the validity of these algorithms is the main issue that this work addresses.

2 The dynamic generalized linear model

The DGLM for the time series $\{y_t\}, (t = 1, 2, ...)$ is defined by the following two components
Observational equation:

$$p(y_t|\lambda_t) \quad \text{and} \quad \lambda_t = g(\eta_t), \eta_t = \mathbf{F}_t^T \boldsymbol{\theta}_t$$

Evolution equation:

$$\boldsymbol{\theta}_t = G_t \boldsymbol{\theta}_{t-1} + \mathbf{w_t} \quad \text{with} \quad \mathbf{w_t} \sim N(\mathbf{0}, W_t)$$

Here the sampling distribution of $\mathbf{y_t}$ given a random variable λ_t belongs to the exponential family possibly non-normal and λ_t is a function of a linear combination of the state vector parameters $\boldsymbol{\theta}_t$ for some known regression vector $\mathbf{F}_t$ and known invertible map $g(.)$ which, in many cases, will be the identity map. The evolution equation is the same as in the normal dynamic linear model. The evolution errors $\mathbf{w_t}$ are assumed uncorrelated over time.

3 An approximate Bayesian analysis

Now in such complex systems, if the sampling distribution of the observations is normal, then fast propagation algorithms over the dynamic system can be used where information can be transmitted through the system by updating the probabilities of each group of states (a clique) sequentially very fast and in closed form Smith and French.(1993). However, when the sampling distribution of the observations given the states is not normal, then the posterior distribution of the states cannot be determined in closed form but the conditional independence relationships among variables still

hold. In this case we deal with a class of models which is a generalization of the standard DLM to non-normal error models for time series. As we mentioned above the observations $y_t|\lambda_t$ are non-normal and the observational mean λ_t is, in general, nonlinear function of $\boldsymbol{\theta}_t$ (and η_t.)The analytical approach to update the states vector $\boldsymbol{\theta}_t$ cannot be adopted and an approximate analysis is needed. A proposed approximate Bayesian analysis, (see, West and Harrison (1997)) develops as follows.

1) Suppose that the posterior distribution of the states at time $t-1$, $(\boldsymbol{\theta}_t|D_{t-1})$ or any linear combination of them is partially specified in terms of the first two moments.

2) Approximate the actual density λ_t by the distribution in the exponential family which is closed under sampling to $y_t|\lambda_t$ by equating the first two moments with those derived from the moments of $g(\eta_t)$.

3) Perform a standard conjugate analysis to calculate the approximate density of $\lambda_t|y_t$ say $\hat{p}(\lambda_t|y_t)$

4) Update the distribution of η_t from $\hat{p}(\lambda_t|y_t)$ as $\eta_t = g^{-1}(\lambda_t)$

5) Estimate the posterior moments of the states $\boldsymbol{\theta}_t$from the moments of the distribution η_t after observing y_t.

Note that if the states are conditionally normal then the posterior of $\boldsymbol{\theta}_t$ is approximated by normal distribution with mean and variance derived from the approximated normal distribution of η_t given y_t.

4 The closeness of dynamic approximation

Here, we choose the Hellinger metric to check the appropriateness of the dynamic approximation. The Hellinger distance between two densities is defined by

$$d_H(f,h) = \left(1 - \int f^{1/2}(\boldsymbol{x})h^{1/2}(\boldsymbol{x})d\boldsymbol{x}\right)^{1/2} \tag{1}$$

Define

$$I(f,g) = 1 - d_H^2(f,g) \tag{2}$$

In fact $d_H^2(f,g)$ can be calculated in closed form for most densities in a standard family. It is also sometimes possible to explicitly write down the Hellinger distance between two densities from different families. For example when f is a normal density with mean μ and variance σ^2 and g a Gamma density with the same mean and variance, then $I(f,g)$ defined above is given by

$$I^2(f,g) = (2\pi)^{-1/2}2^{(\alpha-1)}\alpha^{1/2\alpha}\frac{(\Gamma(1/4[\alpha+1]))^2}{\Gamma(\alpha)}e^{1/2\alpha}$$

where $\alpha = \mu^2/\sigma^2$. It is easily checked that $I^2(f,g)$ is small when α is moderately large.
We note that the two properties listed below also hold true both for the variation metric and the popular Kullback-Leibler separation measure.
Suppose that p and $\hat{p}$ are joint densities on $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ which have different margins p_1 and $\hat{p}_1$ on $\boldsymbol{X}_1$ but whose conditional densities of $\boldsymbol{X}_2|\boldsymbol{X}_1$ agree. Then, directly from (1) we have that

$$d_H(p,\hat{p}) = d_H(p_1,\hat{p}_1) \tag{3}$$

Now within our context we approximate only the distribution of λ, conditional on λ all states are held fixed. It follows that the closeness of the joint density over states depends only on the closeness of our approximation of the one dimensional normal posterior density of λ to the true posterior density of λ. As an example consider the case when $p_0(x)$ is a Gaussian prior density $\boldsymbol{X}$. Let f_1 and f_2 denote the posterior densities on $\boldsymbol{x}$ given the true normalized Gamma likelihood ℓ_1 associated with a Poisson observation $\boldsymbol{Y}$ or a normalized Gaussian approximation ℓ_2 of the DGLM, respectively. Then, by definition, omitting the arguments,

$$f_i = \frac{p\ell_i}{\int p\ell_i} \qquad i = 1,2$$

So

$$I^2(f_1;f_2) = \frac{(\int p\ell_1^{1/2}\ell_2^{1/2})^2}{(\int p\ell_1)(\int p\ell_2)} \tag{4}$$

$$d_H^2(f_1;f_2) = 1 - \sqrt{I^2(f_1;f_2)} \tag{5}$$

$$= 1 - \frac{B}{\sqrt{A}} \tag{6}$$

where $B = \frac{\int p\ell_1^{1/2}\ell_2^{1/2}}{\int p\ell_1}$ and $B = \frac{\int p\ell_2}{\int pl_1}$
Notice that if ℓ_1 and ℓ_2 are very close, then both A and B will be close to 1 and consequently d_H will be close to zero. An upper bound for $d_H^2(f_1;f_2)$ can easily be derived.

5 Conclusion

Quick computational methods -based on DGLM- are discussed for dealing with non-normal data in complex dynamic scenarios. These methods give a closed form updating and provide approximations whose validity need to be checked numerically. In this paper I examined the appropriateness of these dynamic approximations. The Hellinger metric was computed to check the validity of the approximation.

References

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion) JRSS, ser B, **62**, 3-36.

Gargoum, A. S. (2006). An approximate fast Bayesian method for the analysis of the lognormal time series JASS, **15**, 135-143.

Kitagawa, G. (1987). *Non-Gaussian state-space modeling of non stationary time series.* JASA, **82**, 1032-1063.

Settimi, R. and Smith, J. Q. (2000). A comparison of approximate Bayesian forecasting methods for non Gaussian time series, *Journal of Forecasting.* **19**, 135-148.

Smith, J. Q. and French, S. (1993). Bayesian updating of atmospheric dispersion models for use after an accidental release of radiation. *The Statistician.* **42**, 501-511.

West, M., and Harrison, P. J. (1997). *Bayesian forecasting and dynamic models.* 2nd edition, Springer, New York. West, M., Harrison, P.J. and Migon, H. S. (1985). *Dynamic Generalized Linear Models and Bayesian Forecasting.* (with discussion). JASA, **80**, 73-97.

Parameter Estimation in Skills-based Knowledge Space Theory and Cognitive Diagnosis Models: A Comparison

Ann Cathrice George[1], Ali Ünlü[2]

[1] Research School Education and Capabilities, TU Dortmund, Hauert 14a, D-44227 Dortmund, Germany, a.george@educap.de
[2] Department of Statistics, TU Dortmund, Vogelpothsweg 87, D-44227 Dortmund, Germany, uenlue@statistik.tu-dortmund.de

Abstract: Diagnostic psychometric models using categorical latent variables have the potential to provide individualized feedback relevant for instruction and learning. This paper discusses and compares two approaches, the diagnostic knowledge space and cognitive diagnosis models, at the modeling and estimation levels.

Keywords: Psychometric Modeling; Diagnostic Measurement; Knowledge Space Theory; Cognitive Diagnosis Models.

1 Diagnostic Models

In educational testing, over the recent years, there has been an increasing interest in diagnostic inferences about multiple skills, that is, latent cognitive criteria such as reading comprehension, mathematical or non-verbal abilities. Diagnostic models provide personalized information at a high definitional grain size, based on which targeted learning aids can be developed.

1.1 General Purpose of Diagnostic Models

We consider an $N \times J$ data matrix $\mathbf{X}$ containing the binary responses, 0 or 1, of N examinees to J test items. The nth row $\mathbf{X}_n \in \{0,1\}^J$ of this matrix represents the response pattern of examinee n. Moreover, a set of K skills is assumed to underlie the test items. The skills required to master an item are specified by expert panels, whereas this specification is called skill-item assignment. Each examinee n possesses a subset of these K skills described by a latent (i.e., not observed) binary vector $\alpha_n \in \{0,1\}^K$, which is called the skill pattern of examinee n. The aim is to estimate the occurrence probabilities of the skill patterns and the prevalences of the individual skills in the population under reference. Additionally, information is gained about the response error rates (i.e., guessing and slipping effects) that may cause

atypical responses in test items. For each examinee, her or his posterior probability of mastering each of the individual skills, the examinee's skills profile, is also determined.

1.2 Two Examples: Skills-based Knowledge Space Theory and Cognitive Diagnosis Models

In this paper we consider two sorts of diagnostic models: the skills-based knowledge space theory (KST; e.g., Doignon & Falmagne, 1999) and the cognitive diagnosis models (CDMs; e.g., Rupp et al., 2010). Skills-based KST focuses on deterministic ordinal structures, whereas CDMs are statistical parametric models. Subsequently, we restrict our attention to the special case of a bijective skill-item assignment, which assumes that each item loads on exactly one skill. This is for illustration purposes only and can be generalized. Parameter estimation in CDMs is described, for instance, by de la Torre (2009). The approach to parameter estimation in skills-based KST discussed in this paper is proposed for the first time (however, cf. also Schrepp, 1999).

2 The Structure of Diagnostic Knowledge Space and Cognitive Diagnosis Models

2.1 Skills-based KST

The general idea of skills-based KST is that, given a skill-item assignment, only certain response patterns should occur. These model-based response patterns are called delineated knowledge states. For example, if s_1 is the only skill assigned to the items q_1, q_3 and q_5, then for a respondent having mastered this skill her or his delineated knowledge state is $(1, 0, 1, 0, 1, \ldots)$. Deviations of an empirical response pattern from the delineated knowledge states are considered to be response errors. For each empirical response pattern $\mathbf{X}_n$ a set of corresponding delineated knowledge states $\{\mathbf{K}_l\}_{l \in A_n}$ is predicted. This set of predicted states is determined by calculating the minimal Hamming distance δ between the empirical response pattern and the delineated knowledge states. The index set of predicted states for $\mathbf{X}_n$ is $A_n = \left\{l = 1, \ldots, P : \delta(\mathbf{X}_n, \mathbf{K}_l) = \min_{i=1,\ldots,P} \delta(\mathbf{X}_n, \mathbf{K}_i)\right\}$, where P is the number of all delineated knowledge states.

2.2 CDMs

In CDMs endorsement probabilities are modeled based on guessing and slipping parameters, given the different skill patterns. The probability for examinee n to solve item j is calculated as a function of the examinee's

latent response η_{nj}, and the guessing and slipping rates g_j and s_j for item j, respectively, conditional on the examinee's skill pattern α_n:

$$P_j(\alpha_n) = P(X_{nj} = 1|\alpha_n) = g_j^{(1-\eta_{nj})}(1 - s_j)^{\eta_{nj}}.$$

The examinee's latent response η_{nj} is binary, 0 or 1, indicating absence or presence of all required skills for item j, respectively. Assuming conditional independence of the item responses given the skill patterns and the examinees being sampled randomly, the conditional likelihood of the observed data $\mathbf{X}$ is

$$\prod_{n=1}^{N} L(\mathbf{X}_n|\alpha_n) = \prod_{n=1}^{N}\prod_{j=1}^{J} P_j(\alpha_n)^{X_{nj}}[1 - P_j(\alpha_n)]^{1-X_{nj}}.$$

3 Parameter Estimation

3.1 Skills-based KST

Let $\{\mathbf{K}_l\}_{l\in A_m}$ be the set of predicted knowledge states for each empirical response pattern $\mathbf{X}_m$. For any item $j \in \{1, \ldots, J\}$, the guessing and slipping probabilities g_j and s_j can be estimated based on the deviations between the empirical response patterns and their sets of predicted states:

$$\hat{g}_j = \frac{\sum_{m=1}^{M}\left(\sum_{l\in A_m}\frac{h_m}{|A_m|}I_{\{X_{mj}>K_{lj}\}}\right)}{\sum_{m=1}^{M}\left(\sum_{l\in A_m}\frac{h_m}{|A_m|}\ I_{\{K_{lj}=0\}}\right)},$$

where M is the number of different response patterns $\mathbf{X}_m$, observed with absolute frequencies h_m, and I denotes the indicator function. Analogously, $\hat{s}_j$ is defined. The sets of predicted states can be updated, taking into account the different estimated error probabilities for the various items, as these can change the plausibility of the delineated knowledge states for the observed response patterns. This can be iterated yielding new estimates for the guessing and slipping probabilities. Having calculated the occurrence probabilities for the predicted states, the occurrence probabilities of the skill patterns can be derived given the bijective skill-item assignment.

3.2 CDMs

Parameter estimation in CDMs is performed maximizing the marginal likelihood of the data over $\beta = (g_1, \ldots, g_J, s_1, \ldots, s_J)$ (de la Torre, 2009):

$$L(\mathbf{X}) = \prod_{n=1}^{N} L(\mathbf{X}_n) = \prod_{n=1}^{N}\sum_{l=1}^{L} L(\mathbf{X}_n|\alpha_l)p(\alpha_l),$$

where $L(\mathbf{X}_n)$ is the marginal likelihood of the response pattern of examinee n, $p(\alpha_l)$ is the prior (uniform) probability of the skill pattern α_l, and $L = 2^K$. The estimation routine can be implemented using the EM algorithm. The (posterior) occurrence probabilities of the skill patterns are calculated using Bayes' theorem.

4 Simulation Study

For investigating the estimators in KST and CDMs the following steps are performed. Data for a number of items with known slipping and guessing rates are simulated using the R (R Development Core Team, 2010) computing environment. Delineated knowledge states corresponding to a given set of skills are determined and contaminated by slipping and guessing errors. The skill patterns are taken as uniformly distributed, and therefore their occurrence probabilities are $1/2^K$, and the population prevalences of the skills are $1/2$. For the error rates, occurrence probabilities and population prevalences, the bias is taken as a measure between the estimated and the true parameters. The simulation is performed for several parameter settings. Preliminary simulations with $J = 11$ items, $K = 4$ skills, $N = 2000$ examinees and error probabilities varying between 0.01 and 0.30 yield the following results: In KST models, one-step estimates of the response error probabilities, occurrence probabilities of skill patterns and population prevalences of individual skills are of the same magnitude as the true parameter values, with moderate differences for any of the parameter settings. Compared to CDMs, the results obtained for the KST models are neither better nor worse. Some systematic bias is found for the one-step KST estimates, which can be explained by the underlying parameter settings and possibly be reduced by further iterations. In contrast, the bias in the estimates for CDMs seems to be more item specific.
The current simulations are a starting point for more in-depth analyses. Future research may address the effects of variation of sample size (especially small sample sizes), non-uniform prior distributions of skill patterns, and the effects of misspecification of the assignment of skills to items.

References

de la Torre, J. (2009). DINA model parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, **34**, 115-130.

Doignon, J.-P., and Falmagne, J.-Cl. (1999). *Knowledge Spaces* . Berlin: Springer.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rupp, A.A., Templin, J., and Henson, R.A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: The Guilford Press.

Schrepp, M. (1999). Extracting knowledge structures from observed data. *British Journal of Mathematical and Statistical Psychology*, **52**, 213-224.

Forecasting film revenues using GAMLSS

Robert Gilchrist[13], Robert Rigby[1], John Sedgwick[2], Mikis Stasinopoulos[1], Vlasios Voudouris[2]

[1] STORM, London Metropolitan University, Holloway Road, London N7 8DB
[2] Centre for International Business and Sustainability, London Metropolitan Business School
[3] Communicating author: r.gilchrist@londonmet.ac.uk

Abstract: This paper utilises the GAMLSS framework for the statistical modelling of movie box-office revenues. The dominant modelling paradigm of the film industry, traditionally exemplified by the *nobody knows anything principle* is based upon the infinite variance of the Pareto distribution. We here use GAMLSS to show that total box-office revenue can be better modelled by distributions with finite variance contradicting the Paretian hypothesis. Moreover the paper illustrates that the Box-Cox power exponential distribution gives models where the parameters vary smoothly with an important explanatory variable, namely the opening box-office revenue, leading to the substantive conclusion that the post-opening revenue can be explained by the opening box-office revenue.

Keywords: GAMLSS; movies; Pareto distribution; BCPE distribution, semi-parametric regression.

1 Introduction

Film revenues are highly skewed, in such a way that a small number of large revenue films coexist alongside considerably greater numbers of smaller revenue films. Moreover, the skewed nature of these distributions appears to be an empirical regularity, with Pokorny and Sedgwick (2010) dating this phenomenon back to at least the 1930s, making it an early example of a mass market long tail. De Vany and Walls (2004) comment on the consequential difficulty in modelling the dispersion, skewness and kurtosis of film revenues. We here overcome this difficulty using the GAMLSS (Generalized Additive Models for Location Scale and Shape) framework developed in Rigby and Stasinopoulos (2005), using a dataset of box-office revenues in the 1930s.

Our initial approach is to compare many competing models for the total box-office revenue and specifically to compare these with models based on the Pareto-Levy (or stable Paretian or L-stable) distribution, which has dominated the modelling of end-of-run box-office revenues since De Vany and Walls (1996).

The paper compares distributions that best fit the *post-opening box-office revenue*, conditional on the *opening box-office revenue*. The parameters of the distributions are modelled as smooth non-parametric functions of the *opening box-office revenue*; the latter accounts for 32.2% of the end-of-run box-office cumulative revenue. The model can be used in planning post-opening film distribution.

2 The GAMLSS methodology

GAMLSS provides a very general and flexible system for modelling a response variable. The distribution of the response variable is selected by the user from a very wide range of available distributions including highly skewed and kurtotic continuous and discrete distributions. GAMLSS includes distributions with up to four parameters, denoted by μ, σ, ν and τ, which usually represent the location (e.g. mean), scale (e.g. standard deviation), and skewness and kurtosis shape parameters, respectively. All the parameters of the response variable distribution can be modelled using parametric and/or nonparametric smooth functions of explanatory variables, thus allowing modelling of the location, scale and shape parameters. Specifically, a GAMLSS model assumes that, for $i = 1, 2, \ldots, n$, independent observations Y_i have probability (density) function $f_Y(y_i|\theta^i)$ conditional on $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$ a vector of four distribution parameters, each of which can be a function to the explanatory variables. Rigby and Stasinopoulos (2005) define an original formulation of a GAMLSS model as follows. For $k = 1, 2, 3, 4$, let $g_k(.)$ be a known monotonic link function relating the distribution parameter θ_k to predictor η_k. Then we set

$$g_k(\theta_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(x_{jk}), \tag{1}$$

where h_{jk} is a smooth nonparametric function of variable x_{jk}.

3 Description of the film data

Figures 1 (a) and (b) (the plots in the first row) plot the total end-of-run box-office revenues. Figures 1 (c) and (d) (the plots in the second row) plot the post-opening box-office revenue (= total end-of-run box-office revenue minus opening box-office revenue). These univariate and bivariate exploratory plots give an indication of the complexity of the data in terms of skewness and kurtosis. The first two plots, boxplot and histogram, show the extreme long-right tail frequency distribution of the end-of-run box-office revenues. The third plot shows the post-opening box-office revenues plotted against the opening box-office revenues and a nonparametric curve

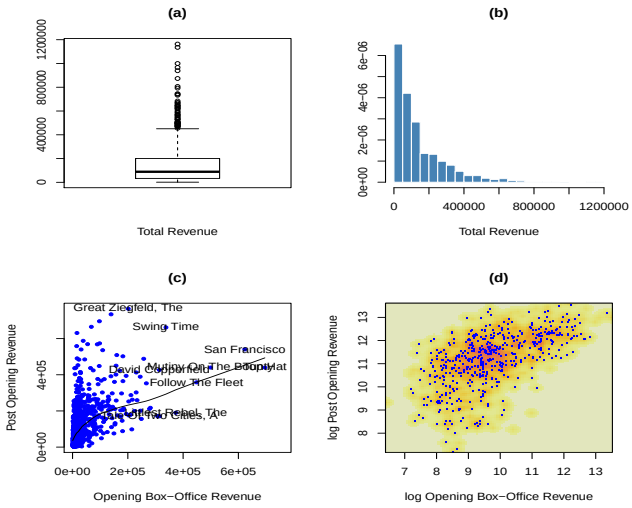


FIGURE 1. a) and (b) total revenue (c) and (d) post opening revenues against opening revenue

using a locally-weighted polynomial regression to aid interpretation. The fourth plot shows the log of post-opening box-office revenues plotted against the log of opening box-office revenues using a two-dimensional histogram smoothing to aid the interpretation of the distribution in dense areas.

4 Model selection strategy

We first analyse the total end-of-run box-office revenues. as given in Figures 1(a) and (b). Given the skewness in this response variable, an initial set of more than 20 distributions is selected for model fitting. The main criterion used is the Generalised Akaike Information Criterion (GAIC), with residual plots and worm plots to add confidence to our selection. We fit the Pareto I and Pareto II distributions as approximations to the L-stable distribution (Mandelbrot, 1997).

5 Analysis of the end-of-run box-office revenues.

Table 1 shows a selected subset of distributions used to model the total end of run box-office revenues. The Pareto I fits poorly. The generalised inverse Gaussian, Weibull and gamma all have lower Schwartz Bayesian Criterion value, GAIC(log(n))=SBC, than the Pareto II distribution. It is clear that the Pareto assumption of infinite variance has little support.

A confirmation of our conclusion is shown in the worm plots (van Buuren and Fredriks, 2001) of Figure 2 where the worm plot (a detrended QQ-plot of the normalised residuals) is shown for selected distributions.

	df	GAIC($k = \log(n)$)
Generalised Inverse Gaussian	3	24919.12
Weibull	2	24938.77
Gamma	2	24942.10
Pareto II	2	24943.94
generalised beta type 2	4	24952.18
Box-Cox t	4	24960.92
Generalised Gamma	3	24981.75
Box-Cox Cole-Green	3	25218.75
Inverse Gaussian	2	25271.45
Pareto I	1	26355.43

TABLE 1. SBC for end-of-run box-office revenues.

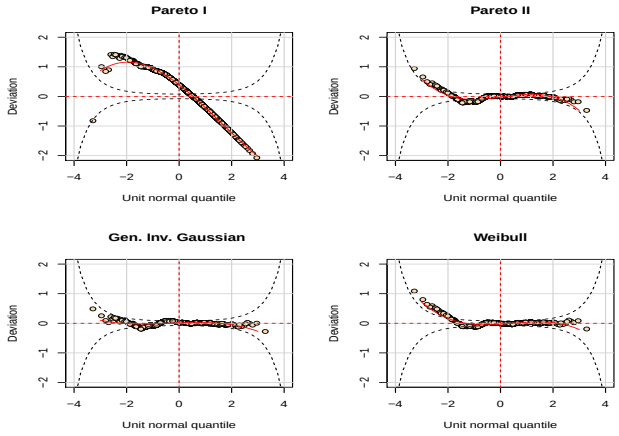


FIGURE 2. Worm plot of a) Pareto 1 b) Pareto 2 c) Generalised Inverse Gaussian, d) Weibull distributions

5.1 Regression-type of modelling of box-office revenues

Here, we model the log of the *post-opening box-office revenue* against an explanatory variable, namely the log of the *opening box-office revenue*. We use different distributions for $\log(Y)$, and we use smooth functions (Eilers and Marx, 1996) of the log of the opening box-office revenue for some or all of the parameters of the distributions. Table 2 shows the AIC for different

Distributions	df	AIC
Box Cox Cole & Green	7.55	2216.33
Box Cox power exp. (ν, τ)	14.30	2216.80
Box Cox power exp.	8.55	2218.32
Box Cox t	8.55	2218.33
Box Cox Cole & Green (ν)	8.54	2218.33
Box Cox t (ν)	10.35	2220.37
Box Cox power exp (ν)	10.35	2220.37
Box Cox t (ν, τ)	12.97	2222.77
Weibull type 3	6.98	2223.90
Generalised Gamma	8.28	2262.20
Generalised Gamma (ν)	9.51	2263.21
Normal	6.32	2267.54
Gamma	7.36	2303.62
Inverse Gaussian	7.34	2325.57

TABLE 2. AIC for analysis on the post-opening box-office revenue,.

fitted models. All models have smooth curves fitted for the location and scale parameters, μ and σ, respectively. The appearance of ν and τ in the table indicates whether or not those parameters have also been modelled using a smooth function of the log of the opening box-office revenue.

The best fitting model appears to be the Box Cox Cole and Green (BCCG) model where (only) μ and σ are modelled as smooth functions of the explanatory variable. A similar fit to this model is given by the Box Cox power exponential (BCPE) model (Rigby and Stasinopoulos, 2004) where all four parameters of this distribution (including ν and τ) are modelled as a function of the explanatory variable. Although not reported in detail here, the BCPE model also fits well to more recent film data from the 1990's. For consistency in comparing different epochs and also because of its flexibility we prefer the BCPE distribution model. The models shown in the first two rows of Table 2, namely the BCCG and BCPE models, show similar residuals and worm plots (not shown here due to space limitation). The data with fitted regression is shown in Figure 3, together with superimposed fitted probability density functions at specific values of the log opening box office revenue.

In conclusion, the model indicates that, given the opening box office income, we can make reasonable predictions of the post-opening box office revenue and, hence, of the total income. This leads us to the substantive conclusion that the *nobody knows anything* paradigm of the film industry is not correct. To the contrary, we can in fact predict the distribution of the total box office revenue of films.

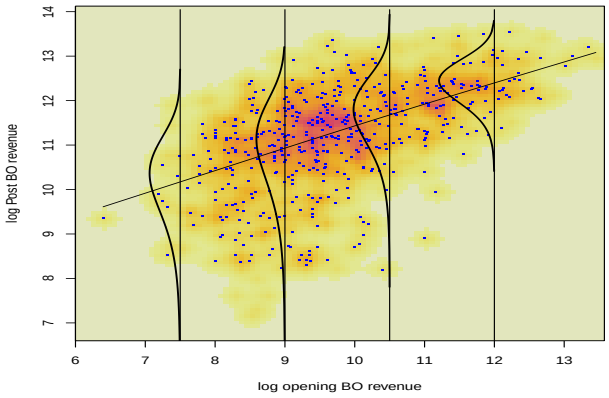


FIGURE 3. The fitted BCPE distribution to the 1930 Film data

References

De Vany, A.D. and and Walls, W.D. (1996) Bose-Einstein dynamics and adaptive contracting in the motion picture industry. *Economic Journal*, **106**, 1493-1514.

De Vany, A.D. and and Walls, W.D. (2004) Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar. *Journal of Economic Dymamics and Control*, **28**, 1035-1057.

Eilers. P.H.C. and Marx, B.H. (1996) Flexible smoothing with b-splines and penalties (with comments and rejoinder). *Statist. Sci*, **11**, 89-121.

Mandelbrot, B. (1997) *Fractals and scaling in finance: Discontinuity, concentration, risk*. New York: Springer-Verlag.

Pokorny, M. and and Sedgwick, J. (2010) Profitability trends in hollywood: 1929 to 1999: somebody must know something. *Economic History Review*, **63**, 56-84.

Rigby, R.A. and Stasinopoulos, D.M. (2004) Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Statistics in Medicine*, **23**, 3053-3076.

Rigby, R.A. and Stasinopoulos, D.M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507-554.

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

Importance of correctly specifying the random structure in growth mixture models

M S Gilthorpe[1], Y-K Tu[12], L D Kubzansky[3], E Goodman[4]

[1] Centre for Epidemiology and Biostatistics, School of Medicine, University of Leeds, LS2 9JT, UK, m.s.gilthorpe@leeds.ac.uk.
[2] Leeds Dental Institute, University of Leeds, Leeds, LS2 9UT, UK.
[3] Department of Society, Human Development and Health, Harvard School of Public Health, Boston, USA.
[4] Center for Child and Adolescent Health Policy, Mass. General Hospital for Children, Boston, USA.

Abstract: To understand developmental processes, health researchers increasingly explore patterns of trajectories in their outcomes using longitudinal data with multiple assessments of each study participant. Commonly used methods include latent growth curve modelling (LGCM) (Bollen and Curran 2006; Duncan et al. 2006) and its extension growth mixture modelling (GMM) (Kreuter and Muthen 2008). GMM aids interpretation if subgroups can be identified that have utility in subsequent analyses. Outcomes are typically modelled as parameterised (i.e. 'smooth') underlying trajectories. For a large part of the lifecourse an individual's growth often tracks this fitted trajectory well, with deviations due to variations in biological, behavioural or environmental factors. Due to similarities amongst successive measures a degree of autocorrelation is generally present, but a fitted smooth trajectory usually accounts for much of this. With GMM, if outcomes exhibit less within-subject than between-subject heterogeneity even greater autocorrelation may be generated as an artefact of the model because individual growth trajectories may then deviate consistently from class mean trajectories. This leads to model-generated autocorrelation amongst the residuals between subject-specific and class-mean trajectories. It is desirable to parameterise this explicitly, thereby capturing the correct underlying random structure of the data, but typically this is not done. The impact on models of not doing so therefore remains unclear.

Keywords: Autocorrelation, Growth Mixture Models, Latent Variable Methods.

1 Data and methods

We investigate model-generated autocorrelation for a growth outcome that typically exhibits greater within-person than between-person homogeneity: body mass index (BMI), a parameter of great interest in understanding normal growth and development, as well as the current worldwide obesity epidemic. BMI has already been considered within the GMM framework

(Goodman et al. 2003; Li et al. 2007; Mustillo et al. 2003; Needham et al. 2010). We used data from a school-based cohort of adolescents from the Cincinnati Ohio, US area with repeated measures of BMI over a 3-year period (Goodman, Adler, Daniels, Morrison, Slap, & Dolan 2003). The study began in the 2001-2002 school year and included students in grades 5-12 at baseline with three further annual waves of data collection. A physical exam measured height and weight. As the cohort was 95% non-Hispanic black and white, analyses were restricted to these two ethnic groups. Analyses focused on measured BMI, as opposed to age-sex standardized z-scores (Berkey and Colditz 2007), because the latter is based on data from studies including some with an almost exclusively non-Hispanic white population (Kuczmarski et al. 2000). We examine cohort trajectories rather than age-specific growth trajectories to reflect the structure of the data (students nested within measurement occasions). BMI trajectories were taken to be quadratic in (centred) time. Outcome variances were constrained to be identical across waves for each class trajectory (homoscedastic) and variances of linear and quadratic terms were constrained to be zero to attain parsimony and improve convergence. Growth trajectory intercepts were conditional on age at measurement, sex, age-sex interaction, and race. Covariate coefficients for each trajectory were constrained to be identical to ensure that parameterisation of underlying BMI growth curves were identical across classes. Trajectory slopes were conditional on age, accommodating trajectory differences in *change* in BMI by age during adolescence. With four repeated measures, autocorrelation was modelled as a 1st-order autoregressive structure. Contrasts focused on models with or without AR(1), the general form of which was:

$$BMI_{ti} = \sum_{c=1}^{C} P(c|age_{ti}, sex_i, race_i)\left(\beta_{0ti}^c + \beta_{1i}^c age_{ti} + \beta_2^c age_{ti}^2\right)$$

with BMI_{ti} at age_{ti} measured at time $t = 1, \ldots, 4$, for individual $i = 1, \ldots, 1528$; c is latent class ($c = 1, \ldots, C$), for which $P(c|age_{ti}, sex_i, race_i)$ is the probability that individual i is in class c, conditioned on age, sex and race; $\beta_{0ti}^c = \beta_0^c + e_{0ti}^c + \gamma_1 age_{ti} + \gamma_2 sex_i + \gamma_3 (age \cdot sex)_{ti} + \gamma_4 race_i$ is the random intercept for class c, conditioned on age, sex, age-sex interaction and race identically across all C classes; $\beta_{1i}^c = \beta_1^c + \gamma_5 age_{ti}$ is the slope for class c, conditioned on age identically across all C classes; β_0^c, β_1^c and β_2^c are the class-dependent marginal mean intercept, slope and acceleration, respectively; $e_{0ti}^c \sim N\left(0, \sigma_{(c)e_0}^2\right)$ is the class-dependent occasion-specific normal residual with zero mean and variance $\sigma_{(c)e_0}^2$, estimated empirically; and γ_m $(m = 1, \ldots, 5)$ are class-independent covariate trajectory coefficients describing the underlying population mean growth for intercept and slope respectively. For models with AR(1) the constraint $Corr(e_{(t)i}, e_{(t+1)i}) = \rho$ $(t = 1, \ldots, 3)$ applies identically across all C classes, else $Corr(e_{pi}, e_{qi}) \equiv 0$ $(\forall p \neq q)$.

Within the Mplus (v6) software we derive BMI trajectories with and without autocorrelation modelled as AR(1) to establish to what degree this affects: (i) model convergence and model-fit, assessed by the BIC; (ii) class size and composition; and (iii) class trajectory variance structure. Since the risk of models converging to local minima increases with increasing number of classes, models were run for 20k random starts, from which the best 10% were used to derive model estimates. The number of classes examined ranged from two to eleven.

2 Results

Nearly all random starts converged for models with no AR(1) structure, though the proportion of the best 10% that settled on the same maximum likelihood (ML) value varied. Amongst models with an AR(1) structure, only 20% of random starts converged, indicating a much smaller solution space for models with the AR(1) parameterisation. Amongst the best 10%, consistency in the optimum ML again varied, but was less than for models without an AR(1) structure. According to the BIC, models with AR(1) consistently fitted better; BIC attained a plateaux around 10 or 11 classes for models without AR(1) and a minimum at 6 classes with AR(1). Under the assumption that relative class sizes are similar (i.e. classes ranked by size corresponded to similar classes across both model types), probabilistically assigned correspondence ranged from 54.1% for the 2-class to 18.7% for the 11-class model, and modally assigned correspondence ranged from 90.7% for the 2-class model to 20.5% for the 9-class model. For models with 3 or more classes there was net 'drift' of membership from smaller to larger classes when AR(1) was incorporated. Intercept residual variances amongst class trajectories with AR(1) was three times greater than amongst models with no AR(1), indicating that individual trajectories had a greater range of intercepts when autocorrelation is modelled than when not. In contrast, residual variances across slopes amongst models with an AR(1) structure were up to five times smaller than amongst models with no autocorrelation, suggesting that individual trajectories had a narrower range of slopes when AR(1) is modelled. Overall, class composition differed depending on whether an AR(1) structure was modelled or not.

3 Discussion

Autocorrelation in these data was mainly model-generated due to variation between individuals within classes at any time point being more marked than variation in individual BMI trajectories over time. BMI typically exhibits less individual than population heterogeneity throughout the life-course, even though this may vary for key growth periods, such as the first few years of life and puberty. Misspecification of the random structure

impacts upon subject classification more than the model's fixed effects, as these are simply the average of individual fitted curves whilst subject classification is based on individual curves. Subject classification is key to the utility of GMMs; models that capture model-generated autocorrelation within the GMM framework are thus preferred. Whilst the exact choice of parameterization remains open, our findings suggest that some kind of explicit modelling of autocorrelation is warranted in these types of models. In any event, correct parameterization of the random structure is needed for growth mixture modelling of outcomes that exhibit less within-subject than between-subject heterogeneity.

Berkey, C.S. & Colditz, G.A. (2007). Adiposity in adolescents: change in actual BMI works better than change in BMI z score for longitudinal studies. *Ann. Epidemiol.*, **17**(1), 44-50.

Bollen, K. & Curran, P. (2006). *Latent curve models*, 2nd ed. New York, Wiley.

Duncan, T.E., Duncan, S.E., & Stryker, L.A. (2006). *An introduction to latent variable growth curve modeling*, 2nd ed. Mahwah, NJ, Laurence Erlbaum Associates Inc.

Goodman, E., Adler, N.E., Daniels, S.R., Morrison, J.A., Slap, G.B., & Dolan, L.M. (2003). Impact of objective and subjective social status on obesity in a biracial cohort of adolescents. *Obes. Res.*, **11**(8), 1018-1026.

Kreuter, F. & Muthen, B. (2008). Analyzing Criminal Trajectory Profiles: Bridging Multilevel and Group-based Approaches Using Growth Mixture Modeling. *Journal of Quantitative Criminology*, **24**(1), 1-31.

Kuczmarski, R.J., Ogden, C.L., Grummer-Strawn, L.M., Flegal, K.M., Guo, S.S., Wei, R., Mei, Z., Curtin, L.R., Roche, A.F., & Johnson, C.L. (2000). CDC growth charts: United States. *Adv. Data.* **314**, 1-27.

Li, C., Goran, M.I., Kaur, H., Nollen, N., & Ahluwalia, J.S. (2007). Developmental trajectories of overweight during childhood: role of early life factors. *Obesity. (Silver. Spring)*, **15**(3), 760-771.

Mustillo, S., Worthman, C., Erkanli, A., Keeler, G., Angold, A., & Costello, E.J. (2003). Obesity and psychiatric disorder: developmental trajectories. *Pediatrics*, **111**(4-1), 851-859.

Needham, B.L., Epel, E.S., Adler, N.E., & Kiefe, C. (2010). Trajectories of change in obesity and symptoms of depression: the CARDIA study. *Am. J. Public Health*, **100**(6), 1040-1046.

Modeling swimming marks through Blocks and POT methods

Dulce Gomes[1], Júlia Teles[2], Luísa Canto e Castro[3]

[1] Departamento de Matemática/CIMA, Escola de Ciências e Tecnologia, Universidade de Évora, Rua Romão Ramalho, 59, 7000-671 Évora, Portugal (email: dmog@uevora.pt)
[2] Departamento de Métodos Matemáticos/CIPER, Faculdade de Motricidade Humana, Universidade Técnica de Lisboa, Estrada da Costa, 1495-688 Cruz Quebrada-Dafundo, Portugal (email: jteles@fmh.utl.pt)
[3] Departamento de Estatística e Investigação Operacional/CEAUL, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Edifício C6, 1749-016 Lisboa, Portugal (email: luisa.loura@deio.fc.pt)

Abstract: The swimming marks in the 100m men's freestyle long course are modelled using extreme value theory. Using the statistical package R, extreme value and generalized Pareto models were adjusted in order to estimate the left endpoint of these models. The left endpoint can be interpreted as the best mark that can ever be reached, admitting that swimming pool conditions, athlete's equipment and training methods remain the same.

Keywords: Extreme value models; Generalized Pareto models; Blocks method; Peaks over threshold method; Swimming marks.

1 Introduction

Extreme value models are frequently used for the analysis of samples of maximum or minimum and generalized Pareto models are commonly used to analysing the samples of exceedances over a high or low threshold.
The swimming marks of 100m men's freestyle (long course), that appear in FINA ("La Federation Internationale de Natation") Website (www.fina.org), are the personal best in a very large sample of marks, so we could say that we are in the presence of extreme value — in this case minima. In this sense, we consider analysing and modeling this type of data by means of extreme value models.
Using two methods of extreme value analysis — the blocks method (De Haan and Ferreira, 2001) and the peaks over threshold (POT) method (Pickands, 1975; Robinson and Tawn, 1995) — we are going to model the swimming marks and estimate their left endpoint. This left endpoint can be interpreted as the best mark that can ever be reached, admitting that swimming pool conditions, athlete's equipment and methods of training remain the same.

TABLE 1. The three best annual marks (in seconds), through 1948 to 2010, of Men's Long Course World Records in 100m freestyle.

year	rank	mark	athlete	nationality
1948	1	57.3	Wally Ris	USA
1948	2	57.6	Keith Carter	USA
1948	3	57.8	Alan Ford	USA
...	...	...	...	...
1999	1	48.35	Pieter van den Hoogenband	NED
1999	2	48.73	Michael Klim	AUS
1999	3	48.82	Alexander Popov	RUS
2000	1	47.84	Pieter van den Hoogenband	NED
2000	2	48.18	Michael Klim	AUS
2000	3	48.27	Alexander Popov	RUS
...	...	...	...	...
2010	1	48.54	Simon Burnett	GBR
2010	2	48.56	William Meynard	FRA
2010	3	48.69	Kyle Richardson	AUS

A draft of the dataset with the three best annual marks (in seconds) of the men's long course world records in 100m freestyle are presented in the Table 1. The information was available from 1948 to 2010, with missing value for 1950 and 1951.
In Figure 1, the marks are plotted against the year and by ranking. As we expected, there is a decreasing trend in the marks. So, the relevant question is: Until when these marks could fall? In order to give answer to this question two different approaches of extreme value theory were used.

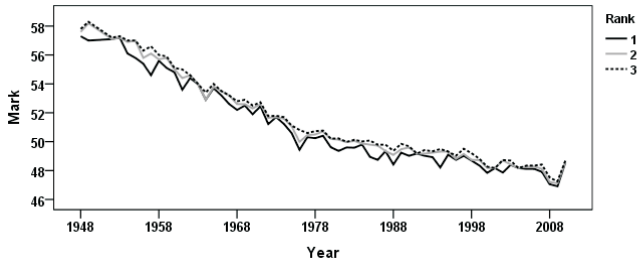


FIGURE 1. Scatter plot of mark against year by ranking.

2 Two different approaches

To apply the extreme value methodology we must have independent and identically distributed observations (Reiss and Thomas, 2001). As we can observe in Table 1, there are some athletes (e.g., Michael Klim and Alexander Popov) which contributed with more than one mark. So to use this type

of analysis we only select the best mark of each athlete. In order to adjust an extreme value or generalized Pareto model, the trend also needs to be removed.
In the POT approach the inference is based in the exceedances over a high threshold that is unknown. Our empirical way of choosing this threshold was through the analysis of the diagram of the shape parameter's estimates. To apply the POT method we adjust a model in the family of generalized Pareto models. For different shape, location and scale parameters we obtain three different submodel families: exponential, Pareto and beta.
In the block method we adjust a model in the family of extreme value models. Also depending on the shape, location and scale parameters we could reach the Gumbel, Fréchet or Weibull submodels.

Acknowledgments: This research was partially support by the Center of Mathematics and Applications, University of Évora, by the Interdisciplinary Centre for the Study of Human Performance, Technical University of Lisbon, and by the Center of Statistics and Applications, University of Lisbon, through the Programs FCT/POCTI, FCT/POCI2010 and POCI/FEDER.

References

De Haan, L., and Ferreira, A. (2006). Extreme Value Theory: An Introduction. Boston: Springer.

Pickands, J. (1995). Statistical inference using extreme value order statistics. *Annals of Statistics*, **3**, 119-131.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria. (Available at http://www.R-project.org/).

Reiss, R.D., and Thomas, M. (2001). *Statistical Analysis of Extreme Values*, 2nd edition. Basel: Birkhäuser.

Robinson, M.E., and Tawn, J.A. (1995). Statistics for expected athletics records. *Applied Statistics*, **44**, 499-511.

Improvement of surface water quality variables modelling that incorporates a hydro-meteorological factor: a state-space approach

A. Manuela Gonçalves[1], Marco Costa[2]

[1] Department of Mathematics and Applications, University of Minho
CMAT-Center of Mathematics, Portugal
mneves@math.uminho.pt

[2] Higher School of Technology and Management of Águeda-University of Aveiro
CMAF-UL, Portugal
marco@ua.pt

Abstract: In this work it is constructed a hydro-meteorological factor to improve the adjustment of statistical time series models, such as state space models, of water quality variables by observing hydrological series (recorded in time and space) in a River basin. The hydro-meteorological factor is incorporated as a covariate in multivariate state space models fitted to homogeneous groups of monitoring sites. Additionally, in the modelling process it is considered a latent variable that allows incorporating a structural component, such as seasonality, in a dynamic way.

Keywords: hydrological basin; water quality, state-space modelling; Kalman filter; hydro-meteorological factor.

1 Introdution

Water quality monitoring is an important tool in the management and assessment of surface water quality. This study focuses on a rather extended data set relative to the River Ave basin (Portugal) and consists mainly of monthly measurements of biochemical variables in a network of monitoring water quality stations.A hydro-meteorological factor is constructed for each monitoring station based on monthly estimates of precipitation obtained by means of a rain gauge network. Through stochastic interpolation (Kriging) it is estimated the mean area rainfall during each month in the area of influence of each water quality monitoring site. These estimates are based on rain gauges located in the respective area of influence. In a recent work, Costa and Gonçalves (2010) show that a set of water quality monitoring sites can be modelled applying cluster techniques that minimize the number of models.

2 Data Set Description

The Northern Regional Directory for the Environment and Natural Resources (DRARN) and the National Institute of Water (INAG) has been collecting various water quality variables (monthly physical-chemical and microbiological analyses) from 16 quality monitoring sites. The data set of the 16 water quality monitoring sites, comprising 11 water quality variables, have been monthly measured between 1988 and 2006. At this time, this work focuses on Dissolved Oxygen (DO) (mg/l) in water because it is one of the most important variables in the evaluation on river water quality. For instance, it is shown the data and the results of one cluster with five water monitoring sites identified in Costa and Gonçalves (2010) as the less polluted cluster.

3 Methods

As starting point, it is constructed a hydro-meteorological factor used as covariate in the modelling process. This covariate will integrate a hydro-meteorological component that is recognized as crucial in any water quality modelling process. This factor is constructed through stochastic interpolation (Kriging) based on an udometric network (Figure 1) with 19 meteorological stations. The model of spatial continuity, which is inferred from monthly precipitation estimates, assumes hypothesis of homogeneity of the process: the process is stationary of 2nd, i.e., intrinsically stationary and isotropic. Under this hypothesis, two observations in the same location but in different times are independent and the spatial variability pattern remains the same (Kyriakidis and Journel, 1999). The empirical semivariogram is given by

$$\hat{\gamma}_z(h \mid l) = \frac{1}{2T|N(h|l)|} \sum_{t=1}^{T} \sum_{(i,j)\in N(h|l)} [(Z_t(s_i) - Z_t(s_j)]^2$$

with $N(h|l) = \{(i,j) : \|s_i - s_j\| - \|h\| \leq l; 1 \leq i \leq j \leq n\}$ and $|N(h|l)| = \#N(h|l)$. The river basin is discretized in 368 points with $2Km$ x $2Km$ (Figure 1) and at each point s_0 the estimate of the monthly mean area precipitation is given by the Kriging estimator, i.e., by a linear combination of the 19 known points s_j, $j = 1, ..., 19$ and $Z_t(s_0) = \sum_{j=1}^{19} \lambda_j Z_t(s_j)$.

3.1 Hydro-meteorological factor

It is constructed one covariate for each water monitoring site based on the estimate of the monthly mean precipitation of its influence region. In this

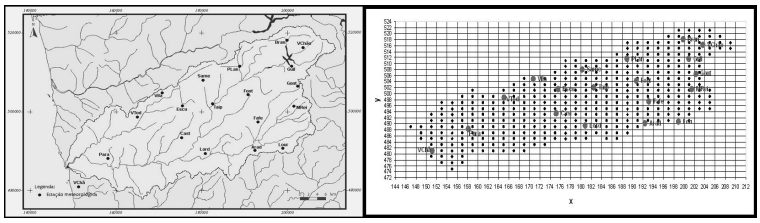


FIGURE 1. Spatial distribution of 19 meteorological monitoring sites in the River Ave basin and discretization of River Ave basin in 368 points.

context, the influence regions of each water monitoring site were defined by technicians of the INAG and they are supported on the region's topography and the land's drainage dynamics. Firstly, for each water monitoring site, it was computed the monthly mean area precipitation in its influence region based on the average of point prediction. Naturally, a large influence region tends to have a greater precipitation amount. Indeed, it is clear that the precipitation amount influences oxygen concentration in water. However, if the goal of this work is to found a prediction model to DO in a month t, the covariate should not incorporate the precipitation amount of the current month, but only the past information. Let $P_t^{(i)}$ be the estimate of the precipitation amount in the influence area of a water monitoring site i at month t. We considered a covariate $H_t^{(i)}$ computed as a weighted average of precipitation amount at months $t-1$ and $t-2$.

3.2 State space model

For each cluster i with homogenous water monitoring site it is fitted a state space model to Dissolved Oxygen concentration incorporating two structural components: the hydro-meteorological factor and a seasonality. In order to simplify, it is considered monthly seasonality assuming 12 known coefficients (for each month it is taken the month mean; Costa and Gonçalves, 2010):

$$\begin{pmatrix} Y_{1,t}^{(i)} \\ Y_{2,t}^{(i)} \\ \vdots \\ Y_{m,t}^{(i)} \end{pmatrix} = \begin{pmatrix} S_t & H_{1,t}^{(i)} \\ S_t & H_{2,t}^{(i)} \\ \vdots & \vdots \\ S_t & H_{m,t}^{(i)} \end{pmatrix} \begin{pmatrix} X_{1,t}^{(i)} \\ X_{2,t}^{(i)} \end{pmatrix} + \begin{pmatrix} \mu_{1,t}^{(i)} \\ \mu_{2,t}^{(i)} \\ \vdots \\ \mu_{m,t}^{(i)} \end{pmatrix},$$

$$\begin{pmatrix} X_{1,t}^{(i)} \\ X_{2,t}^{(i)} \end{pmatrix} = \begin{pmatrix} 1 \\ \mu_{X_2}^{(i)} \end{pmatrix} + \begin{pmatrix} \phi_{11}^{(i)} & \phi_{12}^{(i)} \\ \phi_{21}^{(i)} & \phi_{22}^{(i)} \end{pmatrix} \left[\begin{pmatrix} X_{1,t-1}^{(i)} \\ X_{2,t-1}^{(i)} \end{pmatrix} - \begin{pmatrix} 1 \\ \mu_{X_2}^{(i)} \end{pmatrix} \right] + \\ + \begin{pmatrix} V_{1,t}^{(i)} \\ V_{2,t}^{(i)} \end{pmatrix}.$$

Since normal distribution is not always the best distribution to fit meteorological variables in this work, we adopted consistent distribution-free estimators developed from the original work by Costa and Alpuim (2010). The state space model with these parameters estimates associated to the

TABLE 1. Parameters estimates.

$\hat{\mu}_X$	$\hat{\phi}$		$\hat{\Sigma}_V$		$\hat{\Sigma}_\mu$					*Sites*
1	0.277	−1.045	0.016	−0.005	0.597	0.000	0.000	0.000	0.000	*CANT*
−0.0003	0.038	0.738	−0.005	0.003	0.000	0.265	0.000	0.000	0.000	*GOL*
					0.000	0.000	0.417	0.000	0.000	*FER*
					0.000	0.000	0.000	0.383	0.000	*VSA*
					0.000	0.000	0.000	0.000	0.737	*TAI*

Kalman filter produces monthly one-step predictions for Dissolved Oxygen concentration at each water monitoring site (Table 1). Figure 2 shows observed data and predictions in Vizela Santo Adrião (VSA) and Golães (GOL) monitoring sites.

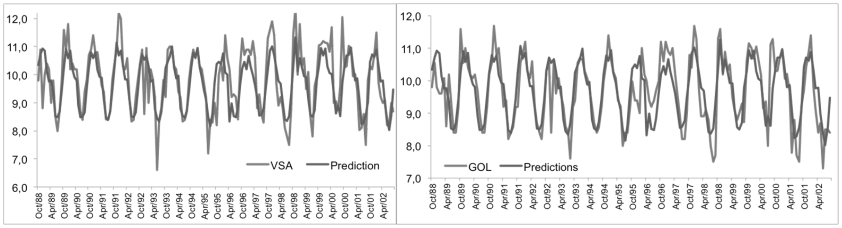


FIGURE 2. Observed and one-step predictions of Dissolved Oxygen concentration in Vizela Santo Adrião (VSA) and in Golães (GOL).

4 Conclusions

It is possible to conclude that the hydro-meteorological factor is an important component adding information beyond the usual seasonality. Moreover, the adoption of the consistent distribution-free estimators for the state space models requires a future comparison with gaussian likelihood estimation, assessing its relative efficiency, and possibly comparing its forecasts mean square error. However, distribution-free estimators are an easy solution without computed problems, nor iterative procedures and neither requires initial values. The next step is to analyse the filtered estimates of states $X_{t|t}^{(i)}$ given by the Kalman filter, which allows an interesting analysis of these latent variables as calibrate factors of the two structural components.

References

Bengtsson, T., Cavanaugh, J. (2008). State-space discrimination and clustering of atmospheric time series data based on Kullback information measures. *Environmetrics*, **10**, 377-394.

Costa, M., Alpuim, T. (2010). Parameter estimation of state space models for univariate observations. *J Stat Plan and Inference*, 140**(7)**, 1889-1902.

Costa M., Gonçalves, A. M. (2010). Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stoch Environ Res Risk Assess*, DOI: 10.1007/s00477-010-0429-5.

Kyriakidis, P.C., Journel, A.G. (1999). Geostatistical space-time models: a review. *Mathematical Geology*, 31**(6)**, 651-684.

Modelling fertility and education in Italy in the presence of time-varying frailty component

Anna Gottard[1], Alessandra Mattei[1], Daniele Vignoli[1]

[1] Department of Statistics of the University of Florence, Viale Morgagni 59, 50134 Firenze, Italy. gottard@ds.unifni.it

Abstract: In this paper, a Bayesian survival model is adapted to study the association between fertility and women education in Italy. The effect of women educational level is studied in the presence of a time-varying unobserved component. This kind of frailty can be interpreted as women's family-orientation, which is sensible to vary over life-course.

Keywords: Bayesian survival analysis; fertility; time-varying frailty.

1 Introduction

The association between fertility and educational achievement is one of the strongest relationships recorded in social science. There is a large agreement among scholars that the level of education represents a pivotal factor that drives differences in fertility choices both in developing and developed countries. Education is in fact a potent marker of individuals' labour market performance and prospects, earnings potential, and social status. For women, higher education also underlines the possibility to behave in autonomy of the male partner and of social norms (Hoem et al. 2001).
The influence of education on fertility developments is essentially ascribable to demographic and socio-economic reasons. From a demographic perspective, women who decide to continue education at higher levels tend to postpone the transition to motherhood, which may have consequences on completed fertility because of the potential room, or lack thereof, that is left for second- or higher-order births. Moreover, delaying the entry into motherhood may in some cases lead to involuntary childlessness. Those women who wish to have more than one child are therefore under a time squeeze and they need to progress to second childbearing more quickly than those who had their first child early in life (Kreyenfeld, 2002). From a socio-economic perspective, child-related career breaks imply income lost due to non-participation and depreciation of human capital as well as lost opportunities for promotion, work-career and independent life.

Overall, empirical studies have shown that the direction of the effect of education on fertility depends on women's parity-specific status: better educated women have lower first birth intensities, even after the time spent in education is taken into account (e.g., Matysiak and Vignoli 2009 for Italy and Poland), whereas the effect of education on second order fertility is found to be positive in many European countries (e.g., Kreyenfeld 2002 for West Germany; Kravdal 2001 for Norway).
Kravdal (2001) and Kreyenfeld (2002) strongly contributed to this debate suggesting the existence of a self-selection effect. They anticipated that some women with tertiary education who gave birth to the first child have a remarkable, unobserved preference for children. Following the methodological framework proposed by Lillard and Panis (2003), they tested this hypothesis adapting a simultaneous-equations survival model that jointly estimates the time-to-event for the first and the second child birth, including a time-constant shared frailty term $U_i \sim N(0, \tau^2)$, shared by both the two possible events for each woman. Controlling for this unobserved component, that they interpreted as women's family-orientation, the significant and positive effect of education on second birth risk vanished.
A possible limit of Kravdal and Kreyenfeld's approach is that they considered family orientation constant over time, using a time-invariant individual level unobserved-heterogeneity component in modelling first and second birth transitions.
The objective of this work is twofold. The role of educational attainment for fertility of Italian women will firstly be explored in the presence of time-invariant heterogeneity. This model can describe a persistent family orientation over the life-course. Secondly, the hypothesis of time-constant heterogeneity will be relaxed to account for possible changes in family orientation during the life-course.

2 Model and data description

The role of educational attainment for fertility of Italian women is here based on retrospective data, stemming from the Household Multipurpose Survey Family and Social Subjects (FSS). The FSS survey was conducted by the Italian National Statistical Office (Istat) in November 2003 on a sample of about 24,000 households and 49,451 individuals of all ages. We selected women aged 20-45 at the time of the interview. Fertility has been measured by means of time to first and second child birth for each woman. Education, together with area of residence, cohort and parents' educational level have been included as explanatory variables. The final sample includes 9,029 women (i.e., cohorts 1958-1983). Time to the interview represents an exogenously fixed censoring time.
Survival models are an ideal framework for studying event occurrence and for modelling the relationship between the risk of an event occurrence and

selected predictors. The interest can be therefore focused on the associate point process $X(t)$, with t representing the time-to-event, $t \in (0, T_c]$, the time origin corresponding to 14 years old age and T_c to age at the interview. Specifically, the fertility process in analysis here admits two kinds of event, the first and the second child birth. Such a process can be viewed as a marked point process $X(t, m)$ (Arjas, 1989), in which the mark $m \in \mathcal{M} = \{1, 2\}$ indicates the kind of event occurred. Notice that the two kinds of event are not competing, but consecutive, as the second child cannot be born before the first child.

The complete description of the finite-dimensional distribution of this kind of process can be formulated in terms of its mark-specific hazard function $h_m(t)$, the instantaneous rate of having in t the m^{th} child. Similarly, the mark-specific survival function can be then specified as

$$S_m(t) = \exp\{-\int_0^t h_m(s)ds\}.$$

A set of explanatory variables can be included by defining a conditional version of the mark-specific hazard function. The likelihood function for the considered fertility process is then

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{m=1}^{2} h_m(t_{im} \mid Z_i, \mathcal{H}(t_{im}^-))^{\delta_{im}} \cdot S_m(t_{im} \mid Z_i, \mathcal{H}(t_{im}^-))^{\zeta_{im}}$$

in which t_{im} represents, occurrence or censoring time for woman i for event m, Z_i is the vector of observed explanatory variables, $\mathcal{H}(t_{im}^-)$ represent the past history of the process and δ_{im} and ζ_{im} are adequately to deal with censured events.

In this work, we assume a parametric model assumed for the fertility process, with a piecewise-constant specification. The conditional mark-specific hazard function has the form

$$h_m(t_{im} \mid Z_i, \mathcal{H}(t_{im}^-)) = \sum_{k=1}^{K} (\lambda_{km} \cdot \mu_{im}) \cdot 1_{\{t_{k-1} < t \leq t_k\}}$$

in which $\log(\mu_{im})$ is assumed as a linear function of the explanatory variables and past history of the process, not depending on k. Here $K = 6$ and $t_K = T_c$.

A time-constant frailty component can be inserted similarly, as

$$\log(\mu_{im}) = \sum_{j=1}^{J} \beta_{jm} Z_{ij} + U_i.$$

with, typically, $U_i \sim N(0, \tau^2)$ and J representing the number of included explanatory variables. This kind of model specification assumes that the

unobserved heterogeneity representing family orientation is constant over time. A time-varying random effect can be viewed as random slopes of time-varying dummy variables. Particularly, in this research, we are assuming a piecewise constant random effects, supposing three fixed time intervals.

$$\log(\mu_{im}) = \sum_{j=1}^{J} \beta_{jm} Z_{ij} + U_{1i}\mathcal{I}_1 + U_{2i}\mathcal{I}_2 + U_{3i}\mathcal{I}_3, \tag{1}$$

where each $\mathcal{I}_r$ equals 1 in the r^{th} time interval, $r = 1, 2, 3$.
Because of the particular application we have in mind, it seems sensible to assume the three random effects to be dependent, so that for example,

$$U_{2i} = \delta_{12}\, U_{1i} + \varepsilon_{2i}.$$

Equivalently, with $\rho_{rs} = \delta_{rs}\tau_r/\tau_s$,

$$\begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 & \rho_{13}\tau_1\tau_3 \\ \rho_{12}\tau_1\tau_2 & \tau_2^2 & \rho_{23}\tau_2\tau_3 \\ \rho_{13}\tau_1\tau_3 & \rho_{23}\tau_2\tau_3 & \tau_3^2 \end{pmatrix}\right) \tag{2}$$

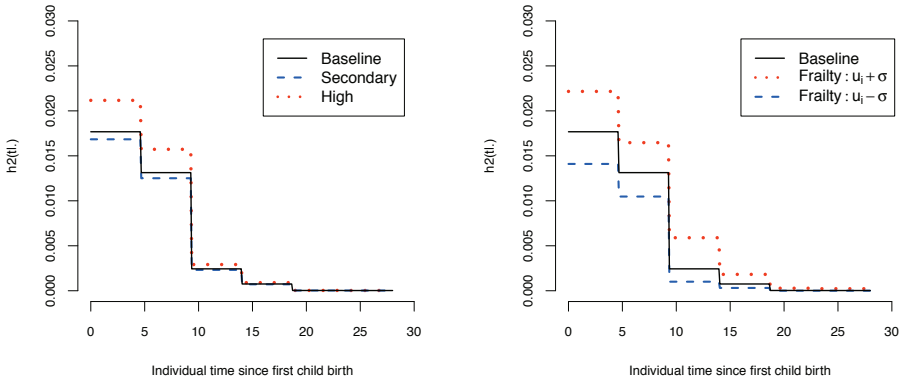
Whenever ρ_{rs} is positive, the individual (unobserved) hazard functions will be more heterogeneous when passing to period s. On the contrary, the hazard functions will be more similar in the s^{th} period, if ρ_{rs} is negative. Calling ρ^{13} the element $(1, 3)$ in the inverse of the variance covariance matrix in (2), whenever $\rho^{13} = 0$, then U_3 is independent of U_1 given U_2, suggesting an AR(1) dependence model among the unobserved components.
To implement the Bayesian survival model (see, for example, Ibrahim et al., 2001), prior distributions for model parameters have been specified. To reflect a vague prior knowledge, we opted for non-informative, although proper, prior distributions. Particularly, denoting $\alpha_{km} = \log \lambda_{km}$ it has been assumed

$$\alpha_{km}|\alpha_{(k-1)m} \sim N\Big(\alpha_{(k-1)m}, \sigma_\alpha^2\Big) \qquad k = 1, \ldots 6, m = 1, 2$$

with $\alpha_{0m} = 0$ and $\sigma_\alpha^{-2} \sim$ Gamma $(0.01, 0.01)$. The inverse of the variance-covariance matrix in (2) for the vector of random effect has been assumed to have a Wishart distribution. Moreover, the coefficients of the explanatory variables in (1) are assumed as $\beta_{jm} \sim N(0, 100)$.
Posterior distributions have been then simulated by using a Markov chain Monte Carlo algorithm. The estimates are based on three chains of 80,000 Monte Carlo replications, after a burn-in stage of 20,000 replications.

3 Some results

At least two crucial findings do emerge from our study. First, the impact of education of fertility is negative on for the transition to the first child,



(a) (b)

FIGURE 1. Effect on the baseline hazard for the transition to second child of (a) education and (b) time-varying frailty.

and positive for the transition to the second child. Namely, our results illustrate that high educated women tend to postpone the birth of the first child, but to anticipate the birth of the second child with respect to low educated women. In Italy, therefore, higher educated women seem to delay the consideration of the *right time* to conceive the first child, which leaves them less time for second and higher order births. As a consequence, better educated women desiring more than one child need to progress to second childbearing more quickly than their least educated counterparts. The effect of education on the hazard function specific for the transition to the second child is reported in Figure 1(a). Point estimates depicted in the Figure are obtained as the mean value of the posterior distributions.

Second, controlling for a common unobserved time-constant or time-varying unobserved heterogeneity component in each fertility transition, the positive and highly significant impact of women's tertiary education on fertility decisions softens. The time-varying frailty component seems to better control for possible changes in women's family-orientation over time. These results suggest that the impact of women's tertiary education on Italian fertility development is at least partially driven by women orientation towards family formation, that is, women who plan to have a child will self-select themselves into family formation prior to childbearing irrespective of their education level. This interpretation applies particularly to the Italian situation, in which women who opt for motherhood are likely to have a high degree of family orientation or low career ambitions, given the unfriendly

institutional setting for balancing work and family life.
Figure 1(*b*) illustrate the effect of the time-varying unobserved heterogeneity component on the baseline hazard function specific for the transition to the second child. In particular, it depicts the baseline hazard function for an average woman (with a zero effect of the frailty component), together with the baseline hazard function for two women having the frailty one standard deviation above and below the average. The figure takes into account the dependence between the unobserved components. It can be seen as the impact of the frailty components seems wider than the education effect. Moreover, posterior distributions for the partial correlation coefficients between the frailty components suggest a persistence of the women's family-orientation over time.

References

Arjas, E. (1989). Survival models and martingale dynamics (with discussion). *Scandinavian Journal of Statistics*, **16**, 177-225.

Kravdal, O. (2001). The high fertility of college educated women in Norway: An artefact of the separate modelling of each parity transition. *Demographic Research*, **5(6)**, 185-216.

Kreyenfeld, M. (2002). Time-squeeze, partner effect or selfselection? An investigation into the positive effect of women's education on second birth risks in West Germany. *Demographic Research*, **7**, 15-48.

Hoem, J., Neyer, G., and Prskawetz, A. (2001). Autonomy or conservative adjustment? The effect of public policies and educational attainment on third births in Austria, 1975-96, *Population Studies*, 55 (2001), 249-261.

Ibrahim, J.G., Chen, M-H., and Sinha, S. (2001). *Bayesian Survival Analysis*. Springer Series in Statistics, Springer, New York.

Lillard, L., and Panis. C.W.A. (2003). aML Multilevel Multiprocess Statistical Software. Release 2.0. EconWare, Los Angeles, California.

Matysiak, A., and Vignoli, D. (2009). Finding the right moment for the first baby to come: A comparison between Italy and Poland. MPIDR Working Paper, 2009-011. Rostock, Germany.

Empirical Bayes models to estimate contextual effects

Laura Grisotto[12], Dolores Catelan[12], Marc Saez[3] , Annibale Biggeri[12]

[1] Department of Statistics "G. Parenti", Viale Morgagni 59, 50134, Florence, Italy
[2] Biostatistics Unit, ISPO Cancer Prevention and Research Institute,Via Cosimo il Vecchio 2, 50134, Florence, Italy
[3] Research Group on Statistics, Applied Economics and Health (GRECS), University of Girona and CIBER of Epidemiology and Public Health (CIBERESP), University of Girona, Campus de Montilivi, 17071 Girona, Spain

Abstract: The association between disease risk and socio-economic indicators such as material deprivation or education-based indexes is often investigated using ecological data. In this kind of analysis a contextual effect has been documented. We developed a series of empirical Bayes models to integrate aggregate data on a discrete response variable (frequency of disease) with a large sample of individual data on risk factors (material deprivation) and to estimate both individual and contextual effects. We found an important effect of material deprivation on mortality which is consistent with epidemiological literature.

Keywords: Empirical Bayes; Ecological regression; Health inequalities.

1 Introduction

The variability of disease occurrence among populations is generally higher than that within population. Notwithstanding, epidemiological studies usually evaluate differences in individual risk of disease within population and may loose power in identifying association with potential risk factors. Hybrid ecological models that integrate aggregate information on the frequency of disease with individual data on risk factors have been proposed (Prentice and Sheppard, 1995) to overcome such difficulties. These models are extensions of Generalized Estimating Equation approach but are not robust and, in some cases, fail to converge (Lancaster et al., 2006). Wakefield e Salway (2001; 2008) provided Bayesian solutions. The association between disease risk and socio-economic indicators such as material deprivation or education-based indexes is often investigated using ecological data. In this kind of analysis a contextual effect has been documented (e.g. Biggeri et al., 2004). The hybrid models previously mentioned do not provide estimates of contextual effects. We developed a series of empirical

Bayes models to integrate aggregate data on a discrete response variable (frequency of disease) with a large sample of individual data on risk factors (e.g. material deprivation) and to estimate both individual and contextual effects. Model comparisons have been addressed through the Expected Predicted Deviance (Gelfand e Ghosh, 1998). The motivating example was given by the assessment of the predictive validity of the Italian deprivation index on all causes mortality (Grisotto, 2009).

2 Data

We considered ISTAT death certificates for all causes of death (ICD IX 001-999) for the period 2000-2004. Total, males plus females, deaths were aggregated by Province (n=103). Expected counts were obtained by internal indirect age-gender standardization. Data on socio-economic factors at individual level come from the Multiscopo survey for the year 1999-2000 (ISTAT, 2002). A material deprivation at individual level has been constructed as the sum of zeta scores of four indicators of adverse events: low education (less then 6 yrs of education), unemployment, being a tenant and crowding index.

3 Methods

We first describe the statistical models for ecological, individual, and contextual effects.We then present their Bayesian specification.

3.1 Basic formulations

The data consist of the number of disease cases by province, Y_j, and a sample of n_j subjects for each province with information on individual material deprivation, X_{ij}.
Let assume that the number of observed of cases Y_j in the j-th province follows a Poisson distribution with parameters $E_j\theta_j$, where E_j represents the expected counts fixed by design, and θ_j the unknown relative risk. X_{ij} is the deprivation index for the i-th individual in the j-th province. Let define μ_{x_j} and $\sigma^2_{x_j}$ the mean and variance of the distribution of the material deprivation index for the generic j-th province. The parametric ecological model of Salway and Wakefield (2008) is defined as:

$$log(\theta_j) = \alpha + \beta\mu_{x_j} + \frac{\beta^2}{2}\sigma^2_{x_j}. \tag{1}$$

To derive a model for both the individual and the contextual effect recall that the ecological effect is given by the sum of the two (Cronbach and Webb, 1975; Firebaugh, 1978). Therefore with simple algebra we get:

$$log(\theta_j) = \alpha + \frac{\beta_I^2}{2}\sigma_{x_j}^2 + \beta_A \mu_{x_j} \quad (2)$$

where $\beta_A = \beta_C + \beta_I$, β_A is the ecological effect, β_C the contextual effect and β_I the individual effect. Since we have infomation on X from a sample of n_j individuals, μ_{x_j} and $\sigma_{x_j}^2$ are model parameters. A Besag York Mollié (1991) spatial convolution model is specified on the α intercept.

3.2 Hierarchical bayesian models

As benchmark, we specifiy a measurements model under a full Bayesian approach to model (1) (Best *et al.*, 2001). We then propose a class of Empirical Bayes models. For model (1) we specify a parametric EB model with the following priors:

$$\mu_{x_i} \sim Normal(x_j, \sigma_j^2/n_j)$$

$$\sigma_{x_i}^2 \sim Chi_{\nu_j}(s_j^2/n_j)$$

where x_j and s_j^2 are the sample mean and variance. For model (2) we specify a plug-in empirical Bayes solution.
Weakly informative prior distributions on β_I, β_A are assumed. We use Expected Predictive Deviance (EPD) (Gelfand and Ghosh, 1998) to compare full Bayesian models and the EB models previously defined. All computations were performed with WinBugs 1.4 (Lunn *et al.*, 2000).

3.3 Simulation study

We use the spatial layout of Italian Provinces with expected counts from internal standardization. We assume covariate values fixed to the observed at Province and Individual level in the ISTAT 2002 Multiscopo survey. The pseudo-data were generated in two steps:
- first, we obtain a baseline count at unit j-th Y_j^P drawing from a $Poisson(E_j\lambda_j)$ with $log(\lambda_j) = \mu + u_j + v_j$;
- second, the additional case due to the extra-risk by covariate effect at individual level i-th Y_{ij}^{PB} drawing from a $Bernoulli(\pi_{ij})$ with $logit(\pi_{ij}) = \beta_I(x_{ij} - x_i) + \beta_A x_i$.

We then generate the counts at unit j-th summing up the two contributions. The model parameter u_j and v_j are fixed to the posterior means obtained fitting a spatial convolution model to the aggregate province data. The simulations plan is to generate 1000 datasets for each combination of: 1) $\beta_A = 0.00$ and $\beta_I = 0.15$; 2) $\beta_A = 0.30$ and $\beta_I = 0.15$.

TABLE 1. All Cause mortality. Italy, male and female, 2000-2004. Regression coefficient (log relative risk and credibility interval 95%) for material deprivation. β_I: individual effect; β_A: ecological effect.

Model	β_I	β_A
FB	$_{0.076}0.108_{\,0.132}$	–
EB	$_{0.057}0.102_{\,0.146}$	–
Cronbach	$_{-0.224}0.035_{\,0.314}$	$_{0.085}0.120_{\,0.151}$

4 Results

Table 1 shows the results of the different fitted models. There is an important effect of material deprivation. The individual effect in the Cronbach model showed a large imprecision. We found that the amount of information on individual effect in hybrid designs is small (Sheppard, 2003). Noticeable, different modelling choices lead to different weight to the individual or contextual component of the ecological effect.

References

Best, N., Cocking, S., Bennett, J., Wakefield, J., Elliott, P. (2001): Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society A*, **164**, 155–174.

Biggeri, A., Dreassi, E., Marchi, M. (2004): A multilevel Bayesian model for contextual effect of material deprivation. *Statistical Methods & Applications*, **13**, 87–101.

Cronbach, L.J., Webb, J. (1975): Between-class and within-class effects in a reported aptitude X treatment interaction. *Journal of Educational Psychology*, **67**, 6–717.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*, **43**, 557–572.

Gelfand, A.E., Ghosh, S.K. (1998): Model choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

Grisotto, L. (2009): *Modelli Bayesiani gerarchici per il controllo della distorsione ecologica*. Phd Thesis in Applied Statistics. Department of Statistics "G. Parenti", University of Florence.

Jackson, C., Best, N., Richardson, S. (2008): Hierarchical related regression for conbining agregate and individual data in studies of socio–economic disease risk factors. *Journal of the Royal Statistical Society A*, **171**, 159–178.

Lancaster, G.A., Green, M., Lane, S. (2006): Reducing bias in ecological studies: an evaluation of different methodologies. *Journal of the Royal Statistical Society A*, **169**, 4, 681–700.

Lunn, D.J., Thomas, A., Best, N,. and Spiegelhalter D. (2000):WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**, 4, 325–337.

Prentice, R.L., Sheppard, L. (1995): Aggregate data studies of disease risk factors. *Biometrika*, **82**, 1, 113–25.

Salway, R., Wakefield, J. (2008): A hybrid model for reducing ecological bias. *Biostatistics*, **9**, 1, 1–17.

Sheppard, L. (2003): Insight on bias and information in group-level studies. *Biostatistics*, **4**, 2, 265–278.

Wakefield, J., Salway, R. (2001): A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society A*, **164** Part 1, 119–137.

Generalized Frailty Model for Comet Assays

Aklilu Habteab Ghebretinsae[1], Christel Faes[1],Geert Molenberghs[1,2],Marlies De Boeck[3],Helena Geys[3,1]

[1] I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium
[2] I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium
[3] Janssen Pharmaceutica, Turnhoutseweg 40, B-2430 Beerse, Belgium

Abstract: This paper proposes a flexible modeling approach for so-called comet assay data regularly encountered in pre-clinical research. While such data consist of non-Gaussian outcomes in a multi-level hierarchical structure, traditional analyses typically completely or partly ignore this hierarchical nature by summarizing measurements within a cluster. Molenberghs *et al* (2010) proposed a broad class of generalized linear models accommodating overdispersion and clustering through two separate sets of random effects. Here, we used this method to model comet assay data that exhibit an extra level of hierarchy. Whereas a conjugate gamma random effect is used for the overdispersion random effect, both gamma and Normal random effects are considered for the hierarchical random effect. Apart from model formulation, we place emphasis on Bayesian estimation.

Keywords: Frailty; Hierarchical model; Random effect; Weibull model.

1 Introduction

The comet assay is a technique used to assess the genotoxic potential of a compound by means of its ability to induce DNA damage in organ cells of male rats. Because the comet assay is quick, sensitive, and cheap, the assay is now widely used and a number of protocols have been developed for use in different types of investigations (Lovell and Omori 2008). However, the statistical analysis of such a comet assay is complicated because of several issues in the data:the multi-level structure of the data, the type of data, and the skewness of the outcome of interest.
In a typical comet assay study, a set of cells from exposed animals are investigated for DNA damage. This is done by considering the migration of DNA fragments out of the nucleus after electropheresism which induces typical comet-like structures. In many protocols, the cells from a single animal are placed on a number of slides. Each cell is then investigated for DNA damage by measuring the tail length and tail intensity of the comet. Because variability is expected between slides and between animals, this needs to be taken into account in the statistical analysis. This results in three-level hierarchies, with clustering at the animal and slide level.

Moreover, exploration of the distribution of the gathered data and previous work in this area indicate that the distribution for the responses (tail length and tail intensity) are asymmetric (Lovell and Omori 2008). The standard approach of modeling non-normal data, such as the tail intensity in the comet assay is using a generalized linear model (e.g., a Weibull model). The generalized linear model framework (McCullagh and Nelder 1989) is a very rich one. Nevertheless, already in the univariate case, it is well known that many standard members of the family may exhibit overdispersion. This results from the fact that various commonly used members prescribe a relationship between mean and variance. For example, in the Poisson model for count data, mean and variance are equal. In the exponential and Weibull cases, there is a quadratic relationship between them, etc. Molenberghs *et al* (2010) proposed an extended framework where two types of random effects are considered simultaneously, so as to deal, at the same time, with overdispersion on the one hand and data hierarchies on the other. Hierarchical random effects are frequently assumed to be normal, but they can take various distributional forms. An illustrious counterexample is time-to-event data where gamma random-effects, usually termed gamma frailties, are in common use. We considered both.

1.1 Data

The data refer to four groups of six male rats that received a daily oral dose of a compound in three dose levels (low, medium, and high) or vehicle control. On the day of necropsy, an extra group of three animals received a single dose of a positive control (200 mg/kg ethyl methanesulfonate, EMS, PC). The animals were sacrificed 3 hours after the last dose administration, their liver was removed and processed for the comet assay. For each animal, a cell suspension is prepared. From each cell suspension, three replicate samples were prepared for scoring. Fifty randomly selected, non-overlapping cells per sample were then scored for DNA damage using a semi-automated scoring system. A total of 150 liver cells were thus scored per animal. DNA damage was assessed by the software system by measuring tail migration, % tail intensity, and tail moment. Tail migration is the distance from the perimeter of the comet head to the last visible point in the tail; % tail intensity is the percentage of DNA fragments present in the tail; and tail moment is the product of the amount of DNA in the tail and the mean distance of migration in the tail.

2 General Frailty Models

For a one level of hierarchy, Molenberghs *et al* (2010) use a combined model with a normal random effect to handle the hierarchy in the data and a conjugate random effect to account for overdisperion in the response.we

propose extending the model to account for an extra level of hierarchy by the use of three random effects of which one is the overdispersion effect. In addition, while typically a normal random effect is included in the linear predictor to account for the clustering, as in Molenberghs *et al* (2010), also a multiplicative factor using a multivariate gamma distribution can be used, similar to the multiplicative factor for the overdispersion random effect. consider a model with a normally distributed random effect for the first hierarchy in the data and a gamma random effect for the second hierarchy in the data. In addition, we allow for the overdispersion in the model via another gamma-random effect.

$$f(y_{ijk}|\theta_{ijk}, b_i, b_{ij}) = \lambda\rho\theta_{ijk}b_{ij}y_{ijk}^{\rho-1}e^{\mathbf{x_{ijk}}'\xi+b_i}e^{-\lambda y_{ijk}^{\rho}\theta_{ijk}b_{ij}e^{\mathbf{x_{ijk}}'\xi+b_i}}, \quad (1)$$

$$f(\theta_{ijk}) = \frac{1}{\left(\frac{1}{\alpha_1}\right)^{\alpha_1}\Gamma(\alpha_1)}\theta_{ijk}^{\alpha_1-1}e^{-\alpha_1\theta_{ijk}}, \quad (2)$$

$$f(b_i) = \frac{1}{(2\pi d)^{1/2}}e^{-\frac{1}{2d}b_i^2}, \quad (3)$$

$$f(b_{ij}) = \frac{1}{\left(\frac{1}{\alpha_2}\right)^{\alpha_2}\Gamma(\alpha_2)}b_{ij}^{\alpha_2-1}e^{-\alpha_2 b_{ij}}, \quad (4)$$

$$\mathbf{x_{ijk}}'\xi = \beta_0 + \beta_1 L_{ijk} + \beta_2 M_{ijk} + \beta_3 H_{ijk} + \beta_4 PC_{ijk}. \quad (5)$$

Y_{ijk} is the Tail Intensity or Tail length measured for cell $k = 1, \ldots, n_{ij}$ of rat $i = 1, \ldots, N$, in slide $j = 1, \ldots, n_i$. The fixed effect β_0 denotes the control (vehicle) effect. The parameters β_1 to β_4 are the contrasts of interest that represent the effect of low dose, medium dose, high dose, and positive control versus vehicle. The random intercept b_i corresponds to the rat-specific effect whereas b_{ij} corresponds to the slide-specific effect j of rat i. θ_{ijk} is the overdispersion random effect. Other models can be defined where either a gamma or a normal random effect is considered. We refer to Table 1 for the overview of the models considered. Models are implemented in R2winbugs.

3 Result and Conclusion

The different models considered are compared using Deviance information criterion (DIC). Weibull Gamma(RE2) was the preferred model followed by Weibull Normal(RE1) Gamma(RE2) model for Tail Intensity. We refer to Table 2 for the summary result. comparison of the classical Weibull model and Weibull Gamma(RE2), the parameters of interest are highly significant in both cases. Yet, the standard errors, likewise the credible intervals of Weibull Gamma(RE2) are twice that of the classical Weibull model. While not the case in this example because of the high toxicity of

TABLE 1. Overview of models considered with DIC for Tail Intensity(TI)and Tail Length(TL)

	Distribution for							
	Response	Overdisp.	RE1(rat)		RE2(slide)			
Model	Weibull	Gamma	Normal	Gamma	Normal	Gamma	DIC(TI)	DIC(TL)
1	√						33869.6	30878.8
2	√		√				33823.9	30421.6
3	√			√			33823.5	30420.2
4	√	√					33895.6	27378.5
5	√	√	√				33853.7	26901.6
6	√	√		√			33852.5	26883
7	√				√		33728.9	29622.6
8	√					√	**33728.5**	29620.8
9	√	√			√		33760.7	26386.9
10	√	√				√	33760.6	26377
11	√		√		√		33728.7	29623.4
12	√		√			√	33728.6	29619.5
13	√			√	√		33730.3	29631.1
14	√			√		√	33729.7	29605.2
15	√	√	√		√		33761.6	26374.4
16	√	√	√			√	33760.5	26333.1
17	√	√		√	√		33760.6	26338
18	√	√		√		√	33758.6	**26209.6**

the compound of interest, this suggests that ignoring the hierarchical structure and overdispersion could have major influence on the final conclusion. Significant estimates in the classical Weibull model may be insignificant in Weibull Gamma(RE2). In other words, a compound might be erroneously declared toxic.

Based on the analysis for tail intensity, more elaborate models did not outperform (not much improvement in terms of DIC). However, this was not the case for the second response, tail length. Based on the DIC, the most complicated model has the best fit, showing the importance of the hierarchical structure as well as overdispersion. Models with one hierarchical random effect were better fitting as compared to the classical Weibull model. Models with two random effect improved the fit further, and models with the complete hierarchical structure and overdispersion random effect appear to be best. Generally, for tail length like for tail intensity, we did not reach a different conclusion, due to high toxicity of the compound; however, inclusion of the hierarchical structure and overdispersion random effect had severe impact on the magnitude, standard errors as well as the credible intervals. Results for the classical Weibull, a model with two hierarchical random effects model and the preferred model with full hierarchical structure and overdispersion are summarized in Table 3.

TABLE 2. Parameter estimates obtained from Models 8 [Weibull-Gamma(RE2)] and 12 [Weibull-Normal(RE1)-Gamma(RE2)] for Tail Intensity.

		Weibull-Gamma(RE2)	
Effect	Parameter	Est.(s.e.)	95% C.I.
Vehicle	β_0	-2.419(0.079)	[-2.57,-2.26]
Low *versus* vehicle	β_1	-2.854(0.097)	[-3.04,-2.66]
Medium *versus* vehicle	β_2	-3.092(0.098)	[-3.29,-2.90]
High *versus* vehicle	β_3	-3.317(0.098)	[-3.51,-3.12]
Pos. control *versus* vehicle	β_4	-1.829(0.115)	[-2.05,-1.60]
Weibull shape	ρ	1.420(0.019)	[1.38,1.46]
RE2 parameter	α_2	18.33(4.036)	[11.68,27.3]
		Weib.-Norm.(RE1)-Gamma(RE2)	
Effect	Parameter	Est.(s.e.)	95% C.I.
Vehicle	β_0	-2.427(0.085)	[-2.59,-2.25]
Low *versus* vehicle	β_1	-2.850(0.104)	[-3.06,-2.65]
Medium *versus* vehicle	β_2	-3.088(0.106)	[-3.30,-2.88]
High *versus* vehicle	β_3	-3.312(0.107)	[-3.53,-3.11]
Pos. control *versus* vehicle	β_4	-1.826(0.124)	[-2.07,-1.58]
Weibull shape	ρ	1.419 (0.019)	[1.38,1.46]
Precision of RE1	$\frac{1}{d}$	114.2(79.29)	[28.60,331.61]
RE2 parameter	α_2	19.99(4.493)	[12.08,29.54]

3.1 Conclusion

In this paper, we proposed a flexible modeling framework for the comet assay data using a Bayesian hierarchical model that takes into account the complete hierarchical nature, the possible overdispersion and the appropriate non-Gaussian probability distribution for the response. The more conventional models with either the overdispersion, or just one hierarchical random effect being submodels.

The method was applied to the comet assay data gathered to assess the toxicity of 1,2-Dimethylhydrazine dihydrochloride at different dose levels. For this particular dataset, a Weibull-gamma(RE2) model seemed adequate for tail intensity, whereas a Weibull-gamma(OD)- gamma(RE1)-gamma(RE2) was better fit for tail length. A comparison of these analysis with the conventional approach, which ignores the overdispersion and the hierarchy in the data, revealed that both models led to the same qualitative conclusion of severe toxicity of the compound at all dose levels. This notwithstanding, estimates, standard errors, and credibility intervals were severely affected, underscoring the risk of using models that are too simple. In general, proper models encompassing at the same time the hierarchical nature in the data, combined with overdispersion effects, need to be adopted. In this case,

TABLE 3. Parameter estimates obtained from Weibull Gamma(OD) Gamma(RE1) Gamma(RE2) and Weibull Normal(RE1) Gamma(RE2) and Weibull Model for Tail Length.

		Weib. G G G	Weib. N G	Weibull
Effect	Parameter	Est.(s.e.)	Est.(s.e.)	Est.(s.e.)
Veh.	β_0	-30.44(0.6646)	-15.26(0.2519)	-12.76(0.1543)
Low *vs.* veh.	β_1	-11.99(0.4977)	-4.79(0.2468)	-3.55(0.0530)
Medium *vs.* veh.	β_2	-12.14(0.5061)	-4.89(0.2479)	-3.65(0.0535)
High *vs* veh.	β_3	-12.57(0.4946)	-5.10(0.2509)	-3.85(0.0550)
Pos. C. *vs* veh.	β_4	-9.75(0.5523)	-3.79(0.3028)	-2.70(0.0590)
Weibull shape	ρ	10.71(0.2727)	4.96(0.0572)	4.01(0.0422)
Precision of RE1	$\frac{1}{d}$	–	45.83(54.60)	–
OD parameter	α_1	0.894(0.0431)	–	–
RE1 parameter	α_2	4.597(3.179)	–	–
RE2 parameter	α_3	1.611(0.2985)	3.031(0.5393)	–

the use of the overdispersion and hierarchical structure improved the fit for one response. Furthermore, even when the more elaborate model does not provide a substantially improved fit, nor alters the inferences drawn, the development is still very useful because it provides further confidence, by way of model specification assessment, on the quality of the purported model.

References

Duchateau, L. and Janssen, P. (2007). *The Frailty Model*. New York: Springer.

Lovell, D. P. and Omori, T. (2008). Statistical issues in the use of the comet assay. *Mutagenesis*, **23**, 171–182.

Molenberghs, G. and Verbeke, G (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **00**, 000-000.

Interval Estimation of Random Effects in Frailty Models

Il Do Ha[1], Florin Vaida[2], Youngjo Lee[3], Maengseok Noh[4]

[1] Department of Asset Management, Daegu Haany University, South Korea,
[2] Department of Family and Preventive Medicine, University of California, San Diego, USA
[3] Department of Statistics, Seoul National University, South Korea
[4] Department of Statistics, Pukyong National University, South Korea

Abstract: Semi-parametric frailty models are widely used to analyze clustered survival data. In this talk, we propose the use of the hierarchical-likelihood (HL) interval for frailties (random effects). We study the relationship between HL, empirical Bayesian, and fully Bayesian intervals for frailties. The proposed HL interval can be interpreted as a frequentist confidence interval and fully Bayesian credible interval under a uniform prior. We also propose an adjustment of the proposed interval to avoid null intervals. The proposed methods are demonstrated using numerical studies based on a data set from the design of a multicenter clinical trial.

Keywords: Empirical Bayes; Hierarchical likelihood; Interval estimator; Random effects; Survival analysis.

1 Introduction

It is important to investigate the potential heterogeneity in survival among clusters in order to understand and interpret the variability in the data (Vaida and Xu, 2000). Multivariate semi-parametric frailty models offer a flexible framework for modeling this heterogeneity. For example, the effects of a treatment can vary substantially across participating centers in a multicenter clinical trial with a censored event-time endpoint (Gray, 1994). Such heterogeneity can be accounted for by using random treatment effects, possibly in addition to random cluster effects on the baseline hazard. In addition to the estimation (or prediction) of random effects, a measure of uncertainty for these point estimates is useful and necessary. The standard methods in use are empirical Bayes (EB) confidence intervals, based on the conditional posterior distribution of random effects given the observed data and the estimated parameter values (Vaida and Xu, 2000). However, the EB interval estimators have been criticized for not maintaining the nominal level (Carlin and Louis, 2000). Gray (1994) and Legrand et al. (2005) developed fully Bayesian methods. Recently, Lee and Ha (2010) used HL

methods to estimate random effects and their confidence intervals for hierarchical generalized linear models (HGLMs, Lee and Nelder, 1996). In this talk, we extend these methods to semi-parametric frailty models. Here, one particular difficulty is that in certain cases, likelihood methods may lead to zero estimates for strictly positive variance components, leading to null confidence intervals. For the non-null interval we also extend the adjustment proposed by Morris (2006) in linear mixed models to general random-effect models, including HGLMs and frailty models. Through numerical studies, we show that the proposed interval improves the empirical Bayes interval by maintaining the stated nominal level.

2 Model formulation

Suppose that the data consist of censored time-to-event observations collected from q clusters (e.g. centers). Let T_{ij} be the survival time for the jth observation in ith cluster, $i = 1, \ldots, q,\ j = 1, \ldots, n_i,\ n = \sum_i n_i$. Denote by v_i an s-dim'l vector of unobserved log-frailties (random effects) associated with the ith cluster. Given v_i, the conditional hazard function of T_{ij} is of the form

$$\lambda_{ij}(t|v_i) = \lambda_0(t)\exp(\eta_{ij}), \tag{1}$$

where $\lambda_0(\cdot)$ is the unknown baseline hazard function, $\eta_{ij} = x_{ij}^T\beta + z_{ij}^T v_i$ is the linear predictor for the log-hazard, and $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ and $z_{ij} = (z_{ij1}, \ldots, z_{ijs})^T$ are $p \times 1$ and $s \times 1$ covariate vectors corresponding to fixed effects $\beta = (\beta_1, \ldots, \beta_p)^T$ and log-frailties v_i, respectively. We assume v_i are independent and follow a multivariate normal distribution,

$$v_i \sim N_s(0, \Sigma) \tag{2}.$$

The covariance matrix $\Sigma = \Sigma(\phi)$ depends on a vector of unknown parameters ϕ. The normal distribution has been used for modelling multicomponent (Ha et al., 2007) and correlated frailties (Rondeau et al., 2008).

3 Interval estimators for random effects

For observations j of cluster i, let T_{ij} and C_{ij} be the event and censoring times, respectively, and response variable $y_{ij} = \min\{T_{ij}, C_{ij}\}$ with event indicator $\delta_{ij} = I(T_{ij} \le C_{ij})$. Since the functional form of $\lambda_0(t)$ in (1) is unknown, following Breslow (1972), we consider the baseline cumulative hazard function $\Lambda_0(t)$ to be a step function with jumps at the observed event times, $\Lambda_0(t) = \sum_{k:y_{(k)} \le t} \lambda_{0k}$, where $y_{(1)} < \ldots < y_{(r)}$ are the ordered distinct event times and $\lambda_{0k} = \lambda_0(y_{(k)})$. Following Lee and Nelder (1996) and Ha et al. (2001), the HL for semi-parametric frailty models (1) is defined by the joint likelihood of (y, δ) and v, which is of the form

$$h = h\{(\beta, \lambda_0, \phi), v\} = \log f(y, \delta|v; \beta, \lambda_0) + \log f(v; \phi).$$

Since (β, λ_0, v) and ϕ in (2) are asymptotically orthogonal as in HGLMs (Lee and Nelder, 1996), we only need to consider the Hessian matrix of v and $\psi = (\beta^T, \lambda_0^T)^T$. Along the lines of Lee and Ha (2010), the interval estimation of v is based on $\hat{v}(\hat{\psi})$ with $\psi = (\lambda_0, \beta)$ and the inverse of Hessian matrix $H(h; \psi, v) = \partial^2 h / \partial \psi^2$ of v and ψ which gives the estimated standard error of $\hat{v} - v$ in sense of Conditional MSE (CMSE) of Booth and Hobert (1998). Here, $\hat{v}(\hat{\psi}) \equiv \hat{v}(\psi)|_{\psi=\hat{\psi}}$, where $\hat{v}(\psi)$ is the solution to $\partial h / \partial v = 0$ for a given ψ. Note that $\hat{v}(\psi) = E_\psi(v|y, \delta)$ asymptotically. Since the number of nuisance parameters λ_{0k} increases with sample size n, $H(h; \psi, v)^{-1}$ requires an inversion of a high-dimensional $(p + q + r)$ matrix. Following Ha et al. (2001), we propose the use of the profiled HL, h^*, that eliminates λ_0:

$$h^* = h|_{\lambda_0 = \widehat{\lambda}_0},$$

where $\widehat{\lambda}_{0k}$ are solutions of $\partial h / \partial \lambda_{0k} = 0$. Again, the covariance estimates for $\hat{v} - v$ are obtained from $H(h^*; \beta, v)^{-1}$, leading to an efficient computation of the confidence interval for v. Thus, we propose that the individual $(1 - \alpha)$-level HL confidence intervals for the components v_k of v are of the form

$$\hat{v}_k \pm z_{\alpha/2} \cdot \mathrm{SE}(\hat{v}_k - v_k),$$

where $z_{\alpha/2}$ is the normal quantile with probability $\alpha/2$ in the right tail and $\mathrm{SE}(\hat{v}_k - v_k)$ is obtained from the square root of lower-right-hand corner of $H(h^*; \hat{\beta}, \hat{v})^{-1}$. For EB confidence intervals, $\mathrm{SE}(\hat{v}_k - v_k)$ is also obtained from the square root of $(-\partial^2 h^* / \partial v \partial v^\top)^{-1}$. Furthermore, for the non-null interval for v, following the Morris method (2006), we propose the use of the adjusted likelihood p_{adj}, defined as

$$p_{\mathrm{adj}} = p_{\beta,v}(h^*) + \log \det(\Sigma).$$

Here $p_{\beta,v}(h^*)$ is an adjusted profile h-likelihood (an extended restricted likelihood) for ϕ and it is the first-order Laplace approximation to a modified marginal likelihood, which becomes exact as $N = \min_{1 \le i \le q} n_i \to \infty$: see Lee et al. (2006) and Ha et al. (2010) for more justifications of asymptotic property.

4 Numerical Study

We conducted a numerical study, based upon 500 replications of simulated data, in order to compare the operating characteristics of the EB and HL intervals. Following the setup of the Vaida and Xu (2000) data analysis of a multicenter lung-cancer clinical trial, we consider the two frailty models, from (1):
M1: $\eta_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + v_{i0}$,
M2: $\eta_{ij} = (\beta_1 + v_{i1}) x_{ij1} + \beta_2 x_{ij2} + v_{i0}$.

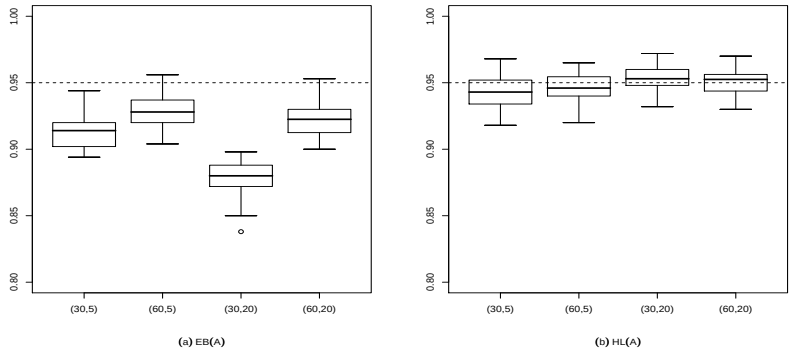


FIGURE 1. Simulation results for coverage probabilities of the nominal 95% (*dotted line*) EB and HL intervals of all random effects (v_{i0}'s) in frailty model (M1) under $\sigma_0^2 = 1$ and 15% censoring.

Here we assume $\lambda_0(t) = 1$, $\beta_1 = -0.5, \beta_2 = 0.5$, $\sigma_0^2 = \sigma_1^2 = 0.2, 1$, and $\rho = -0.5$ for $v_i = (v_{i0}, v_{i1})^T$ in M2. The binary covariates x_{ij1} and x_{ij2} are each generated from a Bernoulli distribution with success probability 0.5. We set the following sample sizes: $n = \sum_{i=}^{q} n_i$ with $n = 150, 300, 600$, and 1200, and (q, n_i) =(30,5), (60,5), (30,20), and $(60, 20)$. The censoring times were each generated from an exponential distribution with 15% and 50% rates. HL(S) and EB(S) denote the HL and EB methods using standard and adjusted REML estimators for variance components, respectively. Though not reported here, the coverage probabilities (CPs) of the nominal HL(S) and EB(S) 95% intervals for all random effects (v_{i0}'s) in M1 are liberal, particularly for a small variance ($\sigma_0^2 = 0.2$) and small sample ($n_i = 5$) which often give null intervals. Figure 1 shows that the adjustment HL(A) adequately corrects this issue. That is, for a large variance $\sigma_0^2 = 1.0$ the EB(A) does not maintain the nominal level, even when n_i is large. In contrast, the HL(A) intervals maintain the nominal 95% level in all cases studied, indicating that it is necessary to correct for the uncertainty in the estimation of β. Although not shown here, the CPs of intervals for all v_i's in M2 were similar to those of M1, and the results for 50% censoring were also to 15% censoring.

Acknowledgments: This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2009-0088978 and No. 2010-0021165).

References

Gray, R. J. (1994). A Bayesian analysis of institutional effects in multicenter cancer clinical trial. *Biometrics* **50**, 244-253.

Ha, I.D., Lee, Y. and MacKenzie, G. (2007). Model selection for multi-component frailty models. *Statistics in Medicine* **26**, 4790-4807.

Ha, I. D., Lee, Y, and Song, J. K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233-243.

Ha I.D., Noh, M. and Lee, Y. (2010). Bias reduction of likelihood estimators in semi-parametric frailty models. *Scandinavian Journal of Statistics* **37**, 307-320.

Lee, Y., and Ha, I. D. (2010). Orthodox BLUP versus HL methods for inferences about random effects in Tweedie mixed models. *Statistics and Computing*, **20**, 295-303.

Lee, Y., and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619-678.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalised Linear Models with Random Effects: Unified Analysis via h-Likelihood.* Chapman and Hall.

Legrand, C., Ducrocq, V., Janssen, P., Sylvester, R. and Duchateau, L. (2005). A Bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model. *Statistics in Medicine* **24**, 3789-3804.

Morris, C. N. (2006). Mixed model prediction and small area estimation. *Test* **15**, 72-76.

Vaida, F., and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, **19**, 3309-3324.

Functional Clustering of Water Quality Data in Scotland

Ruth Haggarty[1], Claire Miller[1], Marian Scott[1], Malcolm Smith[2], Fiona Wyllie[2]

[1] School of Mathematics and Statistics, 15 University Gardens, University of Glasgow, G12 8QQ. Contact: rhaggarty@stats.gla.ac.uk
[2] Scottish Environment Protection Agency, Erskine Court, Stirling, FK9 4TR

Abstract: A functional data analysis (FDA) approach is presented to investigate the grouping structure of water bodies that is used for classification within the EU Water Framework Directive. FDA has been used to compare groups of Scottish standing waters in terms of temporal dynamics of several different chemical determinands of interest with a functional clustering model proposed to examine the existing grouping structure currently used by the Scottish Environment Protection Agency (SEPA).

Keywords: functional data analysis; clustering; water quality.

1 Introduction

1.1 Background

Under the EU Water Framework Directive (European Parliament, 2000), water bodies can be grouped together and classification of all members of the group are based on the classification of a single representative site. Currently, the groups that are used for classification of standing waters are based on typology which is derived from broad categories of alkalinity and altitude. Often, the representative site within each group is determined by logistics and ease of access for sampling purposes. There is some question as to how reliable the current grouping approach is as wrongly specifying either the groups, or the representative site within each group, could potentially result in misclassification of all members. It is therefore of interest to investigate statistical approaches for clustering sites and hence alternative group structures.

1.2 Data

In total there are approximately 104 standing waters in Scotland that are classified within groups for the Water Framework Directive. These lochs make up 30 distinct groups, with the number of sites within a single group

ranging from two to eight. From 2007 onwards, when classification based on groups was first introduced, there is often only data available on the representative loch. Ideally, in order to ensure a reasonable comparison of groups and sites, a dataset is required where there are observations taken over a common period on all sites within each group. For this reason, data from a subset of lochs were provided by SEPA. The dataset used in this paper consists of 21 lochs which make up seven groups. The time period covered by the data is from January 2003 to December 2006. The number of samples, and the dates at which samples were collected varies enormously from site to site. Data were available on 5 different determinands of interest however this paper will focus on alkalinity values measured in micrograms per litre (μg/L). Values have been log transformed to stabilize the variance.

2 Methods - Functional Data Analysis (FDA)

Grouping sites using standard clustering techniques often only utilises annual averages of the variables of interest and so valuable information about the variables' temporal dynamics is lost. FDA is an approach which enables curves, that are constructed from time series collected on individual sites, to be analysed using functional equivalents to many standard statistical techniques. A detailed discussion of FDA techniques is given in Ramsay and Silverman (2003). The functional clustering model based approach (James and Sugar (2003)) not only enables curves to be partitioned into distinct groups but also provides a confidence in classification by quantifying the uncertainty in the partition. In addition, the model accounts for sparse data which is a problem in the loch grouping data. A brief description of the model is given below. Further to this, more details are provided in James and Sugar (2003) and in Pastres *et al.* (2010).
Let there be n individual sites and let the function which represents log alkalinity at site i at time t be written as

$$Y_i(t) = g_i(t) + \epsilon_i(t), \text{ where } i = 1, ..., n \tag{1}$$

then $g_i(t)$ is the true value of the i-th curve at time t and $\epsilon_i(t)$ is the corresponding measurement error. Dropping the time index notation then Equation (1) can be written more simply as $\mathbf{Y_i} = \mathbf{g_i} + \epsilon_\mathbf{i}$. It is assumed $\epsilon_\mathbf{i} \sim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2})$ and are independent. Following this, $\mathbf{g_i}$ can be expressed as the sum of a group effect and a random independent site effect to give the functional clustering model which can be written as

$$\mathbf{Y_i} = \mathbf{S_i}(\lambda_\mathbf{0} + \mathbf{\Lambda}\alpha_\mathbf{k} + \gamma_\mathbf{i}) + \epsilon_\mathbf{i} \tag{2}$$

$$\epsilon_\mathbf{i} \sim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2}\mathbf{I}) \text{ and } \gamma_\mathbf{i} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Gamma})$$

where $i = 1, .., n$ and $\mathbf{S_i}$ is the spline basis matrix for the i-th curve evaluated at time points $t_{i1}, ..., t_{im_i}$. Further details on spline functions are

provided in Green and Silverman (1994). In Equation (2), $\lambda_{\mathbf{0}}$ is a p dimensional vector which represents the overall mean for all sites, $\alpha_{\mathbf{k}}$ is an h dimensional vector which represents the group effect and $\mathbf{\Lambda}$ is a $p \times h$ matrix where $h \leq min(p, G-1)$, where G is the number of groups. It is assumed that all random site effects, γ_i, have a common covariance structure represented by Γ. Subject to certain constraints, detailed in James and Sugar (2003), the unknown model parameters are estimated using a maximum likelihood approach. In addition to these parameters, the probability that the ith curve comes from group k can also be estimated. While h and G have to be specified prior to fitting the model to the data, optimal values of these two parameters can be obtained by minimising Bayes Information Criterion (BIC). Subsequently, the G clusters are formed by allocating each of the n sites into the group for which they have the greatest corresponding membership probability.

3 Results

3.1 Scottish Loch Analysis

Cubic P-splines were used to fit curves to the log transformed alkalinity data at each loch with a different smooth function fitted in each case. Figure 1 shows the smooth function fitted to each of the sites. It is clear from Figure 1 that there is a great deal of overlap in the sites and no clear indication that 7 groups (the number currently used by SEPA) is the most appropriate. After application of the functional clustering model, with a range of different parameter values, the values which minimised the BIC were found to be 3 groups and $h = 1$. Figure 2 shows the estimated cluster mean curve for each of the 3 groups, represented by the heavier solid lines. The dashed lines represent each of the individual sites with the different shading representing the predicted groups. Figures 3 and 4 show the geographical locations of the sites along with both the original SEPA group structure, and the new predicted group structure based on the fitted model for log(Alkalinity) respectively. SEPA groups 4, 5 and 6 become groups B, C and A respectively in the new groups. The remaining sites within SEPA groups 1, 2, 3 and 7 are split and are predicted to fall into groups A and B in the new group structure.

3.2 Summary and Future Work

The functional clustering model is a useful tool for exploring both existing group structures and monitoring networks for classification of water bodies. It has been shown that the model works well in terms of identifying distinct groups where the within group heterogeneity is small. The methods presented here have been applied to several different chemical determinands. However, further work will include investigation of group structures based

on multiple variables of interest and development of methods for application to a set of monitoring sites connected by a network of streams and rivers.

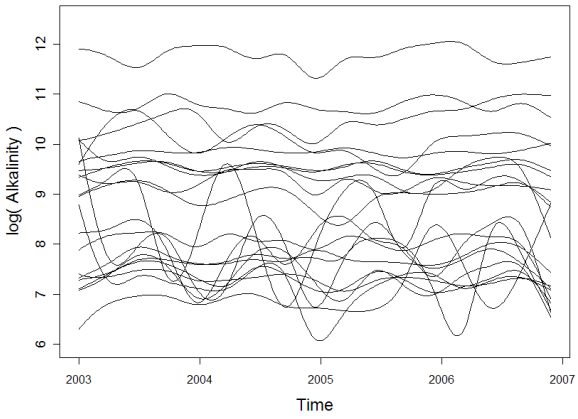


FIGURE 1. Fitted cubic spline functions for log(alkalinity) at each site

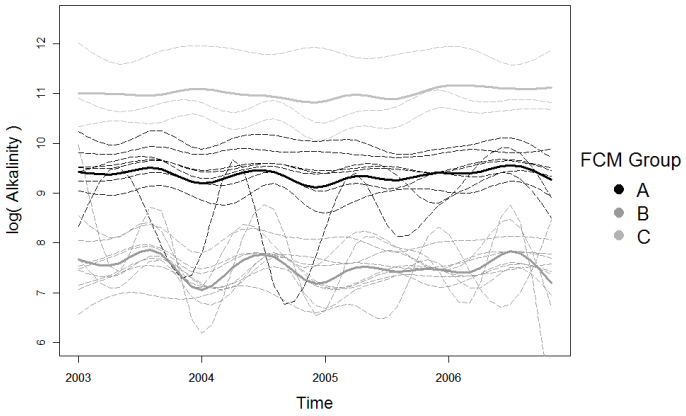


FIGURE 2. Cluster means and predicted group structure for log(alkalinity)

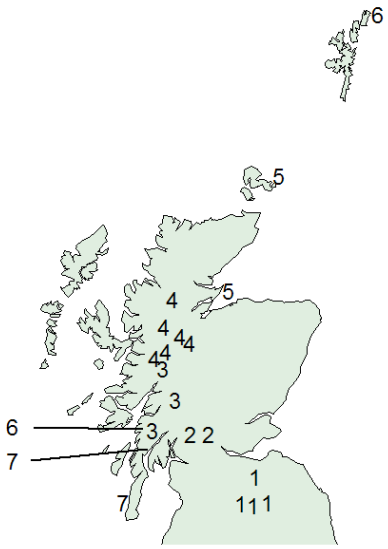


FIGURE 3. Map of Scotland showing original SEPA groups

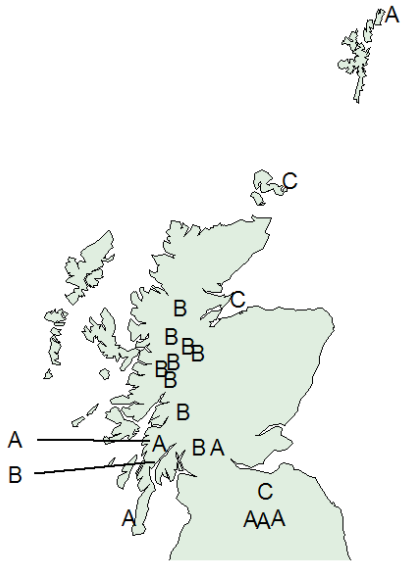


FIGURE 4. Map of Scotland showing predicted groups for log(alkalinity)

References

European Parliament (2000). Directive 2000/60/EC. of the European Parliament, establishing a framework for community action in the field of water policy. *Official Journal of the European Communities*, **327**, 1–72.

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

James, G, M. and Sugar, C. A. (2003). Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Assosciation*, **98**, 397–408.

Pastres, R., Pastore, A., and Tonellato, S. F. (2010). Looking for similar patterns among monitoring stations. Venice Lagoon application. *Environmetrics* (to appear) (**DOI: 10.1002/env.1066.**)

Ramsay, J. and Silverman, B. W. (1997). *Functional Data Analysis (Springer Series in Statistics)*, Springer.

Using Probability Models to Classify Software Patterns

S. Hasso[1] , K. M. Matawie[2]

[1] Wolters Kluwer, Law & Business, 4025 W Peterson Avenue, Chicago, Illinois 60646, USA. email: Sargon.Hasso@wolterskluwer.com
[2] School of Computing and Mathematics, University of Western Sydney, Po Box 10, Kingswood NSW 2747, Australia. email: K.Matawie@uws.edu.au

Abstract: We propose an approach for creating software design patterns classification scheme based on probability models and statistical methods used in information retrieval domain. The approach looks for a set of words, phrases, and topics, i.e. concepts embedded or represented by words and phrases that describe the pattern. We also present a process that generates a list of terms, associate each list with a pattern category, and search the resulting list with user queries to select a particular pattern.

Keywords: Design Patterns, Classification, Probability Models, Design Patterns Catalog, Poisson Filtering, Bayesian Text Classifier

1 Introduction

It has been evident that design patterns are extremely useful design tools for software designers because each design pattern describes both a problem and a design solution. Generally, software design problems tend to be general enough that they surface repeatedly in a variety of design situations. Since their introduction to the software community by Gamma et al. (1993,1995), Coad and Yourdon (1991), Coplien (1992), Buschmann et al (1996), design patterns' acceptance has grown considerably and they have become an important new approach to software design. Originally, Gamma et al. have published 23 design patterns, but it was not very long before that number has increased considerably by the software community. So while hundreds of design patterns have been published, it is not unlikely that many more of those will be discovered. While this is great news for software re-use, one big problem still remains: how do you find a particular design pattern quickly and efficiently?
This paper focuses on addressing one problem area in using software design patterns: the lack of a process to catalog the hundreds of design patterns that have appeared since this concept was introduced to the software development community.

In a previous research Hasso PhD (2007), we relied heavily on manual process in analyzing and selecting terms (keywords) for classification purposes. Ultimately, the classification structure produced resulted in a set of pre-determined and limited set of controlled vocabulary that became the basis for indexing any software pattern. Locating a pattern then becomes a matter of identifying relevant topics we are interested in.
There is an alternative to this manual process and it provides for an automated way to generate indexes based on probability theory that is used for indexing and searching a vast volume of data like the internet. Specifically, we will explore Poisson Distribution as a filtering tool Harter (1975) used to filter out, throw away, unwanted or irrelevant terms, and then using Bayesian Text Classification to classify patterns.

2 Classifying Software Patterns using Probabilistic Tools

Classification is the act of grouping like things together. Classification displays relationships between things and between classes of things Buchanan (1979). The 'things' we classify could be anything. In our research, they are software design patterns. Classification, in general, is an essential tool to find structure and relationships between terms in any document Aitchison et al. (1997). In Information Retrieval systems, used to retrieve relevant documents, two processes are involved to facilitate the retrieval: indexing and retrieval. In indexing, a concise representation of a document is derived based on key terms used in the document title or document description, while retrieval refers to the search method by which relevant document is identified Srinivasan (1992).
We propose a method by which we adapt the use of probabilistic tools specifically to help us extract relevant terms from patterns documents and use these as a basis for pattern classification that, in the end, serves as a patterns search tool. The following describes briefly the steps in our proposed approach, and in the full version we will give the theoretical details and overview of an end-to-end process to prepare, build, classify, and query patterns repository. Figure 1 depicts graphically this end-to-end process.

1. Use Poisson distribution as a filter to throw away any unwanted terms from a description of patterns.

2. We will use the results of previous step (step 1) as an indicative of the important topics that a pattern document is about.

3. We use the list of terms extracted from step 2 as an input to a Bayesian text classification tool Graham-Cumming (2005) that serves as a training set. This text classifier is an automated means by which

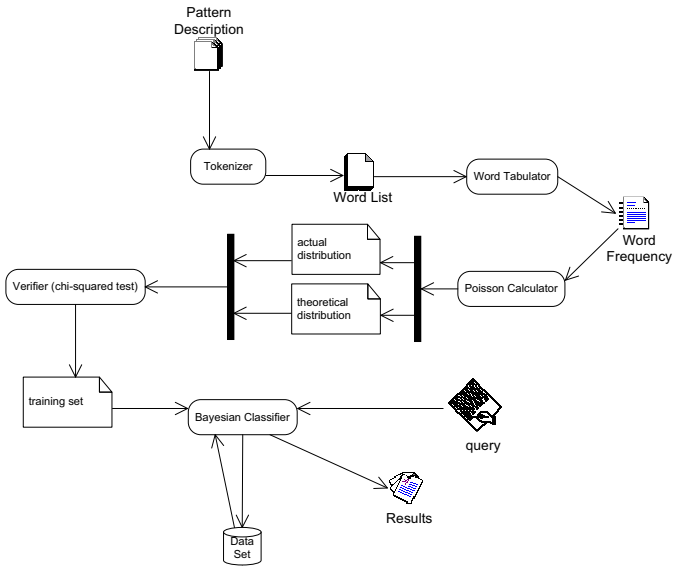


FIGURE 1. A probabilistic pattern classification and query process illustration using UML's (Booch et al., 1999) activity diagram.

we can determine which category a document belongs to. Effectively, the classifier suggests categories for indexing a pattern description. We assign categories and we input the training set to tell the classifier to use as indexing terms. The classifier learns to associate a category with a particular set of terms and we will use this knowledge in predicting the class of a new document.

4. During pattern searching, we collect terms from users and create a search criteria to be submitted to the classifier tool from step 3 to determine approximately the likelihood of a pattern a user is searching for.

3 Conclusion

This paper introduced another method to systematically analyze patterns and create a classification scheme based on tools used successfully to index and classify documents. The statistical approach used here offers promising potential much needed in the area of software engineering to provide a comprehensive, unified, and efficient way to create software patterns catalogs and query tools to retrieve patterns at design time when required.

References

Aitchison, J., A. Gilchrist, and D. Bawden (1997). *Thesaurus Construction and Use: A Practical Manual*. Aslib, Stone House Court, London, 3rd edition.

Booch, G., J. Rumbaugh and I. Jacobson (1999). *The Unified Modeling Language User Guide*. Addison-Wesley, Reading, Massachusetts, USA, 1st edition.

Buchanan, B (1979). *Theory of Library Classification*. Clive Bingley, London, UK.

Buschmann, F., R. Meunier, H. Rohnert, P. Sommerlad, and M. Stal (1996). *Pattern-Oriented Software Architecture: A System of Patterns*. John Wiley & Sons, New York.

Coad, P., and E. Yourdon (1991). *Object-Oriented Analysis*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Coplien, J. O. (1992) *Advanced C++: Programming Styles and Idioms*. Addison Wesley, Reading, MA.

Gamma, E., R. Helm, J. Vlissides and R. E. Johnson (1993). "Design Patterns: Abstraction and Reuse of Object-Oriented Design". In [Nierstrasz, O.], editor, *Proceedings of the ECOOP '93 European Conference on Object-oriented Programming*, LNCS 707, pages 406–431. Springer-Verlag.

Gamma, E., R. Helm, R. Johnson and J. Vlissides (1995). *Design Patterns*. Addison Wesley, Reading, MA.

Graham-Cumming, J (2005). Naïve Bayesian Text Classification. *Dr. Dobb's*, (372):16–20, May 2005.

Harter, S. P. (1975) A Probabilistic Approach to Automatic Keyword Indexing. *Journal of the American Society for Informatyion Science*, 26(4):197–206, Jul/Aug 1975.

Hasso, S. (2007) *A Uniform Approach to Software Patterns Classification and Software Composition*. [Ph.D. Thesis], IIlinois Institute of Technology.

Srinivasan, P (1992). "Thesaurus Construction". In Frakes, B. W. and R. Baeza-Yates, editor, *Information Retrieval: Data Structures & Algorithms*, chapter 9, pages 161–218. Prentice Hall, Englewood Cliffs, NJ.

Linear Model comparison with structured mean and dispersion parameters

Freddy Hernandez[1], Olga Usuga[1], Viviana Giampaoli[1]

[1] fhernanb@gmail.com, ousuga@gmail.com, vivigi08@gmail.com, Mathematics and Statistics Institute, Sao Paulo University

Abstract: Hierarchical Generalized Linear Model (HGLM) and the Generalized Linear Model for the location, scale and shape (GAMLSS) were proposed to model the mean and the dispersion parameters using linear predictors considering random effects and fixed effects using their own set of covariates, where the response variable belongs to families of distributions appropriate in each case. In this work we present the results from a comparison simulation study considering fixed effects and normal random effects for the linear predictor for the mean and fixed effects linear predictor for the dispersion parameter. Two scenarios were considered, response variable distributed normal and gamma. We found that fixed effects estimates obtained by GAMLSS and HGLM were similar in both scenarios.

Keywords: Hierarchical Generalized Linear Model; Generalized Linear Model for the location, scale and shape.

1 Introduction

Generalized Linear Model (GLM) proposed by Nelder & Wedderburn (1972) assumes that the dependent variable y belongs to the Exponential Family (EF) and allows to model the mean μ of the variable y as $\mu = g^{-1}(\eta)$ where $g(\cdot)$ is a known link function and η corresponds to the linear predictor which is a linear function of explanatory variates. GLM considers the variance $V(y)$ of y as a function of the mean through the following relation $V(y) = \phi v(\mu)$ where ϕ corresponds to the dispersion coefficient and $v(\mu)$ is the variance function which is known. For the distributions that belong to the EF, variance, skewness and kurtosis are generally functions of μ and ϕ (Rigby & Stasinopoulos (2005)). Generalized Nonlinear Models (GNLM) are characterized due to the linear predictor η of GLM is replaced by a nonlinear predictor. Generalized Additive Model (GAM) proposes to replace the linear predictor η of GLM by an additive predictor consisting of nonparametric functions of explanatory variables. In Generalized Linear Mixed Model (GLMM) the linear predictor η of GLM is formed by a fixed component (parametric) and a random component (random effects). With each approach GLM, GNLM, GAM and GLMM variance and other mo-

ments of the response variable depend on the estimation of μ and ϕ and it is not possible to model them using a different set of covariates from the used in the estimation of μ. Two proposals are found in the statistical literature that consider the previous problem. The first proposal was presented by Lee & Nelder (1996) and Lee and Nelder & Pawitan (2006) called Hierarchical Generalized Linear Model (HGLM), which allows that μ and ϕ are structured through their own sets of covariates and that eases the distribution of random effects in EF. The second proposal by Rigby & Stasinopoulos (2005) is called Generalized Additive Model for location, scale and shape (GAMLSS) and allows the distribution of the response variable y can be selected from a general family of distributions that include the EF. The systematic part in GAMLSS is expanded to model the mean μ and other parameters associated with the distribution of y.

2 Simulation study

It was considered a normal-normal model with mean and variance structured, the model is based on an example given by Ronnegard et al. (2011). Considering y_i as the answer to the i-th group (with $i = 1, \ldots, n$) the model can be written as:

$$y_i|\beta, u, \beta_d = N\left(X_{i.}\beta + Z_{i.}u, exp\left(X_{d,i.}\beta_d\right)\right) \tag{1}$$

where β and $u \sim MVN(\mathbf{0}, \mathbf{I}\sigma_u^2)$ correspond to the fixed effect and random effect to the mean respectively, β_d to the fixed effects for the variance. The model matrices are denoted by $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{X}_d$; the notation $X_{j.}$ and $X_{.k}$ represent the j row and k column for $\mathbf{X}$ respectively, the same for $\mathbf{Z}$ and $\mathbf{X}_d$. In the simulation study the parameters σ_u^2, the number of groups n and the number of observations by group m varied while β, β_d remained fixed. The values considered in the study were as follows: $\sigma_u^2 = 0.1, 0.5, 1.0, 2.0$, $n = 5, 10, 15$ number of groups, $m = 5, 10, 15, 20, 25$ observations by group, $\beta^{'} = (5, -4, 7)$ and $\beta_d^{'} = (2, -3, 1)$. 10000 iterations were performed for each combination of the above parameters. The model matrix $\mathbf{X}$ is such that $X_{.1} = \mathbf{1}$, $X_{.2} \sim P(\lambda = 2)$ and $X_{.3} \sim Exp(\gamma = 0.5)$. The model matrix $\mathbf{X_d}$ is such that $X_{d.1} = \mathbf{1}$, $X_{d.2} \sim Bi(n = 1, p = 0.5)$ and $X_{d.3} \sim Bi(n = 1, p = 0.7)$.

The criterion used to compare the fits obtained with HGLM and GAMLSS was the multivariate Mean Squared Error (MSE). The MSE for the estimator $\hat{b}$ of b is defined as

$$MSE(b) = tr\left(\Sigma(\hat{b})\right) + \left(\hat{b} - b\right)^{'}\left(\hat{b} - b\right)$$

where $\Sigma(\hat{b})$ corresponds to the variance-covariance matrix for $\hat{b}$. The MSE was calculated for $\hat{\beta}_{MLGH}$, $\hat{\beta}_{GAMLSS}$, $\hat{\beta}_{dMLGH}$ and $\hat{\beta}_{dGAMLSS}$ the estimators for β and β_d with HGLM and GAMLSS.

In the figure (1) are the observed results to the mean MSE for $n = 5$ groups. For any value of σ_u^2 and 5 observations by group there are differences between MSE of $\hat{\beta}_{dMLGH}$ and $\hat{\beta}_{dGAMLSS}$. For 10 or more observations per group the performance with HGLM and GAMLSS are similar. This pattern was also observed in the cases for $n = 10$ and $n = 15$.

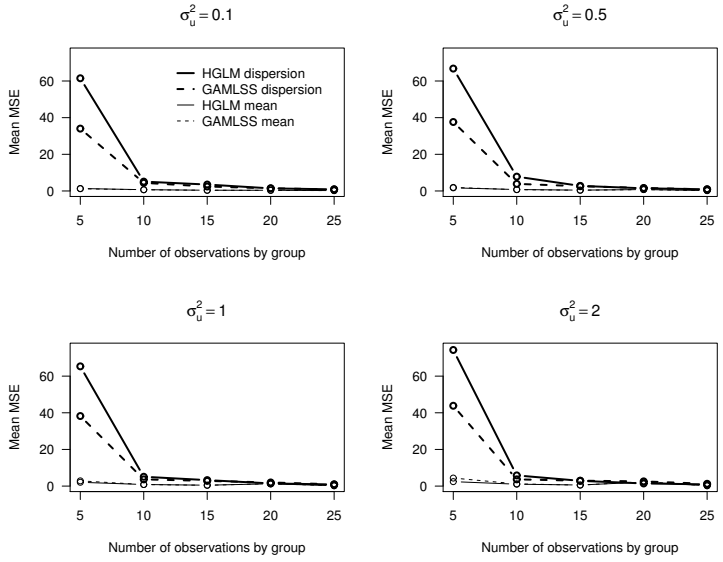


FIGURE 1. Mean MSE for the estimators obtained by HGLM and GAMLSS with $n = 5$ for the normal-normal model

It was also considered a gamma-normal model based on data from the application of semiconductor presented by Myers et al. (2002). The figure (2) presents the results of the mean MSE in this case. The mean MSE with HGLM and GAMLSS for the dispersion parameter is the same regardless of the σ_u^2 value. For a value of $m = 5$, mean MSE of the dispersion parameter is higher with GAMLSS. As the number of observations per group the mean MSE decreases. The mean MSE for the mean is similar for HGLM and GAMLSS and increases with increasing σ_u^2.

3 Conclusions

For the normal-normal scenario was found that the performance for estimating vectors of fixed effects for the mean and dispersion parameter with GAMLSS and HGLM were very similar. As the number of observations, the MSE decreases. For the gama-normal scenario was found again that the performance with GAMLSS and HLGM was similar.

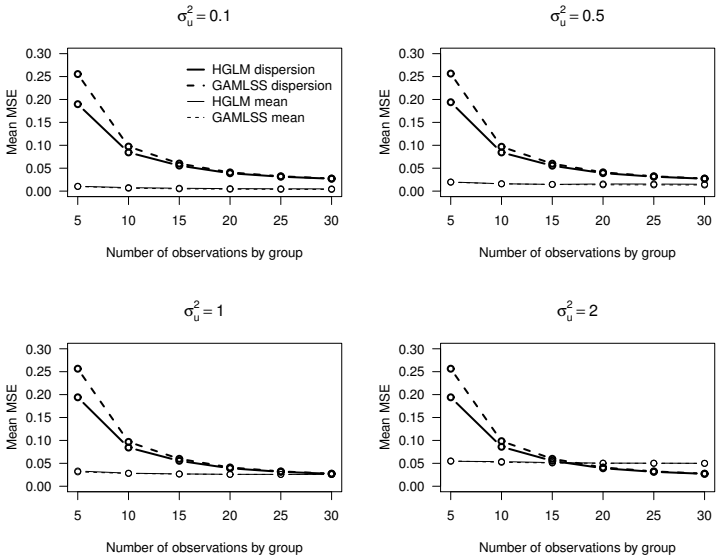


FIGURE 2. Mean MSE for the estimators obtained by HGLM and GAMLSS for the gama-normal model

References

Lee, Y., Nelder, J. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society* **58**(4), 619-678.

Lee, Y., Nelder, J.A., Pawitan, Y. (2006). Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood. Chapman and Hall CRC, London.

Myers, P.H., Montgomery, D.C., Vining, G.G. (2002). Generalized linear models with applications in engineering and the sciences. Jhon Wiley and Sons, New York. **25**, 25-37.

Nelder, J. A., Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society A* **135**, 370-384.

Rigby, R., Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Applied Statistics* **53**(3), 507-554.

Ronnegard, L., Shen, X., Alam, M. (2011). The hglm Package. *R Foundation for Statistical Computing*

Joint Modelling of Two Sequential Times to Events With Longitudinal Information

Jaime-Abel Huertas[1], Guadalupe Gómez[1], Carles Serrat[2]

[1] Dept. Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Edifici C5, Campus Nord, c/ Jordi Girona 1–3, 08034–Barcelona, jaime.abel@upc.edu, lupe.gomez@upc.edu

[2] Dept. Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Avga Dr. Marañón 44–50, 08028–Barcelona, carles.serrat@upc.edu

Abstract: In survival analysis, the lifetimes may be observed in some specified order, where the time to event T_k, cannot be observed until $T_1, ..., T_{k-1}$ have been observed. The present work proposes a joint model of two sequential times to events together with longitudinal information, extending the joint model of Wolfsohn and Tsiatis (1997) for one time to event and one longitudinal variable. We apply the model to the clinical trial called TIBET, in which an intermittent therapeutic strategy has been assigned to each patient. Of special clinical interest is the lifetime that a patient needs before restarting treatment given the progression of biological markers recorded during the followup period.

Keywords: Joint Modelling; Longitudinal Data; Survival Analysis; Sequential Times.

1 Joint Models in the Literature

Likelihood and Bayesian approaches rely on the specification of an appropriate likelihood for the joint model parameters; for both, much of the early literature focuses on models without autocorrelation structure for longitudinal model. Good review can be found in Tsiatis and Davidian (2004). Wulfsohn and Tsiatis (1997) proposed and EM algorithm for a simple joint model, but many proposals for more complex joint models developed recently, have based the estimating procedures in it, among others, the joint model for one time to event with multiple longitudinal variables (Lin *et al.*, 2002), a joint modelling of accelerated failure time and longitudinal data, (Tseng *et al.*, 2005), and a robust joint modelling of longitudinal measurements and competing risks failure time data (Li *et al.* 2009). In Bayesian framework, Chi and Ibrahim (2006) give a model for multivariate longitudinal and multivariate survival data by using MCMC techniques.

2 Notation

The idealized data for each subject $i = 1, ..., n$ followed over an interval $[0, \tau)$ are $\{T_{1i}, T_{2i}, R_i(u), 0 \leq u \leq \tau, X_i\}$, where T_{1i} and T_{2i} are event times, $\{R_i(u), 0 \leq u \leq \tau\}$ is the longitudinal response trajectory for all times $u \geq 0$ and $X_i = [X_{1i}^T \;\; X_{2i}^T]^T$ is a vector of baseline (time 0) covariates, X_{1i} with influence over T_1, and X_{2i} over T_2, which may have elements in common or not.

We will consider only a situation where T_1 and T_2 may be right censored by the censoring times C_1 and C_2 respectively, so instead of T_{ji} we observe (Y_{ji}, δ_{ji}), $j = 1, 2$, where $Y_{ji} = min\{T_{ji}, C_{ji}\}$ and $\delta_{ji} = I(T_{ji} \leq C_{ji})$ which indicates whether Y_{ji} is an uncensored right value of T_{ji}. On the other hand, for some set of times $t_{ij}, j = 1, ..., n_i$, instead of the true values $R_i(t_{ij})$ we observe $Z_i(t_{ij})$, then the observed data for subject i is $O_i = \{X_i, Y_i, \delta_i, Z_i, \tilde{t}_i\}$, where $\tilde{t}_i = (t_{i1}, ..., t_{in_i})^T$, $Z_i = (Z_i(t_{i1}), ..., Z_i(t_{in_i}))^T$, $Y_i = (Y_{1i}, Y_{2i})$, and $\delta_i = (\delta_{1i}, \delta_{2i})$.

3 Joint Modelling of One Time to Event Data and one Longitudinal Variable

For the longitudinal response process, a standard approach is to characterize $R_i(u)$, $u \geq 0$, only in terms of random effects b_{0i} and b_{1i} like

$$R_i(u) = b_{0i} + b_{1i}u. \tag{1}$$

Associations among the longitudinal and time to event processes and covariates, is characterized by the following semi-parametric model for the hazard risk:

$$\begin{aligned} \lambda_i(u) &= \lim_{du \to 0} \Pr(u \leq T_i < u + du \mid T_i \geq u, R_i^H(u), X_i)/du \\ &= \lambda_0(u) \exp(\eta^T X_i + \beta R_i(u)), \end{aligned}$$

where $R_i^H(u) = \{R_i(t), 0 \leq t < u\}$ is the history of the longitudinal process up to time u, and the parameters are represented in β and the η vector. If model takes $\beta R_i(u)$ as $\beta_1 b_{0i} + \beta_2 b_{1i} + \beta_3(b_{0i} + b_{1i}u)$, the parameters β_1, β_2 and β_3 measure the association induced through the intercept, slope and current R value, respectively. Wulfsohn and Tsiatis (1997) give and EM algorithm to estimate the joint model maximizing the resultant log-likelihood.

Zeng and Cai (2005) rigorously prove under the normal assumption for the random effects, among other assumptions, the strong consistency of the maximum likelihood estimators for joint models of repeated measurements and survival time, and derive their asymptotic distributions, which is multivariate normal. Moreover, the asymptotic results hold even if the random effect, has slightly heavier tails than the normal density. The theoretical results further confirm that nonparametric maximum likelihood estimation provides efficient estimation.

4 Joint Modelling of Two Sequential Times to Events and One Longitudinal Variable

We have proposed a joint model for two sequential times to events with one longitudinal variable, as an extension of the Wulfsohn and Tsiatis's model (1997) with a model for two sequential times to events (Lawless 2003, section 11.3). The model permit us to give prognosis for a time to event given covariates, the longitudinal process and the previous event time. Usually the trend of the longitudinal variable changes with the first time to event. If we take the longitudinal variable with two piecewise linear mixed models, the knot where the slope changes is obviously the time to first event T_1, and a particular joint model in which the longitudinal and survival sub-models are linking with the current value may be as:

$$Z_{ij} = b_{0i} + (b_{1i}\ t_{ij} + b_{2i}(t_{ij} - t_{1i})I) + e_i(t_{ij}) \tag{2}$$

$$\lambda(t_1 \mid b_i; \beta_1) = \lambda_{1,0}(t_1) \exp\{\beta_1(b_{0i} + b_{1i}t_1)\} \tag{3}$$

$$\lambda(t_2 \mid t_{1i}, b_i; \beta_2, \gamma) = \lambda_{2,0}(t_2) \exp\{\beta_2(b_{0i}+b_{1i}\cdot t_{1i}+(b_{1i}+b_{2i})t_{2i})+\gamma\ t_{1i}\} \tag{4}$$

where $I = I(t_{ij} \geq t_{1i})$, β_1 and β_2 are parameters of association between the longitudinal and survival process, and γ describes the relation among the times to event. Both baseline risks $\lambda_{1,0}(\cdot)$ and $\lambda_{2,0}(\cdot)$ are left unspecified and different. In the likelihood construction we have the same assumptions made by Wulfsohn and Tsiatis (1997). The assumption of non-informative censoring extend to this case of censoring process. The errors e_i are assumed mutually independent, normally distributed with mean 0 and variance σ_ϵ^2, and independent with b_i and for all other variables conditional on (b_i, X_i). If we may assume that, given random effects and covariates, Z, T_1, and $T_2 \mid T_1$, are all independent, then the observed likelihood is:

$$L(\Omega) = \prod_{i=1}^{n} \int_{b_i} \left\{ \prod_{j=1}^{n_i} f(z_{ij} \mid b_i; \sigma_\epsilon^2) \right\} f(Y_i, \delta_i \mid b_i, X_i; \psi_{T|b}) f(b_i; B, \Gamma) \mathrm{d}b_i, \tag{5}$$

where $\Omega = (\psi_{T|b}, B, \Gamma, \sigma_\epsilon^2)$ and $\psi_{T|b} = (\eta_1, \eta_2, \beta_1, \beta_2, \gamma, \lambda_{1,0}, \lambda_{2,0})$. The vector of random effects $b_i = [b_{0i}\ b_{1i}]^T$ is taken to be normally distributed with mean B and covariance matrix Γ. The function for the survival process is defined as (omitting parameters for simplicity),

$$f(Y_i, \delta_i \mid b_i, X_i) = \left[S(Y_{1i}, \delta_{1i} \mid b_i, X_{1i})\lambda(Y_{1i}, \delta_{1i} \mid b_i, X_{1i})^{\delta_{1i}}\right] \left[S(Y_{2i}, \delta_{2i} \mid b_i, t_{1i}, X_{2i})\lambda(Y_{2i}, \delta_{2i} \mid b_i, t_{1i}, X_{2i})^{\delta_{2i}}\right]^{\delta_{1i}}.$$

We estimate the joint model with an EM algorithm as a natural extension of the algorithm developed by Wulfsohn and Tsiatis (1997). With simulation we can advertise that this EM algorithm may have two convergence points, due to fact that the time to dropout it is not expressed directly in (6) across of the sequential density of T_1 and T_2.

The models for longitudinal data in presence of informative dropout, use the time to dropout to correct bias estimations. In survival point of view for the joint modelling, the observed event time cut the longitudinal process and may be see as the dropout cause. The joint modelling produce proper longitudinal estimations, and of course, good estimations for the survival model. But in this case we do not have a single time to dropout, we have a sequence of times to event.

We might to include properly the time to dropout in the modelling, fitting strategically the models for T_1+T_2, T_1 and $T_2 \mid T_1$, and taking the parameters of the model of T_1+T_2 as nuisance parameters. The suggested method fits the longitudinal process with a unique time to dropout measured by $T = T_1 + T_2$ with $\delta_t = \delta_1 \cdot \delta_2$ as censoring time, afterwards the survival process is fitted, given the parameter estimations of the Z model. Thus, we estimate the parameters for the model of Z and $T_1 + T_2$ as the same form of the Wulfsohn and Tsiatis's method (1997), then, the estimation for the model of T_1 and $T_2 \mid T_1$ are made applying twice and separately for each model the EM algorithm, but having fixed the longitudinal parameter estimates. So this method in essence calculates several times the model by Wulfsohn and Tsiatis (1997).

The alternative in the way that uses a traditional Cox model (1972) to fit the survival models of T_1 and $T_2 \mid T_1$ after and given the model fitting of Z and $T_1 + T_2$, produce some bias estimations in its.

The Cox model for $T_1 + T_2$ may be fitted with baseline covariates and the random effects across of anyone: the current value, the intercept and the slop. Letting $T = T_1 + T_2$, the hazard risk modeled only with the current value is

$$\lambda(t \mid b_i; \beta_t) = \lambda_{t,0}(t) \exp\{\beta_t(b_{0i} + b_{1i}t)\}. \tag{6}$$

Although our problem consist of two times to different events (Restart and suspension of therapy), the proposed model could be used to model data sets where the events are similar, like the problems with disabled recurrences: the first time being the time to some disabled, and the second, the time to the same disabled from the first (having repeated measurements for some marker).

5 Simulation

Simulations are carried out to explore how robust and reliable is our method. In general, our method gives proper estimations for sample sizes bigger than 300, with moderate correlation and variability of the random effects (or less), and with not heavy censoring. Although the results are acceptable with sample sizes n=100 with low censoring.

Because Zeng and Cai (2005) mention that the asymptotic properties of normality and consistency of the maximum likelihood estimators for the joint models with one time to event and one longitudinal variable, are

extensive to the joint models with multivariate survival times, we expect that these properties hold to our maximum likelihood estimators. We have some evidence in that sense supported with simulation.

6 Application

We apply the above described technique to the TIBET clinical trial. The trial contemplates the incorporation of interruption periods in the administration of an intensive therapy *HAART* (Highly Active Antiretroviral Therapy). A cohort of 100 patients enters the study with suspension of the treatment (state *OFF*). Basal and retrospective information is gathered, and every 4 weeks there is registered information of the CD4 cell count. If the patient's conditions deteriorate, the therapy is restarted (state *ON*), and so on. The times to event are T_1: time to first restart of therapy, and T_2: time from the first restart of therapy to the suspension of therapy. The longitudinal variable is the evolution of the CD4 which is not increasing until the first time to event, and then is increasing.
We fit different joint models with two piecewise in the longitudinal part, finding that the best joint model among the analyzed, has for the model of T_1 the viral load pre-therapy (VL), and the effect of the slope and the current value along of T_1. The model of T_2 has the effect of the slope along of T_2 and the effect of T_1 as significative covariates, nevertheless we fit the model also with the current value and the viral load pretherapy, in order to see how is the effect of these variables in the models of T_1 and T_2. We have that $b_1 + b_2$ is the effect of the slope in T_2 and $b_0 + b_1(t_1 + t_2) + b_2 t_2$ is the current value effect. The selected joint model is as follow, and the results are shown in Table 1.

$$Z_{ij} = b_{0i} + (b_{1i}\ t_{ij} + b_{2i}(t_{ij} - t_{1i})I) + e_i(t_{ij}) \tag{7}$$

$$\lambda(t_1 \mid b_i, VL_i; \eta_1, \beta_1) = \lambda_{1,0}(t_1) \exp\{\eta_1 VL_i + \beta_{11}(b_{0i} + b_{1i}t_1) + \beta_{12}b_{1i}\} \tag{8}$$

$$\lambda(t_2 \mid t_{1i}, b_i, VL_i; \eta_2, \beta_2, \gamma) = \lambda_{2,0}(t_2) \exp\{\eta_2 VL_i + \beta_{21}(b_{0i} + b_{1i}(t_{1i} + t_2) + b_{2i}t_2) + \beta_{22}(b_{1i} + b_{2i}) + \gamma\ t_{1i}\}. \tag{9}$$

We refer the following principal findings: 1. the only baseline covariate significative in T_1 was the viral load pre-therapy but this effect is diluted in T_2, 2. The relationship between T_1 and T_2 is inverse, 3. The slope of the longitudinal variable along of T_2 and the observed values of the first time to event T_1, are the only significative covariates in the survival model of T_2, and 3. The influence of the intercepts b_0 and $b_0 + b_1 t_1$ in T_1 and T_2 respectively, is not significative. It is logic since the patients begin the trial without therapy with good and similar conditions, and the restart of therapy is due to the threshold reached in the levels of the CD4 and viral load.

TABLE 1. Joint model for T_1 and T_2 with two piecewise mixed model for the CD4 evolution, based in EM modified algorithms. It is assumed semi parametric form in the hazard risks.

Parameter	*Estimate*	*s.e.*	$p - value$
Mixed			
B_0	25.8263	0.4569	< 0.0001
B_1	-0.0502	0.0041	< 0.0001
B_2	0.1248	0.0080	< 0.0001
σ_{11}	20.8764	2.9524	< 0.0001
σ_{12}	-0.1186	0.0221	< 0.0001
σ_{13}	0.0166	0.0364	0.6484
σ_{22}	0.0017	0.0002	< 0.0001
σ_{23}	-0.0016	0.0004	< 0.0001
σ_{33}	0.0063	0.0009	< 0.0001
σ_{ϵ}^2	6.1463	0.1881	< 0.0001
Survival T_1			
β_{11} $(b_{0i} + b_{1i}t_1)$	-0.2044	0.0405	< 0.0001
β_{12} (b_{1i})	-14.3298	3.5632	< 0.0001
η_1 (VL_i)	0.6521	0.1858	0.0004
Survival T_2			
β_{21} $(b_{0i} + b_{1i}(t_{1i} + t_2) + b_{2i}t_2)$	0.0257	0.0430	0.5500
β_{22} $(b_{1i} + b_{2i})$	6.7904	2.4090	0.0048
η_1 (VL_i)	-0.0169	0.2302	0.9414
γ (t_{1i})	-0.0210	0.0079	0.0079

References

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

Tseng, Y.-K., Hsieh, F., and Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika*, **92**, 587–603.

Lawless, J. F. (2003). Statistical Models and Methods for Lifetime Data, 2nd edition. *Wiley* Hoboken.

Li, N., Elashoff, R. M. and Li, G (2009). Robust Joint Modeling of Longitudinal Measurements and Competing Risks Failure Time Data. *Biometrical Journal*; **51(1)**, 19-30.

Lin, H., McCulloch, C.-E. and Mayne, S.-T. (2002) Maximum likelihood estimation in the join analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, **21**, 2369–2382.

Tsiatis, A.A., and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**, 793–818.

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

Zeng, D. and Cai, J. (2005). Asymptotic Results for Maximum Likelihood Estimators in Joint Analysis of Repeated Measurements and Survival Time. *The Annals of Statistics* **33** 2132-2163.

Elliptical semiparametric mixed models

Germán Ibacache Pulgar[1] , Gilberto A. Paula [1]

[1] Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, Caixa Postal 66281 (Ag. Cidade de São Paulo), CEP 05311-970, São Paulo, Brazil, e-mail: germanp@ime.usp.br and giapaula@ime.usp.br

Abstract: In this work we extend semiparametric mixed models with normal errors to elliptical errors in order to permit distributions with heavier and lighter tails than the normal ones. A reweighed iterative process based on the back-fitting method is proposed for the parameter estimation and the local influence curvatures are derived to study the sensitivity of the estimates. An illustration of the methodology is presented for real data set.

Keywords: Elliptical distributions; Semiparametric models; Local influence.

1 Introduction

Semiparametric mixed models (SMMs) adopt the following relationship:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{N}_i\mathbf{f} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \ , \tag{1}$$

where $\mathbf{y}_i$ is an $(m_i \times 1)$ random vector of observed responses from the ith cluster, $\mathbf{X}_i$ is an $(m_i \times p)$ design matrix, $\boldsymbol{\beta}$ is the $(p \times 1)$ fixed parameter vector, $\mathbf{N}_i$ is an $(m_i \times r)$ incidence matrix with the (j,ℓ)th element equal to the indicator $I(\mathrm{t}_{ij} = \mathrm{t}_\ell^0)$, for $j = 1,\ldots,m_i$ and $\ell = 1,\ldots,r$, $\mathbf{f} = (f(\mathrm{t}_1^0),\ldots,f(\mathrm{t}_r^0))^T$ with $\mathrm{t}_1^0,\ldots,\mathrm{t}_r^0$ being the distinct and ordered values of t_{ij}, $f(\cdot)$ is a smooth function, $\mathbf{Z}_i$ is the $(m_i \times q)$ design matrix associated to the $(q \times 1)$ vector of random effects $\mathbf{b}_i$, and $\boldsymbol{\epsilon}_i$ is an $(m_i \times 1)$ vector of within-cluster errors. In this work we will assume that

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{b}_i \end{pmatrix} \sim \mathrm{El}_{m_i+q}\left\{ \begin{pmatrix} \boldsymbol{\mu}_i \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_i & \mathbf{Z}_i\mathbf{D} \\ \mathbf{D}\mathbf{Z}_i^T & \mathbf{D} \end{pmatrix} \right\},$$

where $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{N}_i\mathbf{f}$, $\boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \phi\mathbf{I}_{m_i}$ and $\mathbf{D} = \mathbf{D}(\boldsymbol{\lambda})$, with $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_d)^T$. Consequently, $\mathbf{y}_i \sim \mathrm{El}_{m_i}(\boldsymbol{\mu}_i\,,\,\boldsymbol{\Sigma}_i)$. Then, the penalized log-likelihood function can be expressed as

$$L_{\mathrm{p}}(\boldsymbol{\theta},\alpha) = \sum_{i=1}^{n}\left[-\frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log g(\delta_i) - \frac{\alpha}{2n}\,\mathbf{f}^T\mathbf{K}\mathbf{f}\right],$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T,\mathbf{f}^T,\boldsymbol{\tau}^T)^T$, with $\boldsymbol{\tau} = (\tau_0,\tau_1,\tau_2,\ldots,\tau_d)^T$, $\tau_0 = \phi$ and $\tau_\ell = \lambda_\ell$ $(\ell = 1,\ldots,d)$, $\delta_i = \mathbf{r}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{r}_i$, $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$, $g(\cdot)$ is a function of $\mathcal{R} \to [0,\infty]$ such that $\int_0^\infty \delta^{m/2-1}g(\delta)\mathrm{d}\delta < \infty$, $\mathbf{K}$ is a smoothing matrix (see Green and Silverman, 1994) and α is the smoothing parameter (for simplicity, fixed).

2 Parameters estimation

Here we consider the maximum penalized likelihood estimate (MPLE) of $\boldsymbol{\theta}$, which leads to a natural cubic spline estimate of $f(\mathrm{t})$ and can be obtained via the following procedure (see also Ibacache-Pulgar and Paula, 2011):

(a) Firstly, we maximize $L_{\mathrm{p}}(\boldsymbol{\beta}, \mathbf{f}, \boldsymbol{\tau}, \alpha)$ over $\boldsymbol{\beta}$ by remaining fixed the parameters $\mathbf{f}$ and $\boldsymbol{\tau}$. The maximum value, $\widehat{\boldsymbol{\beta}}(\mathbf{f}, \boldsymbol{\tau})$, is attained for values of $\boldsymbol{\beta}$ in a set $\mathcal{B}(\mathbf{f}, \boldsymbol{\tau})$ depending on the parameters $\mathbf{f}$ and $\boldsymbol{\tau}$. Thus, if $\boldsymbol{\beta} \in \mathcal{B}(\mathbf{f}, \boldsymbol{\tau})$, the penalized log-likelihood function value is $L_{\mathrm{p}}^{c}(\mathbf{f}, \boldsymbol{\tau}, \alpha) = \max_{\beta} L_{\mathrm{p}}(\boldsymbol{\beta}, \mathbf{f}, \boldsymbol{\tau}, \alpha)$.

(b) Then, in the second step, we maximize the concentrated penalized log-likelihood function $L_{\mathrm{p}}^{c}(\mathbf{f}, \boldsymbol{\tau}, \alpha) = L_{\mathrm{p}}(\widehat{\boldsymbol{\beta}}(\mathbf{f}, \boldsymbol{\tau}), \mathbf{f}, \boldsymbol{\tau}, \alpha)$ over $\mathbf{f}$ by remaining $\boldsymbol{\tau}$ fixed. The maximum value, $\widehat{\mathbf{f}}(\boldsymbol{\tau})$, is attained for values of $\mathbf{f}$ in a set $\mathcal{F}(\boldsymbol{\tau})$ depending on the parameter $\boldsymbol{\tau}$. Therefore, if $\mathbf{f} \in \mathcal{F}(\boldsymbol{\tau})$, the penalized log-likelihood function value is $L_{\mathrm{p}}^{c}(\boldsymbol{\tau}, \alpha) = \max_{f} L_{\mathrm{p}}^{c}(\mathbf{f}, \boldsymbol{\tau}, \alpha)$.

(c) Finally, in the third step, we maximize the concentrated penalized log-likelihood function $L_{\mathrm{p}}^{c}(\boldsymbol{\tau}, \alpha) = L_{\mathrm{p}}(\widehat{\boldsymbol{\beta}}(\mathbf{f}, \boldsymbol{\tau}), \widehat{\mathbf{f}}(\boldsymbol{\tau}), \boldsymbol{\tau}, \alpha)$ over $\boldsymbol{\tau}$. The maximum value, $\widehat{\boldsymbol{\tau}}$, is attained on a set $\mathcal{C}$ of $\boldsymbol{\tau}$ values. Then, three-step procedure (a)-(c) lead to the following iterative process:

Step 1 (Back-fitting algorithm) Let $\mathbf{W}_i = \boldsymbol{\Sigma}_i^{-1}$. For $r, s = 0, 1, \ldots$, repeatedly cycling, until convergence, between the following two equations:

$$\begin{aligned} \boldsymbol{\beta}^{(r+1,s+1)} &= (\mathbf{X}^T\mathbf{W}^{(r)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(r)}\big(\mathbf{y} - \mathbf{N}\mathbf{f}^{(r+1,s)}\big) \quad \text{and} \\ \mathbf{f}^{(r+1,s+1)} &= (\mathbf{N}^T\mathbf{W}^{(r)}\mathbf{N} + \alpha\,\mathbf{K})^{-1}\mathbf{N}^T\mathbf{W}^{(r)}\big(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(r+1,s+1)}\big), \end{aligned}$$

where $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$, with $\mathbf{X}$ and $\mathbf{N}$ being denoted in the same way, and $\mathbf{W}^{(r)} = \mathrm{diag}\{v_1\mathbf{W}_1, \ldots, v_n\mathbf{W}_n\}\big|_{\boldsymbol{\theta}^{(r)}}$, with $v_i = -2\frac{\mathrm{d}\log g(\delta_i)}{\mathrm{d}\delta_i}$.

Step 2 (Concentrate penalized log-likelihood) Update the parameter $\boldsymbol{\tau}$ by $\boldsymbol{\tau}^{(r+1)} = \arg\max_{\tau}\{L_{\mathrm{p}}(\widehat{\boldsymbol{\beta}}^{(r+1,s+1)}, \widehat{\mathbf{f}}^{(r+1,s+1)}, \boldsymbol{\tau}, \alpha)\}$. Thus, alternating between Stages 1 and 2, this iterative process leads approximately to the MPLE of $\boldsymbol{\theta}$.

3 Local influence analysis

Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$ be an $(n \times 1)$ vector of perturbations restricted to some open subset $\Omega \in \mathcal{R}^n$ and the logarithm of the perturbed penalized likelihood denoted by $L_{\mathrm{p}}(\boldsymbol{\theta}, \alpha \,|\, \boldsymbol{\omega})$. Suppose that there is a point $\boldsymbol{\omega}_0 \in \Omega$ that represents no perturbation of the data so that $L_{\mathrm{p}}(\boldsymbol{\theta}, \alpha \,|\, \boldsymbol{\omega}_0) = L_{\mathrm{p}}(\boldsymbol{\theta}, \alpha)$. According to Cook (1986), the normal curvature in the unitary direction $\boldsymbol{\ell}$ is given by $C_{\ell}(\boldsymbol{\theta}) = -2\{\boldsymbol{\ell}^T\boldsymbol{\Delta}_{\mathrm{p}}^T\mathbf{L}_{\mathrm{p}}^{-1}\boldsymbol{\Delta}_{\mathrm{p}}\boldsymbol{\ell}\}$, where $\mathbf{L}_{\mathrm{p}} = \partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \alpha)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T\big|_{\widehat{\boldsymbol{\theta}}}$ and $\boldsymbol{\Delta}_{\mathrm{p}} = \partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \alpha \,|\, \boldsymbol{\omega})/\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^T\big|_{\widehat{\boldsymbol{\theta}},\, \boldsymbol{\omega}_0}$. In this work we to study the normal curvature in the direction $\boldsymbol{\ell} = \mathbf{e}_i \in \mathcal{R}^n$, where $\mathbf{e}_i$ is an vector with 1 in the ith position and zeros in the remaining positions.

4 Application and discussion

The data set used in this work was reported in a medical study conducted with 30 patients to describe the behaviour of the ocular pressure of the right and left eyes on a specific day (see Ibacache-Pulgar et al., 2011). In some patients it was only possible to measure the ocular pressure in one of the eyes. For the purpose of this work we consider all patients with whom it was possible to measure the pressure in the left eye, totaling 29 patients. The response variables correspond to the measurements of ocular pressure registered at three-hour intervals, that is, at 6am, 9am, midday, 15am, 18am, 21am and midnight. We fit the following semiparametric mixed model:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{N}_i\mathbf{f} + \mathbf{Z}_i b_i + \boldsymbol{\epsilon}_i\,,$$

where $\mathbf{y}_i$ is a vector of responses from the ith patient, $\mathbf{X}_i = \mathbf{1}\,\mathrm{x}_i$ (with x_i denoting the age of the ith patient), $\mathbf{N}_i = \mathbf{I}_7$ is the identity matrix of order 7, $\mathbf{f}$ is a vector whose components are the function $f(\cdot)$ evaluated at the time values in the set $\mathbf{t}^0 = \{\ \mathrm{t}_1^0 = 6,\ \mathrm{t}_2^0 = 9,\ \ldots,\ \mathrm{t}_7^0 = 24\ \}$, b_i is the random effect for the ith patient, $\mathbf{Z}_i = \mathbf{1}$ and $\boldsymbol{\epsilon}_i$ is a random error vector for which we will assume normal distribution and Student-t distribution with $\nu = 8$ degrees of freedom. The age is not significant under both models.

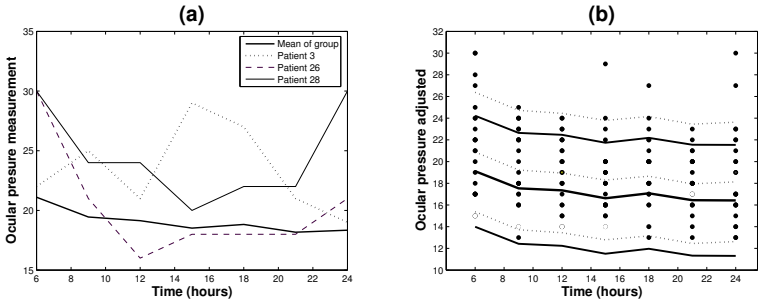


FIGURE 1. Individual profiles of the apparent outliers with the mean profile (a) and the fitted cubic splines ± 1.96 (pointwise) standard errors under Student-t (solid lines) and normal (dotted lines) models (b).

One has in Figure 1a the profiles of some apparent outliers together with the mean profile and in Figure 1b the fitted cubic splines ± 1.96 (pointwise) standard errors are displayed for the two fitted models. We see from the last figure the robust aspects of the MPLEs from the Student-t model whose fitted cubic splines appear to be less sensitive to the apparent outliers than the ones from the normal model. In addition, in order to identify possible influential observations, we will present some local influence graphs. Figure 2 presents the index plots of $C_i = C_{\mathbf{e}_i}(\mathbf{f})$ under the case-weight perturbation scheme. The dotted lines drawn on the graphs correspond to the cutoffs $C_i = 2\bar{\mathcal{C}}$, where $\bar{\mathcal{C}}$ is mean of $\mathcal{C} = \{C_i = C_{\mathbf{e}_i}(\mathbf{f}) : i = 1, \ldots, n\}$. We

see observations $\{3, 26, 28\}$ pointed out under the normal model, but none observation is pointed out under the Student-t model.

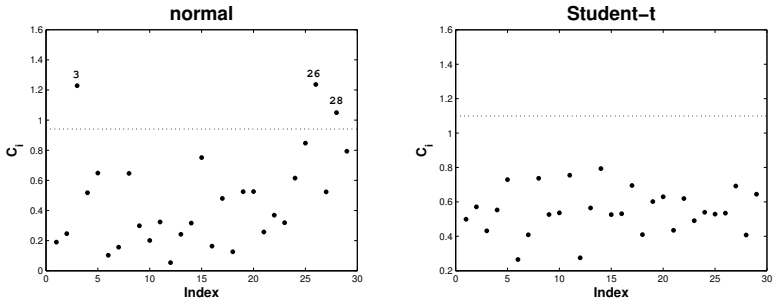


FIGURE 2. Index plots of C_i for assessing local influence on $\widehat{\mathbf{f}}$ under case-weight perturbation scheme under normal and Student-t models.

Thus, we has indication from this example that the well-known robust aspects of the parameter estimates from Student-t model with few degrees of freedom seem to be also extended to the semiparametric mixed case.

Acknowledgments: The authors are grateful to CAPES, CNPq and FAPESP, Brazil.

References

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

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

Ibacache-Pulgar, G., Paula, G. A. and Galea, M. (2011). *Influence diagnostics for elliptical semiparametric mixed models. Statistical Modelling*, Accepted for Publication.

Ibacache-Pulgar, G. and Paula, G. A. (2011). Local influence for Student-t partially linear models. *Computational Statistics and Data Analysis*, **55**, 1462-1478.

The change-point problem in regression with correlated data and change in variance

Gabrielle E. Kelly[1]

[1] School of Mathematical Sciences, University College Dublin, Ireland
email:gabrielle.kelly@ucd.ie

Abstract: Kim and Siegmund (1989) derived an expression for the p-value of the likelihood ratio test for a change-point, where the intercept and slope can change, in simple linear regression. Here we extend their results to correlated data and to where the variance may also change. The accuracy of approximations is assessed using simulations. Results are illustrated with two applications and compared to Bai's (1997) Wald method and to Bayesian methods.

Keywords: Change-point; Regression; Correlated data; Simulation study

1 Introduction

The change-point problem in regression continues to be one of the most interesting and challenging problems in statistics since the paper by Quandt (1958). The interest is partly due to its widespread applicability in many scientific disciplines including epidemiology and economics as shown in Kim and Siegmund (1989). The methodological problems remain challenging because standard maximum likelihood (ml) asymptotic theory does not apply. We consider the model

$$\begin{aligned} y_i &= \alpha_0 + \beta_0 x_i + u_i,\ i = 1, \ldots, \tau \\ y_i &= \alpha_1 + \beta_1 x_i + u_i,\ i = \tau + 1, \ldots, m, \text{ where} \\ u_i &= \rho u_{i-1} + \epsilon_i \text{ and } u_1 = \epsilon_1 \end{aligned} \tag{1}$$

and the ϵ_i are i.i.d. $N(0, \sigma_1^2)$ for $i = 1, \ldots, \tau$ and $N(0, \sigma_2^2)$ for $i = \tau + 1, \ldots, m$ where τ denotes the unknown change-point. We also assume equally spaced x's. The model can also be written in matrix form in the obvious way as

$$Y = X\beta + u, \quad \text{Cov}(Y) = V \tag{2}$$

We consider likelihood ratio tests (lrt's) of the hypothesis of no change, $\text{H}_0 : \beta_0 = \beta_1$ and $\alpha_0 = \alpha_1$, against the alternative H_A: there exists a $j (1 \leq j < m)$ such that $\beta_0 \neq \beta_1$ or $\alpha_0 \neq \alpha_1$.

2 Approximations and Simulations

Firstly, consider the model given in equation (1), where $V = \sigma^2 I$ i.e $\rho = 0.0$ and $\sigma_1^2 = \sigma_2^2$. Let $A_1^T = (1, -1, 0)$, $A_2^T = (0, 1, 0, -1)$, $Y^T = (y_1, \ldots, y_m)$,

$$X_{1i} = \begin{bmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_i \\ 0 & 1 & x_{i+1} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_m \end{bmatrix}, X_{2i} = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_i & 0 & 0 \\ 0 & 0 & 1 & x_{i+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_m \end{bmatrix}$$

Kim and Siegmund (1989) showed -2log(likelihood ratio) statistic for H_0 versus H_A (generalized slightly) is of the form

$$\hat{\sigma^{-2}} \max_{m_0 \le i \le m_1} (U_{1,m}^2(i) + U_{2,m}^2(i)) \tag{3}$$

where $\hat{\sigma^2}$ is the mle of σ^2 under the null model and for $\mu = 1$ or 2

$$U_{\mu,m}(i) \quad = \quad A'_\mu (X'_{\mu,i} X_{\mu,i})^{-1} X'_{\mu,i} Y / [A'_\mu (X'_{\mu,i} X_{\mu,i})^{-1} A_\mu]^{1/2} \tag{4}$$

For $\mu = 1$ or 2, $\lambda = 1$ or 2, let,

$$C_{\lambda,\mu}(i,k) \quad = \quad \sigma^{-2} \quad \text{Cov}[U_{\lambda,m}(i), U_{\mu,m}(k)] \tag{5}$$

Kim (1988) shows that taking limits in this equation gives the covariances of the process $(U_{1,m}([mt]), U_{2,m}([mt]))$, $0 < t < 1$. These will be denoted by $\lambda_{11}(t,s), \lambda_{12}(t,s), \lambda_{21}(t,s), \lambda_{22}(t,s)$. She then shows the probability that the random variable in equation (3) exceeds b^2 is given by

$$p_2 \approx (2\pi)^{-1} b^2 (1 - b^2/m)^{(m-6)/2} \int_{t_0}^{t_1} \int_0^{2\pi} \mu(t,\theta) \nu \left[(\frac{2c^2 \mu(t,\theta)}{(1-c^2)})^{(0.5)} \right] d\theta dt \tag{6}$$

where $c = b/\sqrt{m}$, Φ denotes the standard normal distribution function, $\nu(x) = 2x^{-2} exp\left[-2\Sigma_{n=1}^{\infty} n^{-1} \Phi(-1/(2x\sqrt{n}))\right]$, $x > 0$ and

$$\mu(t,\theta) \quad = \quad \frac{-d}{ds} \lambda_{11}(t,s)|_{s=t} + sin^2(\theta) A_1(t) - cos(\theta) \times sin(\theta) A_2(t), \tag{7}$$

$$A_1(t) = -\left[\frac{d}{ds} \lambda_{22}(t,s)|_{s=t} - \frac{d}{ds} \lambda_{11}(t,s)|_{s=t} \right], A_2(t) = \left[\frac{d}{ds} \lambda_{12}(t,s)|_{s=t} + \frac{d}{ds} \lambda_{21}(t,s)|_{s=t} \right]$$

This simplifies to

$$\mu(t,\theta) \quad = \quad \frac{.5 + [1 - 6t(1-t)] sin^2(\theta) - \sqrt{3}(2t-1) cos(\theta) sin(\theta)}{t(1-t)(1-3t(1-t))} \tag{8}$$

for $V = \sigma^2 I$ and $x_i = i/m$, (i=1,. . . ,m).
Now consider the model (1) where ρ is not necessarily 0 and σ_1^2 not necessarily equal to σ_2^2. It can be shown that it is possible to find a unique non-singular symmetric matrix P such that $P'P = PP = P^2 = V = \mathrm{Cov}(Y)$. Writing $f = P^{-1}u$ then $f \sim N(0, I)$. If we pre-multiply equation (2) by P^{-1} we obtain a new model
$Z = P^{-1}Y = P^{-1}X\beta + P^{-1}u = Q\beta + f$. We apply equations (3-7) to this new model with $\underline{x}$ replaced by $\hat{P_i^{-1}}\underline{x}$ in $Q_{\mu,i}$ where $\hat{P_i^{-1}}$ is the mle of P assuming the change-point is at i, and the degrees of freedom is adjusted to (m-8). To compute the p-value, we evaluate the covariances in equation (5) by sample values, the derivatives in equation (7) by discrete sample approximations and integrals by Riemann sums.
Note using these p-values approximate confidence intervals for the change-point can be found using Worsley's (1986) method. This includes j in a $(1-\alpha)$ confidence region if the lrt's for no change in $[0, j-1]$ and in $[j, m]$ are both accepted at significance levels greater than $1-(1-\alpha)^{0.5} \approx \alpha/2$.

In a 10,000 repetition Monte Carlo experiment, sample sizes m=20 and 40 were considered, with $x_i = i/m (i = 1, \ldots, m)$ and $\rho = 0, .1, .4$, and .7. The 90th, 95th and 99th percentiles of the distribution of the statistic in equation (3) were estimated by means of the experiment and the p-values of equation (6), with (7) and (8) respectively, evaluated at the estimated percentiles. In the case $\rho = 0$ both approximations were similar and for m=40 gave values 0.11,0.05 and 0.01. The approximation (6) with (8) was very poor for correlated data as to be expected while agreement by equation (6) with (7) was good especially for m=40 and smaller probabilities.

3 Examples

3.1 Physiology data

Kelly *et al.* (2001) describe data from healthy subjects undergoing incremental ramp exercise (20 W$\cdot$ min^{-1}) on a bicycle to the limits of tolerance. Oxygen uptake ($\dot{V}O_2$) and carbon dioxide output ($\dot{V}CO_2$) are measured on a breath-by-breath basis. At a point, known as the gas exchange threshold (GET), the linear relationship between $\dot{V}CO_2$ and $\dot{V}O_2$ changes and becomes steeper, as the subject switches from aerobic to a mixture of aerobic and anaerobic metabolism. The GET i.e. change-point is found on (normalised) breath number. For subject 1, the mle of the GET is 198 assuming both correlated observations and a variance change at the change-point and using equation (6) with (7), the p-value of the lrt for a change-point is 0.0006. The associated 95% confidence interval is the single point (198). This is not surprising, as Kelly *et al.* (2001), using a somewhat different model, found the bootstrap distribution of the estimated change-point had a large single mode. Wyse and Kelly (2008) using Bayesian methods with

independent flat priors for the regression parameters, a discrete uniform prior for the change-point and a uniform prior on $(-1, 1)$ for ρ, got a credible interval of (188,200) with an extremely small pseudo-Bayes factor i.e. strong evidence for a change-point. Bai's (1997, Section D, 3., based on asymptotic approximations to Wald type statistics) method gave an interval of (178,205). The differing results indicate that some methodological problems remain even with respect to simple change-point estimation.

3.2 Quandt data

For these simulated independent data, described in Quandt (1958), the exact p-value (by simulation) for the lrt for a change-point is 0.045. Using equation (6) with (8) the p-value is 0.048 and with (7) is 0.047 with 95% confidence interval (5,15) in both cases. Bai's method gave the interval (7,15). Wyse and Kelly (2008) with flat priors reported a credible interval (6,16). As the true p-value of the lrt is close to 0.05 the widest 95% confidence interval is perhaps to be preferred here.

Acknowledgments: This research was partly supported by Science Foundation Ireland grant 06/RFP/MAT024.

References

Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economic Statistics*, **79**, 551-563.

Kelly, G. E., Thin, A. G., Daly, L. and McLoughlin, P. (2001). Estimation of the gas exchange threshold in humans: a time series approach. *European Journal of Applied Physiology*, **85**, 586-592.

Kim, H.-J. (1998). *Change Point Problems in Regression.*. Ph.D. thesis. Department of Statistics, Stanford University.

Kim, H.-J. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression, *Biometrika*, **76**, 409-423.

Quandt, R. E. (1958). The estimation of a parameter of a linear regression system obeying two separate regimes, *Journal of the American Statistical Association*, **53**, 873-880.

Wyse, J. and Kelly, G.E. (2008). A Bayesian analysis of a change-point model with two regimes, *Unpublished manuscript. School of Mathematical Sciences, University College Dublin, Ireland.*

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables, *Biometrika*, **73**, 91-104.

Capabilities of R package `mixAK` for clustering based on multivariate continuous and discrete longitudinal data

Arnošt Komárek[1]

[1] Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic. E-mail: `Arnost.Komarek@mff.cuni.cz`

Abstract: We describe new capabilities of R package `mixAK` designed to perform clustering based on multivariate continuous and discrete longitudinal data using the methodology introduced during the IWSM 2010 presentation. The use of the package will be illustrated on a dataset from a clinical trial on patients with primary biliary cirrhosis.

Keywords: Cluster analysis; Generalized linear mixed model; Functional data; Multivariate longitudinal data; R package.

1 Introduction

Multiple outcomes, both continuous and discrete are routinely gathered on subjects in longitudinal studies. During IWSM 2010 (Komárek, 2010), we introduced a model-based statistical method for clustering (classification) of subjects into a prespecified number of groups with apriori unknown characteristics on basis of repeated measurements of all longitudinal outcomes. The methodology is complemented also by a software implementation, namely extension of the R (R Development Core Team, 2011) package `mixAK` (Komárek, 2009) which was only briefly mentioned in the IWSM 2010 presentation. Hence, it is the main purpose of the poster to show in more details capabilities (extended in the meantime) of the R package `mixAK` for the purpose of clustering based on multivariate continuous and discrete longitudinal data.

2 Methodology

The methodology is based on modelling the evolution of each longitudinal outcome using the classical generalized linear mixed model (GLMM) where we capture possible dependence between the values of different outcomes by specifying a joint distribution of all random effects involved in the GLMM for each response. The basis for subsequent clustering is provided by assuming a heteroscedastic mixture of multivariate normal distributions in

the random effects distribution where each mixture component corresponds to one cluster in subsequent classification. Mainly for computational reasons, the inference is based on a Bayesian specification of the model and simulation based Markov chain Monte Carlo (MCMC) methodology. This allows us to calculate characteristics of the posterior distribution of individual component probabilities (probabilities that a random effects vector for particular subject was sampled from a specific mixture component) which define the classification rule. Not only point estimates represented by posterior means or medians are calculated but also credible intervals which allows us also to evaluate uncertainty in the classification. See Komárek and Komárková (2011) for methodological details.

3 Data and Model

The use of the package will be illustrated on the analysis of the data from a Mayo Clinic trial on 312 patients with primary biliary cirrhosis (PBC) conducted in 1974–1984 (Dickson et al., 1989). We will consider only patients ($i = 1, \ldots, N$, $N = 260$) who survived without liver transplantation the first 910 days of the study and conduct the cluster analysis on basis of the longitudinal measurements of (i) continuous logarithmic serum bilirubin ($Y_{i,1,j}$), (ii) discrete platelet count ($Y_{i,2,j}$), (iii) dichotomous indication of presence of blood vessel malformations in the skin ($Y_{i,3,j}$) available by the pre-specified time point of 910 days. For all markers, $j = 1, \ldots, n_{i,r}$, where $n_{i,r}$, $r = 1, 2, 3$ is the number of available observations for patient i and marker r. The following multivariate GLMM with (i) Gaussian, (ii) Poisson and (iii) Bernoulli distribution, respectively, will be considered to illustrate the use of the clustering procedure based on the observed values $\mathbf{y} = (y_{1,1,1}, \ldots, y_{N,3,n_{N,3}})^\top$ of all outcomes for all patients:

$$\left.\begin{aligned} \mathsf{E}(Y_{i,1,j} \,|\, b_{i,1,1},\, b_{i,1,2}) &= b_{i,1,1} + b_{i,1,2}\, t_{i,1,j}, \\ \log\bigl\{\mathsf{E}(Y_{i,2,j} \,|\, b_{i,2,1},\, b_{i,2,2})\bigr\} &= b_{i,2,1} + b_{i,2,2}\, t_{i,2,j}, \\ \mathrm{logit}\bigl\{\mathsf{P}(Y_{i,3,j} = 1 \,|\, b_{i,3},\, \alpha_3)\bigr\} &= b_{i,3} + \alpha_3\, t_{i,3,j}, \end{aligned}\right\} \tag{1}$$

where $t_{i,r,j}$ is the time in months from the start of follow-up when the value of $Y_{i,r,j}$ was obtained. Further, $\mathbf{b}_i = (b_{i,1,1}, b_{i,1,2}, b_{i,2,1}, b_{i,2,2}, b_{i,3})^\top$ is a vector of patient specific random effects and α_3 is a fixed effect. The model further involves unknown residual variance σ_1^2 from the mixed model on the first line of expression (1), where a Gaussian distribution is assumed. In a sequel, let $\boldsymbol{\psi} = \bigl(\alpha_3,\, \sigma_1^2\bigr)^\top$ be a vector of unknown GLMM related parameters. The random effects vectors $\mathbf{b}_1, \ldots, \mathbf{b}_N$ are assumed to be i.i.d. with a density

$$p(\mathbf{b} \,|\, \boldsymbol{\theta}) = |\mathbf{S}|^{-1} \sum_{k=1}^{K} w_k\, \varphi\bigl(\mathbf{S}^{-1}(\mathbf{b} - \mathbf{s}) \,\big|\, \boldsymbol{\mu}_k,\, \mathbf{D}_k\bigr), \tag{2}$$

where $\varphi(\cdot \,|\, \boldsymbol{\mu}, \mathbf{D})$ is a density of the (multivariate) normal distribution with mean $\boldsymbol{\mu}$ and a covariance matrix $\mathbf{D}$, and $\boldsymbol{\theta} = \big(w_1, \ldots, w_K, \boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_K^\top, \mathrm{vec}(\mathbf{D}_1), \ldots, \mathrm{vec}(\mathbf{D}_K)\big)^\top$ is a vector of unknown mixture related parameters. Finally, $\mathbf{s}$ is a fixed shift vector and $\mathbf{S}$ a fixed diagonal scale matrix which are included in the model mainly due to a possibility of improving the mixing and numerical stability of the MCMC algorithm which is used to obtain a sample from the posterior distribution $p\big(\boldsymbol{\psi}, \boldsymbol{\theta} \,\big|\, \mathbf{y}\big)$ derived from the likelihood and a weakly informative prior distribution $p(\boldsymbol{\psi}, \boldsymbol{\theta})$ for the model parameters.
Mixture model (2) can also be specified hierarchically if we introduce latent component allocations $\mathbf{u} = (u_1, \ldots, u_N)^\top$ and then write $p(\mathbf{b} \,|\, \boldsymbol{\theta}, u = k) = |\mathbf{S}|^{-1}\, \varphi\big(\mathbf{S}^{-1}(\mathbf{b} - \mathbf{s}) \,\big|\, \boldsymbol{\mu}_k, \mathbf{D}_k\big)$, $\mathsf{P}(u = k \,|\, \boldsymbol{\theta}) = w_k$, $k = 1, \ldots, K$. This allows us to develop a clustering procedure which is based on characteristics of the posterior distributions (posterior means, medians, credible intervals) of the individual component probabilities

$$p_{i,k}(\boldsymbol{\psi}, \boldsymbol{\theta}) = \mathsf{P}\big(u_i = k \,\big|\, \boldsymbol{\psi}, \boldsymbol{\theta}, \mathbf{y}_i\big), \qquad i = 1, \ldots, N,\ k = 1, \ldots, K, \tag{3}$$

where $\mathbf{y}_i$ denotes a vector of all observed outcomes for the ith patient.

4 R Package Capabilities

As stated in Introduction, it is the main purpose of the poster to illustrate how the R package `mixAK` can be used to apply the proposed clustering methodology in practice. Suppose that the data are stored in a `data.frame` called `pbc` with columns `lbili`, `platelet` and `spiders` holding the observed values of considered longitudinal markers (one row for each visit), column `id` which identifies the patients and column `month` which gives the time of the visit in months. The sample of size $M = 10\,000$ from the posterior distribution of parameters of model (1) with the mixture distribution (2) for random effects with $K = 2$ components and weakly informative prior distribution with default values for fixed hyperparameters is obtained by running MCMC (1 000 burn-in iterations, 1:100 thinning) using the following command:

```
library("mixAK")
mod <- GLMM_MCMC(y = pbc[, c("lbili", "platelet", "spiders")],
  dist = c("gaussian", "poisson(log)", "binomial(logit)"),
  id = pbc[, "id"],
  x = list("empty", "empty", pbc[, "month"]),
  z = list(pbc[, "month"], pbc[, "month"], "empty"),
  random.intercept = c(TRUE, TRUE, TRUE),
  prior.b = list(Kmax = 2),
  nMCMC = c(burn = 1000, keep = 10000, thin = 100, info = 1000))
```

By extending the `prior.b` argument, the user is able to modify the default values of the parameters of the prior distribution.

It is well known that the posterior distribution is invariant towards $K!$ possible label switching of mixture components which if not taken into account prevents us from using the MCMC sample for clustering. This issue can be solved by applying a suitable re-labelling algorithm (see, e.g., Stephens, 2000) which is provided by running

```
mod  <- NMixRelabel(mod, type="stephens", keep.comp.prob=TRUE)
```

The object `mod` now includes sampled values of model parameters and some derived quantities. These include, among other things, posterior sample of individual component probabilities (3), their posterior means and selected quantiles, all of them needed for clustering, posterior sample of observed data deviances useful for subsequent model selection including selection of a number of mixture components, posterior sample of moments of the mixture distribution (2) which can be used to calculate and plot longitudinal profiles of typical patients, both overally or cluster specific. The package further includes easy to use routines for visualisation and reporting of the results and we illustrate their use on the poster.

Acknowledgments: The work on this paper has been supported by the grant GAČR 201/09/P077, Czech Science Foundation and the grant MSM 0021620839, Ministry of Education, Youth and Sports of the Czech Republic.

References

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989). Prognosis in primary biliary-cirrhosis – Model for decision-making. *Hepatology*, **10**, 1–7.

Komárek, A. (2009). A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Computational Statistics and Data Analysis*, **53**, 3932–3947.

Komárek, A. (2010). Cluster analysis for joint continuous and discrete correlated data. In: *Proceedings of the 25th International Workshop on Statistical Modelling*, Glasgow, Scotland, UK, Bowman, A. W. (Eds.), pp. 291–296.

Komárek, A. and Komárková, L. (2011). Clustering for multivariate continuous and discrete longitudinal data. *Submitted for publication.*

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, **62**, 795–809.

Additive location-scale model when the response and some covariates are interval censored

Philippe Lambert[1]

[1] Institut des sciences humaines, Université de Liège, Boulevard du Rectorat 7 (B31), B-4000 LIEGE. Email: p.lambert@ulg.ac.be

Abstract: An additive model for the location, dispersion and the conditional distribution of a continuous interval-censored response was presented in Lambert (2010). P-splines (Eilers and Marx, 1996) and Bayesian arguments (Jullion and Lambert, 2007) are used to estimate the three components in the location-scale model. Monte-Carlo Markov chains are generated to explore the joint posterior distribution of the spline coefficients and of the penalty parameters controlling the smoothness of the functional components in the model.
We propose to study the impact of Laplace approximations to the posterior distribution of splines coefficients and of the substitution of evidence based estimates for the penalty parameters on the quality of the inference. An extension to deal with interval censored covariates in the additive model will also be presented. We conclude with illustrative examples.

Keywords: Interval censored data ; additive model ; location-scale model ; P-splines ; smooth distribution.

1 Introduction

If Y is a continuous response, X^μ, X^σ a set of continuous covariates (with values on (0,1), say) and Z^μ, Z^σ a set of categorical covariates, the location-scale model assumes that

$$Y = \mu(X^\mu, Z^\mu) + \sigma(X^\sigma, Z^\sigma)\varepsilon \tag{1}$$

where ε is independent of the covariates, $\mu(X^\mu, Z^\mu)$ denotes the unknown regression surface and $\sigma(X^\sigma, Z^\sigma)$ enables to depart from the homoskedastic case.
Assume the following additive model for the conditional location and dispersion of Y_i given $(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)$ $(i = 1, \ldots, n)$:

$$\mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu) \quad = \quad \sum_{j=1}^{J_1} f_j^\mu(x_{ij}^\mu) + \left(\beta_0^\mu + \sum_{j=1}^{p_1} \beta_j^\mu z_{ij}^\mu\right), \tag{2}$$

$$\log \sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma) \quad = \quad \sum_{j=1}^{J_2} f_j^\sigma(x_{ij}^\sigma) + \left(\beta_0^\sigma + \sum_{j=1}^{p_2} \beta_j^\sigma z_{ij}^\sigma \right) \tag{3}$$

Provided that these are smooth, the functional forms in $\mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)$ and $\sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)$ can be approximated using a linear combination of the elements of a (large) B-splines basis $\{s_l(\cdot) : l = 1, \ldots, L\}$ (see Brezger and Lang, 2006, in a GLM setting):

$$f_j^\mu(x_{ij}^\mu) = \sum_{l=1}^{L} s_l(x_{ij}^\mu)\theta_{lj}^\mu \quad ; \quad f_j^\sigma(x_{ij}^\sigma) = \sum_{l=1}^{L} s_l(x_{ij}^\sigma)\theta_{lj}^\sigma.$$

Given $\psi = (\beta^\mu, \Theta^\mu, \beta^\sigma, \Theta^\sigma)$, one can associate to each observation,

$$\{(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu), (\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma), y_i\},$$

the residual $\varepsilon_i(\psi)$ such that

$$\varepsilon_i(\psi) = \frac{Y_i - \mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)}{\sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)}.$$

The location-scale model assumes that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with density f_ε. Using a generous cubic B-splines basis on the support of ε and a partition of that support into a large number (100, say) of consecutive bins $\{\mathcal{J}_j : j = 1, \ldots, J\}$ of equal width Δ with midpoints $u_{j=1}^J$, one can approximate the density through

$$\int_{\mathcal{J}_j} f_\varepsilon(e)de = \pi_j = \frac{\exp([B\phi^*]_j)}{\sum_{\ell=1}^{L} \exp\left([B\phi^*]_\ell\right)} \approx f_\varepsilon(u_j)\Delta$$

(Lambert and Eilers, 2009).
If $n_j = n_j(\psi)$ $(j = 1, \ldots, J)$ denotes the number of observed $\varepsilon_i(\psi)$'s $(i = 1, \ldots, n)$ belonging to bin $\mathcal{J}_j$, then the conditional joint distribution of $(N_1(\psi), \ldots, N_J(\psi))$ is multinomial $\text{Mult}(n; \pi_1(\phi), \ldots, \pi_J(\phi))$. Therefore, the log-likelihood will be

$$\log L(\psi, \phi|\mathcal{D}) = \sum_{j=1}^{J} n_j(\psi) \log \pi_j(\phi).$$

where $\mathcal{D}$ stands for the available data.

2 Penalties and Bayesian formulation

The flexibility provided by the large numbers of B-splines to describe the additive components in location and dispersion as well as the density f_ε can be counterbalanced by a roughness penalty in a frequentist setting

(Eilers and Marx, 1996). In a Bayesian framework, it translates into prior distributions on the spline coefficients:

$$p(\phi|\tau^\phi) \propto (\tau^\phi)^{K/2} \exp\left(-0.5\tau^\phi\ \phi' P^\phi \phi\right), \quad (4)$$

$$p(\theta_j^\mu|\tau_j^\mu) \propto (\tau_j^\mu)^{L/2} \exp\left(-0.5\tau_j^\mu\ (\theta_j^\mu)' P^\mu \theta_j^\mu\right),\ 1 \le j \le J_1 \quad (5)$$

$$p(\theta_j^\sigma|\tau_j^\sigma) \propto (\tau_j^\sigma)^{L/2} \exp\left(-0.5\tau_j^\sigma\ (\theta_j^\sigma)' P^\sigma \theta_j^\sigma\right),\ 1 \le j \le J_2 \quad (6)$$

3 Interval censored responses

Assume that the data take the form $\{(x_i, z_i, (y_i^L, y_i^U)\}$ with an interval for the response. Then, the standardized intervals are $(\varepsilon_i^L, \varepsilon_i^U)$ where

$$\varepsilon_i^L = \frac{y_i^L - \mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)}{\sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)}.\ ;\ \varepsilon_i^U = \frac{y_i^U - \mu(\mathbf{x}_i^\mu, \mathbf{z}_i^\mu)}{\sigma(\mathbf{x}_i^\sigma, \mathbf{z}_i^\sigma)}.$$

If c_{ij} is the proportion of bin $\mathcal{J}_j$ contained in $(\varepsilon_i^L, \varepsilon_i^U)$, then the log-likelihood is

$$\log L(\theta|\text{data}) = \sum_{i=1}^{I} \log\left(\sum_{j=1}^{J} c_{ij}\pi_j\right).$$

A Metropolis-within-Gibbs algorithm can be used to sample the joint posterior. From the generated chain, one can build point estimates and credible regions for the spline parameters and any derived quantity.

4 Further extensions

It also happens that some of the covariates are interval interval censored. Assuming a random design for such a covariate, a B-spline approximation to its density can be set up, see Lambert & Eilers (2009). For each unit of observation, an extra step in the Metropolis algorithm is used to sample a value for the covariate within the reported interval. Conditionally on that quantity, one is back to the setting in Lambert (2010).

The quality of Laplace approximations to the posterior distribution of splines coefficients and the impact of the substitution of evidence based estimates for the penalty parameters on the quality of the inference will also be discussed.

5 Application

The data of interest are the number of marriages in Belgium (in 2006) for given ages of the spouses when the husband already divorced. Ages are reported in one of 11 categories of width 2, 5 or 10 years. Our goal is describe how the distribution of the age of the spouse is changing with that of the partner. The estimated deciles for such a model are shown on Fig. 1 together with the starting contingency table.

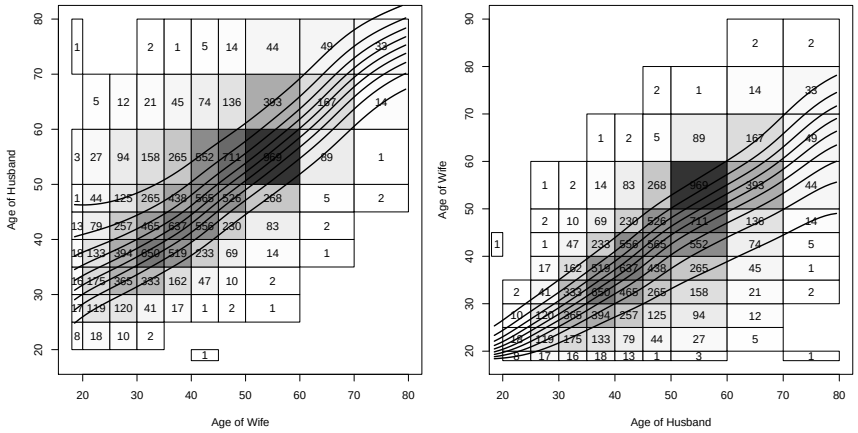


FIGURE 1. Estimated deciles for the age of a spouse conditionally on the age of the partner.

References

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

Jullion, A. and Lambert, P. (2007) Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics and Data Analysis*, **51**: 2542-2558.

Lambert, P. and Eilers, P. H. C. (2009) Bayesian density estimation from grouped continuous data. *Computational Statistics and Data Analysis*, **53**: 1388-1399.

Lambert, P. (2010) Additive model for the conditional location and dispersion of a smooth distribution when the observed data are interval censored, *Proceedings of the 25th International Workshop on Statistical Modelling*, Glasgow, UK, 5-9 July, 2010.

Lambert, P. (2011) Smooth semi- and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data. *Computational Statistics and Data Analysis*, **55**: 429-445.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Second order delta method for estimating the Youden index and optimal threshold

Emilio Letón[1] , Elisa M. Molanes-López[2]

[1] Department of Artificial Intelligence, UNED, C/ Juan del Rosal 16, 28040 Madrid, Spain. E-mail: emilio.leton@dia.uned.es

[2] Department of Statistics, UC3M, Avda. de la Universidad, 30, 28911 Leganés (Madrid), Spain. E-mail: elisamaria.molanes@uc3m.es

Abstract: In medical diagnostics, it is important to measure the effectiveness of a biomarker for classifying individuals in two groups (healthy versus diseased) and to determine the optimal threshold to perform this classification. In order to do so, we propose a second order delta method for estimating the Youden index and its associated threshold value. We also include confidence intervals for both of them. In the simulation study, we compare our new approach with the traditional first order delta method under different scenarios. Finally, the new methodology is illustrated using a real example of prostatic cancer, well-known in the literature.

Keywords: Box-Cox transformation; ROC curve; Youden index.

1 Introduction

The effectiveness of a binary biomarker is described by the sensitivity ('true diseased subjects') and the specificity ('true healthy subjects'). When a continuous biomarker, Y, is used, we need to choose a cut-off ('threshold') value c in order to consider an individual with $Y > c$ as diseased and an individual with $Y \leq c$ as healthy. With the help of a cut-off point c, we can define the sensitivity $q(c)$ and the specificity $p(c)$. Plotting the pairs $(1 - p(c), q(c))$, we construct the 'Receiver Operating Characteristic' (ROC) curve, which is usually summarized with the global index of the area under this curve (AUC) (see, for example, Pepe, 2003). Sometimes there are available several continuous diagnostic variables and it is usual to combine them into a univariate biomarker (see, for example, Pepe et al., 2006, and Ma and Huang, 2007). From here on, we will assume that we are in the univariate case.

A key point in this methodology is to find an optimal threshold, in order to maximize the effectiveness of the biomarker. There are two main methods for identifying the optimal cut-off point: the northwest corner and the Youden index (see Le, 2006, Perkins and Schisterman, 2006, and Letón and Molanes-López, 2009, among others).

The Youden index J, is defined as

$$J = \max\{J(c); c \in \Re\},$$

where

$$J(c) = q(c) + p(c) - 1 = \bar{F}_1(c) + F_0(c) - 1 = F_0(c) - F_1(c),$$

F_0 and F_1 are the cumulative distribution functions (cdf's) of the biomarker Y_0 in the healthy population and of the biomarker Y_1 in the diseased population, respectively, and $\bar{F}_0$ and $\bar{F}_1$ are their complementary ones.
This work is organized as follows. In Section 2, we introduce a second order delta method for estimating J and c, and their confidence intervals. In Section 3, we perform a simulation study under different scenarios, where we compare our approach with the traditional first order delta method. Finally, in Section 4, the new methodology is illustrated using a real example of prostatic cancer, well-known in the literature.

2 Methodology

The main application of the delta method is for constructing approximate confidence intervals (see, for instance, Miller, 1981, Graybill, 1983, and Collet, 2003). In the context of ROC curves, the variance of the Youden index and the associated threshold has been approximated using a first order delta method under the binormal and bigamma models (see Schisterman and Perkins, 2007), providing the following asymptotic $(1-\alpha)100\%$ confidence intervals for J and c:

$$\begin{aligned} CI_{(1-\alpha)100\%}(J) &= \hat{J} \mp z_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}[\hat{J}]}, \\ CI_{(1-\alpha)100\%}(c) &= \hat{c} \mp z_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}[\hat{c}]}, \end{aligned}$$

where $\hat{J}$ and $\hat{c}$ are maximum likelihood estimates of J and c, respectively, and $z_{1-\alpha/2}$ refers to the $(1-\alpha/2)$-quantile of the standard Gaussian distribution, $\mathrm{N}(0,1)$.
In this section, we define a modified version of the delta method based on a second order term on Taylor series expansion. Let $\hat{\theta}$ be the vector of maximum likelihood estimates of the parameters involved in the parametric model assumed for the biomarker. Under the assumption that the distribution of $\hat{\theta}$ is known, the second orden approximated variance of $\hat{J}$ is given by

$$\begin{aligned} \mathbb{V}\mathrm{ar}(\hat{J}) \approx\ & D_{\hat{\theta}}^{J\,T}\Sigma_{\hat{\theta}}D_{\hat{\theta}}^{J} + D_{\hat{\theta}}^{J\,T}E[(\hat{\theta} - E[\hat{\theta}])(\hat{a}_{\hat{\theta}} - b_{\hat{\theta}})] \\ & + \frac{1}{4}E[(\hat{a}_{\hat{\theta}} - b_{\hat{\theta}})^2], \end{aligned} \tag{1}$$

where $\Sigma_{\hat{\theta}}$ is the variance-covariance matrix of $\hat{\theta}$, $D^J_{\hat{\theta}}$ is the Jacobian matrix of J,

$$\hat{a}_{\hat{\theta}} = (\hat{\theta} - E[\hat{\theta}])^T H^J_{\hat{\theta}} (\hat{\theta} - E[\hat{\theta}]), \quad b_{\hat{\theta}} = tr(H^J_{\hat{\theta}} \Sigma_{\hat{\theta}}),$$

with $H^J_{\hat{\theta}}$ denoting the Hessian matrix of J and $tr(A)$ referring to the trace of the matrix A. When $\hat{\theta}$ is normally distributed, (1) can be rewritten in a simple way as follows

$$\mathbb{V}\mathrm{ar}(\hat{J}) \approx {D^J_{\hat{\theta}}}^T \Sigma_{\hat{\theta}} D^J_{\hat{\theta}} + \frac{1}{2} tr\left(\left(H^J_{\hat{\theta}} \Sigma_{\hat{\theta}}\right)^2\right). \tag{2}$$

Analogously, the variance of $\hat{c}$ can be approximated by replacing $D^J_{\hat{\theta}}$ and $H^J_{\hat{\theta}}$ in (1)-(2) by $D^c_{\hat{\theta}}$ and $H^c_{\hat{\theta}}$, the Jacobian and Hessian matrices of c, respectively.

3 Simulation study

A study of interval width, interval coverage and consistency of the point estimates is done through a simulation study based on different sample sizes and several scenarios, previously considered in Fluss et al. (2005). These scenarios cover different real situations, such as symmetry, skewness and distributions outside and inside the Box-Cox transformation family. Details of these scenarios are given in Table 1, where we use the notation $Y = \text{Normal}^{-1/3}$ to indicate that $Y^{-1/3}$ is normally distributed and $Y = \text{Lognormal}$ to indicate that $\ln Y$ is normally distributed. Besides, μ_i, σ_i^2 refer to the mean and variance of a normal distributed population, respectively, and $\alpha_i > 0$ and $\beta_i > 0$ are the shape and scale parameters of a gamma distributed population, respectively, for $i = 0, 1$. The simulations are carried out in MATLAB.

TABLE 1. Parameters under the binormal and bigamma models

Y_0 and Y_1	μ_0	σ_0^2	σ_1^2	μ_1 corresponding to $J(AUC)$		
				0.4 (0.739)	0.6 (0.865)	0.8 (0.958)
Normal	6.5	0.09	0.25	6.873	7.143	7.505
$\text{Normal}^{-1/3}$	3.5	0.09	0.25	3.127	2.857	2.495
Lognormal	2.5	0.09	0.25	2.873	3.143	3.505
	β_0	α_0	α_1	β_1 corresponding to $J(AUC)$		
				0.4 (0.765)	0.6 (0.873)	0.8 (0.956)
Gamma	2	2	2	4.345	7.002	13.828

The second order confidence intervals have good performance in terms of nominal coverage and width, being superior to the first order delta method, recently used by Schisterman and Perkins (2007).

4 Example

The new methodology is illustrated with a real example of 53 patients with prostate cancer: 20 out of them with nodal involvement and 33 without. The biomarker used in this example is the level of acid phosphatase in blood serum (×100). More details of this dataset can be found in Le (2006).

References

Collett, D. (2003). *Modelling survival data in medical research.* Chapman and Hall: Florida.

Fluss, R., Faraggi, D., Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal*, **47**, 458-472.

Graybill, F.A. (1983). *Matrices with applications in statistics.* Duxbury Resource Center: Belmont, CA.

Le, C.T. (2006). A solution for the most basic optimization problem associated with an ROC curve. *Statistical Methods in Medical Research*, **15**, 571-584.

Letón, E. and Molanes-López, E.M. (2009). Adjusted empirical likelihood estimation of the Youden index and associated threshold for the bigamma model. *Statistics and Econometrics Series*, **07**, Working Paper 09-19.

Ma, S. and Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics*, **63**, 751-757.

Miller, R.G.Jr. (1981). *Survival Analysis.* John Wiley & Sons: New York.

Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction.* New York: Oxford University Press.

Pepe, M.S., Cai, T. and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, **62**, 221-229.

Perkins, N.J. and Schisterman, E.F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, **163**, 670-675.

Schisterman, E.F. and Perkins, N.J. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics – Simulation and Computation*, **36**, 549-563.

Modeling growth patterns of the swift tern using nonlinear mixed effect models

Francesca Little[1], Birgit Erni[1], Dismas Ntirampeba[1]

[1] Department of Statistical Sciences, University of Cape Town, Private Bag, Rondebosch 7701, South Africa. *e-mail*:francesca.little@uct.ac.za

Abstract: This paper describes the use of nonlinear mixed effect modeling to fit and compare various growth curves to six different body features of the swift tern.

Keywords: growth curves; nonlinear mixed effect modeling; swift tern.

1 Introduction

The swift tern (*Sterna bergii*) is a nomadic seabird species dispersed around the southern African coastlines(Cooper et al., Hockey et al.,2005). The data used in this study refers to swift tern chicks on Robben Island, off the south-west coast of South Africa. It consists of measurements of body mass(grams), wing length(mm), foot length(mm), head length(mm) and culmen length(mm) taken on several unequally spaced occasions during the period May to June 2001. Chicks were not all measured from day of hatching and thus the time of measurement is not equivalent to age. From a sample of 253 chicks, only 34 chicks were followed from nestling stage and the remainder were first captured when they were already runners (Le Roux,2006).
Of interest was to model the growth patterns of the individual body features and to compare these patterns between features. We fitted and compared several parametric growth functions to the body features, individually and simultaneously.

2 Methodology

Empirical data plots showed that s-shaped and concave growth curves were applicable for our data. Four growth models, Gompertz, logistic, Richards and inverse exponential were fitted using nonlinear mixed effect models to account for the repeated measures within each individual bird using the *nlme* function in R (R Development Core Team, 2006). The choice of best fitting model was based on likelihood ratio tests and Aikaiki's information criterion. Our model building, estimation and validation approach followed

that described by Pinheiro and Bates (2000. In this paper we focus mainly on the methodology and results for the logistic model. Prior to fitting the growth model, we had to estimate age at time of hatching for the individual birds.

2.1 Age determination

For the 34 nestling birds we assumed that age at first capture was two days and fitted a logistic model to the body mass for these nestling birds to obtain estimates of the parameters α, μ and β. We assumed that the growth rate parameters (μ and β) were the same for all birds and that individuals should only differ with respect to their asymptotic weights, thus allowing $\alpha = \hat{\alpha} + \Delta\alpha_i$. The growth curve for the runner birds was thus specified as

$$y_{it} = \frac{\hat{\alpha} + \Delta\alpha_i}{1 + exp(-\frac{t+\Delta t_i - \hat{\mu}}{\hat{\beta}})}.$$

We estimated Δt_i and $\Delta\alpha_i$ by minimizing the sum of absolute residuals, while constraining their values to the following ranges: $\Delta\alpha_i = \hat{\alpha} \pm 55$ and $\Delta t_i = 2 \pm 30$. We used the *optim* function in R (R Development Core Team, 2006).

2.2 Logistic model

We fitted a single logistic growth model for all six body features by adding a categorical covariate to the model that discriminated between the six body features, leading to the following model formulation:

$$y_{ijk} = \frac{\alpha}{1 - exp(-\frac{t_{ijk} - \mu}{\beta})} + \epsilon_{ijk}$$

with

$$\epsilon_{ijk} \sim N(0, R_{ik}),$$

where

$$\alpha = \alpha_1 + \Sigma_{k=2}^{6} \tau_k \alpha_k + b_{1i}$$
$$\mu = \mu_1 + \Sigma_{k=2}^{6} \tau_k \mu_k + b_{2i}$$
$$\beta = \beta_1 + \Sigma_{k=2}^{6} \tau_k \beta k + b_{3i},$$

where τ_k is an indicator variable equal to 1 if feature equals k, zero otherwise (except for k=1 when $\tau_k = 0$) and α_k, μ_k and β_k are differences in parameter values for feature k compared to parameter values for feature 1. We chose body mass as the reference category $k = 1$. Random effects $\mathbf{b}_i$ were assumed to be independent and following a Normal distribution with mean zero and diagonal variance-covariance matrix $\mathbf{D}$. A variance function was fitted to the within bird errors such that the variance increased as a power of the fitted values, $var(\epsilon_{ij}) = \sigma^2 |\mu_{ij}|^{2\delta_k}$ and we imposed a first order autoregressive correlation structure on the within-bird errors.

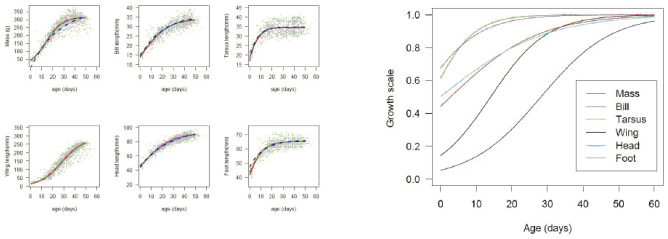


FIGURE 1. (a)Predicted curves obtained from univariate inverse exponential (blue dotted line), univariate (red line) and simultaneous(dotted green line) logistic growth models superimposed on growth data for six body features of Swift terns and (b)Scaled predicted growth curves obtained from the simultaneous logistic growth models for the six body features of Swift terns

3 Results

The results from the simultaneous logistic models are illustrated in Figure 1a. The multivariate logistic model provides an easy and meaningful multiple comparison of growth rates between features as all growth parameters have the same units(days) irrespective of units of the features. Table 1 provide pairwise differences between features with respect to time to reach half of their asymptote values. A feature in a given row is compared to a feature in any column. For instance, 15.07 (second row and first column) indicates that for a wing it took 15.07 days longer to reach half of the maximum wing length than it took for body mass to reach half of its asymptotic value.
For each body feature, the estimates obtained from the logistic model were scaled by dividing the predictions by their asymptotic value. The scaled values were plotted against time to produce Figure 1b, which describes and compares growth of different body features. From this figure it is estimated that ± 20 days after hatching, the foot and tarsus of a tern chick have attained maximum length. For the head, culmen and body mass, it appears that their growth is completed approximately within 50 days. The wing takes longer to reach the maximum length relative to the other body parts and is still not completed at the end of 60 days. From these observations it may be deduced that growth of the swift tern body features follow the following order: (foot,tarsus) - (body mass, bill head) - wing. This growth patterns seems to be justified as it responds to the gradual adaptation of a chick to environmental conditions: adapt to the life in the ceche first (through developed feet and tarsus), followed by developing the capability of getting food on its own(with a developed culmen), and finally the development of wings so that a chick can fly.

TABLE 1. Estimates of differences (in days) (with standard errors) between body features with respect to growth parameter u.

Feature	Mass	Wing	Culmen	Head	Tarsus
Mass					
Wing	14.11(0.30)				
Culmen	-11.08(0.36)	-25.20(0.40)			
Head	-14.13(0.24)	-28.24(0.30)	-2.97(0.39)		
Tarsus	-16.70(0.51)	-30.82(0.54)	-5.66(0.60)	-2.57(0.50)	
Foot	-19.90 0.44)	-34.02(0.48)	-8.89(0.54)	-5.77(0.42)	-3.20(0.59)

4 Discussion

To fit growth curves to multiple responses simultaneously, we have used a different approach from that used in Davidian and Giltinan (1995). Our approach can be used provided the same structural function is valid for all responses. We coped with heteroscedasticity by using feature-specific powers in the variance function. We specified the same within-subject correlation matrix for each feature but we were not able to include estimates of correlations from measurements from different features. The differences in scales for different features is to some extent taken into account by the fixed effect parameters in the model and to some extent by different powers for the variance function.

References

Cooper, J., Crawford, R. J. M. and Williams, A. J. (1990). *Distribution, population size and conservation of the Swift Tern Sterna bergii in southern Africa*. Ostrich 61: 56-65. South Africa: Taylor& Francis

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. New York: Chapman & Hall.

Hockey, P. A. R., Dean, W. R. J. and Ryan, P. G. (2005). *Roberts - Birds of southern Africa*. South Africa:Jacana Media.

Le Roux, J. (2006). *The Swift Tern Sterna bergii in Southern Africa: Growth and Movement*. Unpublished Masters thesis, University of Cape Town.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed Effects Models in S-Plus*. New York:Springer-Verlag.

Zero-Inflated Poisson and Negative Binomial Models Applied to Maternal Mortality Rate in Mozambique

O. Loquiha[12*], M. Aerts[2], L. Chavane[3], M. Temmermans[4]

[1] Department of Mathematics and Informatics, Universidade Eduardo Mondlane, Avenida Julius Nyerere, Campus, 3453 ,P.O. Box 257, Maputo, Mozambique
[2] Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Universiteit Hasselt. Agoralaan 1, B-3590 Diepenbeek, Belgium
[3] Ministry of Health, Mozambique
[4] International Centre for Reproductive Health, Ghent University, De Pintelaan 185 P3, 9000 Ghent, Belgium
* Corresponding author: vadloq06@yahoo.com.br

Abstract: High maternal mortality rate is still one of the main health problems in developing countries. In this study, the objective was to investigate factors related with institutional maternal mortality in Mozambique. We used data from the "Needs in Maternal and Infant Health" survey and applied Zero inflated models for count data to model the mortality rate. Particularly, we compared zero-inflated Poisson with zero-inflated negative binomial and their extensions to account for hierarchy or clustering in the data. Results indicate a better fit for zero-inflated negative binomial with regional differences and rural areas being related to an increase in maternal mortality rate. In addition, the mortality rate tends to increase within health centers with an increase in HIV cases, implying a poor management of these cases within these centers.

Keywords: Maternal mortality, Zero-inflated models, Poisson and Negative binomial distributions

1 Introduction

Since the launch of Safe Motherhood Initiative in 1987 and the addition of maternal mortality in the Millennium Development Goals (MDG 5), maternal mortality has increasingly received a special attention by the various governments worldwide. Mozambique's recent statistics showed a high maternal mortality rate of about 408 deaths per 100 000 live births in 2003, even though the rates tend to decrease since 1990. Obstetric complications are the most common cause of maternal deaths (Romagosa et al., 2007) in the country. This is mostly due to lack of infra-structures and human resources (Cutts et al. 1996), which in many situations requires referrals of patients to larger and better health centers. The transfers to another

centers makes it possible that many other centers will report zero deaths during a period of time, thus the data presenting more excessive zero counts than expected.
In these cases, zero-inflated models have been suggested to model such count outcomes and some extensions (Ridout et al. 1998; Hall, 2000; Lee *et al*, 2006). The most frequently used models are the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB). An inflated model assumes that for each observation, there are two possible data generation processes with different probabilities: one generates the zero and the other the Poisson or negative binomial counts. A Bernoulli model is used to determine which of the two processes is used. The negative binomial distribution is a good alternative to Poisson distribution whenever overdispersion is present. Hall (2000) and Lee *et al* (2006) extended the ZIP models to account for heterogeneity or correlated data, due to multi-level or hierarchical designs by introducing random effects into the ZIP models.
In this study, these models are applied aiming at investigating factors related with the maternal mortality within health centers. Specifically, we investigated the effect of geographical location (region and district), type of health center, existence of emergency obstetric care, waiting house, proportion of HIV and malaria cases (over obstetric admissions), ratio of medical doctors (over total medical staff) on institutional maternal mortality rate defined here as maternal deaths over obstetric admissions. The data used come from the Needs in Maternal and Infant Health survey, with a national wide coverage which included 450 health centers of different types.

2 ZIP and ZINB models

Let Y_{ij} be the number of maternal deaths in the ith province and jth health center ($i = 1, ..., m$; $j = 1, ..., n_i$). In the Poisson model, Y_{ij} is assumed to follow a Poisson distribution with mean and variance $E(Y_{ij}) = V(Y_{ij}) = \mu_i$. From Figure 1, it is clear that the observed zero frequency is more than expected under the Poisson distribution.
An alternative and commonly used approach for modeling excess zero frequency is to assume that Y_{ij} is distributed according to a two component mixture of a Poisson or negative binomial and a degenerated distribution with mass 1 at 0 (Böhning, 1998). The general form of a zero-inflated model is as follows:

$$P(Y_{ij} = y) = \begin{cases} p_i + (1 - p_i)f(y_{ij}) & y_{ij} = 0 \\ (1 - p_i)f(y_{ij}) & y_{ij} > 0 \end{cases}, \qquad (1)$$

with p_i the probability of a zero count and $f(y_{ij})$ the density of either a Poisson or a negative binomial distribution. In this case, for both the Poisson and the negative binomial, the $E(Y_{ij}) = (1 - p_i)\mu_i = \lambda_i$; the Poisson variance equals $V(Y_{ij}) = \lambda_i + (p_i/(1-p_i))\lambda_i^2$ while for the negative

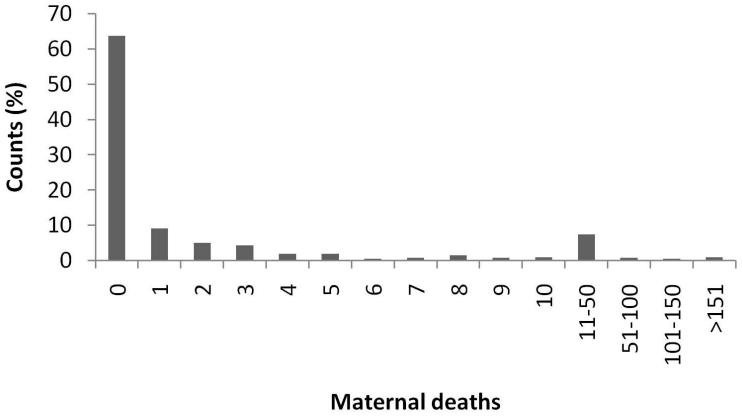


FIGURE 1. Histogram of aggregated number of maternal deaths in health centers in Mozambique from 2006-2007. The counts have been grouped to ease visualization, but only crude counts were used in analysis.

binomial the variance is given by $V(Y_{ij}) = \lambda_i + [(p_i + \rho)/(1 - p_i)]\lambda_i^2$, with $\rho > 0$ the dispersion parameter. The parameters p_i and μ_i can be modeled simultaneously while allowing for covariates effects via a canonical GLM link as:

$$g_1(\boldsymbol{\mu}) = \log(\boldsymbol{\mu}) = \boldsymbol{\eta}_0 + \boldsymbol{X}_1^T\boldsymbol{\beta}, \quad \text{and} \quad g_p(\boldsymbol{p}) = \text{logit}(\boldsymbol{p}) = \boldsymbol{X}_2^T\boldsymbol{\alpha}, \quad (2)$$

where $g(.)$ is a link function linking $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)^T$ and $\boldsymbol{p} = (p_1, p_2, \ldots, p_m)^T$ to the linear predictor and $\boldsymbol{\eta}_0$ is a vector containing the offset effect or $\log(E_{ij})$, with E_{ij} representing the number of obstetric admissions to the health center or women at risk. The design matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ contains the covariates effects which may overlap, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ represents the parameters vectors. The $\boldsymbol{\alpha}$ parameters have interpretations in terms of a covariate's effect on the probability of no intra-health center maternal deaths and the $\boldsymbol{\beta}$'s have interpretations in terms of the effect on the mean maternal mortality rate. The inclusion of covariates in the logit portion of the model is due to the fact that in many applications no prior knowledge about the structural zeros exists.

Due to the hierarchical study design, where hospitals were clustered within provinces, extensions of the above models were considered, with random effects being included in the log portion of the model in accordance with the model suggested by Hall (2000), since no motivations for otherwise could be found. Thus, conditional on the random effects (province effect) $\boldsymbol{b}$, Y_{ij} follows a distribution as in (1) and:

$$g_1(\boldsymbol{\mu}) = \log(\boldsymbol{\mu}) = \boldsymbol{\eta}_0 + \boldsymbol{X}_1^T\boldsymbol{\beta} + \boldsymbol{b}, \quad \text{and} \quad g_p(\boldsymbol{p}) = \text{logit}(\boldsymbol{p}) = \boldsymbol{X}_2^T\boldsymbol{\alpha}, \quad (3)$$

Model	$-2ll$	AIC
Poisson model	2222.87	2260.87
Negative Binomial model	1030.61	1070.61
ZIP model	1634.14	1682.14
ZINB model	**982.0**	**1032.0**
ZIP with random effects (ZIPR)	1469.1	1519.1
ZINB with random effects (ZINBR)	982.0	1034.0

TABLE 1. Model fit comparison for Poisson and Negative Binomial models

where $\boldsymbol{b} = (b_1, b_2, ..., b_m)^T$ is a vector of random effects assumed to be distributed according to a normal distribution with mean 0 and variance σ^2. The model then assumes independence between provinces, but not within.

3 Results

There were 364564 obstetric admissions registered in the sampled health centers (416 excluding missing cases), from which resulted in 2367 maternal deaths (ratio of 649 maternal deaths per 100,000 obstetric admissions). About 68% were due to direct obstetric complications and 32% caused by nonobstetric complications. Only 7.7% of maternal deaths occurred in health centers of class 2 (health centers type II, III and health posts), about 89.9% of all centers sampled, which also included class 1 centers (hospitals and health center type I), much larger and located at the cities or district capital. This low maternal death level may due to the fact that class 2 centers were responsible for approximately 87.5% of referrals due to obstetric complications to class 1 health centers.

Table 1 presents a comparison on the fit for the Poisson and Negative binomials models considered for this analysis. A backward selection procedure was used to select the variables using ZIP for the logistic part of the models, where only significant effects were retained in the model. We also used the proportion of transferred patients (over total obstetric admissions) to correct the estimates of referrals to and from health centers. Models with a zero inflation seemed to improve the fit in both Poisson and Negative binomial models. For ZINB model the inclusion of a province effect as random effect were not significant (p-value=1.0, from $\chi^2_{0,1}$) unlike in ZIP with random effects model (p-value¡0.0001, from $\chi^2_{0,1}$).

Parameter estimates and standard errors were compared for ZINB and ZIP with random effects. The standard errors were generally large in the ZINB as compared to ZIPR though note that the latter model has a hierarchical interpretation. The dispersion parameter in the ZINB model, $\rho = 1.21$, was found to be significant (p-value<0.0001, from $\chi^2_{0,1}$) indicating the adequacy of the model for the overdispersion presented by the data.

or the ZINB model, the odds of reporting no deaths were higher by exp (0.97) 2.6 in the central region as compared to the south region. Also, the odds of reporting no maternal deaths was exp(2.26) 9.5 higher for health centers located outside the district capital than at the capital, and reduced by 17% (exp(-0.19) 0.83) with 1% increase in the malaria cases (p-value=0.09), while controlling for other covariates. For the log part, the expected maternal mortality ratio increased by 10% (exp(0.10)) and 1% (exp(0.01) with 1% increase in the HIV and malaria cases, respectively, for the south region, though the latter was not significant. However, the effect of type of health center was not found to be significant, unlike in the ZIPR model. In addition, there was no significant waiting house or emergency obstetric care effects. A comparison of models fit by plotting the predicted counts versus the observed is presented below for the ZIPR and ZINB models. It can be seen that the ZINB fitted the data considerably better than the ZIPR model, though note that the prediction for the ZIPR model referred to the case where $b_i = 0$ and not population-averaged.

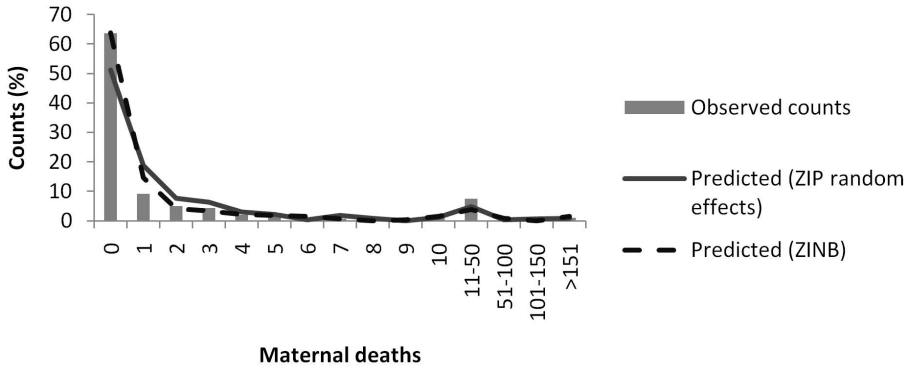


FIGURE 2. Comparison of model fit by plotting observed versus expected counts for the ZIP model with random effects and the ZINB model.

4 Conclusion

This application showed the flexibility of zero-inflated Negative binomial (ZINB) model to handle both overdispersion caused by excess of zero counts as well as lack of independence, compared to zero-inflated Poisson model with and without random effects. An explanation might be the fact that inclusion of a dispersion parameter in the negative binomial model increases the probabilities of both zero counts and non-zero counts, so that inclusion of random effect did not improve the fit. Nevertheless, both models showed

that probability for reporting no maternal deaths depended on geographical location of the health center and proportion of malaria cases. This might be due to a high number of health centers in the sample located outside the district capital which tend to referral most of complicated cases to other facilities. The ZINB model also showed that an increase in the proportion of HIV will tend to increase the maternal mortality rate, reflecting a poor management of patients with this disease. No significant effects for type of health center, existence of emergency obstetric care, ratio of medical doctors and waiting houses were found.

Acknowledgments: This study is part of a Master thesis of Osvaldo Loquiha which was only possible thanks to the financial support of the Flemish Interuniversity Council (VLIR-UOS) in collaboration with Eduardo Mondlane University (UEM) through the DESAFIO Program. The authors would also like to acknowledge the support given by the Mozambican Ministry of Health whom provided the data and research questions.

References

Böhning, D. (1998). Zero-inflated Poisson models and C.A.MAN: A tutorial collection of evidence. *Biometrical Journal*, **40(7)**, 833–843.

Cutts, F. T., dos Santos, C., Novoa, A., David, P., Macassa, G. , and Soares, A. C. (1996). Child and maternal mortality during a period of conflict in Beira city, Mozambique. *International Journal of Epidemiology*, **25(2)**, 349–356.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, **56**, 1030–1039.

Lee, A. H., Wang, K., Scott, J. A., Yau, K.K.W., and McLachlan, G. J. (2006). Multi-level zero-inflated Poisson regression modeling of correlated count data with excess zeros. *Statistical methods in Medical Research*, **15**, 47–61.

Ridout, M., Demétrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. In International Biometric Conference XIX 179–192. Cape Town. Invited papers.

Romagosa, C., Ordi, J., Saute, F., Quintó, L., Machungo, F., Ismail, M. R., Carrilho, C., Osman, N., Alonso, P. L., and Menéndez, C. (2007). Seasonal variations in maternal mortality in Maputo, Mozambique: The role of malaria. *Tropical Medicine and International Health*, **12(1)**, 62–67.

On Bivariate Survival Regression Models

Joseph Lynch[1] , Gilbert MacKenzie[1,2]

[1] Centre for Biostatistics, University of Limerick, Ireland
[2] ENSAI, Rennes, France.

E-mail: joseph.lynch@ul.ie
E-mail: gilbert.mackenzie@ul.ie

Abstract: We compare and contrast the properties of the bivariate Weibull and GTDL regression survival models. An analytic expression for the correlation between times is derived for the Weibull model and a modified Kullback-Leibler distance is proposed for measuring the dependence between times in both models.

Keywords: Weibull, GTDL, Frailty, Dependence between times

1 Introduction

The Weibull model has the Proportional Hazards (PH) property. The Generalised time-dependent logistic model (GTDL), proposed by MacKenzie (1996) is a wholly parametric competitor that can deal with non-PH data. In the univariate time framework, these models were compared by Blagojevic at IWSM in 2003. A cluster is a group of objects sharing a common unobserved characteristic called a frailty. We assume that within the ith bivariate cluster, there is only one unobserved frailty, u_i and that the unobserved frailties follow a common distribution across clusters, ie, $U \sim g(u;\cdot)$. If the frailty term, u_i, were known, the observed bivariate times (t_{1i}, t_{2i}) would be independent, whence $f(t_1, t_2) = f(t_1)f(t_2)$, Hougaard(2000). We further assume that this random frailty effect acts multiplicatively on the hazard functions in the ith cluster.

2 Bivariate Weibull Model

For the jth. object in the ith bivariate cluster, the Weibull frailty hazard function is:

$$\lambda(t_{ij}; u_i, \theta) \quad = \quad u_i \rho \lambda^{\rho} (t_{ij}^{\rho-1}) \exp(x'_{ij}\beta),$$

where $\theta = (\lambda, \rho, \beta)$; $i = 1, \ldots, n$; $j = 1, 2$.

The corresponding survivor function is:

$$S(t_{ij}; u_i, \theta) = \exp\{-u_i \left(t_{ij}^{\rho}\right) \lambda^{\rho} \exp(x'_{ij}\beta).$$

The bivariate density of t_{i1} and t_{i2}, inclusive of frailty, is given by:

$$\begin{aligned} f(t_{i1}t_{i2};u_i,\theta) &= \lambda(t_{i1},t_{i2};u_i,\theta)S(t_{i1},t_{i2};u_i,\theta);\ i=1,\ldots,n; \\ &= \left(u_i\rho\lambda^{\rho}\right)^2(t_{i1}t_{i2})^{\rho-1}\exp\{x'_{i1}+x'_{i2})\beta\} \\ &\quad \times\exp\left[-u_i\{(t_{i1}\lambda)^{\rho}\exp(x'_{i1}\beta_1)+(t_{i2}\lambda)^{\rho})\exp(x'_{i2}\beta)\}\right]. \end{aligned}$$

The frailty density in the ith cluster is given by:

$$g(u_i;\sigma^2)=\frac{u_i^{\frac{1}{\sigma^2}-1}\exp\left(\frac{-u_i}{\sigma^2}\right)}{\Gamma\left(\frac{1}{\sigma^2}\right)(\sigma^2)^{\frac{1}{\sigma^2}}}.$$

This gamma frailty model has shape and scale parameters both equal to $\frac{1}{\sigma^2}$. The expected value of the random effects is $E(U)=\frac{1}{\sigma^2}/\frac{1}{\sigma^2}=1$ and the variance is $var(U)=\frac{1}{\sigma^2}/(\frac{1}{\sigma^2})^2=\sigma^2$. This model was suggested by Clayton (1985) for the analysis of correlation between clustered survival times in genetic epidemiology.

2.1 Marginal Functions

The marginal bivariate Weibull density function of $t_1=t_{i1}$, and $t_2=t_{i2}$, for $i=1,\ldots,n$ is found by integrating the bivariate density over the random effects, $u=u_1,\ldots,u_n$, and is given by:

$$\begin{aligned} f_m(t_1,t_2;\theta) &= (1+\sigma^2)(\rho\lambda^{\rho})^2\exp(x'_1\beta+x'_2\beta)t_1^{\rho-1}t_2^{\rho-1} \\ &\quad \times\left[\sigma^2\lambda^{\rho}\{\exp(x'_1\beta)t_1^{\rho}+\exp(x'_2\beta)t_2^{\rho}\}+1\right]^{-\left(2+\frac{1}{\sigma^2}\right)}. \end{aligned}$$

The marginal bivariate survival and hazard functions are then given by:

$$\begin{aligned} S_m(t_1,t_2;\theta) &= \left[\sigma^2\lambda^{\rho}\{\exp(x'_1\beta)t_1^{\rho}+\exp(x'_2\beta)t_2^{\rho}\}+1\right]^{-\frac{1}{\sigma^2}}. \\ \lambda_m(t_1,t_2;\theta) &= \frac{(1+\sigma^2)(\rho\lambda^{\rho})^2\exp(x'_1\beta+x'_2\beta)t_1^{\rho-1}t_2^{\rho-1}}{\left[\sigma^2\lambda^{\rho}\{\exp(x'_1\beta)t_1^{\rho}+\exp(x'_2\beta)t_2^{\rho}\}+1\right]^2}. \end{aligned}$$

3 Bivariate GTDL Model

For the jth object in the ith bivariate cluster, the GTDL frailty hazard function is:

$$\lambda(t_{ij};u_i,\theta) = \lambda_0 u_i p_{ij},$$

where $p_{ij}=\exp(t_{ij}\alpha+x'_{ij}\beta)\{1+\exp(t_{ij}\alpha+x'_{ij}\beta)\}^{-1}$, $\theta=(\lambda_0>0,\alpha,\beta)$, $i=1,\ldots,n$, and $j=1,2$.

The corresponding survivor function is:

$$S(t_{ij};u_i,\theta) \quad = \quad (q_{ij}g_{ij})^{\frac{u_i\lambda_0}{\alpha}},$$

where $q_{ij} = \{1+\exp(t_{ij}\alpha + x'_{ij}\beta)\}^{-1}$, and $g_{ij} = 1+\exp(x'_{ij}\beta)$.

The bivariate density of t_{i1} and t_{i2}, inclusive of frailty, is given by:

$$f(t_{i1}t_{i2};u_i,\theta) \quad = \quad (u_i\lambda_0)^2\, p_{i1}p_{i2}\,(g_{i1}g_{i2}q_{i1}q_{i2})^{\frac{u_i\lambda_0}{\alpha}}.$$

3.1 Marginal Functions

The marginal bivariate GTDL density function of $t_1 = t_{i1}$, and $t_2 = t_{i2}$, for $i = 1,\ldots,n$ is found by integrating the bivariate density over the random effects, $u = u_1,\ldots,u_n$, and is given by:

$$f_m(t_1,t_2;\theta) \quad = \quad {\lambda_0}^2(1+\sigma^2)p_1p_2\left\{1-\frac{\lambda_0\sigma^2}{\alpha}\log(g_1g_2q_1q_2)\right\}^{-\left(2+\frac{1}{\sigma^2}\right)}.$$

The marginal bivariate survival and hazard functions are then given by:

$$\begin{aligned} S_m(t_1,t_2;\theta) &= \left\{1-\frac{\lambda_0\sigma^2}{\alpha}\log(g_1g_2q_1q_2)\right\}^{-\frac{1}{\sigma^2}}. \\ \lambda_m(t_1,t_2;\theta) &= \frac{\lambda_0^2(1+\sigma^2)p_1p_2}{\left\{1-\frac{\lambda_0\sigma^2}{\alpha}\log(g_1g_2q_1q_2)\right\}^2}. \end{aligned}$$

4 Correlation between times

By calculating the first and second moments of T_1 and T_2, we can evaluate the correlation between the two random variables as a function of the parameters in the Weibull model. Then the correlation ρ^* between T_1 and T_2 is:

$$\rho^*(T_1,T_2) \quad = \quad \frac{\left\{\Gamma\left(1+\frac{1}{\rho}\right)\right\}^2\left[\Gamma\left(\frac{1}{\sigma^2}-\frac{2}{\rho}\right)-\frac{\left\{\Gamma\left(\frac{1}{\sigma^2}-\frac{1}{\rho}\right)\right\}^2}{\Gamma\left(\frac{1}{\sigma^2}\right)}\right]}{\Gamma\left(1+\frac{2}{\rho}\right)\Gamma\left(\frac{1}{\sigma^2}-\frac{2}{\rho}\right)-\frac{\left\{\Gamma\left(1+\frac{1}{\rho}\right)\right\}^2\left\{\Gamma\left(\frac{1}{\sigma^2}-\frac{1}{\rho}\right)\right\}^2}{\Gamma\left(\frac{1}{\sigma^2}\right)}}. \quad (1)$$

The correlation function depends only on the shape parameter, ρ, and the frailty variance, σ^2. It is independent of the scale parameter, λ, and the β parameters of fixed effects.

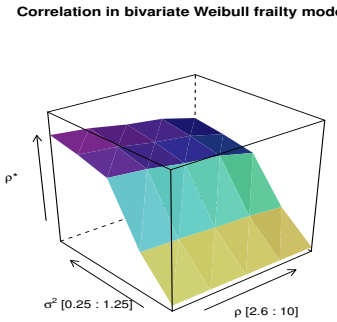


FIGURE 1. Correlation, ρ^*, as a function of frailty variance, σ^2, and of shape parameter, ρ.

In Figure 1, it can be seen, without recourse to simulation, that higher correlation is recorded at lower values of ρ and higher values of σ^2, with variations in the latter having the more potent effect on correlation. It is not possible to find the moments in closed form in the bivariate GTDL frailty model and thus the correlation function cannot be obtained analytically. Figure 2 shows how the sample correlation varies as a function of the α and σ^2 parameters in the bivariate GTDL model with $\lambda_0 = 1$. The correlation increases with increased frailty variance. The time-dependent parameter α has no effect on the correlation.

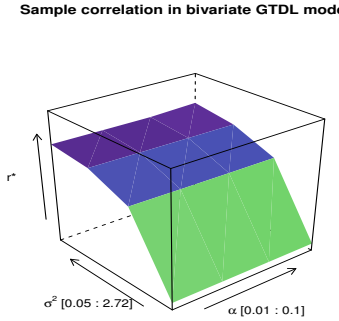


FIGURE 2. 3-d graph of sample estimate of correlation, r^*, between times as a function of parameters in bivariate GTDL model.

5 Dependence

The dependence between times can be measured by a modified Kullback-Leibler (KL) divergence, i.e.:

$$\begin{aligned} \text{KL} &= \log\left\{\frac{\prod_{i=1}^{n} f_m(t_{i1}, t_{i2})}{\prod_{i=1}^{n} f_m(t_{i1}) f_m(t_{i2})}\right\} \\ &= \sum_{i=1}^{n} \log\{f_m(t_{i1}, t_{i2})\} - \sum_{i=1}^{n} \log\{f_m(t_{i1}) f_m(t_{i2})\}. \end{aligned}$$

If times are independent, the KL value is 0, while higher KL values are indicative of dependence between times. Using simulation, we show that this result holds for both the bivariate Weibull and GTDL models. While the analytical correlation in the case of the Weibull model is dependent on ρ and σ^2, the KL value in both models is dependent only on σ^2. In Table 1, we see that as σ^2 increases, the KL divergence and the correlation between bivariate Weibull times increase. The correlation is reduced by higher values of the shape parameter ρ. In Table 2, with $\lambda_0 = 1$, $\gamma_1 =$

TABLE 1. Results for bivariate Weibull frailty model: model parameters (and MLEs), exact Kullback-Leibler(KL) values, sample estimate, r^*, of ρ^* based on simulated data, and MLEs of ρ^* across ten scenarios.

Scenario	$\boldsymbol{\lambda(\hat{\lambda})}$	$\boldsymbol{\rho(\hat{\rho})}$	$\boldsymbol{\sigma^2(\hat{\sigma}^2)}$	**KL**	$\boldsymbol{r^*_{n=200}}$	$\boldsymbol{\hat{\rho}^*_{n=200}}$
1	1.1(1.099)	2.6(2.61)	0.05(0.05)	0.24	0.04	0.05
2	1.1(1.100)	10(10.04)	0.05(0.05)	0.24	0.03	0.03
3	1.1(1.099)	2.6(2.61)	0.37(0.37)	8.87	0.31	0.30
4	1.1(1.100)	10(10.03)	0.37(0.37)	8.89	0.24	0.24
5	1.1(1.100)	2.6(2.61)	1.00(1.00)	38.87	0.73	0.68
6	1.1(1.100)	10(10.00)	1.00(0.99)	38.46	0.57	0.64
7	1.1(1.101)	2.6(2.61)	1.13(1.13)	45.00	0.79	0.84
8	1.1(1.100)	10(10.01)	1.13(1.12)	44.66	0.62	0.61
9	1.1(1.099)	2.6(2.61)	1.24(1.24)	50.78	0.83	0.76
10	1.1(1.099)	10(10.00)	1.24(1.24)	50.54	0.66	0.65

$\gamma_2 = -2$, and sample size $n = 200$, the Kullback-Leibler divergence and sample correlation increase with increased frailty variance.

6 Discussion

The developments above allow us to have PH and non-PH models for bivariate processes and to conduct a number of different types of theoretical and applied comparisons analytically and by simulation. One important

TABLE 2. Simulation of bivariate GTDL frailty model, (with simulation results in brackets), exact Kullback-Leibler divergence and sample Pearson correlation coefficient, r^* with $n = 200$, across ten scenarios.

Scenario	$\alpha(\hat{\alpha})$	$\sigma^2(\hat{\sigma}^2)$	KL	r^*
1	0.01(0.011)	0.05(0.052)	0.27	0.054
2	0.02(0.020)	0.05(0.045)	0.21	0.047
3	0.01(0.012)	0.37(0.375)	8.81	0.339
4	0.02(0.021)	0.37(0.369)	8.83	0.339
5	0.01(0.009)	1.00(0.965)	38.09	0.675
6	0.02(0.017)	1.00(0.935)	39.10	0.680
7	0.01(0.009)	1.13(1.104)	46.34	0.715
8	0.02(0.016)	1.13(1.045)	47.08	0.709
9	0.01(0.009)	1.24(1.190)	51.92	0.727
10	0.02(0.015)	1.24(1.133)	53.23	0.722

area is the nature of the correlation structure supported by the two models. This is relatively easily deduced analytically in the Weibull case, as per equation(1), but simulation is required in the GTDL case.

Acknowledgments: This work was supported, in part, by the SFI's (www.sfi.ie) BIO-SI research programme, grant number, **07MI012**. The first author is an IRCSET Scholar (www.ircset.ie) and the second is the Principal Investigator of BIO-SI (www.ul.ie/bio-si).

References

MacKenzie, G. (1996). Regression models for survival data: the generalised time dependent logistic family. *JRSS Series D*, **45**, 21–34.

Blagojevic M., MacKenzie G. and Ha I.D. (2003). A Comparison of non-PH & PH - Gamma frailty models. *IWSM 2003*, 39–43.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer Verlag, New York.

Clayton, D.G. and Cusick, J. (1985). Multivariate generalisations of the proportional hazards model (with discussion). *J. R. Statist. Soc.A*, **148**, pp 82-117.

Regression graph models: an application to joint modelling of fertility intentions among childless couples

Giovanni M. Marchetti[1], Ilaria Vannini[1], Anna Gottard[1], Daniele Vignoli[1]

[1] Dipartimento di Statistica "G. Parenti", Università di Firenze, viale Morgagni, 59, 50134 Firenze, Italy, giovanni.marchetti@ds.unifni.it

Abstract: We discuss a graphical model for studying the dependence of fertility intentions on several intermediate and background variables. The model, based on a new class of regression chain graphs, suggests a possible generating process and provides simplifications via conditional independencies. With binary responses the models can be parametrized by a sequence of multivariate logistic regression models.

Keywords: Graphical models; chain graphs; multivariate logistic regression.

1 Introduction

Fertility is a major concern for governments in all Western countries. Along with standard determinants of low fertility such as education and labour market situation (cf. Salvini, 2004), housing emerges as a potentially influencing factor. Indeed, some evidence about this relation has been collected in recent years; see Mulder (2006). The Italian situation is an interesting new case of study, being characterized both by a 'lowest low' level of fertility, and by high level of home-ownership; see Vignoli et al. (2010).
In this paper we describe a statistical analysis of fertility intentions and housing using *regression graph models*. These are a new class of chain graph models based on recursive sequences of multivariate regressions that are helpful to understand the development in cohort studies and multi-wave panel data or in cross-sectional data with an assumed ordering of the variables. Regression graph models were introduced by Cox and Wermuth (1993), Wermuth and Cox (2004) and later developed for discrete variables by Drton (2009) and by Marchetti and Lupparelli (2011).

2 Data and variable ordering

The data come from the 2003 Multipurpose Survey "Family and Social Subjects" and concern 710 couples without children. Each partner has

been asked about the fertility intentions and the perceived control over housing conditions, together with several other demographic and socio-economic factors. The main objective is to understand if and how fertility and housing conditions are related using as joint responses the intentions and perceptions of the man and the woman.
As the data are not specifically collected to answer the main question of interest here, but are a subset of a large multipurpose survey, the results obtained are expected to be affected by possible biases. Thus, a certain level of simplification has been used to allow a first understanding of a complex phenomenon, with a model that could be useful to design possible appropriate follow-up studies. The variables selected for the analysis are

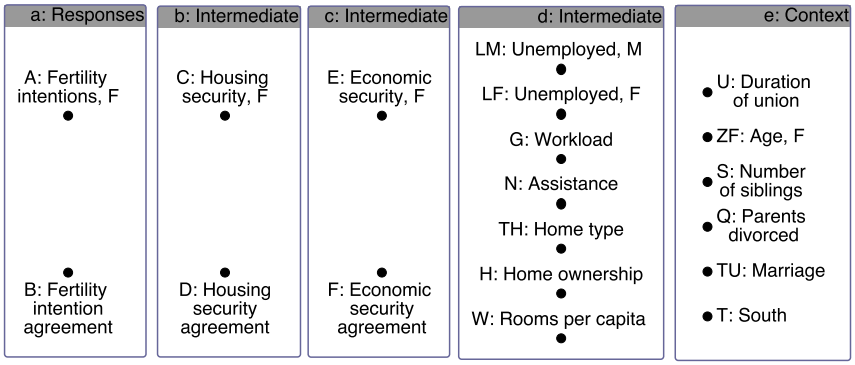


FIGURE 1. Ordering of the variables in a block of joint responses, three blocks of intermediate variables and one block of context variables.

shown in Figure 1 arranged in a series of subsets, called blocks and drawn as boxes. In regression graphs the ordering is an essential part of the model that should agree with time and subject-matter considerations. All the variables within a block are considered to be on an equal standing, while the variables to the right of the block are considered as potential explanatory variables. The first three blocks, a, b and c to the left in Figure 1, contain the main and intermediate responses related to fertility intentions, perception of housing security and economic security, respectively. The original ordinal variables measured for both the woman and the man were transformed in our first analysis into derived variables (see Cox, 1972, p. 117) measuring the opinion of the woman and the agreement with the partner, after dichotomization obtained by aggregating the levels. Thus in block a, A is fertility intention of the woman (1 = surely or probably yes) and B is the partner's agreement (1 = yes). In the last box to the right there are some context variables that have been usually suggested in the literature, like U, the duration of the union, Z the age of the partners, classified as $< 30, 30-40, \geq 40$ years, and T the area of residence in Italy (1 = South). The middle box d contains important intermediate variables that capture

the research hypotheses and permit to trace paths of development. Most of the variables are binary, except age, duration of couple's relationship and number of siblings.

3 Regression graph models

A regression graph is a chain graph defined by a set of nodes $V = \{1, \ldots, d\}$, partitioned into a set of blocks, and by a set of edges E coupling pairs of nodes. The edges within a block are undirected (dashed lines) and between blocks are directed, but always in the same direction, avoiding semi-directed cycles; see Drton (2009). The regression chain graph model is a family of joint distributions of the variables $Y_1, \ldots, Y_d$ associated with the nodes satisfying a set of independencies associated with the missing edges of the graph. Specifically, a missing line uv between two responses in a block means that $Y_u \perp\!\!\!\perp Y_v$ conditional on all variables in preceding blocks, but not on the remaining responses. A missing arrow $u \leftarrow v$ means instead that $Y_u \perp\!\!\!\perp Y_v$ given all other variables in the previous blocks except Y_v. Moreover the joint density of the variables is required to factorize according to the chain graph; see Lauritzen (1996).

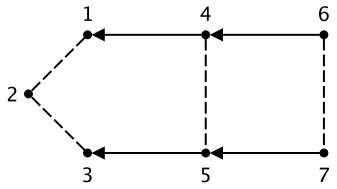


FIGURE 2. A regression graph with three blocks.

For instance the regression graph in Figure 2 defines a model for 7 variables partitioned in 3 blocks $a = \{1, 2, 3\}$, $b = \{4, 5\}$ and $c = \{6, 7\}$ with a factorization

$$f_V = f_{a|b} f_{b|c} f_c \tag{1}$$

thus implying the independence (denoted in a simplified way): $123 \perp\!\!\!\perp 67|45$. Moreover, the structure of the missing edges specifies the independencies $1 \perp\!\!\!\perp 3|45$, $2 \perp\!\!\!\perp 45$, $1 \perp\!\!\!\perp 5|4$, $3 \perp\!\!\!\perp 4|5$, $4 \perp\!\!\!\perp 7|6$ and $5 \perp\!\!\!\perp 6|7$. Notice that this interpretation is unlike that of the classical chain graphs; see Cox and Wermuth (1993) and Drton (2009).

For categorical variables, Marchetti and Lupparelli (2011) have shown that regression graph models can be parametrized by a sequence of *multivariate logistic models*. In this paper we discuss an extension that can incorporate individual discrete and continuous covariates.

For each box of responses we define a linear predictor $\eta_i = X_i\beta$ for the multivariate logistic contrasts $\eta_i = C \log(Mp_i)$ where p_i is the vector of

probabilities for the individual $i = 1, \ldots, n$ and the model matrices X_i are functions of the covariates in the boxes directly influencing the responses. This multivariate link function transforms the vector p_i of probabilities into a set η_i of logits and higher-order log-odds ratios; see Glonek and McCullagh (1995). For instance, with two responses the multivariate logistic contrasts are

$$\eta^{(1)} = \log \frac{p_{2+}}{p_{1+}}, \quad \eta^{(2)} = \log \frac{p_{+2}}{p_{+1}}, \quad \eta^{(12)} = \log \frac{p_{11}p_{22}}{p_{21}p_{12}} \tag{2}$$

where p_{rs}, $r, s = 1, 2$ are the joint probabilities of the responses conditional on the explanatory variables. The independencies specified by the graphical model are obtained by defined appropriate models with special constraints on the βs. Thus, the multivariate logistic model for factor $f_{a|b}$ of (1) is defined by

$$\begin{aligned} Y_1 &: Y_4, \quad Y_2 : 1, \quad Y_3 : Y_5 \\ Y_1Y_2 &: Y_4, \quad Y_1Y_3 : 0, \quad Y_2Y_3 : Y_5, \\ Y_1Y_2Y_3 &: Y_4 * Y_5, \end{aligned}$$

by using an extended model formula notation; see Nelder and McCullagh (1989, section 6.5.5). Therefore, the marginal logits of Y_1 and Y_3 depend on Y_4 and Y_5 respectively, while the logit of Y_2 is constant. On the other hand, the missing edge $(1, 3)$ with associated independence $1 \perp\!\!\!\perp 3|5$, implies that the conditional bivariate logit between X_1 and X_3 is zero. The above equations reflect exactly the independence structure encoded by the multivariate regression Markov property and do not include further simplifying assumptions. Thus, for instance, there is a totally free model for the three variable logistic parameter $\eta^{(123)}$. Usually, however, reduced model are fitted by assuming constant higher order parameters having a more complex interpretation, or omitting the nonsignificant ones.

The full model results by the union of the logistic models defined for each factor of the decomposition (1). These, having variation independent parameters, can be fitted separately by maximum likelihood using appropriate methods. We suggest the efficient algorithm developed by Colombi and Forcina (2001) that can fit general marginal models for mixed ordinal and binary responses, with equality and inequality constraints. In the following section we describe part of the analysis of fertility intentions data, omitting the details concerning the treatment of ordinal data, that will be discussed elsewhere.

4 The analysis

The analysis of data was carried out by recursively fitting multivariate regression models, using individual data, to the responses Y_a given Y_{bcde},

TABLE 1. Maximum likelihood estimates ($\hat{\beta}$) of the bivariate logistic models (3) and (4). The rows labeled with z are the studentized estimates. ZF_2 and ZF_3 are the effects of two categories of age, in baseline coding.

A	const	C	U	ZF_2	ZF_3	Q	TU	G	T	$G \times T$
$\hat{\beta}$	−0.19	0.68	−0.14	−0.41	−1.46	0.66	1.12	0.10	1.46	−1.44
z		2.68	−5.74	−1.46	−3.81	2.11	4.16			−2.30

B	const	TH
$\hat{\beta}$	2.58	−0.66
z		−2.26

AB	const
$\hat{\beta}$	1.51
z	4.56

C	const	E	T	H	LM	U
$\hat{\beta}$	−1.74	3.28	−0.84	1.55	−2.83	−0.05
z		6.43	−3.88	5.58	−2.35	−2.42

D	const	F	H	$F \times H$
$\hat{\beta}$	−0.40	3.00	0.56	−1.77
z				−2.23

Y_b given Y_{cde}, and so on, a, b, c, d, e denoting the blocks in Figure 1. For each regression we used a careful strategy by checking for nonlinear or interactive terms and by retaining significant effects. The fitted models obtained for the first two blocks are:

$$A : C + U + ZF + Q + TU + G * T; \quad B : TH; \qquad AB : 1 \quad (3)$$

$$C : E + T + H + LM + U; \qquad D : F * H; \quad CD : 0 \quad (4)$$

with maximum likelihood estimates and standard errors reported in Table 1. Each term in the formula is associated with an arrow in the graph, i.e., with a substantive effect. One of the research hypotheses of the study, stating that housing, C, has a direct influence on fertility intentions A, after adjusting for duration, age, and area, is confirmed. The marginal model for A includes an interactive effect of area T and presence of heavy workload G: in the regions of southern Italy women's fertility intentions are higher but the effect vanishes if $G = 1$. The complementary part of the model (3) concerning agreement with the partner has a different interpretation, and is related only with home type TH. There is a significant association between the two responses A and B, meaning that the agreement tends to grow with fertility intention.

The bivariate logistic model (4) implies the conditional independence of responses C and D given the explanatory variables, representable by a missing edge between C and D, and is therefore equivalent to two independent logistic models. The likelihood ratio test of the hypothesis of independence is $w = 1.19$ on 1 degree of freedom. Women's perception of

housing security is positively associated with perception of economic security E and with home ownership H and decreases in southern Italy, in the presence of partner's unemployment LM and with a larger duration U of the union. Less clear is the interpretation of the estimates of the logistic model for the agreement about perception of housing security, where there is an interaction implying an inversion of the effect of home ownership H. One interesting consequence of the above discussion is that the regression chain graph model is precisely described by the sequence of multivariate regression model formulae, and that they improve the usual pictorial representation that, with many variables, tends to become rapidly unreadable.

References

Colombi, R. and Forcina, A. (2001). Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, **88**(4), 1007–1019.

Cox, D.R. (1972). The analysis of multivariate binary data. *J. Royal Statist. Soc., C*, **21**(2), 113–120.

Cox, D.R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Science*, **8**, 204–218; 247–277.

Drton, M. (2009). Discrete chain graph models. *Bernoulli*, **15**, 736–753.

Glonek, G.J.N. and McCullagh, P. (1995). Multivariate logistic models. *J. R. Stat. Soc. B*, **57**, 533–546.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.

Marchetti, G.M. and Lupparelli, M. (2011). Chain graph models of multivariate regression type for categorical data. *Bernoulli*, forthcoming. arXiv:0906.2098v2 [stat.ME].

Mulder, C. H. (2006). Population and housing: A two-sided relationship. *Demographic Research*, **15**(13), 401–412.

Salvini, S. (2004). Low Italian fertility: the Bonaccia of the Antilles. *Genus*, **LX**(1), 19–38.

Vignoli, D., Rinesi, F., Mussino, E. (2010). A home to plan the first child? Fertility intentions and housing conditions in Italy. *Conference of the German Society for Demography*, Max Planck Institute, Rostock.

Wermuth, N. and Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *J.R. Stat. Soc. B*, **66**, 687–717.

A spatio-temporal monitoring system for Influenza-Like Illness incidence

M. A. Martínez-Beneito[1], R. Amorós[2], P. Botella-Rocamora[2], D. Gómez-Barroso[3,4], V. Flores-Segovia[3,4] , the Scientific Committee of the GR09/0020 Project

[1] Centro Superior de Investigación en Salud Pública (CSISP), Valencia, Spain
[2] Universidad CEU-Cardenal Herrera, Moncada, Spain
[3] Institute of Health Carlos III, National Centre of Epidemiology, Spain
[4] Ciber Epidemiología y Salud Pública (CIBERESP) Spain

Abstract: In this work we are interested to model the spatio-temporal incidence pattern of Influenza-Like Illness a usual surrogate of Influenza. The main interest of this model has been its application to the information provided by the sentinel surveillance networks integrating the Spanish Influenza Sureveillance System (SISS), which comprises 17 out of 19 Spanish regions. A multi-resolution spatio-temporal kernel process will be used to describe this spatio-temporal pattern. Issues about optimal geographical placement of the nodes of the kernel process will also be discussed at this work. Finally results of the real application of the model to the 2010/2011 Influenza-Like Illness season in Spain will also be shown.

Keywords: Influenza; Spatio-temporal modelling; Kernel smoothing.

1 Introduction

Several proposals have been made for the spatial monitorization of Influenza-Like Illness (ILI) based on sentinel surveillance networks (Carrat & Valeron, 1992; Abellan et al, 2002). The main epidemiological interest on ILI surveillance is motivated because any influenza diagnosis requires a clinical confirmation of the presence of the Influenza virus for all suspected cases. Therefore ILI is a cheap, fast and much less demanding surrogate of influenza incidence surveillance.
Sentinel networks are usually made up of a collection of physicians reporting the number of ILI cases they weekly see at their practices. These physicians submitted reports of all medical visits in their reference populations in accordance with a case definition. Sentinel physicians are geographically spread throughout the whole region of study at some specific locations, therefore we only have information available at several fixed places of the region of study, while we are interested in knowing influenza incidence for every location of that region. For everyone of those locations we have a time

series available with the number of weekly ILI cases reported by everyone of the former physicians.
Spatial monitorization proposals of these kind of data are usually based on geostatistical modelling (Cressie, 1993; Diggle et al., 1998). Nevertheless these approaches show, at least, two important drawbacks. The first of them is that temporal dependence is usually ignored in the modelling process, providing as a consequence highly volatile weekly spatial estimates. That is, spatial estimates varies wildly between consecutive weeks, surely more than would be reasonable to expect. The second one comes from subjective differences among notifiers when classifying a patient either as ILI case or not. These differences make some notifiers systematically to report higher (respectively lower) rates than their colleagues, regardless of the quantity of viruses circulating at every week. This differences can easily distort or mask the spatial pattern of ILI incidence that we are interested to know, indeed this problem has made to question the utility and reliability of previous approaches to the spatial estimation of incidence (Uphoff et al., 2004).
A proposal has been already made in this context circumventing these two problems (Martínez-Beneito et al., 2011). A spatio-temporal model has been proposed for this task where spatial dependence is introduced as a spatial kernel smoothing process and temporal dependence is defined as a first order autoregressive time series on the nodes of the former kernel process. A particular term was also included in the model accounting for the existence of heterogeneity among physicians' notification criteria. This proposal was initially applied to the Valencian Region's Sentinel Network during the winter surveillance seasons corresponding to the 2008/09 and 2009/10 periods. The Valencian Sentinel Network was made up of 48 notifiers during these periods.
The goal of this work is to apply the proposed model, an enhanced version of the Martínez-Beneito et al.'s (2011) model, to the information provided by the surveillance sentinel networks integrated in the SISS during the 2010/2011 season in Spain. The larger number of notifiers at this new context will make it possible to use richer kernel structures than those used for the Valencian surveillance network. Moreover, information provided by the sentinel network of every region will also improve the estimation of the geographic patterns in surrounding regions, therefore this joint study will yield much better estimates than those obtained by separate studies that could be made at everyone of these regions.

2 Methods

A multiresolution spatial kernel process (Higdon, 2002) has been used to induce spatial dependence in our data. This multiresolution kernel process considers smoothing kernels of different spatial ranges therefore this process is able to simultaneously reproduce spatial dependences responding to

very different causations. This multiresolution process has been possible to be implemented due to the amount of data arising from the union of the 17 sentinel networks in the study. All these data make possible to assess the variability corresponding to the different ranges of dependence considered. Temporal dependence will be induced by means of a first order autoregressive process at every one of the nodes of the spatial kernel process. Optimal geographical placement of these nodes will be determined in order to make it the most homogeneous as possible the spatial distribution of the variance of the geographical predictions of the incidence pattern.

3 Results

Results for the ILI surveillance season 2010/2011 will be shown. These results show a high temporal agreement between consecutive weeks, moreover the multiresolution feature of the kernel process used in our model will highlight the presence of different sources of variability in the data of different spatial range. Nevertheless the spatial process of shorter range will be proved to be the spatial component with larger contribution to the whole spatial variability.

4 Conclusions

As described, the Bayesian approach has made possible to implement such a complex model taking into account both spatial and temporal data's dependence and heterogeneity criteria among notifiers. Moreover the proposed kernel approach has made possible to overcome the computational problems that geostatistical Bayesian methods usually show when working with big datasets allowing to make weekly inference (considering all the information for the whole season) in a reasonable amount of time. The multiresolution feature of the proposed kernel structure also makes it possible to define a rich class of spatial processes avoiding to define prior distributions on the parameters of spatial correlation functions. This priors distributions have been stated to be really influential on final results and they are not trivial at all to define, even less in such a complex models as the one introduced at this work.

Acknowledgments: We would like to acknowledge the funding of the Carlos III Health Institute (Project GR09/0020) to carry out this project. We would also like to thank the members of all the sentinel networks taking part in this national project and all the physicians supplying the data without whom this project would not be possible.

References

Abellán, J.J., Zurriaga, O., Martínez-Beneito, M.A., Peñalver, J., and Molins, T. (2002). Incorporación de la metodología geoestadística a la vigilancia de la gripe en una red centinela. *Gaceta Sanitaria*, **16**, 324-333.

Carrat, F., and Valerón A.J. (1992). Epidemiologic mapping using the 'kriging' method: application to an influenza-like illness epidemic in France. *American Journal of Epidemiology*, **135**, 1293-1300.

Cressie, N.A. (1993). *Statistics for Spatial Data. Revised Edition.* New York: John Wiley & Sons.

Diggle, P.J., Tawn, J.A., and Moyeed R.A. (1998). Model based geostatistics (with discussion). *Applied Statistics*, **47**, 299-350.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In: *Quantitative methods for current environmental issues.* 37-54, Springer-Verlag.

Uphoff, H., Stalleicken, I, Bartelds, A., Phiesel, B., and Kistemann, B.T. (2004). Are influenza surveillance data useful for mapping presentations? *Virus Research*, **103**, 35-46.

Bayesian hierarchical modelling for analyzing the efficiency in the European banking system.

Ramón Martínez-Coscollà[1], Carmen Armero[1], David Conesa[1], Emili Tortosa-Ausina[2,3]

[1] Universitat de València
[2] Universitat Jaume I
[3] Instituto Valenciano de Investigaciones Económicas

Abstract: Stochastic metafrontier models are useful models to investigate the technical efficiencies of firms in different groups that may not have the same technology. In this context, efficiencies measured relative to the metafrontier can be decomposed into two components, one measuring the common technical efficiency and another one representing the restrictive nature of the production environment. In this work we propose the use of a Bayesian hierarchical modelling for analyzing these efficiencies. As usual in these models, the estimation of the parameters and hyperparameters is not easy and so techniques such as Monte Carlo Markov Chain (MCMC) methods are needed. We apply these models in a particular banking setting. More precisely, we analyze differences between banks in different countries of the European Union.

Keywords: Hierarchical modelling; Metafrontier; Stochastic frontier analysis.

1 Introduction

Since the mid eighties a great deal of research has been devoted to the study of both the efficiency and productivity of financial institutions all over the world. Despite intense research efforts, though, there is no consensus as to the best method for measuring efficiency in banking. These methods can be broadly classified into parametric and nonparametric methods. Among the former, the most popular is Stochastic Frontier Analysis (SFA); among the latter, the most widely used has been Data Envelopment Analysis (DEA). None of them dominates the other, since both have advantages and disadvantages. They differ in the assumptions they make regarding the shape of the efficient frontier, the existence of random error, and (if random error is allowed) the distributional assumptions imposed on the inefficiencies and random error in order to disentangle one from the other. None of these steams of the literature has stayed still and, in the last few years, there have been appearing new proposals both in the parametric and nonparametric fields.

The basic idea underlying the parametric analyses of efficiency are based on the estimation of a frontier function. In the context of technical efficiency, it is about a production function that indicates the maximum attainable output given some particular inputs. Any lower performances can be traced back to random error -beyond the managers' control- as well as inefficiency. Tipically, after using data on a group of firms to estimate a production frontier, it is common and straightforward to measure the relative performance of firms within the group, but there is often considerable interest in measuring the performance of firms across groups.
O'Donell et al. (2008) developed a formal theoretical framework for making efficiency comparisons across groups of firms using the concept of metafrontier. In this case, efficiencies measured relative to the metafrontier can be decomposed into two components: a component that measures the distance from an input-output point to the group frontier (the common measure of technical efficiency); and a component that measures the distance between the group frontier and the metafrontier (representing the restrictive nature of the production environment). Estimates of technical efficiency are often used to design programs for performance improvement. These programs involve changes to the management and structure of the firm. Estimates of the gap between group frontiers and the metafrontier can also be used to design programs for performance improvement, but these programs involve changes to the production environment. O'Donell et al. (2008) also showed how metafrontiers and group frontiers can be estimated using DEA and SFA techniques.

2 Bayesian modelling of the efficiencies

The Bayesian reasoning is not new in the Stochastic Frontier Analysis literature. Starting with the work of van den Broeck et al. (1994) and continuing with a series of papers from their research group, there has been an increasing interest in this approach (see, for instance, Kim and Smith (2000) for a good review up to that date). The main advantages of its use (easy inference on efficiencies, easy incorporation of prior ideas and restrictions such as regularity conditions and optimal treatment of parameter and model uncertainty) also apply in the context of metafrontier models, specially Bayesian hierarchical modelling.
In particular, the general stochastic production frontier model when there are J different groups can be expressed as:

$$\ln Y_{ij} = \mu_j + \mathbf{x}_{ij}^T \boldsymbol{\beta}_j + V_{ij} - U_{ij} \tag{1}$$

where:

- Y_{ij}: output for the firm i, $i = 1, 2, \ldots, I_j$ in group j, $j = 1, 2, \ldots, J$.
- μ_j: constant term in the regression model.

- $\mathbf{x}_{ij}^T = (x_{ij}^{(1)}, \ldots, x_{ij}^{(N)})$: vector of inputs of firm i in group j.
- $\boldsymbol{\beta}_j = (\beta_j^{(1)}, \ldots, \beta_j^{(N)})^T$: vector of coefficients of regression.
- V_{ij}: random error effect for firm i in group j, $V_{ij} \sim \mathrm{N}(0, \sigma_{V_j}^2)$.
- U_{ij}: nonnegative technical inefficiency component of the error for firm i in group j, $U_{ij} \sim \mathrm{Exp}(\lambda_j)$.

This general modelling can describe different scenarios, depending on the particular characteristics of the different production frontiers. For instance, if we think that all countries have a similar performance, we could consider a pooled model in which $\mu_j = \mu$; $\boldsymbol{\beta}_j = \boldsymbol{\beta}$; $V_{ij} \sim \mathrm{N}(0, \sigma_V^2), \forall i, j$ and $U_{ij} \sim \mathrm{Exp}(\lambda), \forall i, j$. But, if we think that there could be a different behaviour in all them, but that all of them will produce the same output using the same inputs, we could consider then a model in which $\boldsymbol{\beta}_j = \boldsymbol{\beta}$.
As usual in Bayesian statistics, the next step is to assess our prior knowledge about the parameters via their corresponding prior distributions. A good choice when we want to express vague prior knowledge but still use a proper prior is to choose normal distributions for the μ's and $\boldsymbol{\beta}$'s and Gamma distributions for all the σ^2's and λ's. Combining the likelihood and prior information, we obtain the posterior probability distribution of parameters and hyperparameters, which contains all our knowledge about them.
Nevertheless, the resulting expression has not a close form and so numerical simulation from the posterior is needed to perform inference. Among others, the most popular possibility is to use Markov chain Monte Carlo (MCMC) methods, that draw samples from any intractable posterior by running a cleverly constructed Markov chain over a long period, the stationary distribution of which, being the one we want to simulate from. Among the different ways of building these chains, the most popular are Gibbs sampling and Metropolis-Hastings algorithm.
In our case, and taking into account that the conditional distribution of each parameter given the remaining has a close form, the best choice is using Gibb sampling. This algorithm generates an instance from the distribution of each parameter in turn, conditional on the current values of the other parameters. The sequence of samples constitutes a Markov chain, and the stationary distribution of that Markov chain is just the sought-after joint posterior distribution.

3 Bank efficiencies

We apply these models in a particular banking setting. More precisely we analyze differences between banks of the fifteen former countries of the European Union. We analyze the behaviour of the banking system in two particular moments separated by the beginning of the international

financial crisis. Although other options could have been used, our choice for the inputs has been the personnel expenses, the fixed assets and the loanable funds, while for the output the choice has been the loans performed by the banks.

Acknowledgments: Carmen Armero and David Conesa would like to thank the financial support of the Ministerio de Educación y Ciencia (jointly financed with European Regional Development Fund) via the research grant MTM2010-19528 and of the Generalitat Valenciana via the research grant ACOMP11/218.

References

Kim, Y., and P. Schmidt (2000). A review and empirical comparison of Bayesian and classical approaches to inference on efficiency levels in stochastic frontier models with panel data. *Journal of Productivity Analysis*, **14**, 91-118.

O'Donnell, C.J., Prasada-Rao., D.S., and Battese, G.E. (2008). Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. *Journal of Productivity Analysis*, **34**, 231-255.

vanden Broeck, J., Koop, G., Osiewalski, J., and Steel, M. F. (1994). Stochastic Frontier Models: A Bayesian Perspective. *Journal of Econometrics*, **61**, 273-303.

Multidimensional Single-Index Signal Regression

Brian D. Marx[1], Paul H. C. Eilers[2], Bin Li[3]

[1] Dept of Experimental Statistics, Louisiana State University, Baton Rouge, USA
[2] Dept of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands
[3] Dept of Experimental Statistics, Louisiana State University, Baton Rouge, USA

Abstract:
We take a novel approach the signal regression (multivariate calibration) problem, in particular where the signal (spectra) regressors have two dimensional structure. In general linearity is assumed to hold, but this may not be true. Through simultaneous estimation, we parse out and estimate two separate modeling components: (1) a single *smooth regression coefficient surface* associated with the two-dimensional signal, and (2) an unknown, possibly nonlinear, *link function*. Using (tensor product) P-splines for each component, we will see that their combination can lead to a systematic and tractable statistical modeling approach, while having improved external prediction performance when compared to standard signal regression approaches and partial least squares. Optimal tuning will be discussed.

Keywords: Multivariate calibration; P-splines; spectra.

1 Introduction

Our application considers rich and indexed two-dimensional regressor information of UV-VIS spectra taken over several temperatures that are used to predict scalar components of a ternary mixture. We will see that the basic appeal of our particular modelling approach is its explicit estimation of meaningful components: (1) a *smooth regression coefficient surface* associated with the two-dimensional signal (Marx and Eilers, 2005), and (2) an unknown, possibly nonlinear, *link function*. Although the first is linear, the second component explicitly models the nonlinearity, while enhancing insight into the measurement process. Linking the response to the linear predictor is in the spirit of single-index models (Eilers, Li and Marx, 2009).

1.1 First modeling component MPSR

The multidimensional signal regression's (MPSR) goal is to provide a practical solution for functional linear models using the entire two-dimensional signal as regressors. Associated with the regressors is a single overarching

coefficient surface which serves to smoothly weigh each two-dimensional signal digitization. Regularization is needed, and we choose a P-spline approach. Specifically, we take two steps towards smoothness: (a) The coefficient surface (not the signal) is intentionally overfit using two-dimensional tensor product B-splines, making the surface more flexible than needed. (b) Tensor product coefficient estimates are penalized using difference penalties on each of the rows and columns. Given the ith regressor matrix $X_i = [x_{ijk}]$ of dimension $p \times \breve{p}$, signal regressor support on $(v, \breve{v})$, and coefficient surface $\alpha(v, \breve{v})$, express the mean

$$\mu_i = \sum_{j=1}^{p}\sum_{k=1}^{\breve{p}} x_{ijk}\alpha_{jk} = \sum_{j=1}^{p}\sum_{k=1}^{\breve{p}} x_{ijk}\sum_{r=1}^{n}\sum_{s=1}^{\breve{n}} B_{rj}\breve{B}_{sk}\gamma_{rs} = \mathbf{x}_i'\mathbf{T}^\star\gamma, \qquad (1)$$

where $i = 1, \ldots, m$; $j = 1, \ldots, p$; $k = 1, \ldots, \breve{p}$, with tensor product B-splines $\mathbf{T}^\star$, where $\mathbf{x}_i' = \text{vec}(X_i)$. We can further express (1) in matrix form as $\mu = \mathbf{X}\mathbf{T}^\star\gamma = \mathbf{M}\gamma$, where $\mathbf{X}$ is the $m \times p\breve{p}$ matrix of vectorized signals, $\mathbf{M} = \mathbf{X}\mathbf{T}^\star$.
In the P-spline spirit, the objective function is to minimize

$$\begin{aligned} Q_P(\gamma) &= \sum_{i=1}^{m}(y_i - \mathbf{x}_i'\mathbf{T}^\star\gamma)^2 + \lambda\sum_{r=1}^{n}\gamma_{r\bullet}D_d'D_d\gamma_{r\bullet}' + \breve{\lambda}\sum_{s=1}^{\breve{n}}\gamma_{\bullet s}'D_{\breve{d}}'D_{\breve{d}}\gamma_{\bullet s} \\ &= ||y - \mathbf{M}\gamma||^2 + \lambda||P\gamma||^2 + \breve{\lambda}||\breve{P}\gamma||^2, \end{aligned}$$

where $\gamma_{r\bullet}$ ($\gamma_{\bullet s}$) denotes the rth row (the sth column) of Γ. The explicit P-spline solution is

$$\hat{\gamma} = (\mathbf{M}'\mathbf{M} + \lambda P'P + \breve{\lambda}\breve{P}'\breve{P})^{-1}\mathbf{M}'y.$$

Two tuning parameters, associated with the row and column penalties, respectively, allowing continuous control over the surface. The predicted values are $\hat{y} = \mathbf{M}\hat{\gamma}$.

1.2 Second modeling component SISR

The second modeling component is single-index signal regression (SISR), which was presented in Eilers, Li, and Marx (2009) for one-dimensional signals, and is a method that can provide additional insight through the explicit modeling of any nonlinear behavior that may exist with the response. In fact, one could view the standard multivariate calibration problem as using an identity link function, which in actuality may be (slightly) misspecified.In effect, there may exist a true, but "missing link" function (that is nonlinear and monotone) (Cox, 1984), and this approach serves the purpose of estimating this link while improving external prediction. SISR introduces a modification: $\mu_i = f(\sum_{jk} x_{ijk}\alpha_{jk})$. The function $f(\cdot)$

is assumed to be smooth and is estimated from the data using univariate P-splines, having its own additional tuning parameter. This model is generally related to *projection pursuit* (Friedman and Stuetzle, 1981), with additional smoothness demands on α.

Algorithm MSISR

1. Initializations:
   - Choose the tuning parameter values $(\lambda, \breve{\lambda}, \lambda_f)$ for Steps 1 and 2
   - Choose number of knots $(n, \breve{n}, n_f)$
   - Choose penalty order $(d, \breve{d}, d_f)$
   - Set all tuning parameters to λ_0 for the initial Step 1 (default 10^6)
   - Create $\mathbf{M} = X\mathbf{T}^\star$
   - Calculate $\hat{\gamma} = \text{MPSR}(\mathbf{M}, y, (\lambda_0, \lambda_0), (d, \breve{d}), (n, \breve{d}))$
2. Cycle until convergence of $\hat{\gamma}$'s
   - Estimate $\hat{f}$ and the estimate of the derivative $\dot{f}$ from $S(\mathbf{M}\hat{\gamma}, y, \lambda_f, d_f, n_f)$
   - Obtain $y^\star$ and $\mathbf{M}^\star$
   - Update $\hat{\gamma} = \text{MPSR}(\mathbf{M}^\star, y^\star, (\lambda, \breve{\lambda}), (d, \breve{d}), (n, \breve{n}))$
   - Constrain $\hat{\gamma}/||\hat{\gamma}||$
3. Prediction: $\hat{y}^{new} = \hat{f}(x^{new}\mathbf{T}^\star\hat{\gamma})$

end algorithm

1.3 The combined MSISR Methodology

The MSISR model has the form $\mu = f(\mathbf{M}\gamma)$, where the function f and the smooth coefficient surface are unspecified and approximated with P-spline coefficients α and γ. Consequently, the modified MPSR objective can be rewritten as

$$Q_P^\star = ||y - f(\mathbf{M}\gamma)||^2 + \lambda||P\gamma||^2 + \breve{\lambda}||\breve{P}\gamma||^2 + \lambda_f||D_d\alpha||^2. \qquad (2)$$

Given the tensor B-spline coefficient vector γ, the estimation of function f becomes a one-dimensional smoothing problem, and we can apply any scatter-plot smoother to obtain its estimate, which driven by the basis coefficient estimates $\hat{\alpha}$. We estimate f using a (cubic) P-spline scatter smoother (Eilers and Marx, 1996). The penalty on α ensures a smooth f; recall that α is the vector of B-spline coefficients with equally-spaced knots placed along η. Due to the virtue of using B-splines, the first derivative of f (denoted as $\dot{f}$), which is needed in our algorithm, can be easily computed (using a basis with one degree less and first differenced basis coefficients).
Once given an estimate of f, the coefficient vector γ can be estimated using a (first-order) Taylor series approximation of the function f (about

the current estimate, γ_0). Specifically, if γ_0 is the current estimate for γ, then the current estimate of $\mu = f(\mathbf{M}\gamma)$ can be approximated by

$$f(\mathbf{M}\gamma) \approx f(\mathbf{M}\gamma_0) + \dot{f}(\mathbf{M}\gamma_0)\mathbf{M}(\gamma - \gamma_0). \tag{3}$$

Using (3), with fixed f, we have an approximation of $Q_P^\star$

$$\begin{aligned} Q_P^\star &\approx ||y - f(\mathbf{M}\gamma_0) - \dot{f}(\mathbf{M}\gamma_0)\mathbf{M}(\gamma - \gamma_0)||^2 + \lambda||P\gamma||^2 + \breve{\lambda}||\breve{P}\gamma||^2 \\ &= ||y^\star - \mathbf{M}^\star\gamma||^2 + \lambda||P\gamma||^2 + \breve{\lambda}||\breve{P}\gamma||^2, \end{aligned} \tag{4}$$

where $y^\star = y - f(\mathbf{M}\gamma_0) + \dot{f}(\mathbf{M}\gamma_0)\mathbf{M}\gamma_0$ and $\mathbf{M}^\star = \text{diag}\{\dot{f}(\mathbf{M}\gamma_0)\}\mathbf{M}$. Note that (4) implies that given f, the optimal α that minimizes the right-hand side of (4) can be obtained through a MPSR($\mathbf{M}^\star, y^\star, (\lambda, \breve{\lambda}), (D_d, D_{\breve{d}}), (n, \breve{n})$). Hence, in our algorithm, we first carry out a MPSR with the response y on $\mathbf{M}$ (Step 1). Then, given γ, an estimate of f is obtained (Step 2). The two steps, estimation of f and γ, are iterated until convergence of $\hat{\gamma}$.

1.4 Aims and benefits of the combined MSISR approach

The estimation between f and α is iterative and tractable, essentially boiling down to repeated alternate applications of MPSR and P-spline smoothing on "working" responses and regressors. Some additional features of MSISR that are worthy of note include: (a) Although smooth, f can be assumed to be very general, an explicit function can be estimated. (b) Heavy penalization associated with f typically produces low degree polynomial estimates for f. (c) The entire signal can be used as regressors. (d) The number of highly spatially correlated regressors can far exceed the number of observations. (e) The parameterization yields a very manageable system of equations. (f) The candidate coefficient surface can be non-additive. (g) Since the two-dimensional signals and single estimated coefficient surface have a common indexing plane, potentially important regions can be visually identified.

2 Illustration and Optimization

We apply our MSISR to ternary mixture data. The responses are the mole fraction of a mixture, consisting of three components: water, 1,2-ethanediol, and 3-amino-1-propanol. There are 3 pure, 12 edge, and 19 interior (1 center) mixtures. The two-dimensional signal is constructed using the $p \times \breve{p} = 4800$ digitized regressors, X_i, arranged using the (first) differenced UV-spectra, across the temperature levels. The indexing axes that define the support coordinates of X_i are specified as wavelength with $p = 400$ wavelength channels (701 to 1100nm, by 1 nm) and with $\breve{p} = 12$ temperature levels (30, 35, 37.5, 40, 45, 47.5, 50, 55, 60, 62.5, 65, 70^o C). The data were not preprocessed in any other way.

We divided the $m = 34$ observation into three subsets as follows. The training set consisted of $m^{train} = 16$ observations using the 3 pure, 12 edge, and 1 center mixtures. The remaining 18 interior observations were divided into a validation set (to optimize tuning parameters) and a test set (to quantify quality of external prediction). Optimal tuning parameters were determined by minimizing RMSEV in the trained model. Given these optimal tuning parameters, external prediction was evaluated on the test data using RMSEP using the newly trained model that combined both the training and validation data. Table 1 presents the root mean square error of prediction (RMSEP) for the external prediction set, using optimal MSISR, MPSR, and PLS models. For responses water and 1,2-ethanediol, we find an improvement in external prediction for MSISR over both MPSR and PLS, leading to RMSEP reductions that range from 30% to 55%. For MSISR, the external RMSEP values are between 0.0214 and 0.0241, which when multiplied by 100 gives units of percent mixture. Figure 1 displays the optimal MSISR model using the response mixture component 1,2-ethanediol.

Table 1. MSISR, MPSR, PLS external prediction RMSEP, optimal models.

Response	MSISR	MPSR	PLS
Water	0.0214	0.0365	0.0465
1,2-ethanediol	0.0241	0.0338	0.0382
3-amino-1-propanol	0.0306	0.0251	0.0359

3 Discussion

We have shown how to estimate nonlinear relationships in multivariate calibration, by combining the single index model with multidimensional penalized signal regression. We found that the explicit estimation of the nonlinearity can provide some insights into the physical and chemical process underlying the measurements, which we view as a contribution over some of the other more "black box" approaches, while modestly improving external prediction. In the present case the response is assumed to have a normal distribution. Our other current research generalizes SISR, e.g., for binary classification, e.g. a Bernoulli response with probability π_i could be modeled with $\log(\pi/(1-\pi)) = f(X\beta)$. Additionally we are investigating two-dimensional surfaces for f, over another indexing variable, that allows for f to interact with, e.g., temperature.

References

Cox, C. (1984). Generalized linear models- the missing link. *Journal of the Royal Statistical Society, Series C*, **33**, 18-24.

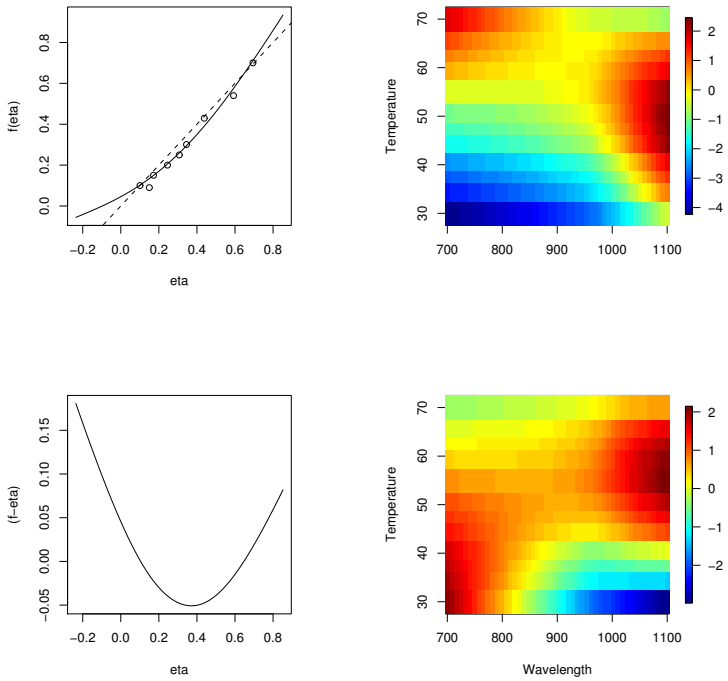


FIGURE 1. 1,2-ethanediol: The estimated $\hat{f}$ function is given (upper, left), along with $(\hat{f} - \hat{\eta})$ (lower, left). The plotted points represent the nine observations in the external test data set. The right panels provide the "optimal" image plots for the estimated coefficient surface (upper) and the coefficient surface difference, MSISR−MPSR (lower).

Eilers, P.H.C., Li, B. and Marx, B.D. (2009). Multivariate calibration with single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, **96**, 196-202.

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

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817-823.

Marx, B.D. and Eilers, P.H.C. (2005). Multidimensional penalized signal regression, *Technometrics* **47**, 13-22.

Multivariate Nonlinear Multi-Level Mixed Effect Models: Techniques and Application to Pharmacokinetic Data

K. Mauff[1], F. Little[1]

[1] University of Cape Town, Department of Statistical Sciences Private Bag, Rondebosch, 7700, South Africa, Katya.Mauff@uct.ac.za

Abstract: This paper discusses the various techniques involved in the fitting of nonlinear mixed effect models, with extensions to the multi-level case and multiple responses. It looks at the application of these techniques to pharmacokinetic data.

Keywords: non-linear mixed effect; pharmacokinetics; multiple responses

1 Introduction

World Health Organization (WHO) statistics indicate that roughly 3.3 billion people are at risk of malaria, with 250 million cases and nearly one million deaths annually, most of which are in Sub-Saharan Africa, (6).
Current recommendations for the prevention of Malaria infection include vector control (through the use of insecticide treated nets and indoor residual spraying) and in pregnant women, intermittent preventative treatment (IPTp) using Sulfadoxine-Pyrimethamine (SP), (2, 6)
Despite this recommendation and widespread use of IPTp treatment with SP, there is limited information regarding the pharmacokinetics of the drugs in pregnancy, information which is necessary to better inform/justify the dosage regimes currently employed.
The objectives of this particular study were thus to characterize the behaviour and disposition of the individual (parent) drug compounds, and because of the simultaneous administration of the drugs, to determine the extent and mechanism of their interaction and inter-dependence on one another. The main purpose in the collection of the data was to characterize the impact of pregnancy induced physiological changes, and pregnancy related factors.

2 Study Design and Data

The analyses described in this paper were conducted using data from a prospective multi-center study comprised of 98 self-matched pregnant

women from four different sites.
Data was initially collected for 31 pregnant women in Mozambique, with the same women returning postpartum to act as their own controls. Using a similar protocol, a study was undertaken in Sudan, with 25 self-matched pregnant women, and again in Mali and Zambia with 18 and 25 self-matched pregnant women respectively. Of the original 98 subjects, 77 returned to complete the postpartum phase of the study, and hence the data is unbalanced.
All subjects were healthy volunteers, undergoing IPTp with SP as part of routine pre-natal care, with maternal age 18-45, and gestational age 15-36 weeks.
There are differing measurement occasions for different sites and pregnancy phases, with sparse sampling for the first 24 hours post dose in pregnant subjects from Mozambique and Sudan, and at most two measurements in the same women postpartum, as concentrations were measured on days 0 and 7 only.
The two drugs under evaluation are simultaneously administered, and we thus have multiple inter-dependent responses. The relationship between the two drugs needs to be correctly quantified in order to accommodate the simultaneous modeling of PK and PD data, which is necessary to determine the concentration-effect relationship (although this is not undertaken here).
Most pertinently, because for each randomly sampled individual we have serial measurements for two observation phases, we have *multiple* levels of grouping, nested within each other. The correlation structure of the data is thus inherently more complicated. This leads to the following levels of random effects:

- Inter-individual variation (IIV), (level 1)
- Intra-individual, inter-occasion (IOV), (level 2)
- Intra-individual, intra-occasion (WIV)- residual measurement error, (level 3: innermost level)

3 Methodology

The technique referred to in this paper is that of non-linear mixed effect (NLME) multi-level models, which are extended to the multivariate platform through the use of sequential and simultaneous modeling procedures using the nlme package in R software.
The structural model forms usually seen in population pharmacokinetic modeling are based on compartmental assumptions: the body is assumed to be made up of a system of compartments, between which the drug is transferred, (5). An alternative approach to this "semi-mechanistic" model (1) is to use an additive series of exponential functions, almost analogous

to the compartmental models, but free of any clinical or pharmacological presuppositions.
The empirical model form describes the concentration-time profiles of specific drug substances reasonably well; assigning an exponential term to each differential phase of the curve as determined by the incline, decline or multiple slopes of decline, (1).
In the individual case, in addition to the quantification of the impact of pregnancy, results from multi-level (nested) random effects were contrasted to those achieved with single level models using an explicitly specified correlation structure for the random effects. Various structures were explored for the modeling of heterogeneous residual variance and we looked at the impact of different parameterizations on the convergence and stability of the models.

3.1 Sequential Model Formulation

The sequential model formulation involves the incorporation of the predicted concentrations of one response in the covariate model of the other as a time-varying covariate. Thus, modeling the impact of predicted Pyrimethamine concentrations on the concentration of Sulfadoxine, (the hypothesized relationship), we could specify, for example:

$$y_{ikj} = \beta_{0ikj} \times [-exp(\beta_{1ikj} \times time_{ikj}) + exp(\beta_{2ikj} \times time_{ikj})] + e_{ikj},$$

where

$$\begin{aligned}
\beta_{0ikj} &= \beta_0 + \beta_3 \times pyr_{ikj} + b_{0i} + b_{0ik}, \\
\beta_{1ikj} &= \beta_1 + \beta_4 \times pyr_{ikj} + b_{1i} + b_{1ik}, \\
\beta_{2ikj} &= \beta_2 + \beta_5 \times preg_{ik.} + b_{2ik},
\end{aligned}$$

and y_{ikj} is the Sulfadoxine concentration at the jth time for the kth observation phase on the ith individual. The variance covariance matrices for the random effects could then be given by:

$$b_i \sim N(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau_0^2 & 0\\0 & \tau_1^2\end{bmatrix}), \quad b_{ik} \sim N(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\tau_{0k}^2 & 0 & 0\\0 & \tau_{1k}^2 & 0\\0 & 0 & \tau_{2k}^2\end{bmatrix})$$

and:

$$e_{ikj} \sim N(0, \mathbf{R_{ik}}), \quad var(e_{ikj}) = \sigma^2 \times ((\theta_1 + {\mu_j}^{\theta_2})^2).$$

3.2 Simultaneous Model Formulation

The simultaneous model formulation which allows for different structural model forms accommodates the association among factors corresponding to the two responses, and allows for greater precision in the estimation of

common elements, (3, 4). The variance-covariance matrices for the random effects and the residuals allow for correlations among group-specific regression parameters for the different response types and within-individual within-phase correlations respectively, (3, 4). Bi- and triple-exponential models were specified for Sulfadoxine and Pyrimethamine respectively, where the response type was indicated by a binary variable.
Additionally, an alternative approach was followed whereby the same structural form was applied to both responses, and the binary response type was included as a covariate. The model formulation presented below is an example for the latter case.

$$\begin{aligned} y_{ikj} = \beta_{2ikj} \times [-exp(-\beta_{1ikj} \times time_{ikj}) + exp(-\beta_{3ikj} \times time_{ikj})] \\ + \beta_{4ikj} \times [-exp(-\beta_{1ikj} \times time_{ikj}) + exp(-\beta_{5ikj} \times time_{ikj})], \end{aligned}$$

where:

$$\begin{aligned} \beta_{1ikj} &= \beta_1 + \beta_6 \times \delta_{ikj}, \\ \beta_{2ikj} &= \beta_2 + \beta_7 \times \delta_{ikj} + \beta_8 \times preg_{ik.} + b_{2i} + b_{2ik}, \\ \beta_{3ikj} &= \beta_3 + \beta_9 \times \delta_{ikj} + b_{3i}, \\ \beta_{4ikj} &= \beta_4 + \beta_{10} \times \delta_{ikj}, \\ \beta_{5ikj} &= \beta_5 + \beta_{11} \times \delta_{ikj}, \end{aligned}$$

and y_{ikj} is the concentration at the jth time for the kth observation phase on the ith individual, and δ is an indicator variable coded as 0 for Pyrimethamine and 1 for Sulfadoxine. The variance-covariance matrices for the random effects $\mathbf{b_i}$ and $\mathbf{b_{ik}}$ for this case would be positive-definite block-diagonal matrices, for subject-specific random effects $\mathbf{b_i} = [b_{2i}, b_{3i}]'$ and occasion-specific random effects $\mathbf{b_{ik}} = b_{2ik}$.

4 Results

Preliminary results for the particular parameterization employed indicate that physiological changes during pregnancy play a differing role in determining both the range of concentrations reached with the same dosing regimen, and in the elimination of the drug from the system for the different compounds.
For the separately fitted models, bi- and triple-exponential models were deemed appropriate for Sulfadoxine and Pyrimethamine respectively. Both single level correlated random effects and multi-level nested models achieved the same results, and the parameterization of the random effects (fit in a linear/non-linear context, the latter using logged parameters) appeared to play a significant role in the ease of convergence.
Results from a model adjusted for pregnancy only indicate that pregnant subjects have a higher range of Sulfadoxine concentrations and a faster rate

of decline. The overall effect of this is a reduction in the total exposure to the drug, as measured by the area under the concentration-time curve. This is illustrated in Figure 1.
The impact of pregnancy on Pyrimethamine concentrations is to similarly increase the range, although no effect could be determined for the various rate constants in this model. Pregnancy was modeled in terms of trimester (postpartum vs. trimesters 2 and 3) for Pyrimethamine.

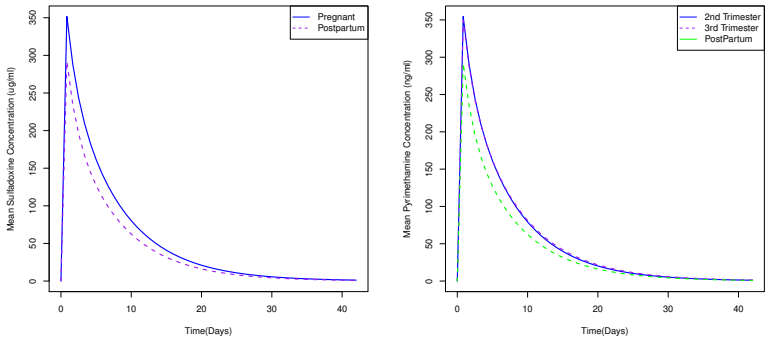


FIGURE 1. Mean Predicted Concentrations over Time Stratified by Pregnancy/Trimester

In the sequential modeling Pyrimethamine appears to influence the absorption properties of Sulfadoxine (predominantly increasing the overall range of concentration reached), rather than any effect vice versa.
In the simultaneous model in which the same structural formulation was applied to both responses, a triple-exponential model was successfully fitted to both Sulfadoxine and Pyrimethamine, where before the data for Sulfadoxine alone did not support the fitting of the additional exponential term.
In this model specification, the parameter estimates of the indicator variable δ are interpreted as effect modifications specifying the change to the Pyrimethamine model, which is the reference category. The model looking at the effect of pregnancy alone indicates significant differences between the Sulfadoxine and Pyrimethamine curves, and we note particularly that the parameters defining the final exponential term are much reduced for Sulfadoxine. Pregnancy again appears to impact the range of concentrations achieved, for both drugs, with no significant difference in the size of the effect for the two drugs. The impact of pregnancy on the rate of decline observed in the individual Sulfadoxine model is no longer indicated.

5 Discussion

We have demonstrated that the use of non linear mixed effect modeling techniques provides a flexible framework for the estimation of separate, sequential and simultaneous drug concentration-time models. We have been able to draw conclusions regarding both the impact of pregnancy on the concentration-time profiles of the individual compounds, as well as examine the interaction between the two drugs.
We chose to fit these nonlinear relationships using a general formulation based on sums of exponentials, rather than the traditional PK parameterization. Relevant PK parameters were retrospectively determined via back-transformation, and their respective standard errors calculated using the Delta method. These clinical parameters were compared with those obtained from a traditional PK analysis on the same data set (where population parameters are obtained from the averaging of parameters from individual-specific models). The mean parameters from the different analyses were similar, and the standard errors resulting from the nlme models significantly reduced. The models were found to be very sensitive to starting values: stable results were obtained using estimates from a curve-stripping procedure. Model building was largely hypothesis driven, since concerns arose regarding the calculation of the degrees of freedom, and an all-subsets procedure was considered computationally intensive and infeasible.

Acknowledgments: Special thanks to the SEACAT project and Professor Karen Barnes (UCT) for the provision of the data and clinical input.

References

Aarons, L. (2005). Editor's View. Physiologically based pharmacokinetic modelling: a sound mechanistic basis is needed. *British Journal of Clinical Pharmacology*, **60:6**, 581-583.

CDC Malaria During Pregnancy [Online] [Cited: May 30, 2009.], *http://www.cdc.gov/malaria/pregnancy.htm*

Davidian, M., and Giltinan, D. (1998). *Non Linear Models for Repeated Measurement Data*. London: Chapman & Hall.

Pinheiro, J., and Bates, D. (2002). *Mixed Effect Models in S and S-Plus*. Statistics and Computing. New York: Springer-Verlag.

Wagner, J.G. (1975). *Fundamentals of Clinical Pharmacokinetics*. Illinois: The Hamilton Press, Inc., Hamilton.

WHO Malaria [Online] [Cited: May 30, 2009.] *http://www.who.int/topics/malaria/en/*

Boosting Generalized Additive Models for Location, Scale and Shape

Andreas Mayr[1], Nora Fenske[2], Benjamin Hofner[1], Thomas Kneib[3] , Matthias Schmid[1]

[1] Department of Medical Informatics, Biometry and Epidemiology, FAU Erlangen-Nürnberg, Waldstr. 6, 91054 Erlangen, Germany
andreas.mayr@imbe.med.uni-erlangen.de
[2] Department of Statistics, Ludwig-Maximilians-Universität München, Germany
[3] Department of Mathematics, Carl von Ossietzky Universität Oldenburg, Germany

Abstract: Generalized additive models for location, scale and shape (GAMLSS) are a popular semi-parametric modelling approach that, in contrast to conventional GAMs, regress not only the mean but every parameter of a conditional response distribution (e.g. location, scale and shape) to a set of covariates. Current fitting procedures for GAMLSS are infeasible for high-dimensional data setups and require variable selection based on (potentially problematic) information criteria. The present work describes a boosting algorithm for high-dimensional GAMLSS that was developed to overcome these limitations. The proposed algorithm was applied to data of the Munich Rental Guide. The net-rent predictions that resulted from high-dimensional GAMLSS were found to be highly competitive while covariate-specific prediction intervals showed a major improvement compared to classical GAMs.

Keywords: GAMLSS, high-dimensional data, gradient boosting.

1 GAMLSS

Generalized additive models for location, scale and shape (GAMLSS) were introduced by Rigby and Stasinopoulos (2005) as a class of statistical models for regression problems with univariate response. GAMLSS can be seen as a flexible alternative to generalized additive models (GAMs) as they extend the traditional GAM framework through a variety of modelling options. Every parameter of the conditional response distribution is modelled by its own predictor and an associated link function. While traditional GAMs are typically restricted to modelling the conditional *mean* of the response variable (treating other distributional parameters as fixed), the GAMLSS approach allows for the regression of each distribution parameter on the covariates.

A GAMLSS is given by the set of equations

$$g_k(\theta_k) = \eta_{\theta_k} = \beta_{0\theta_k} + \sum_{j=1}^{p_k} f_{j\theta_k}(x_{kj}), \quad k = 1, \ldots, 4, \tag{1}$$

where $\beta_{0\theta_k}$, $k = 1, \ldots, 4$ are the intercept values of the four submodels. The function $f_{j\theta_k}$ for $j = 1, ..., p_k$ represents the effect of covariate j on the distribution parameter θ_k. Examples include non-parametric terms based on penalized splines, varying coefficient terms, spatial and subject-specific terms for repeated measurements. The estimation of GAMLSS is usually based on penalized likelihood maximization available with the R add-on package **gamlss** (Rigby and Stasinopoulos, 2005). For variable selection, the authors propose the Generalized Akaike Information Criterion (GAIC). This approach, however, has several shortcomings that are partially inherited from problems associated with the traditional AIC. Additionally, the conventional fitting algorithm cannot include spatial effects and it is not feasible in high-dimensional data settings.

2 The gamboostLSS algorithm

To address these issues, we developed and subsequently applied a boosting technique (*gamboostLSS*) for estimating and selecting the predictor effects in GAMLSS. We propose a component-wise gradient descent algorithm (Bühlmann and Hothorn, 2007) that circles between the different prediction functions of the distribution parameters for GAMLSS. Analogously to the classical gradient descent algorithm, gamboostLSS can handle high-dimensional data settings ($p >> n$) and includes intrinsic variable selection. To extend the classical boosting approach to the GAMLSS framework, we adopted a strategy recently proposed by Schmid et al. (2010): In each iteration, gamboostLSS calculates the negative partial derivatives of the negative log-likelihood function of a GAMLSS with respect to each of the four predictors η_{θ_k}. These four predictors are updated successively in each iteration, in which the current estimates of the other distribution parameters are used as offset values. A schematic overview of the updating process of gamboostLSS in iteration $m + 1$ is as follows:

$$\begin{aligned}
(\hat{\mu}^{[m]}, \hat{\sigma}^{[m]}, \hat{\nu}^{[m]}, \hat{\tau}^{[m]}) &\xrightarrow{\text{update}} \hat{\eta}_{\mu}^{[m+1]} \Longrightarrow \hat{\mu}^{[m+1]}\,, \\
(\hat{\mu}^{[m+1]}, \hat{\sigma}^{[m]}, \hat{\nu}^{[m]}, \hat{\tau}^{[m]}) &\xrightarrow{\text{update}} \hat{\eta}_{\sigma}^{[m+1]} \Longrightarrow \hat{\sigma}^{[m+1]}\,, \\
(\hat{\mu}^{[m+1]}, \hat{\sigma}^{[m+1]}, \hat{\nu}^{[m]}, \hat{\tau}^{[m]}) &\xrightarrow{\text{update}} \hat{\eta}_{\nu}^{[m+1]} \Longrightarrow \hat{\nu}^{[m+1]}\,, \\
(\hat{\mu}^{[m+1]}, \hat{\sigma}^{[m+1]}, \hat{\nu}^{[m+1]}, \hat{\tau}^{[m]}) &\xrightarrow{\text{update}} \hat{\eta}_{\tau}^{[m+1]} \Longrightarrow \hat{\tau}^{[m+1]}\,.
\end{aligned}$$

In every step the negative gradient of the loss from every distribution parameter is fitted to pre-defined base-learners, one for each covariate. The

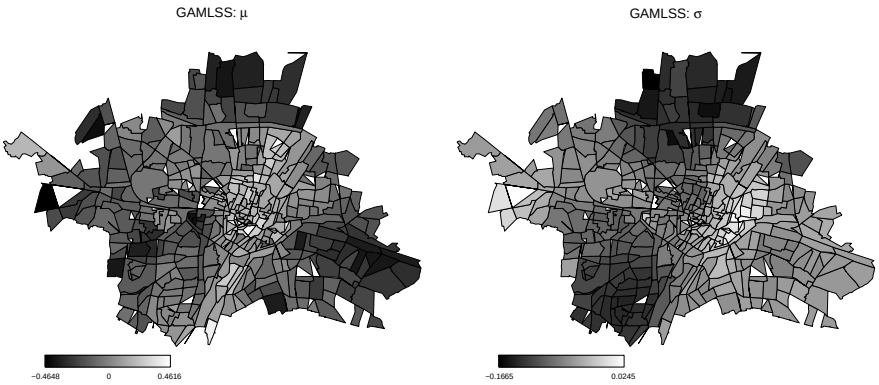


FIGURE 1. Estimated spatial effects obtained for the high-dimensional GAMLSS for distribution parameters μ and σ. For the third parameter `df`, the corresponding spatial variable for the neighbourhoods was not selected by the algorithm.

best performing base-learner is added to the current predictor. Due to additive updates in each iteration, every resulting predictor follows an additive structure (1). The type of the predictor functions $f_{j\theta_k}$ corresponds to the base-learner used for this covariate. Typical examples of base-learners are classification and regression trees, linear models or penalized regression splines.

3 Munich Rental Guide

Most larger German cities publish rental guides as a reference to 'average rents' for both landlords and tenants. We analyse data collected for the 2007 rental guide for the German city of Munich. Earlier modelling approaches identified variance heteroscedasticity, motivating the use of GAMLSS. The main objective of the analysis is to obtain point predictions for the net rent per square metre and to construct prediction intervals holding a prespecified coverage probability for the net-rent. Our sample comprises data obtained from $n = 3016$ flats within Munich, with detailed information on these flats in terms of 238 categorical covariates, two continuous covariates (the size of the flat and the year of the building's construction) and spatial information regarding in which of the 411 neighbourhoods the flat is located. As a response distribution for the net rent per square metre, we consider the three-parametric t-distribution with location parameter

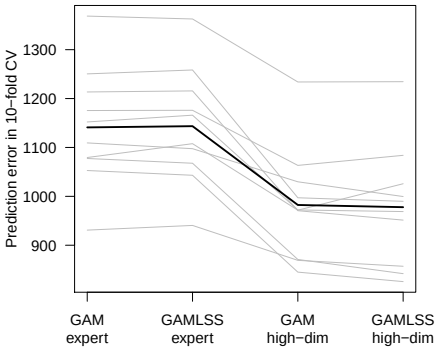


FIGURE 2. Parallel coordinate plot containing the average mean squared prediction errors obtained from the four models (high-dimensional GAMLSS/expert GAMLSS/ high-dimensional GAM/expert GAM) by cross-validation.

$\theta_1 = \eta_\mu =: \mu$, scale parameter $\theta_2 = \exp(\eta_\sigma) =: \sigma$ and degrees of freedom $\theta_3 = \exp(\eta_{\mathrm{df}}) =: \mathrm{df}$. The probability density function of the net rent per square metre conditional on a set of predictor variables is thus given by

$$f(y_i|\mu_i, \sigma_i, \mathrm{df}_i) = \frac{\Gamma(\frac{\mathrm{df}_i+1}{2})}{\sigma_i \Gamma(\frac{1}{2})\Gamma(\frac{\mathrm{df}_i}{2})\sqrt{\mathrm{df}_i}} \left(1 + \frac{(y_i - \mu_i)^2}{(\sigma_i^2 \cdot \mathrm{df}_i)}\right)^{(-(\mathrm{df}_i+1)/2)} .$$

For each of the parameters μ, σ^2, and df, we consider the predictors

$$\begin{aligned}
\eta_{\mu_i} &= \beta_{0\mu} + x_i^\top \beta_\mu + f_{1\mu}(\mathrm{size}_i) + f_{2\mu}(\mathrm{year}_i) + f_{\mathrm{spat}\mu}(s_i) \ , \\
\eta_{\sigma_i} &= \beta_{0\sigma} + x_i^\top \beta_\sigma + f_{1\sigma}(\mathrm{size}_i) + f_{2\sigma}(\mathrm{year}_i) + f_{\mathrm{spat}\sigma}(s_i) \ , \\
\eta_{\mathrm{df}_i} &= \beta_{0\mathtt{df}} + x_i^\top \beta_{\mathrm{df}} + f_{1\mathrm{df}}(\mathrm{size}_i) + f_{2\mathrm{df}}(\mathrm{year}_i) + f_{\mathrm{spatdf}}(s_i) \ ,
\end{aligned}$$

$i = 1, \ldots, n$, where $\beta_{0\theta_k}$ and β_{θ_k} correspond to the intercept and parametric effects of the categorical covariates (denoted by $x_i^\top$), $f_{1\theta_k}(\mathrm{size})$ and $f_{2\theta_k}(\mathrm{year})$ are non-linear effects for the continuous variables and $f_{\mathrm{spat}\theta_k}(s)$ is a spatial effect based on the neighbourhood within Munich. Figure 1 presents the effect estimates for the spatial variable, revealing a high spatial variability with respect to μ and σ.

To evaluate the results of the GAMLSS approach, we compared it to a conventional GAM, which models only the conditional mean of a gaussian distribution. Additionally, we fitted both methods once including the whole set of available covariates and once for a reduced set of categorical covariates, including only an expert selection of 28 effects (see Kneib et al.,

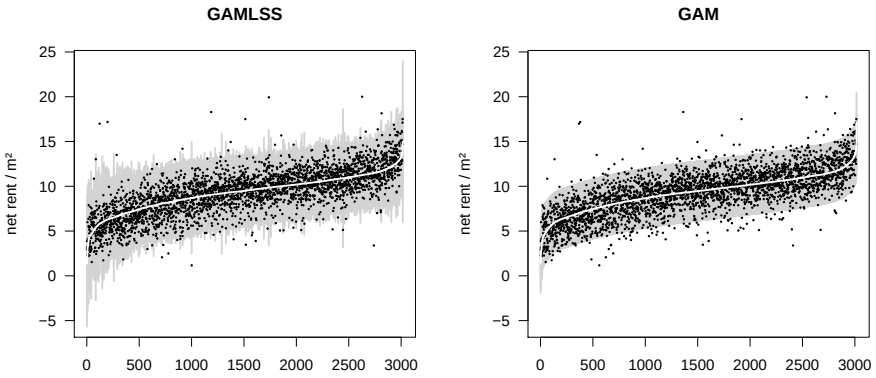


FIGURE 3. 95% prediction intervals based on the quantiles of the modelled conditional distribution from GAMLSS (left) and GAM (right). The solid white line represents point predictions; intervals are shaded grey. The dark points correspond to the observed net rents per square metre.

2011). All models were compared regarding their predictive accuracy using 10-fold cross validation. While Figure 2 clearly suggests that the accuracy of point predictions obtained from classical GAMs carries over to those obtained from GAMLSS, the inclusion of covariate effects on parameters such as σ^2 and df additionally allows for an improved accuracy of the prediction intervals (Figure 3), resulting from the conditional quantiles of the respective distributions. As GAMLSS not only regress the expected mean but all parameters of a distribution, also the size of the intervals – and not only their centre – explicitly depends on the characteristics of a flat, resulting in more accurate intervals with higher coverage.

4 Conclusion

As a natural extension of the well-established GAM framework, GAMLSS have gained increasing popularity in recent years and their use has expanded to include many different fields of application (see for example the information provided at `http://gamlss.org`). For the analysis of the Munich Rental Guide data, we developed the gamboostLSS algorithm, thereby extending the GAMLSS methodology to the analysis of high-dimensional data with potentially large numbers of covariates. Since estimation and selection of predictor effects are carried out simultaneously in gamboostLSS, the new algorithm addresses one of the remaining problems of the classical

fitting methods currently available. Conversely, gamboostLSS can be considered as a natural extension of the boosting framework to include regression models with multiple predictors. Consequently, the classical features of boosting, such as shrinkage, variable selection and additive prediction functions (and thus the interpretability of estimates) carry over to each of the distribution parameters of a GAMLSS.
A limitation of gamboostLSS is its computationally expensive tuning procedure based on multi-dimensional cross-validation. Further research is warranted on the topic of stopping procedures for this class of models.
In summary, the advantages offered by gamboostLSS are the following: (i) Variable selection is accomplished automatically. (ii) The algorithm can be applied to high-dimensional data sets in which the number of predictor variables exceeds the number of observations. (iii) Stopping the algorithm before convergence yields a built-in mechanism for the shrinkage of effect estimates, thereby decreasing the variability of predictor effects and improving the prediction accuracy of the obtained GAMLSS solution.
The algorithm presented here is implemented in the R add-on package **gamboostLSS**, available currently at R-Forge (Hofner et al., 2010).

Acknowledgments: The authors thank Ludwig Fahrmeir for sharing the Munich Rental Guide data. The work of AM and MS was supported by the Interdisciplinary Center for Clinical Research (IZKF) at the University Hospital of the FAU Erlangen-Nürnberg (Project J11).

References

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477-522.

B. Hofner, A. Mayr, N. Fenske, and M. Schmid (2010). gamboostLSS: Boosting Methods for GAMLSS Models. R package version 0.5-0.

Kneib, T., S. Konrath and L. Fahrmeir (2011). High-dimensional structured additive regression models: Bayesian regularisation, smoothing and predictive performance. *Applied Statistics*, **60**, 51-70.

Mayr, A., N. Fenske, B. Hofner, T. Kneib and M. Schmid (2010). GAMLSS for high-dimensional data - a flexible approach based on boosting. *Department of Statistics: Technical Reports*, **98**.

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

Schmid, M., S. Potapov, A. Pfahlberg and T. Hothorn (2010). Estimation and regularization techniques for regression models with multidimensional prediction functions . *Statistics and Computing*, **20**, 139-150.

Estimation of Infection Rates from Repeated ELISA Optical Density Data using Hidden Markov Models

Joris Menten[12], Marleen Boelaert[1], Emmanuel Lesaffre[23]

[1] Institute of Tropical Medicine, Nationalestraat 155, 2000 Antwerp, Belgium
[2] L-Biostat, KULeuven, Kapucijnenvoer 35,3000 Leuven, Belgium
[3] Department of Biostatistics, Erasmus University Rotterdam, 3000 CA Rotterdam, the Netherlands

Abstract: We present a mixture model for the estimation of the incidence of *Leishmania* infection from repeated serosurveys without the need to use explicit thresholds for seropositivity. The data is analyzed through the joint estimation of a response model of the observed outcomes given the unobserved infection status and a Hidden Markov Model for the latent infection status.

Keywords: Mixture Modelling; Hidden Markov Models; Joint Modelling; Latent Variable Models; ELISA tests.

1 Introduction

Serological tests based on the detection of antibodies to infectious agents through the ELISA technique can be used to estimate the prevalence or incidence of infection. These tests are often measured as an optical density (OD) on a continuous scale but reported as binary data (above/below a certain threshold value). An alternative approach is to use mixture modelling of the quantitative OD results (Gay 1996). We expand on this approach by jointly estimating a response model for the observed OD values over time given the infection status and a Hidden Markov Model for the unobserved infection status over time. This model avoids the need to choose a cut-off for the OD values to separate seropositive versus seronegative individuals. Using this model, we estimated infection incidences and the effects of distributing impregnated bednets in a cluster randomized study.

2 Motivating Example

The motivating data were collected during a pair-matched intervention study in 16 villages in India on the use of impregnated bednets for the prevention of leishmaniasis infection (Picado et al., 2010). In each village, all consenting individuals were assessed for leishmaniasis using the RK39

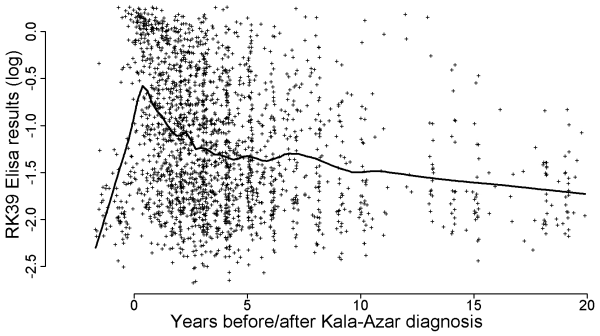


FIGURE 1. Evolution of RK39 Elisa results before and after kala-azar diagnosis. Points are actual observations. Bold line is a lowess scatterplot smoother.

ELISA test, a secondary endpoint of the study. In addition, information on prior or incident Kala-azar (KA), the clinical manifestation of *Leishmania donovani* infection, was collected. Subjects with prior or incident KA are known to be infected with *Leishmania*, for the remainder of subjects the infection status is unknown. Three annual surveys were performed during the study. The first survey consisted of a baseline survey during which bednets were distributed in the intervention villages. The effect of the intervention was then assessed by the two subsequent serosurveys. The log-transformed ELISA data are distributed asymmetrically which may indicate a mixture of two populations: low ELISA (-3 to -1) results for non-infected individuals, and high ELISA results for infected individuals. In patients diagnosed and treated for KA, the average RK39 ELISA results reach a peak (of approximately -0.5) around the time of diagnosis to quickly decrease to -1.2 after 3 years and decrease more slowly afterwards. The average decrease over time appears linear as a function of log time from diagnosis (Figure 1).

3 Statistical Modelling

Let y_{ti} be the observed result for the ELISA test on the ith subject at time t, where $i = 1, ..., N$ and $t_{i,I}$ is the time of infection for subject i. Estimating the infection rates, requires the joint estimation of y_{ti} and $t_{i,I}$, i.e. estimating the model

$$P(y_{ti}, t_{i,I}|\mathbf{x_{ti}}, \mathbf{z_{ti}}, \mathbf{b_i}, \mathbf{e_i}) = P(y_{ti}|t_{i,I}, \mathbf{x_{ti}}, \mathbf{b_i}) \times P(t_{i,I}|\mathbf{z_{ti}}, \mathbf{e_i}), \qquad (1)$$

with $\mathbf{x_{ti}}$, $\mathbf{z_{ti}}$ covariate vectors and $\mathbf{b_i}$, $\mathbf{e_i}$ independent subject random effects. Consequently, we need to jointly estimate a *response model* for the observed data, conditional on the latent infection status and a *structural model* for the underlying infection status over time.

3.1 Response Model

The true infection status of subject i at time t is defined as $d_{ti} = I(t_{i,I} \leq t)$. A subject is presumed to show a sudden increase in the outcome measurement y_{ti} immediately after infection, after which y_{ti} decreases over time, as described by the following regression model:

$$y_{ti} = \alpha + I(t_{i,I} \leq t) \times \beta_{1i} + \Delta_{ti} \times \beta_{2i} + \epsilon_{ti}, \tag{2}$$

with α the average ELISA test result for not-infected subjects, β_{1i} and β_{2i} random intercept and slope terms, respectively, for infected subjects with $\beta_{\mathbf{i}} = (\beta_{1i}, \beta_{2i}) \sim MVN(\mu_\beta, \mathbf{\Omega}_\beta)$ and representing the $\mathbf{b_i}$ terms in Equation 1, $\Delta_{ti} = ln(max(1, (t - t_{i,I})))$ the log time since infection, and $\epsilon_{ti} \sim N\left(0, \sigma^2\right)$ the test measurement error. Given $\beta_{\mathbf{i}}$, the y_{it} are assumed to be independently distributed.

3.2 Structural Model

The underlying infection status at the 3 discrete observation time points j ($j = 1, 2, 3$) is modelled using an inhomogeneous, first-order Hidden Markov Model (Cook, 2000). At each time point j, the infection status d_{ji} of subject i in village $v(i)$ and village-pair $pair(i)$ is Bernoulli distributed: $d_{ji} \sim Bernoulli(p_{ji})$ with at the first time point:

$$logit(p_{1i}) = \pi_{v(i)} + \gamma_{v(i)} \times log(age_i), \tag{3}$$

with age_i the age of subject i at the first serosurvey and $\pi_{v(i)}$ and $\gamma_{v(i)}$ village specific intercept and slope terms, respectively. At subsequent time points:

$$p_{ji} = d_{j-1\ i} + (1 - d_{j-1,i}) \times \kappa_{j-1,i},$$

with:

$$logit(\kappa_{1i}) = \delta_{1,pair(i)} + \delta_{3,pair(i)} \times Int_{v(i)},$$

$$logit(\kappa_{2i}) = \delta_{1,pair(i)} + \delta_{2,pair(i)} + \delta_{3,pair(i)} \times Int_{v(i)},$$

with $\delta_{1,pair(i)}$ the log-odds infection probability in the control village from $pair(i)$ between serosurveys 1 and 2 and $\delta_{2,pair(i)}$ a period effect (log odds-ratio) and $\delta_{3,pair(i)}$ the effect (log odds-ratio) of intervention in $pair(i)$ with $Int_{v(i)}$ an indicator variable for inclusion of the village $v(i)$ in the intervention group. The π, γ and δ terms are modelled as random effects and correspond to the $\mathbf{e_i}$ terms in Equation 1.

3.3 Model Estimation and Priors

The model is estimated using Markov-Chain Monte-Carlo methods in OpenBUGS called from within R. We use informative priors on the measurement model parameters to ensure an identifiable model. These priors are based

on the log RK39 results for the observed KA cases. In addition, to avoid a non-identifiable model, we constrain the random intercepts β_{1i} to be strictly positive and slopes β_{2i} to be strictly negative. In the structural model, we use weakly informative default prior distributions (Gelman et al., 2008).

4 Application to the Kalanet data

Based on this model, the average log RK39 result in non-infected individuals is estimated at -1.99. Immediately after infection, the RK39 increases by 1.78 and decreases subsequently by 0.28/log-month. The analysis shows no significant reduction in infection incidences with intervention: OR = 0.77 (95%CI: 0.45 to 1.55). The estimated intervention effects vary strongly across matched pairs. Comparing the model-based estimates with other estimates of the intervention effect (KA OR, relative risk based on another marker of infection [DAT, primary study analysis], and OR based on Elisa with a fixed cut-off), the model-based estimates are generally in line with the other estimates.

5 Conclusion

We estimated, through joint modelling of a response and a structural model, infection incidences from a continuous measure of infection without using explicit cut-offs for seropositivity. Given the event of interest is not observed for the majority of subjects, substantial subject matter knowledge is needed to ensure an estimable model which is clinically appropriate. In our analysis, both the functional form and the priors and constraints for the parameters of the measurement model were based on subject matter knowledge and data from the subset of subjects who showed clinical disease.

References

Cook R.J., Ng E.T.M., and Meade, M.O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent markov models. 2000; 56: *Biometrics*, **56**:1109-1171.

Gay, N.J. (1996). Analysis of serological surveys using mixture models: application to a survey of parvovirus B19. *Statistics in Medicine*, **15**:1567-1573.

Gelman, A., Jakulin A., Pittau M.G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **4**:1360-1383.

Picado, A., et al. (2010). Longlasting insecticidal nets for prevention of Leishmania donovani infection in India and Nepal: paired cluster randomised trial. *BMJ*, **342 (7788)**:92, c6760.

Nonlinear and Spline Regression Models for Forecasting Gas Flow on Exits of Gas Transmission Networks

Radoslava Mirkov[1] , Herwig Friedl[2]

[1] Humboldt Universität zu Berlin, Department of Mathematics, Unter den Linden 6, 10999 Berlin, Germany, mirkov@math.hu-berlin.de
[2] Graz University of Technology, Institute of Statistics, Münzgrabenstraße 11, 8010 Graz, Austria

Abstract: The flow of natural gas within a gas transmission network is studied with the aim to predict gas loads for very low temperatures. Two models for describing dependence between the maximal daily gas flow and the temperature on network exits are presented. A Brain-Cousens regression model is chosen from the class of parametric models. As an alternative, a semi-parametric logistic regression based on penalized splines is considered. The comparison of prediction based on both models is included.

Keywords: nonlinear regression; penalized splines; gas flow, design temperature.

1 Introduction and Model Motivation

We study historical data of the flow of gas transported in networks in order to support a reliable and realistic prediction of the future gas flow. The forecast of gas loads at the so-called design temperature is of particular interest. The design temperature is the lowest temperature at which the gas operator is still obliged to supply gas without failure, and lies between $-12°$C and $-16°$C. Such low mean daily temperatures are very uncommon in Germany, and there is no gas flow data available at the design temperature. For this reason gas operators are forced to use the predicted gas loads at the design temperature, and we present here two models useful for the forecast.

Data is obtained from measuring stations within the German pipeline network operated by Open Grid Europe GmbH, one of the largest German gas transporters. It contains hourly gas flow for the period between January 2004 and June 2009, and the corresponding mean daily temperatures. We study the dependence of gas loads and air temperature on all exits along the pipelines. Typical exits in such networks are public utilities, industrial consumers and storages, as well as exits on border and regional crossings. Since we want to maximize the transportation capacity through the pipelines, we concentrate on the daily maximum flows y_i^{max}, $i = 1, \ldots, n$ $(n = 2005)$,

at each exit, for every exit in the network. We consider the standardized daily maximum flows $y_i = y_i^{max}/\bar{y}$, where $\bar{y}$ denotes the empirical mean of all maximal daily gas flows at one specific measuring station.
The following model to describe the dependence of the standardized maximal daily gas loads y_i on temperature t_i is studied:

$$y_i = S(t_i) + \varepsilon_i, \tag{1}$$

where t_i stands for the weighted four-day-mean temperature with the weights $(0.53, 0.27, 0.13, 0.07)$, and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ are error terms, for $i = 1, \ldots, n$, as suggested in Cerbe (2008).
Friedl et al. (2011) explore different modeling possibilities for this problem, and suggest several appropriate variants for the function $S(t_i)$. They also compare advantages and disadvantages of both approaches.

2 Nonlinear and P-Splines Regression Models

We fit a parametric as well as a semi-parametric nonlinear logistic regression model and analyze the properties of the gas flow through the pipelines in dependence of the temperature and the forecast based on these models. The so-called Brain-Cousens model (BC-model) is proposed by Ritz and Streibig (2008) for this kind of problems, while many authors propose some variant of spline regression, see e.g. Jones et al. (2009), Jarrow et al. (2004), Eilers and Marx (1996). A comparison of both approaches for a duck growth problem is presented in Vitezica et al. (2010).
In the class of parametric models, we consider the BC-model, which is defined by

$$S(t_i) = \theta_4 + \frac{\theta_1 + \theta_6 \left(\frac{\theta_2}{t_i - 40^\circ\mathrm{C}} + d_i\theta_5 \right) - \theta_4}{1 + \left(\dfrac{\theta_2}{t_i - 40^\circ\mathrm{C}} + d_i\theta_5 \right)^{\theta_3}}, \tag{2}$$

where

$$d_i = \begin{cases} 1 & \text{if day } i \text{ is a working day,} \\ 0 & \text{if day } i \text{ is a holiday or at weekends,} \end{cases}$$

indicates whether the gas loads occurred on working days or on weekends and holidays.
The parameters in model (2) are used as follows: θ_1, θ_6 and θ_4 define the upper and lower asymptotes, θ_5 indicates the type of the day, while θ_2 and θ_3 describe the shape of the decrease of the curve. We use initial values provided in Friedl et al. (2011). The results of the evaluation are given in Table 1. Parameters θ_5 and θ_6 in the model are significant, implying that the that the expected gas loads differ during the week (W) and on weekends and holidays (H), and the upper asymptote is a line with slope θ_6. For low temperatures, the modified upper asymptote in the BC-model implies the

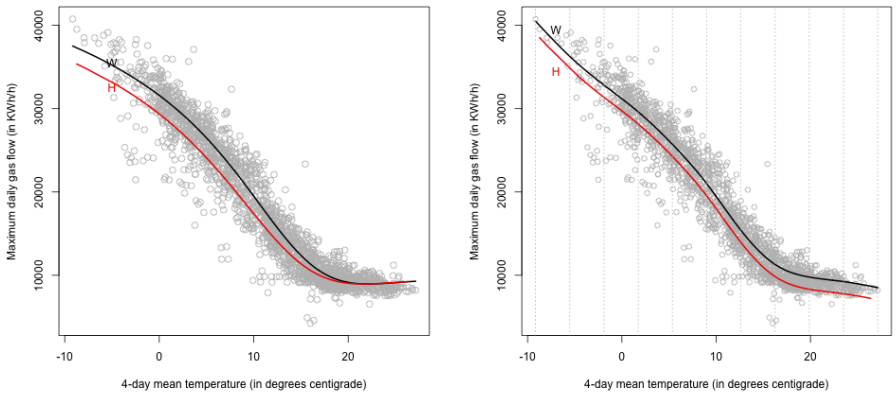


FIGURE 1. Fitted BC-model with indicator day (left) and penalized splines regression with indicator day (right) based on cubic B-splines on the mesh with 10 segments and the second order penalty $\lambda = 2.51$.

day-specific increase of the mean gas flow for approximately 2 times scaled $\bar{y}$ when the temperature decreases for 1°C. The graphical representation of model (2) is shown in Figure 1 (left).

TABLE 1. MLEs (std. errors) of the BC-model.

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
3.1805	−28.0417	6.5713	0.5229	−0.0447	−2.0807
(0.1062)	(0.7960)	(0.3431)	(0.0197)	(0.0031)	(0.2269)

Alternatively, Friedl et al. (2011) suggest the penalized splines (P-splines) approach, and assume that the function $S(t_i)$ is the linear combination of basis functions $B_j, j = 1, \ldots, m$, on the mesh Δ, given by

$$S(t_i) = \sum_{j=1}^{m} a_j B_j(t_i) + a_{m+1} d_i, \tag{3}$$

and B_j are basis functions of the B-spline of degree q, and the mesh Δ is an equidistant grid over $m - q$ segments, i.e. with $m - q + 1$ inner knots. The regression coefficients are obtained taking into account the smoothing penalty λ. We refer to Figure 1 (right) for a graphical representation of the fitted P-splines model and the position of the inner knots based on cubic B-splines on the mesh with 10 segments and the second order penalty $\lambda = 10^{0.4} = 2.51$.

3 Prediction

The models presented in Section 2 are now utilized for the prediction of gas loads at the design temperature. Recall, the design temperature lies outside of the domain of the predictor variable $t_i, i = 1, \ldots, 2005$. To this end, we replace the existing temperature t_i by a new predictor variable $\tilde{t}_k, k = 1, \ldots, \tilde{n}$, generated as an equidistant grid of temperatures, which includes low temperatures of interest. In particular, we generate $\tilde{t}_k$ starting from the lowest possible design temperature, i.e. $-16°$C, and go up to $35°$C with step size 1. Based on the new data and the fitted models (2), and (3), the predictions

$$\tilde{y}_k = S(\tilde{t}_k) + \varepsilon_k, \qquad \varepsilon_k \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \qquad k = 1, \ldots, 52,$$

are calculated.

The predicted values based on the BC-model are obtained using the `predict` method in R, as described in Ritz and Streibig (2008). P-splines allow straightforward smooth extrapolation, and we exploit this property to forecast gas loads at the design temperature. The second order penalty implies the extrapolation by a linear sequence, cf. Eilers and Marx (2010).

Figure 2 illustrates the prediction based on the BC-model and P-splines regression. At the design temperature of $-12°$C the predicted gas loads on working days based on models (2), and (3), are 38817, and 43048 KWh/h, respectively.

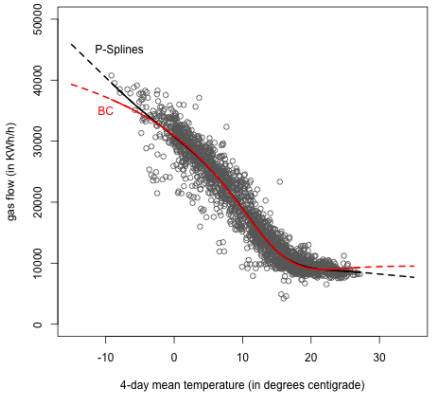


FIGURE 2. Prediction for working days based on the BC and P-splines model.

In the case of nonlinear regression models, the standard error estimates do not change substantially when we leave the domain of the predictor

variable. Figure 3 (left) represents the predicted values for working days based on the BC-model (2), and the corresponding standard error bands. The naïve method based on the assumptions of the normality of error terms and the variance homogeneity is employed to determine standard errors of parameters. Some other methods for constructing prediction intervals for nonlinear regression can be found in Gauchi et al. (2010), and Ritz and Streibig (2008).

It is well known that extrapolation in the case of splines can be unsafe for the prediction, although the model provides a good fit for gas loads. This fact is reflected in the shape of the error bands for the P-splines model (3). Due to the local smoothing, the fit is very good with a small error band width within the domain of the predictor variable, while the increase in the width of error bands is large as soon as we extrapolate. The Bayesian estimate of the standard error bands for the fitted P-splines model for working days are shown in Figure 3 (right).

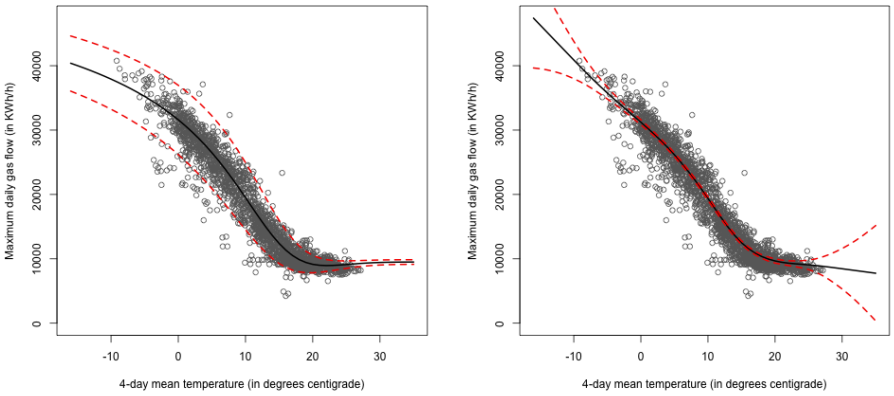


FIGURE 3. Prediction for working days based on the BC model (left) and P-splines (right) with the corresponding standard error bands.

4 Conclusions

We investigate prediction based on the nonlinear BC-model and on the semi-parametric P-splines regression. Both the BC-model and the P-splines reflect the behavior of gas flow for low temperatures in a realistic way. We note that the nonlinear regression models are generally more difficult to handle than the local smoothers like the P-splines, because of their numerical properties. Contrary to them, the P-splines methodology is a very

flexible simpler alternative, but it does not support the multiple regression techniques, and one cannot exploit the desirable flexible temperature effects. The forecast of gas loads based on the BC-model is safer than the one relying on the P-splines, due to the numerical construction of models.

References

Cerbe, G. (2008). *Grundlagen der Gastechnik*. Hanser Verlag.

Eilers, H.C., and Marx, B.D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, **11(2)**, 89-121.

Eilers, H.C., and Marx, B.D. (2010). Splines, Knots, and Penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2(6)**, 637-653.

Gauchi, J-P., Vila, J-P., and Coroller, L. (2010). New Prediction Interval and Band in the Nonlinear Regression Model: Application to Predictive Modeling in Foods. *Communications in Statistics - Simulation and Computation*, **39(2)**, 322-334.

Jarrow, R., Ruppert, D., and Yu, Y. (2004). Estimating the Interest Rate Term Structure of Corporate Debt With a Semiparametric Penalized Spline Model. *Journal of the American Statistical Association*, **99(465)**, 57-66.

Jones, G., Leung, Y., and Robertson, H. (2009). A Mixed Model for Investigating a Population of Asymptotic Growth Curves Using Restricted B-Splines. *Journal of Agricultural, Biological, and Environmental Statistics*, **14(1)**, 66-78.

Friedl, H., Mirkov, R., and Steinkamp, A. (2011). Modeling Gas Flow on Exits of Gas Transmission Networks. *Submitted to a Special Issue of the International Statistical Review on Energy Statistics*.

Ritz, C., and Streibig, J.C. (2008). *Nonlinear Regression with R*. Springer.

Vitezica, Z.G., Marie-Etancelin, C., Bernadet, M.D., Fernandez, X., and Robert-Granie, C. (2010). Comparison of nonlinear and spline regression models for describing mule duck growth curves. *Poultry Science*, **89**, 1778-1784.

Prediction of the rheumatoid arthritis activity score: a joint modeling approach

Siti Haslinda Mohd Din[1,2], Marek Molas[1], Jolanda Luime[3], Emmanuel Lesaffre[1,4]

[1] Department of Biostatistics, Erasmus Medical Centre, P.O.Box 2040, 3000 CA Rotterdam, The Netherlands
[2] Department of Statistics, Block C6, Federal Government Administration Centre, 60588 Putrajaya, Malaysia
[3] Department of Rheumatology, Erasmus Medical Centre, P.O.Box 2040, 3000 CA Rotterdam, The Netherlands
[4] Catholic University of Leuven, I-Biostat, 3000 Leuven, Belgium

Abstract: The aim of this paper is to determine a predictive model for the rheumatoid arthritis patients disease activity using a joint modeling approach.

Keywords: joint modeling; Bayesian methods; rheumatology.

1 Introduction

Joint modeling is a statistical technique to estimate common parameters of two or more models jointly. Here joint modeling is applied to extract random effects from longitudinal profiles to predict a rheumatoid arthritis activity score in a linear regression modeling taking into account that the random effects are not observed. This approach has several advantages, e.g Guo et al.(2004) stated that joint modeling deals better with the missing longitudinal covariates than the standard multiple regression approaches. The longitudinal responses are bounded on an interval and have a discrete nature. For this reason we made use of the approach of Lesaffre et al. (2007) for bounded outcome score. We look at a likelihood and a Bayesian approach and make use of SAS PROC NLMIXED and WinBUGS.

1.1 Motivating data set

The *Rheumatoid Arthritis Patients rePort Onset Reactivation sTudy (Rapport study)* is a longitudinal study that aims to identify an increase in disease activity by self-reported questionnaires in the 3 months preceding the clinical assessment. Between September and December 2008, 159 patients of aged 18 years and older with RA or polyarthritis using Disease Modifying Drugs(DMARDs) for at least 3 months were recruited. Patients'

disease activities were evaluated using the Disease Activity Score of 28 joint counts (DAS28) at every three months during one year follow-up. In this study, four disease-related self reported instruments also called patient's reported outcomes (PROs), i.e. Health Assessment Questionnaires (HAQ), the Rheumatoid Arthritis Disease Activity Index (RADAI), Visual Analogue Scale (VAS global) of the patient's global assessment of disease activity and VAS fatigue were measured at months 0, 1 and 2 and were used to predict DAS28 at month 3. In addition, we included age, gender and the arthritis self efficacy and coping with rheumatic stressors (CORS) as covariates.

2 Statistical Approaches

In order to bypass problems with classical approaches such as classical and ridge regression, we applied a two-stage and a joint modeling approach to build the prediction model.

2.1 Two stage approach

- **First stage model**

  In the first stage, the evolutions of the four measurements are summarized by latent variables, i.e. a random intercept and a random slope. There is, however, a complication in that the PROs are bounded on an interval and are therefore examples of a bounded outcome score (BOS). Typically BOSs have a peculiar distribution that ranges from unimodal symmetric to J- or U-shaped distributions. In addition many BOSs have a discrete nature. Lesaffre et al. (2007) assumed that the observed BOS is obtained from rounding procedure of a latent score measured on a continuous scale.

  For K markers followed up in time, let $\tilde{U}_{ijk}$ denote the true latent score for patient $i = 1, \ldots, N$ at time $t_j = 1, \ldots, n_i$ with marker $k = 1, \ldots, K$ of the observed BOS $\tilde{Y}_{ijk}$. The latent score $\tilde{U}_{ijk}$ is assumed to lie in the interval $\tilde{Y}^L_{ijk}$ and $\tilde{Y}^U_{ijk}$. Then let $U_{ijk} = \log(\tilde{U}_{ijk}/(1-\tilde{U}_{ijk}))$ be modeled longitudinally as a linear mixed effects model:

$$U_{ijk} = \mathbf{x}^T_{ijk}\beta_k + b_{0ik} + b_{1ik}t_j + \varepsilon_{ijk}, \tag{1}$$

  with $\mathbf{x}_{ijk}$ a vector of covariates, b_{0ik}, b_{1ik} are the random intercept and slope, respectively assumed to have a bivariate normal distribution with mean zero and covariance matrix $\Sigma_{k,b}$, ε_{ijk} measurement error independent of the random effect with a normal distribution $\mathrm{N}(0, \sigma^2_{k,\varepsilon})$. The vector of latent transformed scores $\mathbf{U}_{ik} =$

$(U_{i1k}, \ldots, U_{in_ik})$ follows marginally a multivariate normal distribution. The likelihood for the i^{th} individual is:

$$L_{ik}(\theta_k|\mathbf{X}_{ik}) = \left\{ \int_{Y^L_{i1k}}^{Y^U_{i1k}} \ldots \int_{Y^L_{in_ik}}^{Y^U_{in_ik}} \mathcal{MVN}(\mathbf{U}_{ik}|\mathbf{x}^T_{ijk}\beta_k, \Sigma_{ik}) d\mathbf{U}_{ik} \right\},$$

with θ_k the stacked vector of mean parameters β_k and variance parameters $\Sigma_{k,b}, \sigma^2_{k,\varepsilon}$ and $\mathbf{b}_{ik} = (b_{0ik}, b_{1ik})$. In our application the $\boldsymbol{b}_{ik}$ are of interest. These random effects can be estimated using empirical Bayes methodology resulting in $\hat{\boldsymbol{b}}_{ik}$.

- **Second stage model**

  In the second stage a classical multiple regression is used where the response is regressed on the estimated random effects of the first stage, i.e. the predicted random intercept and slope $\hat{b}_{0ik}$ and $\hat{b}_{1ik}$, and possibly some extra covariates $\mathbf{z}$. In our case, there were 8 random effects included in the regression model to predict DAS28 at month 3.

However, the two stage approach ignores the fact that the random effects are estimated and hence prone to measurement error which causes the regression coefficients of these random effects in the second regression model to be distorted.

2.2 Joint modeling approach

In the joint approach the likelihoods of the first and second stage are handled simultaneously. We have applied two joint modeling approaches. The first approach is based on maximizing the joint marginal likelihood which is the marginal likelihood combining the first and second stage models into one encompassing integrated likelihood. This is an integrated likelihood whereby the random effects appearing in both likelihoods are integrated out. For this a Laplace approximation has been used. In the second approach the posterior is sampled with Bayesian Markov Chain Monte Carlo (MCMC) techniques to arrive at the parameter estimates. In both ways the uncertainty about the true values of the random effects is naturally taken into account.

3 Application to motivating data set

For the two stage approach, we used the SAS procedures PROC NLMIXED and PROC REG in the first and the second stage, respectively. In the joint modeling approach, we use the SAS procedure PROC NLMIXED of SAS for the likelihood approach and WinBUGS software for the Bayesian approach. Weakly informative priors were chosen for all parameters in the Bayesian approach. Table 1 shows the results of applying the two stage and the joint modeling approaches on the motivating data set.

TABLE 1. Results of the two stage and joint model approach on the Rapport data set

Parameter		2 stage model			Joint model					
		Model 1(BOS*-NLM†+MLR*)			Model 2 (BOS*-Bayesian)			Model 3(BOS*-NLM†)		
		Estimate	S.E.	p-Value	Posterior Mean	S.E.	95% CI	Estimate	S.E.	p-Value
Intercept	β_0	3.74	0.65	<0.01	3.17	0.58	(2.04,4.30)	3.79	0.65	<0.01
HAQ										
Intercept	β_1	0.34	0.11	<0.01	0.20	0.09	(0.04,0.38)	0.32	0.11	<0.01
Slope	β_2	-0.01	0.74	0.99	0.07	0.94	(-1.86,2.07)	-0.05	0.76	0.95
RADAI										
Intercept	β_3	0.36	0.17	0.03	0.41	0.13	(0.16,0.66)	0.36	0.15	0.02
Slope	β_4	0.72	0.43	0.10	0.77	0.62	(-0.23, 2.20)	0.67	0.40	0.10
VAS global										
Intercept	β_5	-0.21	0.22	0.35	-0.16	0.17	(-0.49,0.18)	-0.19	0.21	0.38
Slope	β_6	0.56	0.79	0.48	0.98	1.16	(-0.81,3.94)	0.59	0.92	0.52
VAS fatigue										
Intercept	β_7	-0.29	0.29	0.32	0.08	0.12	(-0.17,0.32)	-0.15	0.54	0.79
Slope	β_8	-4.25	2.77	0.13	-1.54	1.45	(-4.93,0.81)	-2.60	5.10	0.61
Age	β_9	0.01	0.01	0.12	0.01	0.01	(-0.01,0.03)	0.01	0.01	0.17
Sex (Male)	β_{10}	-0.86	0.22	<0.01	-0.84	0.28	(-1.37,-0.29)	-0.78	0.24	<0.01
Self efficacy	β_{11}	-0.01	0.01	0.34	-0.01	0.01	(-0.03,0.01)	-0.01	0.01	0.27
Coping pain	β_{12}	-0.12	0.08	0.13	-0.08	0.08	(-0.25,0.08)	-0.13	0.08	0.11
Coping limitation	β_{13}	0.02	0.09	0.85	0.03	0.09	(-0.14,0.20)	0.01	0.08	0.86
Coping dependence	β_{14}	-0.05	0.05	0.32	-0.03	0.05	(-0.12,0.07)	-0.04	0.05	0.35

S.E.:estimated standard error

* BOS: Bounded outcome score for HAQ, RADAI, VAS global and VAS fatigue

† NLM: Non linear mixed model

* MLR: Multiple linear regression

4 Conclusion

We have used a two stage and joint modeling approach to predict DAS28 for the rheumatoid arthritis patients. However, the two stage approach is not recommended as it results in biased inference due to the unaccounted error in the estimation of the random coefficients (Wang et al. (2000)). We applied both ways of estimation: likelihood based and Bayesian MCMC sampling to obtain the final model results. The Bayesian approach is simple to program and offers more flexibility in distributional assumptions than the likelihood approach (because of the WinBUGS implementation).

References

Guo, X. and Carlin, B.P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, **58**, 1-7.

Lesaffre, E., Rizopoulos, D. and Tsonaka, R. (2007). The logistic-transform for bounded outcome scores. *Biostatistics*, **8**, 72-85.

Wang, C.Y., Wang, N. and Wang, S. (2000). Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, **56**, 487-495.

Covariate-adjusted inference for the Youden index and associated classification threshold

Elisa M. Molanes-López[1], Juan Carlos Pardo-Fernández[2] , Emilio Letón[3]

[1] Department of Statistics, UC3M, Avda. de la Universidad 30, 28911 Leganés (Madrid), Spain. E-mail: elisamaria.molanes@uc3m.es
[2] Departament of Statistics and OR, UVigo, Fac. Económicas, Campus As Lagoas-Marcosende, 36310, Vigo, Spain. E-mail: juancp@uvigo.es
[3] Department of Artificial Intelligence, UNED, C/ Juan del Rosal 16, 28040 Madrid, Spain. E-mail: emilio.leton@dia.uned.es

Abstract: In many medical studies, continuous variables or biomarkers are used to classify patients into diseased or healthy populations. The classification rule is based on a threshold value, which should be properly chosen. In some cases, a covariate (or vector of covariates) can be used in order to increase the performance of the classification procedure. Taking into account the effect of the covariates, in this work we propose a new nonparametric approach for estimating the Youden index and the associated optimal threshold value in the biomarker scale.

Keywords: location-scale regression models; ROC curves; Youden index.

1 Introduction

In medical studies, a continuous diagnostic test, Y, is commonly used for classifying subjects into diseased and healthy populations. Without loss of generality, it is usually assumed that for a given threshold, c, a subject with $Y \geq c$ is classified as diseased and as healthy otherwise. This kind of classification procedure will lead to some classification errors, which are usually calibrated on the basis of two indicators depending on c: sensitivity (probability of diagnosing a diseased person as diseased or 'true positive fraction', denoted by $q(c)$) and specificity (probability of diagnosing a healthy person as healthy or 'true negative fraction', denoted by $p(c)$).
The accuracy of a continuous classifier is usually described graphically by using the 'Receiver Operating Characteristic' (ROC) curve, which is obtained by plotting the pairs $(1-p(c), q(c))$, with $-\infty < c < \infty$. In practice, a crucial point in this context is to find the optimal threshold value to classify the individuals. Depending on the definition of optimality, different methods have been proposed in the literature to identify that threshold

value. In this paper, we will focus on the Youden index,

$$J = \max\{J(c); -\infty < c < \infty\}, \text{ where } J(c) = q(c) + p(c) - 1,$$

and its associated threshold value, c_J, studied recently by Le (2006), Schisterman and Perkins (2007) and Letón and Molanes-López (2009), among others.

In many studies, a covariate (or vector of covariates), X, is available and can be used in order to increase the classification performance of Y. In this work, we propose a new nonparametric approach for estimating the Youden index and the corresponding optimal threshold value in the biomarker scale, taking into account the effect of the covariates. The proposed method is described in Section 2. In Section 3, we study its practical performance in a small simulation study. Finally, the new method is illustrated in Section 4 using a real example.

2 New method

The information of a covariate along with the classifier can be incorporated in a general framework given by the location-scale regression models studied by Pepe (1997, 1998, 2003), Faraggi (2003) and more recently by González-Manteiga et al. (2011).

In this article we work with fully nonparametric location-scale regression models. This means that the relationship between the covariate and the classifier in each population (healthy, denoted by 0, or diseased, denoted by 1) is given by the following models:

$$Y_0 = \mu_0(X_0) + \sigma_0(X_0)\varepsilon_0,$$
$$Y_1 = \mu_1(X_1) + \sigma_1(X_1)\varepsilon_1,$$

where, for $j = 0, 1$, $\mu_j(\cdot) = E(Y_j \mid X_j = \cdot)$ and $\sigma_j^2(\cdot) = Var(Y_j \mid X_j = \cdot)$ are nonparametric functions representing, respectively, the conditional mean and the conditional variance of the response Y_j given the covariate X_j in each population, and ε_j is the regression error, which we assume independent of X_j.

For a fixed value of the covariate, x, the covariate-adjusted ROC curve is defined by

$$ROC_x(p) = 1 - F_1(F_0^{-1}(1 - p \mid x) \mid x),$$

where $F_j(y \mid x) = \Pr(Y_j \leq y \mid X_j = x)$ is the conditional cumulative distribution function of Y_j given that $X_j = x$, and $F_j^{-1}(p \mid x)$ is the corresponding conditional quantile function. Under the location-scale regression models given above, it is easy to check that, for $j = 0, 1$, the conditional distribution function and the conditional quantile function of the variable Y_j given $X_j = x$ can be written as

$$F_j(y \mid x) = G_j\left(\frac{y - \mu_j(x)}{\sigma_j(x)}\right) \qquad \text{and} \qquad F_j^{-1}(p \mid x) = \mu_j(x) + \sigma_j(x)G_j^{-1}(p),$$

where $G_j(y) = \Pr(\varepsilon_j \leq y)$ is the error distribution, and $G_j^{-1}(p)$ is the corresponding quantile function. Hence the covariate-adjusted ROC curve becomes

$$ROC_x(p) = 1 - G_1(G_0^{-1}(1-p)b(x) - a(x)),$$

where $a(x) = (\mu_1(x) - \mu_0(x))/\sigma_1(x)$ and $b(x) = \sigma_0(x)/\sigma_1(x)$.
Under this model, for a given value of the covariate, x, we define the covariate-adjusted Youden index by $J_x = \max_p |ROC_x(p) - p|$. For this index, let $p_{J_x} = \arg\max_p |ROC_x(p) - p|$. Then, the associated covariate-adjusted threshold, denoted by c_{J_x}, can be seen as both the conditional $(1-p_{J_x})$-quantile of Y_0 given $X_0 = x$ or the conditional $(1 - ROC_x(p_{J_x}))$-quantile of Y_1 given $X_1 = x$.
In practice, we assume that two samples of sizes n_0 and n_1 of i.i.d. data are observed from populations (X_0, Y_0) and (X_1, Y_1). Given a value of the covariate, x, we propose to estimate the covariate-adjusted Youden index by $\hat{J}_x = \max_p |\widehat{ROC}_x(p) - p|$, where $\widehat{ROC}_x(p)$ denotes the estimator of the covariate-adjusted ROC curve proposed and studied in González-Manteiga et al. (2011). The associated conditional optimal threshold, c_{J_x}, is estimated by any of the following two quantities

$$\hat{c}_{J_x 0} = \hat{Q}_{x0}(1 - \hat{p}_{J_x}),$$

$$\hat{c}_{J_x 1} = \hat{Q}_{x1}(1 - \widehat{ROC}_x(\hat{p}_{J_x})),$$

or by the weighted average of the previous two estimators

$$\hat{c}_{J_x} = \frac{n_0}{n_0 + n_1}\hat{c}_{J_x 0} + \frac{n_1}{n_0 + n_1}\hat{c}_{J_x 1},$$

where $\hat{p}_{J_x} = \arg\max |\widehat{ROC}_x(p) - p|$, $\hat{Q}_{xj}(p) = \hat{\mu}_j(x) + \hat{\sigma}_j(x)\hat{G}_j^{-1}(p)$, for $j = 0, 1$, and $\hat{\mu}_j$, $\hat{\sigma}_j$ and $\hat{G}_j$ denote the estimates given in González-Manteiga et al. (2011).

3 Simulation study

In this section, we study the practical behaviour of the estimators of the covariate-specific Youden index and the corresponding threshold. We have simulated data from two scenarios (S1 and S2) as detailed below:
S1. $\mu_0(x) = 0$, $\mu_1(x) = x$, $\sigma_0^2(x) = \sigma_1^2(x) = 0.5^2$.
S2. $\mu_0(x) = 0.5, \sin(2\pi x)$, $\mu_1(x) = \sin(\pi x)$, $\sigma_0^2(x) = \sigma_1^2(x) = (0.25 + 0.5x)^2$.
In both scenarios, the covariates X_0 and X_1 are uniformly distributed in $[0, 1]$, and the regression errors ε_0 and ε_1 have standard normal distribution. We will focus on three values of the covariate: $x = 0.25, 0.50, 0.75$.
The estimators of the regression curves, $\mu_0(\cdot)$ and $\mu_1(\cdot)$, and variance curves, $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$, which are needed in the construction of the estimator of the conditional ROC curve, are Nadaraya-Watson estimators based on cross-validation bandwidths.

The estimator of the ROC curve proposed and studied in González-Manteiga et al. (2011) depends on a smoothing parameter, h. When $h = 0$, the estimator of the ROC curve is based on the empirical distribution function and the empirical quantile function of the regression errors, and so the obtained estimator is a stepwise function. In order to obtain a continuous estimator of the ROC curve, some smoothing can be added to the empirical estimator by means of the parameter h. In these simulations, we will show results for $h = 0$ (empirical estimator) and $h = 0.05, 0.10$ (smooth estimators).

Table 1 shows the mean square error (MSE) of the estimator of the Youden index and the estimators of the associated threshold (times 100) averaged over 1000 data sets simulated according to scenarios S1 or S2. The sample sizes are $(n_0, n_1) = (100, 100)$, $(100, 200)$ and $(200, 200)$.

TABLE 1. Estimated MSE ($\times 100$) under S1 and S2.

			S1				S2			
x	(n_0, n_1)	h	$\hat{J}_x$	$\hat{c}_{J_x 0}$	$\hat{c}_{J_x 1}$	$\hat{c}_{J_x}$	$\hat{J}_x$	$\hat{c}^0_{J_x}$	$\hat{c}^1_{J_x}$	$\hat{c}_{J_x}$
0.25	(100,100)	0.00	1.41	5.75	5.62	5.52	2.33	3.47	3.61	3.53
		0.05	1.15	5.91	5.56	5.71	1.97	3.58	3.38	3.47
		0.10	1.04	5.98	5.24	5.57	1.81	3.65	3.37	3.49
	(100,200)	0.00	0.94	4.73	4.79	4.77	1.87	2.77	2.81	2.79
		0.05	0.77	4.65	4.57	4.59	1.56	3.09	3.00	3.03
		0.10	0.70	4.80	4.53	4.61	1.43	3.10	2.87	2.94
	(200,200)	0.00	0.60	3.33	3.37	3.35	1.08	2.18	2.20	2.19
		0.05	0.49	3.19	3.17	3.17	0.92	2.13	2.09	2.11
		0.10	0.45	2.99	2.95	2.96	0.85	2.00	1.94	1.97
0.50	(100,100)	0.00	0.97	2.52	2.67	2.59	0.87	1.62	1.70	1.64
		0.05	0.84	2.58	2.47	2.51	0.83	1.71	1.57	1.62
		0.10	0.79	2.24	2.11	2.15	0.81	1.65	1.35	1.46
	(100,200)	0.00	0.77	2.04	2.09	2.07	0.65	1.37	1.42	1.40
		0.05	0.65	1.99	1.96	1.97	0.59	1.32	1.19	1.22
		0.10	0.61	1.74	1.66	1.68	0.57	1.25	0.99	1.05
	(200,200)	0.00	0.52	1.46	1.49	1.47	0.49	0.95	1.01	0.98
		0.05	0.46	1.36	1.35	1.35	0.46	0.92	0.87	0.89
		0.10	0.44	1.17	1.14	1.15	0.45	0.90	0.70	0.78
0.75	(100,100)	0.00	0.75	1.69	1.82	1.74	0.88	2.59	2.71	2.63
		0.05	0.66	1.57	1.51	1.53	0.90	2.26	2.18	2.20
		0.10	0.63	1.30	1.21	1.23	0.91	2.05	1.84	1.89
	(100,200)	0.00	0.59	1.38	1.41	1.40	0.72	2.16	2.22	2.20
		0.05	0.51	1.24	1.18	1.20	0.75	1.96	1.86	1.88
		0.10	0.48	1.12	0.96	1.00	0.77	1.76	1.52	1.57
	(200,200)	0.00	0.38	1.08	1.10	1.09	0.48	1.71	1.77	1.74
		0.05	0.34	0.96	0.94	0.94	0.50	1.44	1.43	1.42
		0.10	0.32	0.76	0.71	0.73	0.52	1.20	1.09	1.12

The table shows that the MSE decreases as the sample sizes increase for both the estimator of the covariate-specific Youden index and the estimators of the associated threshold value in both scenarios and for the three chosen values of the covariate.

In general, the estimator of the Youden index behaves better when some smoothing is applied to the ROC curve, that is, when $h > 0$. Note that in most cases, the values of the MSE are very similar for the two positive

values considered for h.
Concerning the estimators of the associated threshold, the observed MSE suggests that smoothing in the ROC curve produces better results. Among the three estimators of c_{J_x}, the estimator $\hat{c}_{J_x 1}$ gives better results when smoothing is applied to the ROC curve.

4 Example

An example of 286 individuals, analyzed in Smith and Thompson (1996), is used here to illustrate the new approach. The study is related to detection of diabetes, and how the glucose concentration in blood can be used as a classifier. Besides, the age of the patient is also available and used as a covariate. Due to medical reasons, it is known that glucose levels are expected to be higher for older persons even when they are not diabetic. Therefore, in order to check the ability of this classifier, it is necessary to take into account the age of the subject. This dataset has 88 individuals diagnosed as diabetic and 198 as not diabetic. It has been previously analyzed in Faraggi (2003) and González-Manteiga et al. (2011), among others.

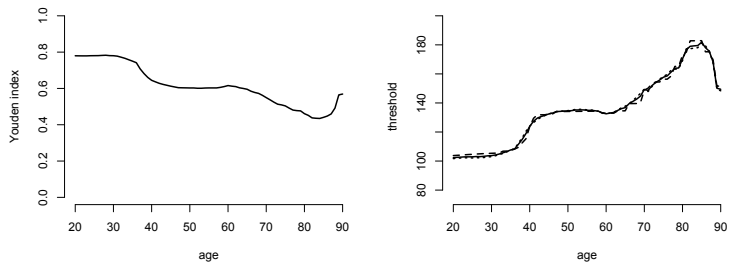


FIGURE 1. Left panel: estimated covariate-specific Youden index. Right panel: estimated covariate-specific threshold associated to the Youden index (solid line: $\hat{c}_{J_x}$; dotted line: $\hat{c}_{J_x 0}$; dashed line: $\hat{c}_{J_x 1}$).

The purpose of the study is twofold: firstly, we want to know if the glucose concentration is a good biomarker to detect diabetes; and secondly, we want to give threshold values associated to the Youden index for each value of the covariate.
The left panel of Figure 1 shows the covariate-specific Youden index estimated for the values of the covariate ranging from 20 to 90. The plot shows that the capability of the glucose concentration to discriminate between the healthy and the diseased populations decreases with the age of the subject. The right panel of the figure shows the estimated thresholds

for each value of the age. The three proposed estimators produce almost the same results. The threshold changes along with the age of the patient. All shown estimators are based on smooth estimators of the ROC curves, with $h = 0.10$.

References

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician*, **52**, 179-192.

González-Manteiga, W., Pardo-Fernández, J.C. & Van Keilegom, I. (2011). ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics*, **38**, 169-184.

Le, C.T. (2006). A solution for the most basic optimization problem associated with an ROC curve. *Statistical Methods in Medical Research*, **15**, 571-584.

Letón, E. and Molanes-López, E.M. (2009). Adjusted empirical likelihood estimation of the Youden index and associated threshold for the bigamma model. *Statistics and Econometrics Series*, 07, Working Paper 09-19.

Pepe, M.S. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, **84**, 595-608.

Pepe, M.S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, **54**, 124-135.

Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press.

Schisterman, E.F. and Perkins, N.J. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics - Simulation and Computation*, **36**, 549-563.

Smith, P.J. and Thompson, T.J. (1996). Correcting for confounding in analyzing receiver operating characteristic curves. *Biometrical Journal*, **38**, 857-863.

An R Package for the Estimation of the Bivariate Distribution for Censored Gap Times

Ana Moreira[1], Luis Machado[2]

[1] Department of Mathematics and Applications, University of Minho, 4810-058 Azurem, Guimaraes. Portugal, Telephone: +351/ 253510400, E-mail: id2809@alunos.uminho.pt ,
[2] E-mail: lmachado@math.uminho.pt

Abstract: In many medical studies, patients can experience several events. The times between consecutive events (gap times) are often of interest and lead to problems that have received much attention recently. In this work we consider the estimation of the bivariate distribution function for censored gap times, using survivalBIV a software application for R. Some related problems such as the estimation of the marginal distribution of the second gap time is also discussed. It describes the capabilities of the program for estimating these quantities using four different approaches, all using the Kaplan-Meier estimator of survival. One of these estimators is based on Bayes' theorem and Kaplan-Meier survival function. Two estimators were recently proposed using the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data (de Uña-Álvarez and Meira-Machado (2008) and de Uña-Álvarez and Amorim (2011)). The software can also be used to implement the estimator proposed in Lin, Sun, and Ying (1999), which is based on inverse probability of censoring weighted. The software is illustrated using data from a bladder cancer study.

Keywords: censoring;Kaplan-Meier;multi-state model;gap times;inverse censoring.

1 Introduction

Let (T_1, T_2) be a pair of gap times of successive events, which are observed subjected to random right-censoring. Let C be the right-censoring variable, assumed to be independent of (T_1, T_2) and let $Y = T_1 + T_2$ be the total time. Because of this, we only observe $(\widetilde{T}_{1i}, \widetilde{T}_{2i}, \Delta_{1i}, \Delta_{2i})$, $1 \leq i \leq n$, which are n independent replications of $(\widetilde{T}_1, \widetilde{T}_2, \Delta_1, \Delta_2)$, where $\widetilde{T}_1 = T_1 \wedge C$, $\Delta_1 = I(T_1 \leq C)$, and $\widetilde{T}_2 = T_2 \wedge C_2$, $\Delta_2 = I(T_2 \leq C_2)$ with $C_2 = (C - T_1)I(T_1 \leq C)$ the censoring variable of the second gap time. Define $\widetilde{Y} = Y \wedge C$ and let F_1 and G denote the distribution functions of T_1 and C, respectively.
This paper describes the R-based `survivalBIV` package's capabilities for implementing nonparametric and semiparametric estimators for the bivari-

ate distribution function for censored gap times. In this work we present four methods (estimators) for the bivariate distribution function of the gap times. One simple estimator is based on Bayes' theorem and Kaplan-Meier survival function. This estimator is related to that proposed in Lin, Sun and Ying (1999) and with estimators proposed by de Uña-Álvarez since all use (in different ways) the Kaplan-Meier estimator (Kaplan and Meier (1958)). The estimator proposed by Lin in 1999 uses Inverse Probability of Censoring Weighted (IPCW) based on the Kaplan-Meier estimator. On the other hand, the idea behind both estimators proposed by de Uña-Álvarez is using the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. Difference between these two methods is that the more recent paper uses a presmoothed version of the Kaplan-Meier estimator.

2 Methodological background

A simple estimator for the bivariate distribution function of the gap times is based on Bayes' theorem and Kaplan-Meier survival function (conditional Kaplan-Meier, CKM). One simple estimator for the bivariate distribution is given by

$$\widehat{F}_{12}(x,y) = \widehat{F}_1(x)\widehat{F}_{KM}(y|T_1 \leq x, \Delta_1 = 1) \tag{1}$$

where $\widehat{F}_1(x)$ is the Kaplan-Meier product-limit estimator based on the pairs $(\widetilde{T}_{1i}, \Delta_{1i})$'s and $\widehat{F}_{KM}(y)$ is the Kaplan-Meier estimator based on the pairs $(\widetilde{T}_{2i}, \Delta_{2i})$'s. The $\widehat{F}_{KM}(y|T_1 \leq x, \Delta_1 = 1)$ is the conditional distribution function for the subset of $T_1 \leq x$ and $\Delta_1 = 1$ (the Kaplan-Meier estimator based on the pairs $(\widetilde{T}_{2i}, \Delta_{2i})$'s such that $\widetilde{T}_{1i} \leq x$ and $\Delta_{1i} = 1$). Another simple estimator was recently proposed by de Uña-Álvarez (2008). The idea behind the estimator is using the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. The proposed estimator (Kaplan-Meier Weighted Estimator, KMW) is given by

$$\widetilde{F}_{12}(x,y) = \sum_{i=1}^{n} W_i I(\widetilde{T}_{1i} \leq x, \widetilde{T}_{2i} \leq y) \tag{2}$$

where $W_i = \frac{\Delta_{2i}}{n-R_i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{2j}}{n-R_j+1}\right]$ is the Kaplan-Meier weight attached to $\widetilde{Y}_i$ when estimating the marginal distribution of Y from $(\widetilde{Y}_i, \Delta_{2i})$'s, and for which the ranks of the censored $\widetilde{Y}_i$'s, R_i, are higher than those for uncensored values in the case of ties. Recently, de Uña-Álvarez propose a modification of estimator (2) based on presmoothing, which allows for a variance reduction in the presence of censoring. By "presmoothing" it is meant that each censoring indicator is replaced by a smooth fit of a binary regression of the indicator on observables. This estimator (Kaplan-Meier

Presmooth Weighted Estimator, KMPW) is expressed as

$$\widetilde{F}_{12}^{\star}(x,y) = \sum_{i=1}^{n} W_i^{\star} I(\widetilde{T}_{1i} \leq x, \widetilde{T}_{2i} \leq y) \tag{3}$$

where $W_i^{\star} = \frac{m(\widetilde{T}_{1i},\widetilde{Y}_i)}{n-R_i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{1j},\widetilde{Y}_j)}{n-R_j+1}\right]$ are the presmoothed Kaplan-Meier weights. Here, $m(x,y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y, \Delta_1 = 1)$, belongs to a parametric (smooth) family of binary regression curves, e.g. logistic. Our package provide the results assuming that m: (a) denotes a logistic regression model (KMPW_glm); (b) denotes an additive logistic regression model (KMPW_gam). Another estimator for the bivariate distribution function was proposed by Lin, Sun and Ying (1999). This estimator is based on inverse probability of censoring weighted (IPCW) and is expressed as

$$\overline{F}_{12}(x,y) = \overline{H}(x,0) - \overline{H}(x,y) \tag{4}$$

where $\overline{H}(x,y) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(\widetilde{T}_{1i} \leq x, \widetilde{T}_{2i} > y)}{1-\widehat{G}((\widetilde{T}_{1i}+y)^-)}$. From (1), (2), (3) and (4) we may obtain an estimator for the marginal distribution of the second gap time.

3 Package Description and Application

The `survivalBIV` software contains functions that calculate estimates for the bivariate distribution function. This software is intended to be used with the R statistical program (R Development Core Team 2010). Our package is composed of 9 functions that allow users to obtain numerical and graphical output for all four methods (CKM, KMW, KMPW and IPCW). In addition, users may generate bivariate survival data from two of the most known copula functions: Gumbel's bivariate exponential distribution, also known as the Farlie-Gumbel-Morgenstern distribution and the bivariate Weibull distribution.

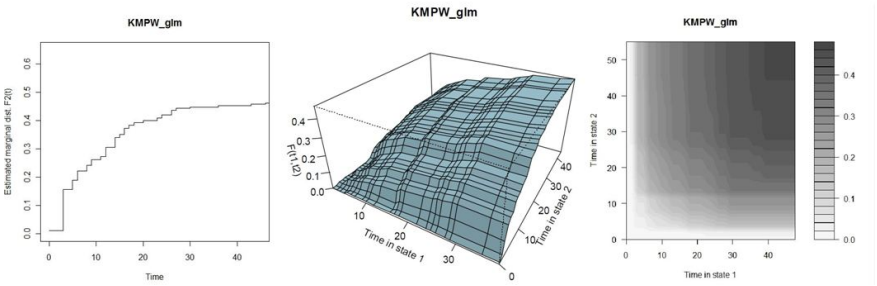
The methods described in Section 2 are illustrated using data from a bladder cancer study (Byar (1980))conducted by the Veterans Administration Cooperative Urological Research Group. In this study, many patients had multiple recurrences (up to a maximum of 9) of tumors during the study. These data are available as part of the R survival package. Here, only the first two recurrence times (in months) and the corresponding gap times, T_1 and T_2, are considered. In the following, we will demonstrate the package capabilities using data from the bladder cancer study (called bladder2). Details about the input data and the functions in the package are reported elsewhere. For illustration purposes we report the estimated values of $F_{12}(29, 17.5)$ for all methods.

```
R>library("survivalBIV")
R>data("bladder2")
R>bladderBIV <- adapt(data = bladder2)
R>summary(bladderBIV,t1=29,t2=17.5,method="all")
```

	CKM	IPCW	KMW	KMPW_glm	KMPW_gam
F(29,17.5)=	0.368028	0.3829919	0.3671296	0.3607358	0.3607358

In this case it is clearly seen that the four methods can provide quite different results. The outputs for the bivariate distribution function and for the marginal distribution of the second gap time are useful displays that greatly helps to understand the patients course over time. Plots for these quantities can easily be obtained. The following input command provides the graphical output for all methods.

```
R>plot(bladderBIV, plot.marginal = TRUE,
plot.bivariate = TRUE, method = "KMPW")
```



4 Conclusion

This paper discusses implementation in R of some newly developed methods for the bivariate distribution function for censored gap times. The `survivalBIV` package uses four nonparametric and semiparametric estimators. One of these estimators is the conditional Kaplan-Meier, based on Bayes' theorem and Kaplan-Meier estimator; also, two recent estimators based on the Kaplan-Meier weights pertaining to the distribution of the total time (time to the second or final event of interest). It also implements the inverse probability of censoring weighted estimator proposed by Lin (1999). Numerical results as well as graphics are easily obtained. We mention two important topics that we shall consider in future versions of the package. First, covariates have not been included in our methods. Another topic of much practical interest is that of providing pointwise confidence bands for these quantities.

Acknowledgments: The authors acknowledge receiving financial support from the Portuguese Ministry of Science, Technology and Higher Education in the form of grants PTDC/MAT/104879/2008 and SFRH/BD/62284/2009. The research was also partially funded by CMAT and FCT under the POCI 2010 program.

References

Byar D (1980). Veterans Administration Study of Chemoprophylaxis for Recurrent Stage I Bladder Tumors: Comparisons of Placebo, Pyridoxine and Topical Thiotepa. *Bladder Tumors and Other Topics in Urological Oncology*, **18**, 363-370.

de Uña-Álvarez J, Amorim AP (2011). A Semiparametric Estimator of the Bivariate Distribution Function for Censored Gap Times. *Biometrical Journal*.

de Uña-Álvarez J, Meira-Machado LF (2008). A Simple Estimator of the Bivariate Distribution Function for Censored Gap Times. *Statistics and Probability Letters*, **78**, 2440-2445.

Kaplan E, Meier P (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Lin D, Sun W, Ying Z (1999). Nonparametric estimation of the time distributions for serial events with censored data. *Biometrika*, **86**, 59-70.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Testing for a breakpoint in segmented regression: a pseudo-score approach

Vito M. R. Muggeo, Gianfranco Lovison

[1] Dipartimento Scienze Statistiche e Matematiche 'S. Vianelli', Università di Palermo, ITALY - email: vito.muggeo@unipa.it, lovison@unipa.it

Abstract: To overcome the well known oddities in testing for the existence of a breakpoint in segmented regression models, we discuss a novel approach based on the generalized Pearson X^2 statistic, which can be considered as an approximation of the Score statistic. We describe the method and present results from some simulations.

Keywords: segmented regression; break-point; hypothesis testing; Pearson chi-squared; non-standard inference.

1 Introduction

The segmented regression model for a response variable Y and a covariate X postulates that the relationship between X and the conditional mean $E[Y|x] = \mu$ is piecewise linear, i.e. two straight lines connected at an unknown point to be estimated. More broadly we can assume the response belongs to the exponential family with link function $g(\cdot)$ leading to the regression equation

$$g(\mu_i) = z_i^T\gamma + \beta(x_i - \psi)_+ \qquad i = 1, 2, \ldots, n \tag{1}$$

where $(x_i - \psi)_+ = (x_i - \psi)I(x_i > \psi)$ and $z_i^T\gamma$ may include additional linear terms, such as other covariates, the model intercept, and the linear term for the segmented variable that represents the 'left slope' of the piecewise relationship. The choice of a variance function $V[Y|x_i] = \phi v(\mu_i)$ completes the specification of the GLM. This paper deals with testing for the existence of ψ in model (1). When ψ does not exist, model (1) reduces to a 'simple' GLM with linear effects. Roughly speaking, estimation and inference in the segmented regression model are difficult and challenging for several reasons. In particular, testing for the existence of a breakpoint is a non-regular problem which makes the usual statistical tests invalid and involves a lot of theoretical issues, see Feder (1975) for an early work on the topic. The traditional tests are far from being helpful in this context: for instance, the null distribution of the likelihood ratio statistic is bimodal with a zero mean, but its analytical density is unknown. At the best of our

knowledge two approaches have been suggested in the literature. Davies (1987) proposed an approach based on the theory of stochastic processes; the test is currently implemented by the `davies.test()` function in the R package segmented (Muggeo, 2008). The other approach by Kim et al. (2000) uses permutations to obtain the null distribution and to compute the p-value accordingly. However, both approaches provide sub-optimal solutions in some contexts: the permutation test has been discussed only for continuous responses using permutations of the residuals of the null fit and therefore generalizations to other responses, e.g. binary, are not immediate; moreover this approach may become computationally cumbersome for large samples. The Davies test may also be hard to use with large datasets, as several fits (about ten) are needed; furthermore it does not generalize to multiple breakpoints. We discuss a simple and very intuitive approach based on a Pearson-type statistic which performs reasonably well under different models and is simple to implement.

2 Methods

We are interested in testing for the existence of the breakpoint in model (1). Without loss of generality, let $\hat{\mu}_{0i}$ be the fitted values for the 'null' (i.e. no breakpoints) model and $\hat{\mu}_i$ the fitted values under the alternative, namely for the segmented regression fit. The link function $g(\cdot)$ and possible presence of additional covariates do not matter. A generalized form of the Pearson statistic which can be used to compare the two models is

$$X^2_{1|0} = \sum_{i=1}^{n} \frac{(\hat{\mu}_i - \hat{\mu}_{0i})^2}{\phi v(\hat{\mu}_{0i})}, \tag{2}$$

where the dispersion parameter, if unknown, is usually replaced by a corresponding consistent estimate. Notice that, when the alternative model is the saturated model, i.e. $y_i = \hat{\mu}_i$, $X^2_{1|0}$ is the usual Pearson goodness of fit statistic which is equivalent to the score statistic for any GLM. Lovison (2005) showed that for canonical GLM $X^2_{1|0}$ is greater than the equivalent score statistic and he also gave an X^2-like formula for the score statistic. Motivated by these connections, the Pearson-type statistic (2) is referred to as *pseudo*-score statistic, and Agresti and Ryu (2010) used it to build confidence intervals in discrete statistical models. Here we use it for testing for a breakpoint in segmented GLMs, where the usual asymptotic tests fail and current proposals do not appear to be fully satisfactory.
To perform hypothesis testing we need to known the null distribution of $X^2_{1|0}$. With respect to the null linear model, the segmented 'alternative' model has two additional parameters, the difference in slope parameter and the breakpoint, therefore it seems reasonable that under H_0 $X^2_{1|0} \stackrel{d}{\rightarrow} \chi^2_2$. Like for interval estimation problems in Agresti and Ryu (2010), we do not

yet have formal arguments to show that the chi-squared distribution holds under H_0, but we show its performance via simulations.
Table 1 reports the actual sizes of the pseudo score statistic $X^2_{1|0}$ to test for the existence of the breakpoint. We consider different scenarios, with four sample sizes and three densities for the responses: Gaussian, $Y_i \sim \mathcal{N}(\mu_i = 0.15x_i, 0.01^2)$; Poisson, $Y_i \sim \mathcal{P}(\mu_i = e^{2+0.5x_i})$; Negative Binomial, $Y_i \sim \mathcal{NB}(\mu_i = e^{2+0.5x_i}, \mu_i + \mu_i^2/2)$, where $x_i = i/n$ in every scenario. The Negative Binomial family has been selected to assess the performance of the pseudo-score statistic when the model is estimated via a quasi-likelihood approach; for this and the Gaussian example, the dispersion parameter is assumed unknown and it is replaced by a corresponding method-of-moments estimate under the null hypothesis (i.e. from the linear model).

TABLE 1. Empirical sizes (based on 2,000 replicates) of the pseudo score test testing for a breakpoint in different scenarios.

	Gaussian			Poisson			Negative Binomial		
n	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
50	0.007	0.045	0.099	0.015	0.058	0.104	0.009	0.050	0.106
100	0.012	0.055	0.112	0.013	0.059	0.109	0.016	0.061	0.111
500	0.011	0.056	0.105	0.014	0.057	0.118	0.013	0.050	0.110
1000	0.010	0.047	0.093	0.010	0.057	0.112	0.012	0.061	0.117

We observe that the pseudo score test for a breakpoint in segmented regression performs reasonably well by providing empirical sizes close enough to the corresponding nominal values.
Table 2 shows the power of the proposed X^2-type test and the Davies test in detecting a changepoint: we consider two sample sizes ($n = 50, 100$), $\mu_i = 0.05(x_i - \psi)_+$ for Gaussian responses and $\mu_i = e^{2+(x_i-\psi)_+}$ for Poisson and Negative Binomial responses; we also assess the effect of the location of the breakpoint by considering $\psi = 0.50$ and $\psi = 0.75$.

TABLE 2. Empirical power at level 0.05 (based on 1,000 replicates) of the pseudo score test and the Davies test in different scenarios.

			Family		
ψ	n		Gaussian	Poisson	Neg Binom
0.50	50	X^2	0.593	0.269	0.099
		Davies	0.555	0.227	0.096
	100	X^2	0.910	0.457	0.135
		Davies	0.879	0.374	0.119
0.75	50	X^2	0.303	0.125	0.070
		Davies	0.282	0.100	0.091
	100	X^2	0.568	0.227	0.098
		Davies	0.482	0.170	0.088

As expected both tests perform better when ψ is in the middle of the

range of the segmented variable and with larger sample sizes. Although the differences are moderate, generally $X^2_{1|0}$ outperforms the Davies test, and moreover it is actually much simpler to compute, since it requires only two fits.

3 Conclusions

The estimation problem for GLMs involving segmented relationships appears to have received much attention by several authors in the literature, and different solutions are available; see for instance Muggeo (2003). On the other hand, hypothesis testing problems currently present open research questions. In this paper, we have illustrated a very simple, intuitive, and general approach to the problem of testing for a breakpoint in GLMs. Results from some simulation studies show that the Pearson-type statistic provides satisfactory results, at least in the simple case of testing '1 vs. 0' breakpoints. Possible further uses of the Pearson-type statistic concern testing with multiple breakpoints, e.g. 2 vs. 1 or 0 breakpoints, and testing under model misspecification, e.g. in the presence of heteroscedasticity and autocorrelation with continuous responses. These topics need further investigation.

References

Agresti, A., and Ryu, E. (2010) Pseudo-Score Confidence Intervals for Parameters in Discrete Statistical Models *Biometrika*, **97**, 215-222.

Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33-43.

Feder, P.I. (1975). The log likelihood ratio in segmented regression. *Annals of Statistics* **3**, 84-97.

Kim, H.-J., Fay, M.P., Feuer, E.J., and Midthune, D.N. (2000). Permutation Tests for Joinpoint Regression with Applications to Cancer Rates. *Statistics in Medicine* **19**, 335-351.

Lovison, G. (2005) On Rao score and Pearson X^2 statistics in generalized linear models. *Statistical Papers* **46**, 555-574

Muggeo, V.M.R. (2003). Estimating regression models with unknown breakpoints. *Statistics in Medicine*, **22**, 3055-3071.

Muggeo, V.M.R. (2008). Segmented: an R package to fit regression models with broken-line relationships. *R News* **8**(1), 20-25.

Geostatistical modelling with non-Euclidean distances

Facundo Muñoz[1], Antonio López-Quílez[1]

[1] Universitat de València, Spain

Keywords: Metric; Distances; Positive-definiteness; Geostatistics

1 Motivation

Geostatistics provides a set of statistical tools specifically designed for spatial problems, in which prediction is required over a region of interest where some observations have been taken. Predictions are based on an underlying statistical model that can take additional information into account as explanatory variables. In addition, the prediction error can be derived.
Specifically, let $Z(\boldsymbol{s}_i)$, $i = 1, \ldots, n$ be a set of measurements at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ in a (typically) 2-dimensional region D. These measurements are often assumed to be one realisation of a random process $Z(\cdot)$ such that:

- $E(Z(\cdot)) = \mu$,
- $C(\boldsymbol{s}_i - \boldsymbol{s}_j) = \mathrm{cov}(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j))$, $\forall \boldsymbol{s}_i, \boldsymbol{s}_j \in D$, exists and only depends on the vector $\boldsymbol{s}_i - \boldsymbol{s}_j$.

These assumptions form the *second order* (or weak or wide-sense) *stationarity hypothesis* (Cressie 1993, p.53). Additionally, *isotropy* is often assumed, where $C(\boldsymbol{s}_i - \boldsymbol{s}_j) = C(|\boldsymbol{s}_i - \boldsymbol{s}_j|)$; that is, the covariogram depends only on the length of the vector $\boldsymbol{s}_i - \boldsymbol{s}_j$ and not on its direction.
The stationarity and isotropy assumptions make no sense when the region of interest is very heterogeneous, or has irregularities affecting the structure of correlations. The covariogram depends not only on the distance between locations, but also on the geographical configuration of the environment. These are clearly non-stationary situations. One extreme case is the presence of barriers in the region of interest.
One way of dealing with this is using *Cost-based distances* (Krivoruchko and Gribov 2002, López-Quílez and Muñoz 2009). With this approach, a cost-surface must be defined, representing how difficult is the flow of information at every location. For instance, a barrier gets an infinite cost, while regular medium gets cost 1. The Cost-based distance between two locations is defined as the length of the minimum-cost path between them. In this way, we have incorporated the geographical information into the

distances between locations, and once again, we are able to talk about stationarity and isotropy with respect to the Cost-based distances.
However, a fundamental problem arises. Note that for every set of locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$, the corresponding covariance matrix $(\mathrm{cov}(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)))$ must be positive definite. This sets up a condition over the covariance function called *positive definiteness*. When the distances are Euclidean, the family of (isotropic) positive definite functions is fully characterised by Schoenberg's (1938) theorem. This led to a number of parametric families of valid functions with varying properties that are available to the user of geostatistical methods. However, when the distances are not Euclidean there is no guarantee of positive-definiteness, and there are no general results characterising the family of positive definite functions.
It is our goal to investigate the properties of the family of positive definite functions with respect to Cost-based distances.

2 Approaches

2.1 Riemannian manifolds

Mathematically, we can think of the region D as a Riemannian manifold. Let $\mathfrak{c}(\boldsymbol{s})$ be the cost surface. We can define the Riemannian metric as

$$g_{\boldsymbol{s}}(\boldsymbol{u}, \boldsymbol{v}) = \mathfrak{c}(\boldsymbol{s})^2 \langle \boldsymbol{u}, \boldsymbol{v} \rangle, \tag{1}$$

for all $\boldsymbol{u}, \boldsymbol{v}$ in the tangent space $T_{\boldsymbol{s}}D$, where $\langle \cdot, \cdot \rangle$ represent the Euclidean scalar product. With this setting, a metric is naturally induced as

$$\tau_g(\boldsymbol{s}, \boldsymbol{t}) = \inf \left\{ L(c) : \ c \in D^1([0,1]; D)_{(\boldsymbol{s}, \boldsymbol{t})} \right\}, \tag{2}$$

where

$$L(c) = \int_0^1 \sqrt{g_{c_x}(c'_x, c'_x)}\, dx = \int_0^1 \mathfrak{c}(c_x)|c'_x|\, dx, \tag{3}$$

is the length of the curve c, and $D^1([0,1]; D)_{(\boldsymbol{s}, \boldsymbol{t})}$ is the set of all piecewise continuously differentiable maps $c : [0,1] \to D$ with $c(0) = \boldsymbol{s}$ and $c(1) = \boldsymbol{t}$. In other words, the *distance* between points $\boldsymbol{s}$ and $\boldsymbol{t}$ is the infimum of the lengths of the (continuous) paths connecting $\boldsymbol{s}$ and $\boldsymbol{t}$.
Characterising the positive definite functions over D in the spirit of Shoenberg's theorem involves developing Fourier and spectral analysis in this (much) more general context. This is an open line of work.

2.2 Pseudo-Euclidean spaces

One approach that has been considered is the use of Multidimensional Scaling (MDS) for *representing* the set of locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ and their Cost-based distances in a Euclidean space, where standard covariance functions

can be used. This is an unsatisfying approximation, though. However, the locations can be *exactly represented* in a pseudo-Euclidean space.
A pseudo-Euclidean space is a vector space of dimension d, say $\mathbb{R}^d$, with a non-degenerate symmetric bilinear form

$$\begin{aligned}(\cdot,\cdot) &: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \\ (\boldsymbol{x},\boldsymbol{y}) &= (x_1 y_1 + \cdots + x_k y_k) - (x_{k+1} y_{k+1} + \cdots + x_d y_d),\end{aligned} \tag{4}$$

where k is called the *index*, while the pair $(k, d-k)$ is called the *signature* of the space. The space is denoted $E_{(k,d-k)}$.
This was promising, since a pseudo-Euclidean space provides enough structure to try to characterise the family of positive definite functions following the ideas of Schoenberg. The procedure involves assuming that a given isotropic and stationary correlation function is positive definite, and integrating it out over a *sphere* of radius r.
In the pseudo-Euclidean space, the sphere (as a surface of constant radius) becomes a hyperboloid. In contrast to the sphere, a hyperboloid has an infinite area, and this fact causes divergence in most integrals. This leaves little hope of characterising its positive definite functions.
Possibly, this space is much bigger than needed. The pseudo-Euclidean space is actually able to represent any set of points with prescribed *dissimilarities*; i.e., not necessarily satisfying the triangle inequality. However, the Cost-based distance is a (full) *metric*. This means that the family of positive definite functions in the pseudo-Euclidean space (which includes the trivial constant function 1) is a subset of those in the space D,

$$1 \in \mathfrak{P}(E_{(k,d-k)}) \subset \mathfrak{P}(D). \tag{5}$$

Here we have more open questions. Are there more functions (apart from the trivial) in $\mathfrak{P}(E_{(k,d-k)})$? Is there a way of characterising them all? Are there functions in $\mathfrak{P}(D)$ not positive definite in the pseudo-Euclidean space? We believe there are. Specifically, we have shown that the exponential correlation function is not positive definite in the pseudo-Euclidean space, while it has produced positive definite covariance matrices in all the examples we have run.

2.3 Bayesian simulation

While we don't know any family of positive-definite functions, we can use some sort of an accept-reject method in a Bayesian hierarchical spatial model like the following.
In model 1, the locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ determine the matrix $\mathbf{D}_{\text{cb}}$ of Cost-based distances. This matrix is non-negative, symmetric and has zeroes in its diagonal. The corresponding measurements $y_1, \ldots, y_n$ are (say) Gaussian and a spatial effect $\boldsymbol{\omega}$ is introduced in their mean. This effect is given a

$$\begin{aligned} \boldsymbol{s}_1,\ldots,\boldsymbol{s}_n &\rightsquigarrow \mathbf{D}_{\text{cb}} = (r_{ij}); \quad r_{ii}=0,\, r_{ij}\geq 0,\, r_{ij}=r_{ji} \\ y_1,\ldots,y_n &\rightsquigarrow \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \tau^2\mathbf{I}); \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\omega} \\ &\quad\; \boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{P}) \\ &\quad\; \mathbf{P} = f(\mathbf{D}_{\text{cb}}) \\ &\quad\; f \sim \cdots; \quad f(0)=1, |f(r)|\leq 1, f(\mathbf{D}_{\text{cb}})\text{p.d.} \end{aligned} \tag{6}$$

correlation matrix which is a (unknown) transformation f of the Cost-based distances, with some constraints, including the positive-definite condition. We could simulate f from a family of functions (maybe a combination of a base of functions), rejecting those giving rise to non positive-definite matrices, and perform inference on the parameters of the model.
This procedure still lacks theoretical foundation. Although this specific matrix is positive-definite, if any other location $\boldsymbol{s}$ was added, the extended transformed matrix might not be. Hence, the theoretical stochastic process is not necessarily valid.

3 Conclusions

This is pure ongoing work. Although it started as a very applied project, it turned quickly into deep mathematics. There are various lines of work, involving diverse areas of mathematics, such as Topology, Measure Theory, Geometry, Fourier analysis, Algebra, etc. It has been difficult to cover all these fields at once and combine them coherently to achieve results.

Acknowledgments: This research has been partially supported by the Ministerio de Ciencia e Innovación grant MTM2010-19528.

References

Cressie, N. (1993). *Statistics for Spatial Data,* Wiley, New York.

Krivoruchko K., Gribov, A. (2002). Geostatistical interpolation and simulation with non-Euclidean distances. In Sanchez-Villa, Carrera, & Gomez-Hernandez (eds.), *geoENV IV*, pp. 331-342. Kluwer Academic.

López-Quílez, A., Muñoz, F. (2009). Geostatistical computing of acoustic maps in the presence of barriers *Mathematical and Computer Modelling*, **50**(5-6): 929–938.

Schoenberg, I. J. (1938). Metric Spaces and Completely Monotone Functions *The Annals of Mathematics*, **39**(4): 811–841.

Multi-state models for non Markov process

Magdalena Murawska[1*], Dimitris Rizopoulos[1], Emmanuel Lesaffre[12]

[1] Department of Biostatistics, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands,*e-mail: m.murawska@erasmusmc.nl
[2] I-Biostat, Catholic University of Leuven, Belgium

Abstract: In multi-state models the interest is typically in modeling the transition probabilities using baseline covariates. The Aalen-Johansen (A-J) estimator is the standard nonparametric estimator to estimate transition probabilities of a time-inhomogeneous Markov multi-state model. However when the Markov assumption is violated, A-J estimators may be systematically biased. One can still estimate stage occupation probabilities and the A-J method provides consistent estimators in this case. In this paper we use the pseudo-values approach for direct modeling of the state probabilities using covariates. The approach is exemplified in a study on heart transplant data.

Keywords: Heart Transplant Data; Aalen-Johansen; Pseudo-values.

1 Introduction

In transplantation studies often categorical longitudinal measurements are collected for patients waiting for an organ transplant. It is of primary interest to assess whether available history of the patient can be used for predicting patient survival as well as further performance on the waiting list.
In this work we suggest to use a multi-state models approach to handle this aim. Typically in the multi-state models framework Markov models are used, mainly due to the availability of software. However sometimes the model assumptions are not supported by the data (as for our data set), and alternative approaches are required. Here we propose to use the pseudo-value approach introduced by Andersen et al (2003) in the context of a competing risks problem. We apply this approach for the A-J estimator for state occupation probabilities. This approach can be used for any general multi-state model and with some additional programming it can be applied using standard software.

1.1 Motivating Data Set

The data come from the Eurotransplant heart recipients list, which contains 2921 recipients who entered the waiting list from 01.01.2006 to 31.12.2008.

Each recipient was classified to one of the following states: Transplantable (T), Non-Transplantable (NT), Urgent (U) and High Urgent (HU). The first evaluation took place at entry and additional evaluations were performed while the patient remained on the waiting list. The follow-up was censored at 31.03.2010. By that date 528 patients had died (D) without receiving a transplant, 1565 patients received a transplant (TT) and 239 patients had been removed (R) because of other reasons. Additionally at entry the list the following baseline information was recorded, namely: age, height, weight, country (7 countries), blood group, panel reactive antibodies level (PRA) (in percentages), cardiovascular disease (categorized into Dilated Cardiomyopathy (DCM), Coronary Artery Disease (CAP) and others). Additionally information about having Ventricular Assist Device (VAD), a mechanical pump that supports heart (Y/N), was collected.
The purpose of the study was to predict the state of the patient based on the history on the waiting list and to examine the effect of baseline covariates on this prediction.

2 Background Methodology

A multi-state process is defined as a stochastic process $\{X(t), t \in \mathcal{T}\}$ with a finite space state $S = \{1, \ldots, N\}$ and time interval $\mathcal{T} = [0, \tau]$, with $\tau < \infty$. For a particular time t, $X(t)$ is the state occupied at that time. The process is characterized by the transition probabilities between states h and r, defined as:

$$p_{hr}(s,t) = P(X(t) = r \mid X(s) = h, H_{s-}), h, r \in S, s, t \in \mathcal{T}, s \leq t, \qquad (1)$$

where H_{s-} denotes the history of the process up to time s. The process could be alternatively characterized by its transition intensities:

$$q_{hr}(s) = \lim_{\Delta s \to 0} \frac{p_{hr}(s, s + \Delta s)}{\Delta s}, \qquad (2)$$

which are the instantaneous hazard of progression from state h to state r, conditionally on being at state h. Both $p_{hr}(s,t)$ and $q_{hr}(s)$ may in principle depend on the history H_{s-}. For Markov models the transition intensities depend only on the current state. Moreover we can additionally assume that the intensities are constant over time (Time Homogenous Models) or depend on time (Non-Homogenous Models).
For the Heart Data we have the four possible states (T, NT, U and HU) while remaining on the waiting list, and three states corresponding to a removal of a patient from the list (D, R, TT). Therefore, there are 3 absorbing states $h \in \{D, R, TT\}$, for which $p_{hh} = 1$ and 4 transient states $r \in \{T, NT, U, HU\}$, for which $p_{rr} < 1$.

3 Proposed Methodology

Initial analyses of the heart transplant data with a Markov model revealed (even after adjusting for covariates) considerable discrepancies between the fitted and observed probabilities. There might be two main causes for these discrepancies. First, the transition rates may vary within time or another omitted covariate (non homogenous model). Secondly, the Markov assumption may be violated. In particular, the transition intensities may depend on time spent in current state (semi-Markov process) or the history of the process in general. To investigate this we relaxed the homogeneity assumption by allowing the intensities to be piecewise constant (PCI model) and changing only at arbitrary chosen times. Nevertheless the PCI model did not lead to a substantial improvement of the model fit, and therefore in the next step we estimated each of the transition probabilities as a function of time using the nonparametric Aalen-Johanson estimator (Aalen and Johansen 1978), defined as:

$$\hat{P}(s,t) = \prod_{s<t_j\leq t} (I + \hat{Q}_j), \tag{3}$$

where t_j is assumed to be an exact time of transition from state h to r, I is the $N \times N$ identity matrix and $\hat{Q}_j$ is the $N \times N$ intensity matrix with (h,r) element $\hat{Q}_{hrj} = d_{hrj}/r_{hrj}$, where d_{hrj} is the number of individuals who experience a transition from state h to r at time t_j and r_{hrj} is the number of individuals in state h just prior to time t_j.
The A-J estimates were examined with respect to different baseline covariates. Also dependence of the A-J estimates on the history of the process was considered. A-J estimates for transition probabilities are consistent if the Markov assumption is fulfilled. However, since the previous analysis revealed dependence on the previous state, in the final analysis we use the A-J estimates for the state occupation probability $p_h(t)$, that is the probability of occupying state h at time t :

$$p_h(t) = \sum_{k=1}^{N} p_k(0) \cdot p_{kh}(0,t), \tag{4}$$

which is consistent regardless of the Markov assumption being satisfied. To measure the effect of covariates in these occupation probabilities we employed the pseudo-values approach proposed by Andersen et al (2003). To introduce this approach, let $\hat{F}^h(t)$ be the A-J estimator for $p_h(t)$ calculated for all n individuals. Denote $\hat{F}^h_{-i}(t)$ the A-J estimate excluding individual i. Then the pseudo-value for subject i at time t is defined as:

$$\hat{\theta}_i(t) = n\hat{F}^h(t) - (n-1)\hat{F}^h_{-i}(t). \tag{5}$$

The pseudo-values are calculated for arbitrary chosen time points $t_1, t_2, \ldots, t_k$ for each individual, obtaining thus k pseudo-values per subject. Next we

fit a regression model on the pseudo-values using a GEE (Generalized Estimating Equations) model:

$$g(\hat{\theta}_i(t)) = \beta^T Z_i, \tag{6}$$

with $g(\cdot)$ denoting a link function and Z_i - the design matrix of the covariates of interest. Estimates of β are based on unbiased estimating equations and a sandwich estimator is used to estimate the variance of $\hat{\beta}$.

4 Results and Conclusions

We applied the pseudo-values approach to the Heart Data Set. We chose 7 time points, the logit link and an unstructured correlation matrix. As covariates we included time and the baseline characteristics. Results from the regression on pseudo-values revealed that AB blood group increased the probability of getting a transplant. Patients from countries with informed consent (permission for organ donation required) had lower transplantations rate compared to patients from countries with presumed consent law. Current state HU increased the probability of getting a transplant. When previous state was T, patients with blood group 0 as well as patients with CAD or DCM disease were less likely to have a transplant from HU. From the current states other than HU patients were most likely to go back to the previous state. The exception was the current state NT and previous HU, when the risk of death was the highest. Longer history as well as the length of time spent in the previous state were not found to be important. The theoretical justification for the pseudo-values approach was given for the simple multi-state models (Graw and Gerds 2009) and the impact of the number of time grid points in GEE model was confirmed based on simulation studies. Further research is needed to investigate the approach for more complicated models under different scenarios.

References

Aalen, O.O., Johansen, J.P. (1978). An Empirical Transition Matrix for Non-Homogenous Markov Chains Based on Censored Observations *Scand. J. Statist.*, **5**,141-150

Andersen, P.K., Klein, J.P. (2007). Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies. *Scand. J. Statist.*, **34**,3-16

Andersen, P.K., Klein, J.P., Rosthøj, S. (2003). Generalized linear models for correlated pseudo-observations with applications to multi-state models. *Biometrika*, **90**,15-27

Graw, F., Gerds, T.A. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Anal.*, **15**,241-255

Some approaches to correct for misclassification in the absence of an internal validation data set

Timothy Mutsvari[1], Dominique Declerck[2], Emmanuel Lesaffre[13]

[1] I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium,
[2] School of Dentistry, Katholieke Universiteit Leuven, Leuven, Belgium,
[3] Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

Abstract: The relationship between covariates and a binary outcome can be distorted in the presence of misclassification errors in the response. To correct for misclassification an internal validation data set is needed, i.e. a random sample of the main data. However, it may be challenging to obtain internal validation data in practice. Rather, external validation data sets are often obtained. External validation data may differ in many ways from internal validation data. Therefore different approaches may be necessary in order to make use of the obtained external validation sample to correct for misclassification in the main data. We focus here on the approach which resembles best what happened in our motivating data set obtained from the Signal Tandmobiel® (ST) study. The approach is to correct for differential misclassification in the main data by conditioning the misclassification probabilities on a rich structure of covariates such that the external validation data come closer to the internal one. We explore the relationship of various factors and caries experience on children of age twelve by a multilevel model.
Keywords: Misclassified binary data; Differential; Non-differential; Validation.

Keywords: Misclassified binary data; Differential; Non-differential; Validation.

1 Introduction

Many epidemiologic studies attempt to characterize associations between risk factors and disease occurrence based on data from the main study. However, they often rely on disease exposure assessments by diagnostic tests that are subject to misclassification error and this may introduce bias into the study results. Research on misclassification revealed that, under non-differential misclassification (misclassification not depending on covariates), the regression coefficients are attenuated towards the null, see e.g. Bross (1954). The effect of differential misclassification is, however, less clear.

Caries experience (CE) studies suffer from misclassification errors, see e.g. Lesaffre et al. (2004). In order to correct for misclassification an internal validation data set is needed, i.e. a random sample of the main data. However, internal validation data may be challenging to obtain in practice due to several constraints. Instead, external validation data are often obtained under different scenarios. External validation data differ from internal validation data by (i) being a sample but not a random sample from the population of interest, taken under identical conditions as in the main data, (ii) being a sample but not a random sample from the population of interest but now taken under different conditions than in the main data and (iii) being a sample taken from a different population. In the absence of an internal validation data, one can assume informative priors of the misclassification parameters and proceed with Bayesian approach in estimating the model parameters. Yet another approach is to make use of the available external validation data of type (i), (ii) or (iii) by using survey random sampling techniques. We focus here on the external validation sample of the first scenario. In order to correct for differential misclassification, we propose to condition misclassification probabilities on a rich set of covariates, thereby coming close (hopefully) to internal validation data.
The proposed approach is applied to investigate the factors affecting CE in the Signal Tandmobiel® (ST) study, which is a longitudinal oral health study conducted in Flanders (Belgium). For this project, 16 trained dentists (examiners) conducted annual examinations of children. In the ST study, children that participated in the calibration exercises provided the validation data. However, it was not possible to take the validation data set at random from the main data. Rather a school was selected with a presumed high prevalence. The outcome of interest is the binary score CE (CE=1 if the surface shows CE and 0 if not) subject to misclassification. We will also account for the multilevel structure of CE data.

2 Statistical modeling approach

Multilevel model for cross-sectional true CE data

Let Y_{stme} be the true CE score of surface s, $(s = 1, \ldots, n_t)$ nested in tooth $t = 1, \ldots, n_m$, which is nested in child/mouth $m = 1, \ldots, N$ according to examiner e, $(e = 1, \ldots, n_e)$ in the main study. The model uses $\pi_{stme} = Pr(Y_{stme} = 1|\boldsymbol{\beta}, \mathbf{x}_{stme}, \mathbf{u}_m)$, which is the true conditional probability for CE on surface s nested in tooth t in mouth m from the main data set. The multilevel logistic model for the true main data is given by:

$$logit(\pi_{stme}) = \mathbf{x}_{stme}^T\boldsymbol{\beta} + u_m + u_{tm} + u_e, \tag{1}$$

where $\mathbf{u}_m = (u_m, u_{tm}, u_e)$ is a set of random effects assumed to be independently distributed with mean zero and variances $\sigma_m^2, \sigma_{tm}^2, \sigma_e^2$ at mouth,

tooth (nested in mouth) and examiner level respectively. $\mathbf{x}_{stme}$ is a vector of covariates associated to the regression coefficients $\boldsymbol{\beta}$.

Models for the validation data

Let $\tau_{11} = Pr(Y^*_{stme} = 1|Y_{stm} = 1, \boldsymbol{\alpha}, \mathbf{z}_{stme})$ and $\tau_{00} = Pr(Y^*_{stme} = 0|Y_{stm} = 0, \boldsymbol{\eta}, \mathbf{z}_{stme})$ be the differential sensitivity (SE) and specificity (SP) with $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ as the regression coefficients respectively. $\mathbf{z}_{stme}$ is a set of covariates for both SE and SP. The logistic models (which can be extended to contain random effects) for SE and SP are given by:

$$logit(\tau_{11}) = \mathbf{z}^T_{stme}\boldsymbol{\alpha}, \tag{2}$$

$$logit(\tau_{00}) = \mathbf{z}^T_{stme}\boldsymbol{\eta}. \tag{3}$$

Multilevel model for cross-sectional observed CE data

Let $Y^*_{M,stme}$ be the observed CE score in the main data. Using models for τ_{11} and τ_{00} above, the corrected multilevel logistic model for the observed main data based on the approach of Neuhaus (2002) with *link* function g is given by:

$$Pr(Y^*_{M,stme} = 1|\mathbf{u}_{tme}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{x}_{stme}) = (1-\tau_{00})+[\tau_{11}+\tau_{00}-1][g^{-1}(\mathbf{x}^T_{stme}\boldsymbol{\beta}+u_m+u_{tm}+u_e)],$$

3 Results

The variables gender, dentition type, tooth type and surface type were considered as risk factors in the model for validation data. Similar risk factors were considered for the main data in addition to age and the geographical location (represented by the standardized (x,y) coordinate of the municipality of the school to which the child belongs). The results of the validation data model (SE and SP) are not shown here. Table 1 shows the results of the three multilevel models (no correction, non-differential correction and differential correction) for the main data. The parameter estimates for the differential correction are generally higher than those of non-differential correction. Also, the 95% credible intervals for differential correction are wider than for the non-differential one. We note a significant effect of the x-coordinate under differential correction which was not the case in the non-differential one. This change was also reported in previous research by our group, see e.g. Lesaffre et al. (2004). The positive effect of tooth type (canine versus incisor) changed to a negative one under differential correction. Larger estimates of random effects were observed for the differential correction compared to the non-differential.

TABLE 1. Parameter estimates of the main model for cross-sectional data with no-correction, non-differential and differential correction

Parameter	No Correction (NC) Estimate[2.5% , 97.5%]	Non-Differential (ND) Estimate[2.5% , 97.5%]	Differential (D) Estimate[2.5% , 97.5%]
FIXED EFFECTS:			
Intercept	-8.86[-9.79 ; -8.12]	-10.86[-12.31 ; -9.53]	-10.73[-12.78 ;-9.23]
Gender			
Girls	-0.07[-0.63 ; 0.48]	-0.18[-0.85 ; 0.57]	-0.07[-1.00 ; 0.87]
Boys	...	...	...
Age	1.27[0.56 ; 2.04]	1.67[0.61 ; 2.68]	2.12[0.95 ; 3.29]
Geographical location			
x-coordinate	0.27[-0.05 ; 0.59]	0.37[-0.01 ; 0.77]	0.50[0.02 ; 1.00]
y-coordinate	-0.18[-0.46 ; 0.10]	-0.30[-0.71 ; 0.14]	-0.35[-0.84 ; 0.14]
Type			
Permanent	-2.12[-2.38 ; -1.86]	-2.79[-3.20 ; -2.42]	-3.26[-3.81 ; -2.76]
Deciduous	...	...	...
Tooth-type			
Canine	-0.14[-0.83 ; 0.56]	0.32[-0.65 ; 1.31]	-0.69[-2.05 ; 0.54]
Molar	3.98[3.51 ; 4.60]	5.46[4.54 ; 6.43]	3.99[2.88 ; 5.09]
Premolar	-1.95[-2.70 ; -1.18]	-2.14[-2.80 ; -1.48]	-4.17[-5.71 ; -2.84]
Incisor	...	...	...
Surface-type			
Distal	0.86[0.58 ; 1.12]	1.17[0.81 ; 1.55]	1.42[0.89 ; 2.05]
Mesial	1.55[1.29 ; 1.80]	2.13[1.74 ; 2.55]	2.96[2.35 ; 3.82]
Lingual	0.13[-0.16 ; 0.40]	0.19[-0.20 ; 0.56]	0.47[-0.05 ; 0.99]
Occlusal	3.64[3.37 ; 3.90]	4.65[4.15 ; 5.20]	5.17[4.60 ; 5.91]
Buccal	...	...	...
RANDOM EFFECTS:			
σ^2_{mouth}	6.50[5.20 ; 8.18]	10.82[8.12 ; 14.44]	15.92[11.97 ; 21.90]
σ^2_{tooth}	3.35[2.86 ; 3.96]	4.75[3.42; 6.45]	6.50[4.97 ; 9.36]
$\sigma^2_{examiner}$	0.13[0.0004 ; 0.72]	0.27[0.0001 ; 1.44]	0.30[0.002 ; 1.69]

4 Discussion

Misclassification of a disease outcome is common in studies that involve multiple examiners since different examiners exhibit different scoring behaviors. In order to correct for misclassification in the main data, information about the misclassification probabilities is needed. The most common being the use of validation data or via a double sampling procedure. Another approach is to elicit prior information of misclassification from experts and proceed with Bayesian methods for estimating the model parameters. However informative prior information can be subjective. Generally, external validation data are often gathered rather than internal. However, the consequence of using external validation to correct for misclassification has not been widely discussed. Several approaches such as survey sampling techniques are useful tools to remedy this problem. Here we opted to condition misclassification probabilities on a rich set of covariates, thereby coming close (hopefully) to internal validation data.

References

Bross, I. (1954). Misclassification in 2 × 2 tables. *Biometrics*, **6**, 478-486.

Lesaffre, E., Mwalili, S. and Declerck, D. (2004). Analysis of caries experience taking inter-observer bias and variability into account. *Journal of Dental Research*, **83**, 951-955.

J. M. Neuhaus, J.M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*, **58**, 675-683.

Phylogenetic models for Semitic vocabulary.

Geoff K Nicholls[1] , Robin J. Ryder[2]

[1] Statistics Department, University of Oxford, 1 South Parks Road, Oxford OX13TG, UK, nicholls@stats.ox.ac.uk
[2] Centre De Recherche Économie et Statistique, ENSAE

Abstract: Kitchen *et al.* (2009) analyze a data set of lexical trait data for twenty five Semitic languages, including ancient languages Hebrew, Aramaic and Akkadian, modern South Arabian and Arabic languages and fifteen ethiosemitic languages. They estimate a phylogenetic tree for the diversification of lexical traits using tree and trait models and methods set up for genetic sequence data. We reanalyze the data in a homplasy-free model for lexical trait data. We use a prior on phylogenies which is non-informative with respect to some of the key scientific hypotheses (concerning topology and root time). Our results are in broad agreement with those of Kitchen *et al.* (2009), though our 95% HPD for the root of the Semitic tree (the branching of Akkadian) is $[4400, 5100]$BP and we place Moroccan and Ogaden Arabic in the Modern South Arabian Group.

Keywords: Bayesian Phylogenetics, Cultural traits, Languages

1 Data and problem statement

Kitchen *et al.* (2009) give a Bayesian phylogenetic analysis of lexical trait data for $L = 25$ Semitic languages: Ugaritic, Ge'ez, and the languages shown in Figure 2. The data are homology classes of words from the core vocabulary, allowing just a small variation in meaning within a class. Thus the English 'all' and Dutch 'alle' meaning *all* are homologous, but in a distinct class from Spanish 'todas' and Italian 'tutte'. They gathered words in $K = 96$ meaning categories and grouped these words in $N = 673$ homology classes. They find evidence that Akkadian is an outgroup. This supports an independent hypothesis that these languages diversified from a 'homeland' in the north west of modern Syria. Our analysis is consistent with this result. However, the uncertainty is substantial.

Bayesian phylogenetic studies of this data type (Gray *et al.* (2003)) use models and software from genetics. Model assumptions, including the tree itself, are rejected by historical linguists (McMahon *et al.* (2005)). Criteria related to parsimony are applied in tree and network visualisation tools (Ringe *et al.* (2002), Bowern (2010)). These tools support the comparative method, allowing the user to intervene in the analysis, and are assumed to be free from modeling assumptions. There are few attempts to quantify uncertainty numerically. They work with heterogeneous data

types, including traits for word phonology and morphology. Most Bayesian analyses (including our own) model just the lexical traits.
Kitchen *et al.* (2009) register the data as a 25×673 binary matrix D, with $D_{i,j} = 1(0)$ if language i possesses (lacks) a word in homology class j, and $D_{i,j} =?$ if this is not known. Obvious loan words have been removed from the data. The published data fill empty meaning categories with a missing value. Ringe *et al.* (2002) register loan words as isolated cognates, while Bowern (2010) leaves identified loan words in the data. This is preferred.
The reconstructed phylogeny is constrained to fit historically known dates (calibration data). The Akkadian vocabulary data come from Assyrian texts from 2700-2900 years Before Present. The biblical Aramaic is 1700-1900BP, Ge'ez is 1600-1800BP, ancient Hebrew 2500-2700BP and Ugaritic 3300-3500BP. The times at which some vocabularies branched from their parent is fixed: the origin of ancient Hebrew is 3200-4200BP, the origin of Ugaritic 3400-4400BP, Aramaic 2850-3850BP and Amharic 700-1700BP. Kitchen *et al.* (2009) cite sources. Modern languages have age zero.
The substitution model which Kitchen *et al.* (2009) fit allows a single word to come into existence with the same meaning independently in several locations, and ancient words to be revived, at relative rates which are not controlled by the data. It is a finite sites model for character substitution developed as a model for character substitution in DNA base character sequences, adapted for generic traits by Lewis (2001). We check their results using a homoplasy-free model for trait evolution and check goodness of fit.

2 Models and Methods

We model the core vocabularies as sets, and the tree as a branching process of sets, with set elements (words) undergoing a birth and death process. The stochastic Dollo model of Nicholls *et al* (2008) has word birth according to a Poisson process of constant rate λ. Words are copied into child languages when a language branches. Each word in each language dies at constant rate μ. Ryder *et al* (2011) add rate heterogeneity via a catastrophe process. Point-like catastrophes are realized on the tree in a Poisson process with rate ρ. When a vocabulary enters a catastrophe, each word in the set dies with probability κ. A Poisson number of words with mean ν are born. If $\nu = \kappa\lambda/\mu$, then one catastrophe equals $-\log(1-\kappa)/\mu$ years in the birth death process. Ryder *et al* (2011) show how to sum over missing data. In this model, the probability that we cannot determine whether language $i = 1, 2, ..., L$ contains a word in homology class $j = 1, 2, ..., N$ is ξ_i. This parameter varies from one language to another.
The parameters are the tree $g = (E, t, k)$ (edge set E, node ages $t = (t_1, ..., t_{2L-1})$, and $k = (k_1, ..., k_{2L-2})$ the number of catastrophes on each edge), the rates λ, μ and ρ, and the probabilities $\xi_i, i = 1, 2, ..., L$ and κ. The prior age t_R of the tree root node (label R say) is approximately

uniformly distributed up to U a fixed maximum (our $U = 16000$BP is very conservative). The distribution over topologies is approximately uniform. This weighting is available in MrBayes (Huelsenbeck *et al.* (2001)) also. Let Γ be the set of all trees g consistent with the calibration data. Fix a tree $g = (E, t, k)$, and let $T_g = \{t'; (E', t') \in \Gamma, E' = E\}$ be the set of admissible node age vectors. For ancestral node i, $s_i^+(g) = \sup\{t_i; t \in T_g\}$ and $s_i^-(g) = \inf\{t_i; t \in T_g\}$ give the greatest and least ages node i can take given g. Let $F(g) = \{i \in 1, 2, ..., 2L-1; s_i^+(g) = U, i \neq R\}$. These are the 'free' nodes in g with ages in g bounded only by U. Let $Z(g)$ be the number of distinct complete orderings $t_{i_1} < t_{i_2} < ... < t_{i_{2L-1}}$ achievable for $t \in T_g$. The probability density on trees $g \in \Gamma$ given by

$$f_G(g) \propto \left[Z(g) \prod_{i \in F(g)} \frac{t_R - s_i^-(g)}{U - s_i^-(g)} \right]^{-1}$$

has marginal distributions on topologies and root age that are approximately marginally uniform. Topology is conditionally approximately uniform given root age and vis versa. These results are exact if all leaves have equal fixed time and there are no calibration constraints. Probability parameters have $U(0, 1)$ priors. The catastrophe rate ρ has a Gamma prior. It varies from 1/1000 (the scale of edge length), and 1/25000 years (the scale of tree length), in the prior 90% interval. The λ- and μ-priors are proportional to $1/\mu\lambda$. The unknown birth and death times of words on the tree, and λ, are integrated analytically, and the remainder using MCMC.
We check for model misspecification. First, we simulate posterior predictive distributions for 'singleton' columns of the data. These are cognates displayed in just a single language. We remove singletons and correct the likelihood, fit the remaining data and then predict singletons and compare predictions with the reserved data. Secondly, we remove historically attested constraints and check that we can recover them, using the Bayes factor to compare models with and without the constraint. Ryder *et al* (2011) give a stable estimator related to the Savage-Dickey ratio. Thirdly, we check that results are insensitive to omitting leaves. The model error arising where language i has loan words from language j is removed if language j is removed. We fit data simulated out of model (including loan words). We found date estimation to be fairly robust, tree topology less so.

3 Results and Conclusions

A Bayesian cross validation analysis of the ten calibration constraints on the full data (KEAM-25) showed problems with the fit. The historically attested constraints on the branching of Biblical Aramaic and the leaf ages for Ugaritic and Ge'ez were rejected. There was strong support for catastrophe events on the branches above Ugaritic. There was very little rate

heterogeneity elsewhere on the tree (except above Ge'ez). These catastrophes are artifacts of model misfit. We treat the Ugaritic and Ge'ez data as outliers and remove them (KEAM-23). This improves the fit. We found little evidence for rate heterogeneity in these data. The posterior probability for zero catastrophes is 0.33 against 0.01 in the prior. As part of our goodness of fit we drop eight more languages from the tree (KEAM-15, with Tigre, Tigrinya, Amharic, Argobba, Geto, Chaha, Zway, Walani, Hebrew, Aramaic, Akkadian, Moroccan Arabic, Ogaden Arabic, Jibbali and Soqotri) and check results are robust.
Cross-validation of the KEAM-23 data gave Bayes factors in favor of the constraint as follows: 'All' 3.9, 'Akkadian' 0.5, 'Amharic branching' 2, 'Aramaic' 0.3, 'Aramaic branching' 6, 'Hebrew' 1.8, 'Hebrew branching' 1.8. See Figure 1. The least Bayes factor is 0.3 so we reject no historically attested

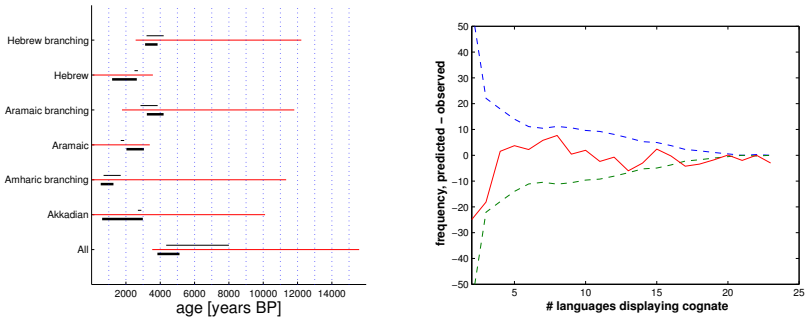


FIGURE 1. (Left) Bayesian cross-validation check on the model for KEAM-23 data. (top thin bars) Calibration constraint. (bottom thick bars) 95% HPD interval for constrained age estimated in an analysis with the single constraint removed. (centre long bars) 95% highest prior density interval estimated in a prior simulation with the single constraint removed. (Right) Posterior predictive distributions (predicted-observed, with 95% envelope) for the number of traits displayed at two, three up to twenty three leaves.

constraint. The bottom bar for 'All' gives the posterior HPD interval for the age of the root. The 95% HPD for the root age in Semitic (KEAM-23) is $[3800, 5100]$BP. Kitchen *et al.* (2009) report $[4400, 7400]$BP. There is an extra bound of $[4350, 8000]$BP. With this we have $[4400, 5100]$.
Posterior predictive 95% HPD intervals for the data for traits at single leaves show that 11 of the 23 reserved singleton counts fall below the 95% HPD predictive interval. The conflation of loan words with unidentified missing data depletes the number of singletons. We remove the singleton data, as it is unreliable. The fit to the frequency distribution of the more commonly occurring cognates, in Figure 1 (Right), is good. There is a small excess of high frequency words: a small number of words evolve at rates lower than the bulk rate. Unidentified loan words inflate the number of

frequently occurring words and must be rare.
The consensus tree for KEAM-23 (Figure 2) is very like the consensus tree

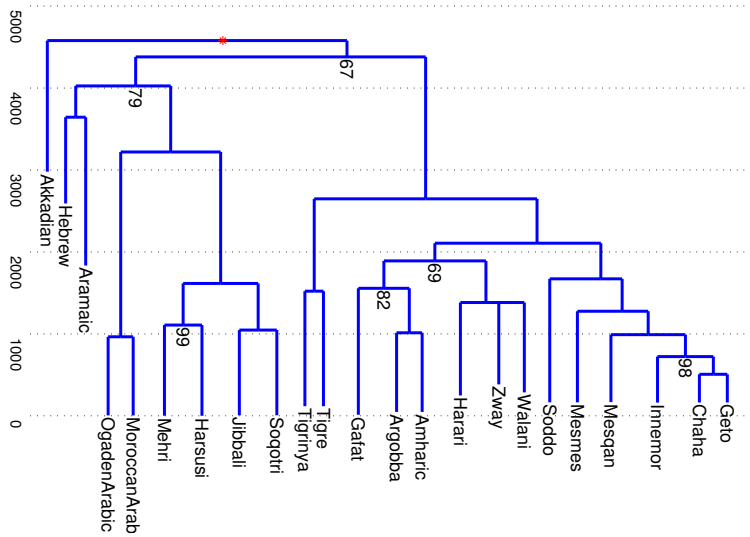


FIGURE 2. Consensus tree for the KEAM-23 data. Edge lengths are proportional to posterior mean time to branching. Edges thresholded at support 50% posterior probability. Numbers on nodes give posterior probability for the edges above. Unnumbered edges have posterior support equal one.

in Kitchen *et al.* (2009). Akkadian is an outgroup with posterior probability 0.67 and prior probability 0.04. Figure 3 shows the posterior probabilities for a few clades of interest. There is evidence for an Akkadian outgroup (Akkadian.Out) in KEAM-22/15. The Arabic languages group with Modern-South-Arabian (MS.Arabian). The evidence for a Modern-South Arabian outgroup (MS.Arabian.Out) is at a similar level to Akkadian in KEAM-25 and KEAM-15, but these are dominated by bias and variance respectively. Hebrew and Aramaic are split by Ugaritic in the unreliable KEAM-25 analysis (Heb.Ara). Posterior distributions for ages and topology are in agreement between KEAM-23 and KEAM-15.
To conclude, the overall tree structure in Figure 2 is very close to that reported in Kitchen *et al.* (2009). It is supported by our goodness-of-fit tests. The main point of difference is in the position of the two Arabic languages and the narrowed posterior distribution of the root time.

References

Bowern, C. (2010). Historical linguistics in Australia: trees, networks and their implications, *Phil. Trans. R. Soc. B* **365**, 3845–3854

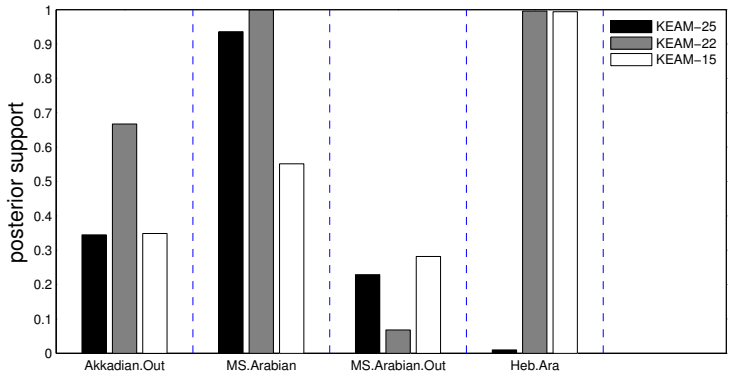


FIGURE 3. Posterior probabilities for selected clades in the three analyses.

Gray, R. and Atkinson, Q. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature* **426** 435–439.

Huelsenbeck, J.P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogeny, *Bioinformatics* **17** 754–755.

Kitchen, A., Ehret, C., Assefa, S., and Mulligan, C.J. (2009). Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East, *Proc. Roy. Soc. B*, **276**, 2703–2710.

Lewis, P.O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data, *Systematic Biol.* **50** 913–925.

McMahon, A. and McMahon, R. (2005). Language Classification by Numbers, Oxford University Press

Nicholls, G.K., and Gray, R.D. (2008). Dated ancestral trees from binary trait data and its application to the diversification of languages, *J. Roy. Statist. Soc. B*, **70**, 545–566.

Ringe, D., Warnow, T. and Taylor, A. (2002). Indo-European and Computational Cladistics, *Trans. Philological Soc.* **100** 59–129.

Ryder, R.J., and Nicholls, G.K. (2011). Missing data in a stochastic Dollo model for binary trait data, and its application to the dating of Proto-Indo-European, *Applied Statistics*, **60**, 71–92.

Partial Order Models for Episcopal Social Status in 12th Century England

Geoff K Nicholls[1,2] , Alexis Muir Watt[2]

[1] nicholls@stats.ox.ac.uk
[2] Statistics Department, University of Oxford, 1 South Parks Road, Oxford OX13TG, UK

Abstract: Our data are lists of bishops signed in the 12th Century, in an order which respects the relative importance of the individual bishops. We model the underlying social order as a partial order, and the list data as a random complete order which respects this underlying partial order. We give static and dynamical models for the partial order. We summarize the posterior distribution using MCMC samples and a particle filter. We fit the models and find evidence for significant order, and for significant change in the order, over time.

Keywords: Social Hierarchy, Partial Orders, Bayesian

1 Data and Questions

Witness lists are ordered lists of the signatories to historical legal documents called acta. We have a large collection of "Royal Acta" from twelfth century England (these were provided by Dr David Johnson of St Peter's College, University of Oxford and Dr Nicholas E Karn, History, School of Humanities, University of Southampton). Witnesses generally signed in order of importance. The different social classes signed in groups in an order which is very obvious. What order relations existed between the bishops who appear in the lists? How did they change? Changes in this hierarchy reflect political events of the time. In the period 1070AD to 1150AD there are $m = 511$ lists with two or more bishops. The time at which a list was signed is known to within an interval (mean interval length 4 years, 90% less than 11 years). Approximately one half of the lists have length just two bishops, and the mean length is 3.5. Each bishop is given a numerical index. For $i = 1, 2, ..., m$, let $y_i = (y_{i,1}, y_{i,2}, ..., y_{i,n_i}), i = 1, 2, ..., m$ give the ordered set of indices of the n_i bishops who witnessed the ith list.

2 Models and Inference

We represent the unknown true order relation in the data as a partial order, that is, as a transitively closed directed acyclic graph $h = h[1 : n]$ with n

nodes, one node for each bishop in the analysis. We can think of h as a binary matrix with $h_{a,b} = 1$ (or 0) if bishop a ranks above b (or not). The model reflects a rigid social hierarchy, which is respected by all, but subject to occasional upheaval. Any particular witness list y_i is modeled as a random total order (a linear extension) respecting the suborder $h[o_i]$ for the bishops who attended that signing. This observation model arises if the signing order is a snapshot of a rapidly evolving total order on the individuals present. However, individuals may jump the queue. Before the j'th person signs in the ith list, there are $n_i - j + 1$ individuals remaining. With probability p the next to sign is chosen at random, ignoring any order constraints, and otherwise, the next person is the first person in a random linear extension of the suborder for the remaining individuals. Let $C(h)$ be the number of total orders consistent with partial order h, and let $C_i(h) = C(h[-i])$ be the number of linear extensions headed by bishop i. The likelihood $L(h, p; y_i) = \Pr(Y_i = y_i | H = h, O = o_i, P = p)$ for partial order h is

$$L(h,p;y_i) = \prod_{j=1}^{n_i-1} \left(\frac{p}{n_i - j + 1} + (1-p)\frac{C_{y_j}(h[y_{j:n_i}])}{C(h[y_{j:n_i}])} \right)$$

so that $L(h, p; y_i) = 1/C(h[y_i])$ at $p = 0$. We can compute the count $C[h]$ quickly for partial orders on up to about $n = 15$ bishops. There were in the period of interest just over twenty bishops at any given time, just a subset of whom are active, so our algorithms are just adequate. There is recent work in Beerenwinkel *et al.* (2007) on maximum likelihood partial orders for conjunctive Bayesian networks, applications of Bayesian inference for 'bucket' orders can be found in Mannila (2008), and on Bayesian inference for generalized Bradley-Terry models in Caron *et al.* (2010). However, we know of no well-developed Bayesian framework for partial orders.

We describe a family of prior distributions for partial orders. These prior models for partial orders are derived from k-dimensional random orders, reviewed in Brightwell (1993). They are marginally consistent for suborders. The prior probability for a suborder is the marginal probability for that order in the prior for any superset of its nodes. Latent variables $Z = (Z_1, Z_2, ...Z_n)$ determine the partial order h on n bishops. The i'th bishop has K real-valued traits $Z_i = (Z_{i,1}, ..., Z_{i,K})$. These traits are not physical, but act as measures of status. Bishop a beats Bishop b ($h_{a,b} = 1 = 1 - h_{b,a}$ if $Z_{a,j} > Z_{b,j}$ for all $j = 1, 2, ..., K$ so that $h = h(Z)$. If the variables overlap then $h_{a,b} = h_{b,a} = 0$. Prior elicitation informed the partial order depth, so we have parameterized the prior to control depth by correlating the latent variables for a given Bishop. Let $Z_i \sim MVN(0, \Sigma)$ with $\Sigma_{i,j} = \rho$ for $i \neq j$ and $\Sigma_{i,i} = 1$. The hyperprior for $R = \rho$ is $R \sim \text{Beta}(1, 1/6)$. The joint latent variable prior, $f(z, \rho)$ say, gives a prior on partial orders (through $h = h(Z)$) which is roughly uniform on depth.

We have extended $h(Z), R$ to a process $h(Z(\tau)), R(\tau)$ in time. At the event

times $\phi = (\phi_1, \phi_2, ...)$ of a Poisson process (the catastrophe process) of rate λ_C, there is a change point where $(Z(\phi_j), R(\phi_j)) \sim f$ independent of all history. At the event times $\psi = (\psi_1, \psi_2, ...)$ of a Poisson process (the singleton process) of rate λ_S the latent variables of a single Bishop $i \sim U\{1, 2, ..., n\}$ are (independently) renewed $Z_i(\psi_j) \sim MVN(0, \Sigma)$ at fixed R. This process has equilibrium $f(z, \rho)$.

We fit the static model to m witness lists from short intervals of time (it does not allow for evolution in the order). This analysis treats the uncertain list-dates as fixed (some are dated, and the uncertainty is often small). We use MCMC to simulate the posterior distribution $\pi(z, p, \rho|y) \propto L(h(z), p; y) f(z, r)$. The prior for p is uniform in $[0, 1]$. We use a hybrid MCMC/particle filtering approach as in Andrieu *et al.* (2010) to simulate the posterior distribution for the dynamical model. Let $t = (t_1, t_2, ..., t_m)$ parameterize the unknown true dates associated with the m lists, and let $t_{[i]}$ be the i'th date in an ordered list of the dates. We carry out MCMC for t, λ_S, λ_C. We estimate $p(y|t, \lambda_S, \lambda_C)$ using paths from a particle filter. The filter integrates the $Z(\tau), R(\tau)$ process using a discrete time HMM with hidden states $(Z(t_{[i]}), R(t_{[i]}), p)$ and emitted states $y_{[i]}$ (index ordered on t), so the HMM states are maintained at the m list times only.

3 Results and Conclusions

We present results for both the static and dynamical models. We illustrate the static model using lists taken from two windows of time. The lists are given below. What partial orders on status constrain the lists at each time? Do the orders differ from one window to the next?

```
Witness lists 1119-1121          Witness lists 1127-1129
[1119]   5   6   4   7           [1127]   9  10
[1120]   3   4                   [1127]   2   9  10
[1121]   1   2                   [1127]   2   1   6   5   8  10
[1121]  10   1   2   5   6   8   [1127]   2   6
[1121]   1  10   2   5   6   9   [1127]   2   9   6  10
[1121]   1   2                   [1127]   2   9   6  10
[1121]  10   1   2               [1129]   7  10
                                 [1129]   6   7   4  10
                                 [1129]   3   4   2
```

We make a separate static Bayesian analysis for each window. Figure 1 displays a graphical summary of the posterior distribution on partial orders in each time window. The posterior probability for each edge is estimated via MCMC, and thresholded at one half. We illustrate the dynamical model on the 1119-21 data, conditioning on $\lambda_S = 1$ and $\lambda_C = 0.1$, just to show consistency. The consensus order for the year 1120 is show in Figure 1. It agrees well with the adjacent result for the corresponding static analysis.

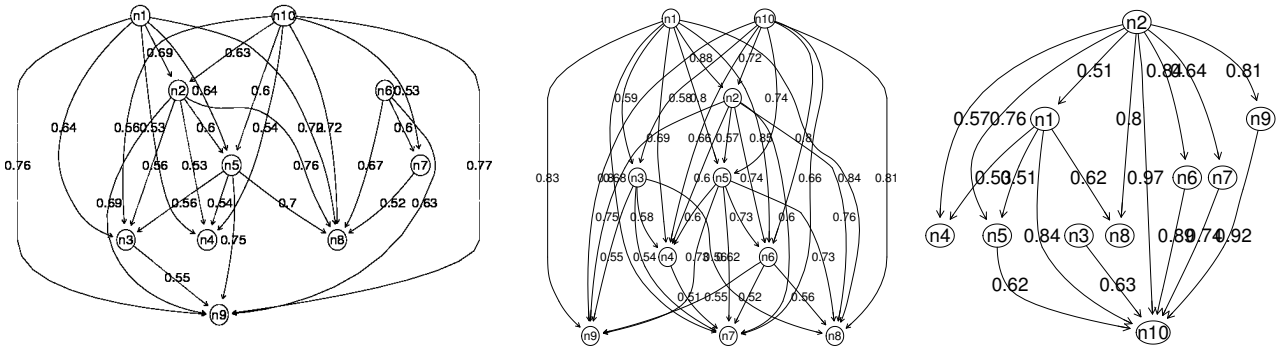


FIGURE 1. Consensus partial orders show marginal posterior support for each directed edge. Edge labels are marginal posterior probabilities. (Left) Dynamical model/MCMC-Particle filter, 1120. (Mid) Static model/MCMC 1119-21 (Right) 1127-29

We see from Figure 1 that there is evidence for significant order (edges supported with high marginal posterior probability). There is evidence for change. The postholder of the position of Bishop of London changes from the left to right graph, and node 10 (Bishop of London) correspondingly moves from the top to the bottom of the figure with some strongly supported edges changing direction. The probability for a catastrophe in the short interval 1119-21 was low (10%). The linear extensions in 1119-21 are well explained by the higher rate singleton change process.

Acknowledgments: Thanks to Prof Bernard Silverman, Dr David Johnson and Dr Nicholas E Karn for their help.

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods. J. Roy. Statist. Soc. B, **72**, 269-342.

Beerenwinkel, N., Eriksson, N., and Sturmfels, B. (2007) Conjunctive Bayesian networks. *Bernoulli*, **13**, 893-909.

Brightwell, G. (1993) Models of random partial orders. In: *Surveys in Combinatorics* 5383, ed. K. Walker, Cambridge Univ. Press

Caron, F., and Doucet, A. (2010). Efficient Bayesian Inference for Generalized Bradley-Terry Model. *J. Comp. Graphical Stat* to appear.

Mannila, H. (2008) Finding Total and Partial Orders from Data for Seriation. *Lecture Notes in Comp. Sci.*, **5255**, 16-25.

Testing Goodness-of-Fit of Parametric Models for Censored Data

R. Nysen[1], M. Aerts[1], C. Faes[1]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Center for Statistics, Universiteit Hasselt. Agoralaan 1, B-3590 Diepenbeek, Belgium, ruth.nysen@uhasselt.be, marc.aerts@uhasselt.be, christel.faes@uhasselt.be

Abstract: A goodness-of-fit test for left-, right- and interval-censored data, assuming random censorship is proposed and studied. In the first step of the test, the null model is extended to a series of nested alternative models for censored data as in Zhang and Davidian (2008). Then a modified AIC model selection is used to select the best model to describe the data. If a model with one or more extra parameters is selected, then the null hypothesis is rejected. This new goodness-of-fit test procedure is based on the order selection test as described in Aerts, Claeskens and Hart (1999). The applicability of the test is illustrated in the context of microbial agents, and its performance characteristics are demonstrated through simulation studies.

Keywords: Goodness-of-fit test; Censored data; SNP estimator; Order selection test

1 Introduction

Censored data are often encountered in medical and public health studies. In survival studies, time to death can be right censored due to end-of-study or loss to follow-up. In infectious diseases, seroconversion time might only be known to fall in some interval, leading to interval-censored data. Within the framework of chemical risk assessment, the handling of concentration data reported to be below the limit of detection (left-censored) or between the limit of detection and the limit of quantification (interval-censored) present challenges to the statistical analysis of chemical occurrence data. When using parametric models, the choice of the distribution for such censored data is an important step in the analysis.

Goodness-of-fit tests for censored data have not been studied extensively. Hollander and Proschan (1979) present a test for a simple null hypothesis for right-censored data. This test can be applied for left-censored data by reversing the order of the observations. A test for interval censored data, based on the Cramér-von Mises statistic and a leveraged bootstrap, was introduced by Ren (2003). Bayesian tests were proposed by Yin (2009), Cao et al. (2010) and Calle and Gómez (2008, Chap. 21).

In this paper we propose and study a new goodness-of-fit test for left-, right- and interval-censored data, assuming random censorship. The test is based on the order selection test as described by Aerts, Claeskens and Hart (1999), which requires a series of nested alternative models in which the null model is nested. For censored data, such a family of densities can be described by the SNP (SemiNonParametric) representation of Zhang and Davidian (2008). The combination of the order selection test and the SNP representation results in a goodness-of-fit test for censored data.

2 Methodology

The test is based on the order selection test as described by Aerts, Claeskens and Hart (1999). They use a modified AIC criterion (MAIC) and accept the null hypothesis if and only if the prescribed distribution is chosen by the criterion

$$\text{MAIC}(r; C_n) = 2(\mathcal{L}_r - \mathcal{L}_0) - C_n r, \quad r = 0, 1, \ldots,$$

where C_n is some constant larger than 1. By appropriate choice of C_n, the asymptotic type I error probability of the test can be any number between 0 and 1. To determine C_n, a statistic T_n is defined as

$$T_n = \max_{1 \le r \le R_n} \{2(\mathcal{L}_r - \mathcal{L}_0)/r\},$$

for which the asymptotic distribution is known. The rejection of the null hypothesis is equivalent to $T_n > C_n$. For example, a test of asymptotic level .05 is obtained by $C_n = 4.18$. The P-value corresponding to an observed T_n can also be approximated by a bootstrap.
The procedure of Aerts, Claeskens and Hart (1999) requires that the null model is nested within the family of alternative models, which in turn form a sequence of nested models having more and more parameters. For censored data, such a broad class of densities can be described by the SNP (SemiNonParametric) representation of Zhang and Davidian (2008). In this representation, the density function under the null hypothesis is extended by multiplying with a polynomial of fixed degree r, introducing r new parameters in the model:

$$\gamma_r(z) = P_r^2(z)\psi(z),$$

where $P_r(z) = a_0 + a_1 z + \cdots + a_r z^r$ and $\int \gamma_r(z)dz = 1$. In case the null hypothesis states that data come from a lognormal distribution, the logarithm of the data can be written as $\log(T_0) = \theta_1 + \theta_2 Z$, where Z follows the standard normal distribution. The density $\psi(x)$ of the standard normal distribution is then multiplied with the square of a polynomial of degree r, such that r parameters are added to the model. The proposed goodness-of-fit test will reject the null hypothesis if the MAIC criterion selects a model

with $r > 0$. However, it is not guaranteed that the limiting distribution of T_n still holds when using censored data. Therefore the bootstrap offers an alternative.
The test can be used for different types of censoring. However, in the data analysis and the simulation study we focus on left- and interval-censored data.

3 Data analysis

The applicability of the test is illustrated through the analysis of some real data. The data under consideration consist of measurements of the cadmium level in some food category. 99 observations are available of which 42 are censored by the Limit of Detection (LOD). These limits of detection are in the range $[0.001, 0.01]$. The truly observed values are in between 0.0015 and 4.14. Some of the truly observed values are smaller than some of the LOD, because the data come from different laboratories, where different LOD's are applied.

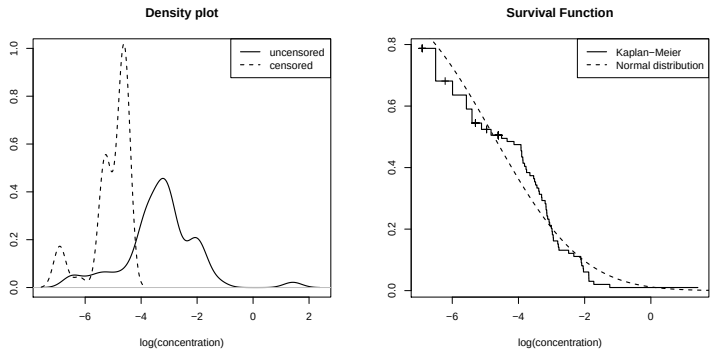


FIGURE 1. Cadmium data: Kernel density estimate and estimated survival functions of log(concentration).

A visual representation of the data is given in Figure 1. The left panel shows a kernel density of the logarithm of the concentrations. The right panel shows the Kaplan-Meier estimate, a frequently used estimator of the survival function in censored data. The LOD's are concentrated to the left and these values are denoted by a plus-sign (+) in the KM estimate. The fit for the normal distribution is represented by the dashed line.
In this situation we are interested in testing whether the concentrations are lognormally distribution. The proposed test was applied and the maximum MAIC was reached for $r = 3$, meaning that the null hypothesis of the lognormal distribution is rejected at significance level 5%.

4 Simulation Study and Discussion

The simulation study shows good performance of the test. Data are drawn from different distributions, with different sample sizes and different percentages of censoring. Focus is on left- and interval-censored data, where the null hypothesis states that the data come from the lognormal distribution.
Under the null hypothesis, the achieved percentage of rejected null hypotheses approximates the significance level. For example, a data set of size 100 is simulated with 12% left censoring, induced by five limits of detection. At 5% (respectively 10%) significance level, in 3% (respectively 8%) of the simulations, the hypothesis is rejected. The power of the test is high, especially for large sample sizes. For example, data are drawn from a mixture of two lognormal distributions. At 5% (10%) significance level, 55% (85%) of the tests are rejected.
A bootstrap is used to further investigate the distribution of the test statistic when data are censored. The bootstrapped sample is simulated under the null hypothesis and censoring is imposed by two different principles. Both principles try to resemble the censoring as good as possible. The p-values from both methods are close to the theoretical p-value.

References

Aerts, M., Claeskens, G. and Hart, J. (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association*, **94**, 869-879.

Calle, M. L. and Gómez, G. (2008) *Statistical models and methods for biomedical and technical systems* Birkhäuser Boston

Cao, J., Moosman, A. and Johnson, V. E. (2010). A Bayesian Chi-Squared Goodness-of-Fit Test for Censored Data Models. *Biometrics*, **66**, 426-434.

Hollander, M. and Proschan, F. (1979). Testing to determine the underlying distribution using randomly censored data. *Biometrics*, **35**(2), 393-401.

Ren, J. (2003). Goodness of fit tests with interval censored data, *Scandinavian Journal of Statistics. Theory and Applications*, **30**(1), 211-226.

Yin, G. (2009). Bayesian goodness-of-fit test for censored data. *Journal of Statistical Planning and Inference*, **139**(4), 1474-1483.

Zhang, M., Davidian, M. (2008). Smooth semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics*, **64**, 567-669.

Testing against ordered alternatives with interval-censored data

Ramon Oller[1], Guadalupe Gómez[2]

[1] Departament d'Economia i Empresa, Universitat de Vic, Sagrada Família 7, 08500 Vic, Spain
[2] Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain

Abstract: In many K sample problems is common to test against ordered alternatives. Although many statistical methods have been proposed for uncensored and right-censored data, there is a small number of methods for interval-censored data. Abel (1986) gives one of the few generalizations of the well-known Jonckheere-Terpstra test to interval censored data. In this paper we propose some extensions of the Jonckheere-Terpstra test. We use permutational and Monte Carlo approaches for making inferences. This work is motivated by the analysis of a dataset from a study of the benefits of zidovudine in patients in the early stages of the HIV infection (Volberding et al., 1995).

Keywords: Jonckheere-Terpstra test; order restricted inference; nonparametric test; interval-censored data.

1 Introduction

An important issue that arises in survival studies is to establish an ordering alternative in the k-sample problem. For instance, the effect of increasing dose levels of a drug either can involve increasing survival times (simple increasing ordering) or increasing up to a certain optimal point and then decreasing (umbrella ordering). In studies of new treatments with no negative effects, the survival times of the control group can be expected to be lower than those of the treatments (simple tree ordering).
Jonckheere (1954) and Terpstra (1952) were among the first to develop a nonparametric statistic to test for ordered alternatives. The Jonckheere-Terpstra (JT) test for trend has received much attention and discussion in the literature (Terpstra and Magel, 2003; Alonzo *et al.*, 2010; Davidov and Herman, 2010). Abel (1986) generalizes the JT test for interval censored data. In this paper we propose some extensions of the JT test.

2 Notation

Interval-censoring mechanisms arise when the event $\mathcal{E}$ cannot be directly observed and it is only known to have occurred during a random interval of time. We denote by T the lifetime variable. Based on a sample of N individuals with potential times (responses) $T_1, \ldots, T_N$, we observe $(l_1, r_1], \ldots, (l_N, r_N]$ censoring intervals.
We assume that we have K groups, $G_1, \ldots, G_K$, with respective sample sizes $N_1, \ldots, N_k$. We define $S_1, \ldots, S_K$ and $F_1, \ldots, F_K$, respectively, as the survival and the distribution functions of T under each group. We denote by $\hat{F}$ the Turnbull's NPMLE of F from the pooled sample. Then, $\hat{F}_i(t) = P_{\hat{F}}([0,t)|(l_i, r_i])$ is an estimate of the distribution function of the i-*th* individual (Fay and Shih, 1998) which holds that $\hat{F}(t) = \frac{1}{N}\sum_{i=1}^{N} \hat{F}_i(t)$. In this paper we focus on the simple increasing ordering. Our goal is to determine whether $H_0 : S_1 = \cdots = S_k = S$ or $H_1 : S_1 \leq \cdots \leq S_k$.

3 A class of Kendall-type tests

As extension of the JT test for interval censored data we consider a class of test statistics which are based on a weighted sum of two-sample statistics:

$$WJT = \sum_{\substack{r,\, s\, =\, 1 \\ r\, <\, s}}^{K} \sum_{i,j=1}^{N} w_{r,s}\, \alpha_i^s\, \alpha_j^r\, \Phi(\hat{F}_i, \hat{F}_j)$$

where α_i^r is an indicator function that is equal to 1 if the i-*th* individual belongs to group G_r and 0 otherwise, Φ is one of the functionals defined in Fay and Shih (1998) and $w_{r,s}$ is a weighting function.
The WJT test is a Kendall's correlation coefficient $\frac{1}{2}\sum_{i,j=1}^{N} a_{ij}\, b_{ij}$ with $a_{ij} = \Phi(\hat{F}_i, \hat{F}_j)$ and $b_{ij} = \sum_{\substack{r,\, s\, =\, 1 \\ r\, <\, s}}^{K} w_{r,s}\, \alpha_i^s\, \alpha_j^r - \sum_{\substack{r,\, s\, =\, 1 \\ r\, >\, s}}^{L} w_{r,s}\, \alpha_i^s\, \alpha_j^r$. Under the null hypothesis, the permutational distribution of WJT is asymptotically normal with zero mean and variance given by $V = \frac{1}{2N(N-1)} \left[\sum_{i,j=1}^{N} a_{ij}^2\right]\left[\sum_{i,j=1}^{n} b_{ij}^2\right]$
$+ \frac{1}{N(N-1)(N-2)} \left[\sum_{i,j_1,j_2=1}^{N} a_{ij_1} a_{ij_2} - \sum_{i,j=1}^{N} a_{ij}^2\right]\left[\sum_{i,j_1,j_2=1}^{N} b_{ij_1} b_{ij_2} - \sum_{i,j=1}^{N} b_{ij}^2\right]$.
When $w_{r,s} = 1$ and $\Phi(\hat{F}_i, \hat{F}_j) = \int \hat{F}_j(s) d\hat{F}_i(s) - \int \hat{F}_i(s) d\hat{F}_j(s)$, the WJT test is a natural extension for interval-censored data of the JT test. When $w_{r,s} = s - r$, $w_{r,s} = s$ or $w_{r,s} = K - r$ the statistic WJT is specially adequate for linear, convex or concave trends respectively. In the case $w_{r,s} = s - r$, WJT is equivalent to the linear tests studied in Gómez and Oller (2008).

4 An extension of the Terpstra and Magel test

Terpstra and Magel (2003) propose a test based on comparing measurements from all the groups at the same time, rather than performing pairwise comparisons as the JT test does. As extension of this test we propose the following test statistic:

$$TM = \sum_{i_1,\ldots,i_K=1}^{n} \alpha_{i_1}^1 \cdots \alpha_{i_K}^K \int \mathbb{1}_{\{t_1 \leq \cdots \leq t_K\}} \, d\hat{F}_{i_1}(t_1) \ldots d\hat{F}_{i_K}(t_K)$$

The permutational distribution of the TM test statistic under the null hypothesis is obtained by a Monte Carlo approach. The main virtue of the TM test is that we expect poor power whenever the configuration in the alternative does not hold. This is an important property which is not usually satisfied by most trend tests.

5 Data analysis

We analyze the data in an AIDS Clinical Trial designed to study the benefits of zidovudine therapy in patients in the early stages of the HIV infection, see Volberding *et al.* (1995). The lifetime variable is the number of months from randomization until the CD4 count first reaches 400 cells per cubic millimeter. The design compares three groups. The first group, G_1, corresponds to those patients who started zidovudine monotherapy after their CD4 cell count fell below 500 per cubic millimeter. In the second and third groups, G_2 and G_3, two different dosages of zidovudine were given immediately after randomization. Among the 1607 subjects who could be evaluated, 541 were in the deferred-therapy group, 538 in the 500–mg group and 528 in the 1500–mg group. Figure 2 shows the probabilities of keeping CD4 values larger than a certain number of months. The three groups show an increasing ordering between survival functions. The resulting p-values for the TM test and different configurations of the WJT test are highly significant (p-value $< 10^{-4}$).

6 Concluding remarks

In this paper we propose the WJT and TM tests as extensions of the Jonckheere-Terpstra test for interval-censored data. We expect the WJT test to have higher power than the TM test but it needs prior evidence to know which Φ and $w_{r,s}$ to choose. On the other hand, we expect poor power for the TM test when the configuration in the alternative does not hold. This good property is not necessarily true for the WJT test.

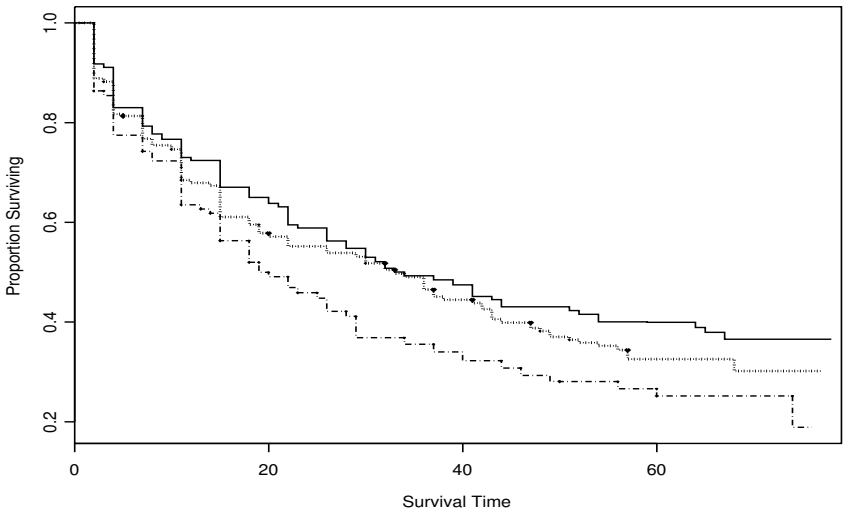


FIGURE 1. Probabilities of keeping CD4 values larger than 400 for group G1 (dashed curve), G2 (thick dotted curve) and G3 (solid curve).

Acknowledgments: The authors are grateful to the GRASS group for their fruitful discussions. This research was partially supported by Grant MTM2008-06747-C02-00 from the Ministerio de Educación y Ciencia.

References

Abel, U. (1986). A nonparametric test against ordered alternatives for data defined by intervals. *Statistica Neerlandica*, **40**, 87-91.

Alonzo, T.A., Nakas, C.T., Yiannoutsos, C.T., and Bucher, S. (2010). A comparison of tests for restricted orderings in the three-class case. *Statistics in Medicine*, **28**, 1144-1158.

Davidov, O., and Herman, A. (2010). Testing for order among K populations: theory and examples. *The Canadian Journal of Statistics*, **38**, 97-115.

Fay, M.P., and Shih, J.H. (1998). Permutation tests using estimated distribution functions. *Journal of the American Statistical Association*, **93**, 387-396.

Gómez, G., and Oller, R. (2008). *A new class of rank tests for interval-censored data*. Harvard University Biostatistics Working Paper Series, Working Paper 93.

Jonckheere, A.R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, **41**, 133-145.

Terpstra, T. (1952). The asymptotic normality and consistency of Kendall's test against trend when ties are present in one ranking. *Indagationes Mathematica*, **14**, 327-333.

Terpstra, J.T., and Magel, R.C. (2003). A new nonparametric test for the ordered alternative problem. *Nonparametric Statistics*, **15**, 289-301.

Volberding, P.A., Lagakos, S.W., Grimes, J.M., et *al.* (1995). A comparison of immediate with deferred zidovudine therapy for asymptomatic HIV-infected adults with CD4 cell counts of 500 or more per cubic millimiter. *New England Journal of Medicine*, **333**, 401-451.

Examining distance-based grouping on the simplex sample space: the fuzzy clustering case

J. Palarea-Albaladejo[1], J. A. Martín-Fernández[2]

[1] Biomathematics & Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ, UK. E-mail: javier.palarea@bioss.ac.uk
[2] Dept. Informàtica i Matemàtica Aplicada, UdG, Campus Montilivi, Edifici P-IV. E-17071, Girona, Spain. E-mail: josepantoni.martin@udg.edu

Abstract: Most fuzzy clustering techniques aim at minimizing an objective function which measures the overall dissimilarity within clusters. Then, a measure of dissimilarity coherent with the geometrical particularities of the sample space is required. Here we focus on the simplex, whose elements are characterised by non-negativity and constant-sum constraints. The fuzzy c-means (FCM) is not well-behaved on the simplex. Drawbacks are pointed out and simulation results are used to comparing with an FCM approach based on log-ratios.

Keywords: Fuzzy clustering; FCM; simplex; log-ratio analysis.

1 Introduction

Fuzzy clustering techniques allows gradual memberships of objects to clusters. Applying any procedure of clustering, the underlying geometrical structure of the sample space must be considered. Otherwise, misleading or inconsistent conclusions may be drawn. Here we focus on the simplex sample space $\mathcal{S}^D = \{\mathbf{x} = [x_1, \ldots, x_D] : x_1 > 0, \ldots, x_D > 0; \sum_{i=1}^{D} x_i = t\}$, whose elements $\mathbf{x}$ represent parts of a whole. Typically, chemical compositions; household or time budgets, election vote shares, and so on. Aitchison (1986) introduced the log-ratio methodology for their statistical analysis. Nowadays, the simplex has been well characterised as an Euclidean space (*e.g.* Pawlowsky-Glahn and Egozcue, 2001). Then elements of the simplex can be expressed on real coordinates with respect to an orthonormal basis. This fact was exploited by Egozcue et al. (2003) by defining the isometric log-ratio (ilr) transformations

$$y_i = \frac{1}{\sqrt{i(i+1)}} \log \frac{\prod_{j=1}^{i} x_j}{x_{i+1}^{i}}, \quad i = 1, \ldots, D-1. \tag{1}$$

As a result, a distance $d(\mathbf{x}, \mathbf{x}^*)$ in $\mathcal{S}^D$ can be analogously worked out in the real space R^{D-1} as $d_e(\mathbf{y}, \mathbf{y}^*)$, where d_e refers to the Euclidean distance.

2 The probabilistic fuzzy c-means algorithm on ilr coordinates

The probabilistic FCM algorithm (Bezdek, 1980) distributes the total membership of the objects among all the clusters. It recognises a number of c hyper-spherical clouds of points in a data set, each one of them represented by its centre $\nu_k \in R^d$, $k = 1, \ldots, c$. The problem to be solved is

$$\min_{\mathbf{U},\nu} \left\{ J_m = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ki}^m d_e^2(\mathbf{y}_i, \nu_i) \right\}, \tag{2}$$

where the elements of $\mathbf{U}$ satisfy $\sum_{i=1}^{n} u_{ki} > 0,\ k = 1, \ldots, c$ and $\sum_{k=1}^{c} u_{ki} = 1,\ i = 1, \ldots, n$. The parameter $m \geq 1$ is the degree of fuzzification (usually $m = 2$). Both the cluster centres, ν_k, and the membership probabilities, u_{ki}, are obtained by an iterative process. At the t-step these are calculated as follows:

$$\nu_k^{(t)} = \frac{\sum_{i=1}^{n} u_{ki}^{m,(t-1)} \mathbf{y}_i}{\sum_{i=1}^{n} u_{ki}^{m,(t-1)}},\ \ k = 1, \ldots, c, \text{ and } u_{ki}^{(t)} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_e(\mathbf{y}_i, \nu_k^{(t)})}{d_e(\mathbf{y}_i, \nu_k^{(t)})} \right)^{2/(m-1)}}. \tag{3}$$

Once the algorithm converges, an object i is usually classified into the cluster k with highest membership probability u_{ki}.

3 Simulation results

The classic FCM directly applied on the simplex is compared with the FCM algorithm applied on the ilr space (Eq. 1). For ease reference, we will refer the former as FCM and the latter as FCM-C. Two opposite scenarios have been considered. In the Scenario A clusters are very close each other and denser than in Scenario B, where clusters are less homogeneous and located near the corners (except for one around the barycentre). Accordingly, two data sets in $\mathcal{S}^3$ were simulated with a 4-cluster structure embedded into them (see Figs. 1A-B).

After applying FCM and FCM-C, the total percentage of wrong allocations is computed. In Scenario A (the most complex) it is 19.87% for FCM and 13.87% for FCM-C. In Scenario B these percentages are 9.87% and 4.20%, respectively. Table 1 summarises the erroneous allocations by cluster. FCM-C provides in general less errors. Globally, both algorithms make fewer errors when allocating objects belonging to cluster 4. The FCM works better here because it is the more spherical one. Even so, the FCM-C provides smaller errors, specially in scenario B.

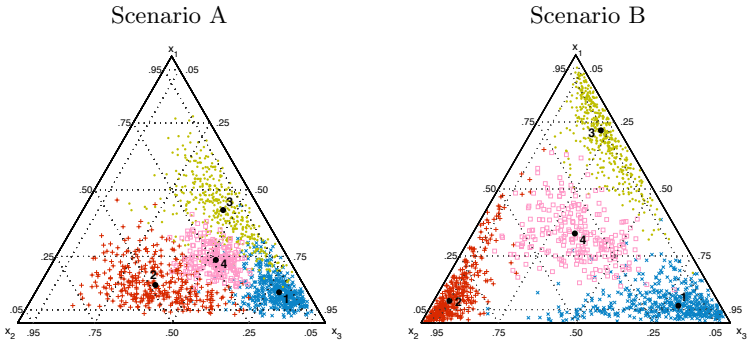


FIGURE 1. Simulated data sets and clusters for two scenarios in $\mathcal{S}^3$: (A) four close and dense clusters, (B) four scattered clusters. Solid points (•) represent the respective centres.

TABLE 1. Wrong allocations for Scenario A and Scenario B by cluster (values are percentages).

	FCM		FCM-C	
Cluster	A	B	A	B
1	9.79	14.37	13.75	5.83
2	18.95	4.12	10.43	2.19
3	39.90	11.26	21.83	5.39
4	5.21	6.95	4.78	1.73

3.1 Inter-cluster distribution of errors

Tables 2 and 3 show the percentage of points belonging to a cluster that have been allocated to any of the others. The main confusion is that points belonging to clusters 1, 2 or 3 are grouped into cluster 4. FCM yields wrong assignments among all the clusters.

4 Concluding remarks

In order to clustering a set of objets in the simplex, a measure of dissimilarity coherent with its geometrical particularities is required. Numerical results illustrate that the FCM-C algorithm produces a lower number of wrong allocations than the common FCM approach. Only when points are located near the barycentre of the simplex, both may provide similar results. For points near the corners, the FCM does not take into account that a small variation involves a great relative change.

TABLE 2. Scenario A: distribution of errors by clustering method (values are percentages). True belonging cluster is in rows. Allocated cluster is in columns.

	FCM				FCM-C			
True cluster	1	2	3	4	1	2	3	4
1	-	2.13	2	95.74	-	15.15	31.82	53
2	7.25	-	1.45	91.30	0	-	0	100
3	18.82	2	-	79.41	31.18	0	-	68.82
4	8.33	41.67	50	-	0	64	36	-

TABLE 3. Scenario B: distribution of errors by clustering method (values are percentages). True belonging cluster is in rows. Allocated cluster is in columns.

	FCM				FCM-C			
True cluster	1	2	3	4	1	2	3	4
1	-	5.80	0	94.20	-	7.14	17.86	75
2	0	-	13.33	86.67	0	-	0	100
3	6.25	0	-	93.75	4.35	0	-	95.65
4	18.75	37.50	44	-	0	25	25	-

Acknowledgments: This research has been supported by the Spanish Ministry of Science and Innovation under the project "CODA-RSS" Ref. MTM2009-13272; by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref: 2009SGR424.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Bezdek, J. (1980). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35**, 279-300.

Pawlowsky-Glahn, V., and Egozcue, J.J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, **15**, 384-398.

The use of GEE for analyzing housing prices

M. C. Pardo[1], T. Pérez[2]

Keywords: Housing price; Longitudinal data; GEE methodology.

1 Introduction

Several studies have been made to explain housing prices evolution in Spain. Maza and Peñalosa (2010) show how from 2008 began a decline in levels that still remains. However, they have observed that in recent quarters the fall of housing prices have moderated, with smaller quarterly decreases. They have found substantial differences according to type of housing, location or within cities. In this paper we have performed the same analysis but using statistical inferences methodology, so that conclusions can be extending to the whole population from where the sample has been selected. On building econometric approaches, there are two aspects which play a crucial role in determining the accuracy of modeling results, the first is the selection of the variables, and the second the statistical method used to construct the model. The ordinary least squares (OLS) is the commonly used technique, but it assumes that observations are independent random variables, identically distributed, with normal distribution and common variance. Therefore if you have time-dependent covariates, missing data, or non-normality, then this approach may not be adequate.
In our case the data are not independent, we have repeated measurements in the same municipalities at different times. Furthermore, the normality assumption does not hold. The aim of this work is to explore the use of the generalized estimating equations (GEE) as a potential alternative to the classical methods usually considered.

2 A review of the GEE method

In recent years, researchers have begun to use a new method for the repeated measurements analysis, the generalized estimating equation approach (Liang and Zeger, 1986). It provides a semiparametric approach to longitudinal data analysis with univariate outcomes and allows the response probability distribution to be any member of an exponential family of distributions.

The GEE model is an extension of the generalized linear model. The extension is presented in the subscript j time, which indicates that the same individuals can be measures repeatedly over time.
Let t be the maximum number of time points at which data are collected. Let $y_i = (y_{i1}, ..., y_{it_i})^t$ be the $t_i \times 1$ vector of responses and $x_i = \left(x_{i1}^t, ..., x_{it_i}^t\right)^t$ be the $t_i \times p$ matrix of covariate values for the ith subject $(i = 1, ..., n)$. Here t_i is the number of time points for the ith subject and may be less than t because of missing observations. The marginal density of y_{ij} is

$$f(y_{ij}) = \exp\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right].$$

The mean and variance of y_{ij} are given by $E(y_{ij}) = \mu_{ij} = b'(\theta_{ij})$, $var(y_{ij}) = V(\mu_{ij})a(\phi)$ where $V(\mu_{ij}) = b''(\theta_{ij})$ is the variance function and ϕ is a possibly unknown dispersion parameter. The regression model for the mean is $\eta_i = x_i\beta$, where $\eta_i = (\eta_{i1}, ..., \eta_{it_i})^t$ with $\eta_{ij} = g(\mu_{ij})$ and $\eta_{ij} = x_{ij}\beta$. Here $\beta = (\beta_1, ..., \beta_p)^t$ is the $p \times 1$ vector of unknown parameters to be estimated and $g(.)$ is a link function.
Let $R_i(\alpha)$ be the $t_i \times t_i$ "working" correlation matrix of y_i and let α be an $s \times 1$ vector which fully characterizes $R_i(\alpha)$. Let A be a $t \times t$ diagonal matrix with $V(\mu_{ij})$ as the jth diagonal element. Then, for the ith subject the $t_i \times t_i$ working matrix of y_i is given by

$$V_i = \phi (A_i)^{1/2} R_i(\alpha) (A_i)^{1/2},$$

where A_i is a $t_i \times t_i$ submatrix of A. If $R_i(\alpha)$ is indeed the true correlation matrix for the $y_i^{'}s$, V_i is equal to $cov(y_i)$. The generalized estimating equations for estimating the vector of parameters β is given by

$$\sum_{i=1}^{n} D_i^t V_i^{-1} S_i = 0$$

where $S_i = y_i - \mu_i$ with $\mu_i = (\mu_{i1}, ..., \mu_{it_i})^t$ and $D_i = \partial\mu_i/\partial\beta$.
Estimation requires iterating between this method for estimating β and a robust method for estimating α and ϕ as it is explained in Liang and Zeger (1986). There are several possibilities for the working correlation structure, see Hardin et al (2003). The GEE generally produces consistent estimators of the true variance of the estimated parameters, even when the working correlation has been misspecified.

3 Application

3.1 Dataset

The dataset used have been downloaded from official website of the Spanish Housing Ministry for the years 2005 to 2010, a sample size of 150 municipalities has been selected.

As dependent variable we consider the mean price per square meter of housing and as exploratory variables we have *region, province, type of housing* with two levels: housings with two years old or less and housings with more than two years old, *size of the municipality,* categorized into 4 levels: level 1 (less than 50.000 inhabitants); level 2 (between 50.000 and 150.000 inhabitants); level 3 (between 150.000 and 250.000 inhabitants) and level 4 (more than 250.000 inhabitants). A longitudinal study of these data was carried out using GEE. We have checked that the dependent variable follows a gamma distribution.

3.2 Conclusions

The results show a significant influence of the variables introduced into the model, *region*, *province*, *type of housing*, *size*, *quarter* and *year*, plus an interaction effect between *region* and *year* and between *quarter* and *year*. Using its effects estimated, we can conclude that the mean prices are bigger in municipalities with more number of inhabitants and new housing, as expected.

As we can observe in Figure 1, from 2005 to 2007, the average price of housing grew continuously, this situation changed in the second quarter of 2007, starting a period of flat growth until third quarter 2008 when it began a continuous decreasing up to the last period considered, the first quarter of 2010. Our results are consistent with those obtained by Maza and Peñalosa (2010).

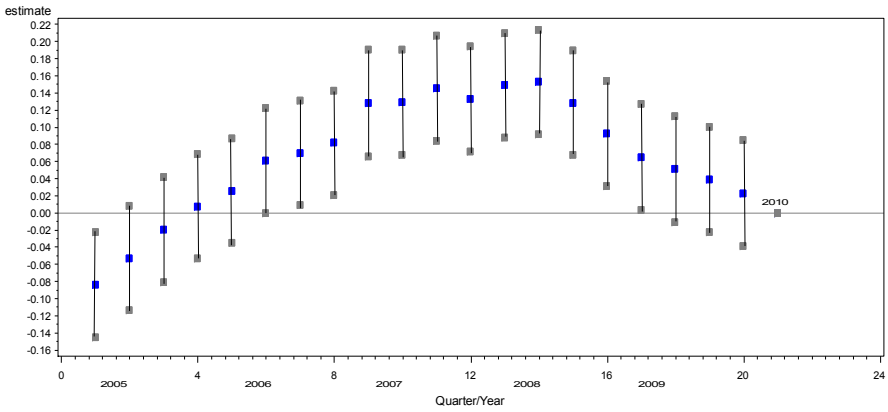


FIGURE 1. Coefficients and 95% CI estimated for time.

In order to make regression a useful and meaningful statistical tool, emphasis should be put not only on inference or fitting but also on diagnosing

potential data problem. For checking the systematic departure, graphical methods such as scatter plots of residuals against fitted values and/or prognostic factors are helpful. The residual values should reflect only random fluctuation, when all the corresponding requirements of model fitting are met.

In Figure 2 we inspect the Pearson residuals against the predicted values, this plot indicates that the model is satisfactory for the data, the residuals appear to be randomly distributed around the line ε=0. Data do not contain any outlier since no observations is far away from the rest of points. Then GEE approach is a good alternative to model this type of data.

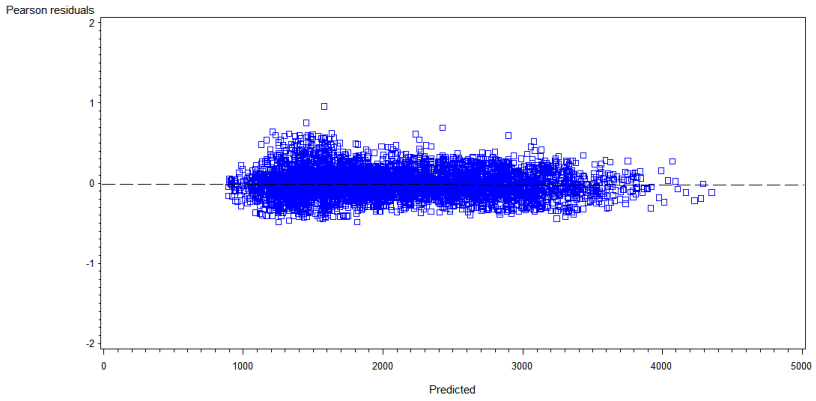


FIGURE 2. Plot of Predicted versus Residual.

References

Hardin,J.W., and Hilbe, J.M. (2003) *Generalized Estimating Equations*. Chapman and Hall / CRC.

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

Maza, L.A., and Peñalosa, J.M. (2010). La situación actual del ajuste de la inversión residencial en España. *Boletín Económico Diciembre*, Banco de España.

Precision of estimators in interval censored parametric survival models

Defen Peng[1] , Gilbert MacKenzie [1,2]

[1] Centre of Biostatistics, University of Limerick, Ireland & Department of Statistics, Zhongnan University of Economics and Law, China
[2] ENSAI, France & Centre of Biostatistics, University of Limerick, Ireland

Abstract: Recently, several advances have been made in the analysis of interval censored (IC) data mainly in relation to semi-parametric proportional hazard (PH) models (Gómez et al., 2009, Lesaffre et al, 2005). It is arguable, however, that the parametric case has been somewhat neglected, overall, and that more can be learned, especially in relation to non-PH models. Accordingly, we focus on simple parametric models for interval censored survival data arising in longitudinal RCTs. For the exponential regression model we compare the performance of a general likelihood with commonly used proxy likelihoods, which ignore the interval censoring by treating the interval censored times to events as if they were exact. We show *analytically* that use of proxy likelihoods leads to estimators which are artificially precise and we quantify the extent of the resulting biases in a simulation study and by analyzing real data. We also compare the likelihoods using non-PH models and obtain different findings.

Keywords: Artificial precision, Interval Censoring, Longitudinal RCTs; PH & non-PH Survival Models, Proxy likelihoods.

1 Introduction

In longitudinal settings where the response variable, $Y(t)$, is binary typically we observe the ith patient at baseline in a healthy state, ie, $Y_i(t_0) = 0$. As the process evolves an adverse event may occur, i.e., $Y_i(t_s) = 1$ where $t_s > t_0$. Finkelstien (1986) and Collett (1994) elected to adopt a "time to event" analysis in order to recover information on the treatment effect in the LDA-RCT setting. Moreover, clinicians (Bergink *et al.*, 1998) have adopted a similar approach in which interval censored follow-up times, to the loss of 3 lines of visual acuity (Bailey-Lovie, 1976), were treated as if they were exact times to events. Intuitively, this simple expedient seems sub-optimal and this note investigates the extent of any penalty incurred by comparing a proxy likelihood with the IC likelihood which arises in longitudinal data (MacKenzie, 1999) . Here we focus on the use of proxy times (beginning, midpoint and endpoint of intervals) to construct the likelihood rather than treating the lack of exact times as missing data to be imputed. We also focus on simple parametric survival models.

2 Likelihood Construction

Suppose there are $m+1$ *fixed*, scheduled, inspection times, t_0^*, t_1^*,..., t_m^* at which continuous or ordinal responses Y_0, Y_1,..., Y_m, are measured. This arrangement implies m+1 time intervals: $I_1 = (t_0, t_1^*], I_2 = (t_1^*, t_2^*], ..., I_k = (t_{k-1}^*, t_k^*], ..., I_m = (t_{m-1}^*, t_m^*]$ and $I_{m+1} = (t_m^*, \infty]$. Typically, $t_0 = 0$, especially in RCTs where, $t_0 = 0$ represents time of randomization. Hence, let T be a non-negative random variable denoting the time to some outcome of interest defined on the Ys. Let $S(t;\theta)$ and $\lambda(t;\theta)$ be the corresponding survival and hazard functions, respectively, depending on the unknown possibly vector-valued parameter $\theta \in \Theta$. Then, for a sample of n independent subjects subject to non-informative censoring the usual likelihood for the unknown parameters is

$$L_2(\theta) = \Pi_{i=1}^n [\lambda(t_i;\theta)S(t_i;\theta)]^{\delta_i}[S(t_{ic};\theta)]^{1-\delta_i}, \tag{1}$$

where $\lambda(t_i;\theta)S(t_i;\theta) = f(t_i;\theta)$, δ_i is the censoring indicator ($\delta_i = 1$ for an event and 0 otherwise) and t_{ic} is a right censored survival time. Substituting, one of: (a) the beginning point of the interval, t_{ib}, or (b) the interval mid-point, t_{im} or, (c) the interval end-point, t_{ie}, $\forall\, i$, as if it were the exact time at which failure occurred in $L_2(\theta)$ yields the proxy likelihood.
Typically each individual ($i = 1, \ldots, n$) defines their own trajectory over the course of the longitudinal study, thereby generating a person-specific set of intervals. Accordingly, we obtain the following interval censored likelihood

$$L_1(\theta) = \Pi_{i=1}^n \{S(t_{i,k-1};\theta)[1 - S(t_{i,k-1}, t_{ik};\theta)]\}^{\delta_i}[S(t_{ic};\theta)]^{1-\delta_i}. \tag{2}$$

Now, $L_1(\theta)$ and $L_2(\theta)$ may be used for comparative inference. Other authors have reached similar conclusions about the structure of the likelihood in the so-called Case II censoring situation; see Yu et al.(2000) and Schick and Yu (2000), for further details of likelihood construction in related contexts. Note, however, it is unusual to have any exact times to events in a longitudinal study.

3 The Exponential Regression Model

MacKenzie (1999) showed analytically that estimators obtained from the proxy likelihood were artificially precise in the simple Exponential case. Here we extend his results to the Exponential Regression case.

3.1 Likelihoods

Armed with these general formulae we investigate the Exponential Regression model. Let T follow the exponential regression model defined by

$$\lambda_{i2} = \lambda(t_i;\alpha_2,\beta_2) = \exp(\alpha_2 + x_i'\beta_2),$$

where $S(t_i;\alpha_2,\beta_2) = \exp[-\lambda_{i2}t_i]$ and α_2 is an unconstrained parameter, β_2 is $p \times 1$ vector of regression coefficients and x_i is a $p \times 1$ vector of fixed covariates. The corresponding proxy likelihood is

$$L_2(\alpha_2,\beta_2) \quad = \quad \prod_{i=1}^{n} \left\{\lambda_{i2}\mathrm{e}^{-\lambda_{i2}t_i}\right\}^{\delta_i} \left\{\mathrm{e}^{-\lambda_{i2}t_{ic}}\right\}^{1-\delta_i}, \tag{3}$$

For the IC likelihood we have

$$\lambda_{i1} = \lambda(t_i;\alpha_1,\beta_1) = \exp(\alpha_1 + x_i'\beta_1),$$

where $S(t_{i,k-1},t_{ik};\alpha_1,\beta_1) = \exp[-\lambda_{i1}d_i(t_k)]$, and $d_i(t_k) = t_{ik} - t_{i,k-1}$ is the width of the kth interval. Then,

$$L_1(\alpha_1,\beta_1) = \prod_{i=1}^{n} \left\{\mathrm{e}^{-\lambda_{i1}t_{i,k-1}}\left[1-\mathrm{e}^{-\lambda_{i1}d_i(t_k)}\right]\right\}^{\delta_i} \left\{\mathrm{e}^{-\lambda_{i1}t_{ic}}\right\}^{1-\delta_i}, \tag{4}$$

3.2 Comparison of IC and Proxy Approaches

Comparing the Proxy and IC approaches we find that approximate IC mles are identical to those estimated at $t_{ie} = t_{ik}$, the end points of the interval using the proxy likelihood (ie, $\hat{\alpha}_1 = \hat{\alpha}_2$ and $\hat{\beta}_{1r} = \hat{\beta}_{2r}$) with proxy t_{ie}.

We compared the relative efficiency of the two estimators by examining $V_2(\hat{\alpha}_2)/V_1(\hat{\alpha}_1)$ and $V_2(\hat{\beta}_{2r})/V_1(\hat{\beta}_{1r})$, $r = 1,2,\cdots,p$. The details are too lengthy to reproduce here. Analytical results are available only for categorical covariates. We have proved the following result for a categorical covariate with $p+1$ categories, modelled by p binary dummy variables, i.e.

$$\begin{aligned} V_2(\hat{\alpha}_{2e})/V_1(\hat{\alpha}_1) \quad &< \quad 1 \\ V_2(\hat{\beta}_{2er})/V_1(\hat{\beta}_{1r}) \quad &< \quad 1 \end{aligned} \tag{5}$$

so that the conjecture that the proxy mles are artificially precise holds, under the first order conditions invoked above, for a single categorical covariate.

We have also proved a similar result for two correlated binary covariates. For higher numbers of correlated binary covariates and for continuous covariates the matrix algebra rapidly becomes intractable. We conjecture that the results hold for two or more categorical variables, but must resort to simulation.

We note in passing that any continuous covariate may be represented in $p \leq n$ distinct categories and hence for such a representation of a continuous covariate the above conjecture holds.

3.3 Information Matrices

The Fisher information matrix based on the IC likelihood for the Exponential regression model is

$$\mathcal{I}(\alpha,\beta) = \begin{bmatrix} \sum_{i=1}^{n} \mathrm{e}^{\alpha+x_i^T\beta} E(t_i') & \sum_{i=1}^{n} x_i^T \mathrm{e}^{\alpha+x_i^T\beta} E(t_i') \\ \sum_{i=1}^{n} x_i \mathrm{e}^{\alpha+x_i^T\beta} E(t_i') & \sum_{i=1}^{n} x_i x_i^T \mathrm{e}^{\alpha+x_i^T\beta} E(t_i') \end{bmatrix} \tag{6}$$

where $E(t_i') = E\big[\delta_i t_{i,k-1} + (1-\delta_i)t_{ci}\big]$. In general, we have

$$\mathcal{I}(\alpha,\beta) = \mathcal{I}_b(\alpha,\beta) + \mathcal{I}_c(\alpha,\beta)$$

where the subscripts represent the beginning of the interval $(t_{i,k-1})$ and right censored (t_{ci}) components respectively.
Fisher Information involves taking the expectation of the negative of the hessian matrix with respect to the random variable T. In this sense it is an averaging or centering operation. Accordingly, in this spirit we may define "general" Fisher information for the IC case by replacing $t_{i,k-1}$ with t_k^* and replacing t_{ci} with its future expectation, as in Buckley & James (1979) yielding

$$\mathcal{I}_{\mathrm{gen}}(\alpha,\beta) = \mathcal{I}_{t_k^*}(\alpha,\beta) + \mathcal{I}_c(\alpha,\beta).$$

Looking at the structure of (6) it is tempting to simplify further by choosing $E(t_i') = E(T_i) = \lambda(t_i)^{-1} = \mathrm{e}^{-(\alpha+x_i^T\beta)}$, whence

$$\mathcal{I}_{\mathrm{ideal}}(\alpha,\beta) = \begin{bmatrix} n & \sum_{i=1}^{n} x_i^T \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i x_i^T \end{bmatrix},$$

an "idealized" form, which is identical to the uncensored solution.
In simulation studies we conduct the exact survival times are known and in these circumstances it is possible to compute an information matrix of the form

$$\mathcal{I}_{\mathrm{rc}}(\alpha,\beta) = \mathcal{I}_u(\alpha,\beta) + \mathcal{I}_c(\alpha,\beta)$$

which we refer to as the "right-censored" version.
In the simulation section we evaluate the performance of all of the above and compare it with the observed information from the IC likelihood which, broadly, we consider should be regarded as the "truth". In the simulation study we found the following relationship between the generalized variances:

$$\det[I_o^{-1}(\hat\alpha,\hat\beta)] > \det[\mathcal{I}_{\mathrm{gen}}^{-1}(\hat\alpha,\hat\beta)] > \det[\mathcal{I}_{\mathrm{ideal}}^{-1}(\alpha,\beta)] > \det[\mathcal{I}_{\mathrm{rc}}^{-1}(\hat\alpha,\hat\beta)].$$

where we have assumed throughout that the δ_i are known.

4 Simulation Study

We conducted a data-directed simulation study mimicking the conduct of a RCT with two arms and a follow-up period of 2 years (Hart et al., 2002). We generated failure times from the Exponential regression model with two covariates: x_1, a binary covariate mimicking the treatment effect (1 = New(50%) and 0 = Old(50%) and x_2 a continuous baseline covariate distributed, $N(0, \sigma^2_{x_2})$, where $\sigma_{x_2} \leq 1$ ($\sigma_{x_2} = 0.5$ in our simulation study). The trajectories for each individual in the study were constructed according to two schedules: an irregular schedule (0.25, 0.5, 1 and 2 years) and a regular schedule (0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75 and 2 years), respectively. Censoring rates of 20% (normal) and 50% (heavy) were considered. The method of creating intervals is non-informative about the survival distribution.

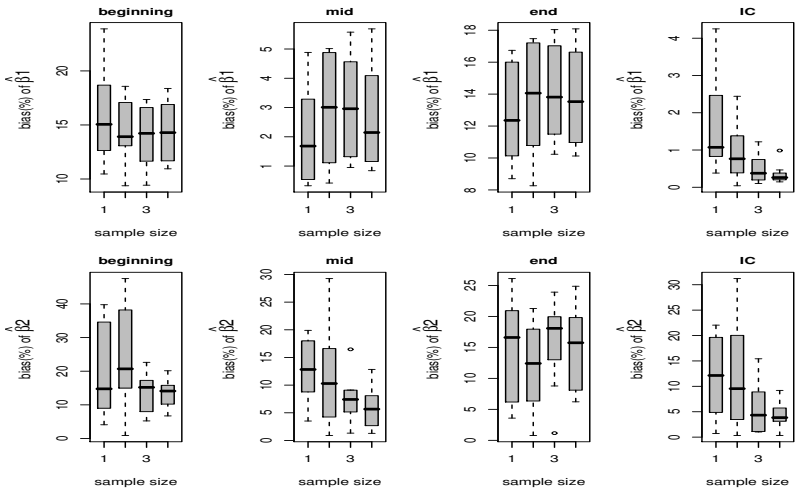


FIGURE 1. Percentage bias for 32 scenarios by sample size for β_1 and β_2. The x-axis labels 1-4 represent sample sizes n=100, 200, 500 and 1000 respectively. Boxplot titles: estimates obtained at the beginning, mid, and end points (identical to the first order approximation) by proxy likelihoods and by IC likelihood (NR).

Figure 1 shows the average percentage bias in an exponential regression model by sample size (n=100, 200, 500 and 1000), likelihood method and for two covariates (β_1, β_2) using the irregular schedule. These results show that the estimators from the IC likelihood (NR) have minimum bias and that the bias is asymptotically consistent. However, for the proxy likelihoods this is not the case. Only the mid-point estimator has acceptable levels of bias, but the box-plots (which depict the variation over scenarios) suggest a lack of consistency for β_1. The findings are similar for the regular schedule. We also considered the Weibull PH model and two non-PH models - the log-logistic the canonical time dependent logistic.

For PH models the results showed that, among the IC and proxy likelihoods considered, the estimators in the IC likelihood had the largest variances. This was re-assuring, as á priori, one might reasonably expect the IC likelihood to represent the most uncertainty. Accordingly, this demonstrates that the estimators in all of the proxy likelihoods are artificially precise. However, surprisingly, for non-PH models this finding did not hold. We were able to find immediate contradictions in the non-PH models. The results will be described in detail at the Workshop together with the analysis of two published data sets.

5 Discussion

The analysis of IC data has been reviewed recently by Gomez et al (2009). Here, we tried to develop an analytical approach to the analysis of precision of the regression estimator. This was successful in the Exponential Regression model for simple cases. However, for more complicated cases, the algebra rapidly becomes intractable and one must resort to simulation. Our findings support the conjecture that the estimators based on the proxy likelihoods are artificially precise in the PH models studied. Hence proxy approaches should be avoided, especially in RCTs, when the data obey the PH assumption. However, this is apparently not true of non-PH models, a finding which warrants further investigation.

Acknowledgments: The work was supported by the Science Foundation Ireland (SFI, www.sfi.ie) Mathematics Initiative, II, via the BIO-SI (www.ul.ie/bio-si) research programme in the Centre of Biostatistics, University of Limerick, Ireland: grant number 07/MI/012. In addition, Professor Peng is also supported by SFI via a Research Frontiers Programme award, grant number 05/RF/MAT 026.

References

Gómez et al (2009). Tutorial on methods for interval censored data *Statistical Modelling* 9, 4, 259-298.

Lawless J and Babineau (2006). Models for interval censoring and simulated-based inference for lifetime distributions. *Biometrika*, **93**, 671-686.

MacKenzie G (1999). Survival analysis for longitudinal data. Proceedings of the 14th International Workshop on Statistical Modelling, July, Graz, Austria. July 1999, pages 259-264.

Peng D (2009). *Inferences in the Interval Censored Exponential Regression Model*. Masters Thesis MacMaster University, Ontario, Canada.

A Bayesian spatial approach to modelling fish species occurrence.

Maria G. Pennino[1], José M. Bellido[1], David Conesa[2], Antonio López-Quílez[2], Facundo Muñoz[2]

[1] Instituto Español de Oceanografía. Centro Oceanográfico de Murcia. C/Varadero 1. San Pedro del Pinatar. 30740. Murcia. Spain.
[2] Geeitema. Dept. Estadística i Investigació Operativa. Universitat de València. C/Dr. Moliner 50. Burjassot. 46100. Valencia. Spain.

Abstract: A methodological approach for modelling species occurrence patterns for fisheries management purposes is here proposed. The presence/absence of the fish species is modelled with a hierarchical Bayesian model using the geographical and environmental characteristics of each fishing spot. Maps of predicted probabilities of presence are generated using Bayesian kriging.

Keywords: Bayesian kriging; Bayesian hierarchical models; fisheries; modelling distribution species.

1 Introduction

Modelling species presence/absence patterns using local environmental factors has been a growing matter in Ecology in the last few years. The distribution models have been extensively used to address several issues, which include identifying Essential Fish Habitat (EFH) in order to classify and manage conservation areas, and predicting the response of species to environmental features. Different approaches and methodologies have been proposed for modelling species distribution. Generalized linear and additive models (GLM and GAM), species envelope models such as BIOCLIM, and the multivariate adaptive regression splines (MARS) are some of them. Most of these applications are only explanatory models that seek to assess the relationship between a species presence and a suite of one or more explanatory variables (e.g. precipitation, bathymetry, etc.). But, as these models are based on the use of independent variables, its application to fishery data often ignores the spatial autocorrelation. On the other hand, few works have been developed for predictive models, although these models, in addition to offer an estimate of the processes that drive the species distribution, also seek to provide one or more useful factors to predicting the probability of species occurrence in unsampled areas (Chakraborty et al., 2010).

Our interest here is to propose a hierarchical Bayesian model to predict species occurrence while incorporating the environmental and spatial features of each fishing spot. Bayesian approach is appropriate to spatial hierarchical model analysis of fisheries because it allows both the observed data and model parameters to be random variables (Banerjee et al., 2004), resulting in a more realistic and accurate estimation of uncertainty. Also, implementation of Markov chain Monte Carlo (MCMC) avoids asymptotic inference and the computational problems encountered in likelihood-based fitting. Moreover, incorporating prior information can usually be very helpful in discriminating spatial autocorrelative effects from ordinary non-spatial linear effects. In our proposal, maps of predicted probabilities of presence in unsampled areas are generated using Bayesian kriging.
In particular, we have applied our approach to describe the distribution of the Mediterranean mackerel (*Trachurus mediterraneus*), in the Western Mediterranean. Environmental satellite data, such as the monthly data on precipitation, sea surface temperature (SST) and chlorophylla-a concentration have also been included into the analysis.

2 Modelling fish presence

Point-referenced spatial models are very suitable for situations in which we have observations made at continuous locations happening within a defined spatial domain. This particular case of spatial models has also the appealing characteristic that the spatial domain is unchanging, even though the spots locations will change over time. In fisheries, this resolves the dimensional control guarantying that the inference is realized in relation to the domain instead of the current observed positions, which can change over the years. When analyzing the distribution of fish species, data observed at each location could be used as a measure of relative species occurrence at those precise locations. When absolute abundance information is not available, the spatial distribution of a species based on presence/absence could be used to assess their status and distribution. In other words, our response variable of interest will be the presence/absence (with Bernoulli distribution) instead of the absolute abundance (with normal distribution). Then, assuming that the probability of catching a species is related to its presence, we can model the presence/absence using a point-referenced spatial hierarchical bayesian model. In particular, if Z_i and π_i represents respectively the presence (1) or absence (0) and the probability of presence at location i, then:

$$Z_i|W_i, U_i \sim Ber(\pi_i)\,, \tag{1}$$

where W_i and U_i represent the spatial random effects and the non-spatial random effects respectively, and the relation between π_i and the covariates of interest is the usual logit link:

$$logit(\pi_i) = X\beta + U_i + W_i\,,$$

where $X\beta$ represents the usual linear predictor.
In order to complete the model we have to specify the prior distributions of W_i, which is responsible of the spatial dependence through the correlation matrix with an exponential variogram:

$$W|\sigma^2, \phi \sim N(0, \sigma^2 H(\phi)), \tag{2}$$

and U_i, which provides the nugget effect:

$$U_i|\tau^2 \sim N(0, \tau^2). \tag{3}$$

Once the model is determined, the next step is to estimate its parameters. As we are using the Bayesian paradigm, we have to specify the (hyper) prior distributions of each parameter involved in the model. We have considered rather noninformative prior distributions, with the aim of expressing our initial vague knowledge about the parameters. Expressions above jointly with the priors of all the parameters contain all our knowledge of the system but they do not yield to analytical estimates. Therefore, we have had to resort to numerical methods in order to obtain the posterior distributions of all the parameters and also to make prediction about the presence/absence in a series of unsampled locations. In particular, MCMC inference have been carried out using WinBUGS (Spiegelhalter et al., 1999).

3 Predictive model

Bayesian kriging has been used to predict the species occurrence where fishery data is not available. This approach treats the species occurrence at a new location as a random variable and calculates, in addition to the estimation, a range of likely values together with their probabilities to be the true values at a specific location. Furthermore, it takes into account the uncertainty over the hyperparameters, a substantial advantage over the classical kriging methods. We have computed the predictive posterior distributions at new locations using `INLA` (Rue et al., 2009).

4 Results and Conclusions

To demonstrate the usefulness of our approach, we have applied the model discussed above to the Mediterranean mackerel fishery data, in the Almería Bay. This species is not a targeted species of the commercial fishery, so its occurrence is a good indicator of this presence/absence. We have used bathymetry and environmental factors, such as precipitation, SST, and chlorophylla-a concentration as covariates. In order to compare competing models, we have used the deviance information criterion (DIC) introduced by Spiegelhalter et al. (2002). Our analysis demonstrates that both the bathymetry and the chlorophylla-a concentration play an important role in the mackerel distribution. Figure 1 shows the estimated spatial effect for the pelagic fishery unit.

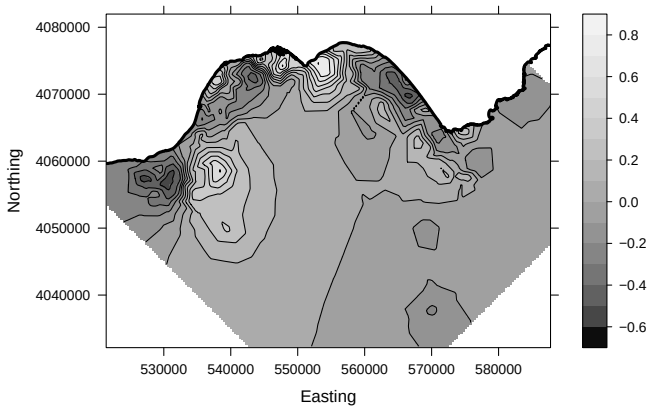


FIGURE 1. Mediterranean mackerel (*Trachurus mediterraneus*) mean of the spatial effect in the Almería Bay.

Acknowledgments: David Conesa, Antonio López-Quílez and Facundo Muñoz would like to thank the financial support of the Ministerio de Educación y Ciencia (jointly financed with European Regional Development Fund) via the research Grant MTM2010-19528 and of the Generalitat Valenciana via the research Grant ACOMP11/218.

References

Chakraborty, A., Gelfand, A.E., Wilson, A.M., Latimer, A.M., Silander, J.A. (2010). Modeling large scale species abundance with latent spatial processes. *The Annals of Applied Statistics*, **4**, 3, 1403-1429.

Banerjee, S., Carlin, B.,Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC Press.

Spiegelhalter, D.J., Myles, J. P., Jones, D.R., Abrams, K.R. (1999). Methods in health service research. An introduction to Bayesian methods in health technology assessment. *British Medical Journal*, 319, 508-612

Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, **6**, 71, 319-392.

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

The truncated inflated beta regression

Gustavo H. A. Pereira[1], Denise A. Botter[1], Mônica C. Sandoval[1]

[1] Department of Statistics, University of São Paulo, Caixa Postal 66281, São Paulo/SP, 05311970, Brazil. E-mail: ghapereira@terra.com.br

Abstract: The beta regression or the inflated beta regression may be a reasonable choice to fit a proportion in most situations. However, they do not fit well variables that do not assume values in the open interval $(0, c)$, $0 < c < 1$. Variables related to a kind of double bounded payment amount when studied as a proportion of the maximum payment amount have this feature. For these variables, Pereira et al. (2011) introduced the truncated inflated beta distribution (TBEINF). This work proposes a regression model where the response variable is TBEINF distributed. The model allows all the unknown parameters of the conditional distribution of the response variable to be modelled as a function of explanatory variables. Moreover, the model allows nonconstant known parameter c across population units. For these model, maximum likelihood estimation is discussed and closed-form expressions for the score function and for Fisher's information matrix are provided. In addition, some results when c is not constant are obtained, Monte Carlo simulation studies are performed and an application to credit card data is presented.

Keywords: Beta regression; Inflated distributions; Proportions.

1 Introduction

The beta regression (Ferrari and Cribari, 2004) or the inflated beta regression (Ospina, 2010; Ospina and Ferrari, 2008) may be a reasonable choice to fit a proportion in most situations. However, they do not fit well variables that do not assume values in the open interval $(0, c)$, $0 < c < 1$. Variables related to a kind of double bounded payment amount when studied as a proportion of the maximum payment amount have this feature. An example of these variables is the ratio between the payment amount and the total amount owed in the credit card (proportional payment amount, PPA). A credit card holder receives a monthly statement indicating the minimum payment and, hence, PPA can not assume values in the closed interval $(0, c)$, where c is a known value. Therefore, the variable has positive probability at points zero, c and one because many credit card holders do not have enough money to pay anything, and many others can pay only the minimum due, whereas there are many others who pay the entire amount

owed to avoid interest charges. In addition, it can assume any real number in the interval $(c, 1)$. Another variable with the same features is the ratio between the amount of the unemployment insurance benefit and the maximum allowable benefit paid to unemployed workers in Brazil.
Pereira et al. (2011) introduced the truncated inflated beta distribution (TBEINF) for variables that assume values at zero, at one and at a known value c with positive probability and at any real number in the open interval $(c, 1)$. This work proposes a regression model where the response variable is TBEINF distributed. The model allows all the unknown parameters of the conditional distribution of the response variable to be modelled as a function of explanatory variables. For this model, maximum likelihood estimation is discussed and closed-form expressions for the score function and for Fisher's information matrix are provided. In addition, some results when c is not constant are obtained, Monte Carlo simulation studies are performed and an application to credit card is presented.

2 The TBEINF regression

The model proposed in this work allows nonconstant known parameter c across population units. For PPA in the credit card, c is not constant across population units and in the unemployment insurance example, c is constant. The properties of the model are similar when c is constant or nonconstant. For this reason, we propose a single model that allows nonconstant c. Thus, we propose the following definition for the truncated inflated beta regression (TBEINF regression).

Definition 1 *Let $Y_1, Y_2, \ldots, Y_n$ independent random variables, where $Y_i \sim$ TBEINF$(\delta_{i0}, \delta_{i1}, \delta_{ic}, \mu_i, \phi_i, c_i)$ as defined in Pereira et al. (2011). The truncated inflated beta regression are defined by the TBEINF distribution and the following functional relations:*

$$g_1(\mu_i) = \eta_{i1},\ g_2(\phi_i) = \eta_{i2},\ H(\delta_{i0}, \delta_{i1}, \delta_{ic}) = (\zeta_{i0}, \zeta_{i1}, \zeta_{ic}), \tag{1}$$

where $\eta_{i1} = x_{i1}^\top \beta_1$, $\eta_{i2} = x_{i2}^\top \beta_2$, $\zeta_{i0} = z_{i0}^\top \gamma_0$, $\zeta_{i1} = z_{i1}^\top \gamma_1$ and $\zeta_{ic} = z_{ic}^\top \gamma_c$ are linear predictors, $\beta_1 = (\beta_{11}, \beta_{21}, \ldots, \beta_{p_\mu 1})^\top$, $\beta_2 = (\beta_{12}, \beta_{22}, \ldots, \beta_{p_\phi 2})^\top$, $\gamma_0 = (\gamma_{10}, \gamma_{20}, \ldots, \gamma_{p_0 0})^\top$, $\gamma_1 = (\gamma_{11}, \gamma_{21}, \ldots, \gamma_{p_1 1})^\top$ and $\gamma_c = (\gamma_{1c}, \gamma_{2c}, \ldots, \gamma_{p_c c})^\top$ are vectors of unknown parameters, $x_{i1} = (x_{i11}, x_{i21}, \ldots, x_{ip_\mu 1})^\top$, $x_{i2} = (x_{i12}, x_{i22}, \ldots, x_{ip_\phi 2})^\top$, $z_{i0} = (z_{i10}, z_{i20}, \ldots, z_{ip_0 0})^\top$, $z_{i1} = (z_{i11}, z_{i21}, \ldots, z_{ip_1 1})^\top$ and $z_{ic} = (z_{i1c}, z_{i2c}, \ldots, z_{ip_c c})^\top$ are known explanatory variables, $g_1 \colon (c_i, 1) \to \mathbb{R}$ and $g_2 \colon \mathbb{R}^+ \to \mathbb{R}$ are link functions strictly monotonic and twice differentiable, and $H \colon C \to \mathbb{R}^3$ is a bijective link function and twice differentiable, where C is a subspace of $\mathbb{R}^3$ defined as $C = \{(\delta_{i0}, \delta_{i1}, \delta_{ic}) : 0 < \delta_{i0} < 1, 0 < \delta_{i1} < 1 - \delta_{i0}, 0 < \delta_{ic} < 1 - \delta_{i0} - \delta_{ii}\}$.

Parameters can be estimated by maximum likelihood using a numerical nonlinear optimization algorithm. We find closed-form expressions for

Fisher's information matrix and conclude that $(\gamma_0^\top, \gamma_1^\top, \gamma_c^\top)$ and $(\beta_1^\top, \beta_2^\top)$ are orthogonal parameters. We also obtain expressions for confidence intervals, confidence regions and four test statistics.
An interesting feature of the TBEINF regression model is that it allows nonconstant known parameter c across population units. This means that the model allows the support of the distribution of the response variable varies across population units, since y_i can assume values in $0 \cup [c_i, 1]$. As mentioned earlier, the properties of the model are similar when c is constant or nonconstant. However, some practical issues should be discussed when c is not constant. In these cases, as c increases, the width of the interval $(c, 1)$ decreases. Therefore, it is reasonable to expect that even if two population units have the same values for the explanatory variables, they will probably have different μ if they have different c. For this reason, it may be reasonable to include c_i or a function of c_i as an explanatory variable for μ_i. For the link function $g_1(\mu_i) = \log[(\mu_i - c_i)/(1 - \mu_i)]$, some results are obtained when c is not constant.

3 Simulation studies

The finite-sample behavior of the estimators are studied through Monte Carlo simulation studies. The obtained results suggest that maximum likelihood estimators in TBEINF regression model do not have large bias even in small samples. In addition, the performances of the estimators of the vector of parameters β_1 seem to be better than the estimators of the vectors β_2, γ_0, γ_1 and γ_c. Some others conclusions are obtained and discussed.

4 Application

This section contains an application to credit card data. Sample includes 5000 credit cards for non-account customers of a financial institution. The dependent variable is the PPA in the credit card. Length of time as customer, balance-to-limit ratio, a dummy variable for installment purchases, among others are used as independent variables.
For all variables, the signs of the estimated parameters are in agreement with what was expected from the descriptive analysis performed. For parameters δ_0, δ_1, δ_c and μ of the distribution of the response variable, we split the sample into five groups based on quintiles of the fitted values for these parameters and obtain for each group the average fitted value, the relative frequency (for δ_0, δ_1 e δ_c) and the sample average (for μ). Table 1 presents the results and suggests that the model fits relatively well.

5 Conclusion

We introduced in this work the truncated inflated beta regression. Inferential results were obtained, Monte Carlo simulation studies were performed

TABLE 1. Empirical and fitted values for the TBENF regression for the variable PPA in the credit card.

	δ_0			δ_1	
Fitted value	Aver. fit	Relat. freq.	Fitted value	Aver. fit	Relat. freq.
]0.0000;0.0088]	0.006	0.006	]0.0000;0.1450]	0.070	0.075
]0.0088;0.0131]	0.011	0.009	]0.1450;0.7900]	0.392	0.404
]0.0131;0.0222]	0.016	0.018	]0.7900;0.8920]	0.860	0.840
]0.0222;0.0746]	0.046	0.044	]0.8920;0.9159]	0.906	0.899
]0.0746;1.0000[	0.157	0.147	]0.9159;1.0000[	0.924	0.937
	δ_c			μ	
Fitted value	Aver. fit	Relat. freq.	Fitted value	Aver. fit	Sample aver.
]0.0000;0.0195]	0.016	0.007	]0.0000;0.3460]	0.306	0.302
]0.0195;0.0245]	0.022	0.021	]0.3460;0.4465]	0.395	0.386
]0.0245;0.0374]	0.029	0.035	]0.4465;0.5800]	0.507	0.512
]0.0374;0.1728]	0.084	0.088	]0.5800;0.7168]	0.670	0.709
]0.1728;1.0000[	0.353	0.357	]0.7168;1.0000[	0.766	0.767

and two applications were presented. The results suggest that TBEINF regression is useful to fit some variables in the field of econometrics.

Acknowledgments: The authors gratefully acknowledge partial financial support from CNPq and FAPESP. We also thank the financial institution that provided credit card data.

References

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

Ospina, R. (2008). Modelos de regressão beta inflacionados. Phd thesis. Universidade de São Paulo.

Ospina, R., and Ferrari, S.L.P. (2010). Inflated beta distributions. *Statistical Papers*, **51**, 111-126.

Pereira, G.H.A., Botter, D.A., and Sandoval, M.C. (2011). The truncated inflated beta distribution. *Communications in Statistics - Theory and Methods*, **to appear**.

A Bayesian analysis of survival times for stage IV non-small cells lung cancer

S. Perra[1], A. Quirós[2], C. Armero[3], S. Cabras[1], M. E. Castellanos[2], M. J. Oruezábal[4], J. Sánchez-Rubio[4]

[1] Università degli Studi di Cagliari
[2] Universidad Rey Juan Carlos
[3] Universitat de València
[4] Hospital Infanta Cristina de Madrid

Abstract: We present a Bayesian Weibull regression analysis of survival times for stage IV non-small cells lung cancer and propose a methodology to select the prognostic variables for overall and progression-free survival.

Keywords: BIC; Prognostic factors; Weibull regression.

1 Introduction

Lung cancer is the second incident malignant neoplasia in Spain. Despite all new advances in its treatment, five-year absolute survival rate is currently only 10.2% (Sant *et al.*, 2009), and consequently, searching for new therapeutic strategies becomes essential. One of the most important facts which restricts the election of the appropriate treatment is the impossibility of personalizing the most adequate therapeutic option for each patient.
Although usual procedures to select the most appropriate treatment for a patient suffering cancer are based on clinical trials, there are a wide scientific consensus (National Cancer Institute, National Institute for Health and Clinical Excellence) for promoting alternative methodologies, such as observational studies, to help improving therapy decision making. In particular, understanding the role and significance of prognostic factors in cancer will be very useful to gain insights about the different and complex elements of this disease.
This work discusses a Bayesian Weibul regression analysis for survival times of stage IV non-small cells lung cancer (NSCLC), which is the most common type of lung cancer, and proposes a methodology to select the most related predictor variables to the overall and the progression-free survival.

2 The data

Data are provided by the Infanta Cristina Hospital of Madrid and consist of survival times for stage IV NSCLC and several covariates that may be

related to the disease, observed along the time period from January 2008 to December 2010. A total of 35 patients have been observed, for which we annotate 14 variables related to: patient information (age, sex, smoking habit, body mass index, baseline state, previous complications), tumour characteristics (location, number of affected organs, histologic type) and baseline analytics (cea: carcinoembryonic antigen, ldh: lactate dehydrogenase, anaemia, calcaemia, albumin). From the 35 patients present in the study, there are 19 (54%) censored overall survival, while the number of censored observations is 22 (63%) for the progression-free survival.

3 Survival Analysis

Consider $(t_1, \boldsymbol{x}_1, \delta_1), \ldots, (t_n, \boldsymbol{x}_n, \delta_n)$, where t_i denotes the time in which even occur or a censored time for individual i, $\boldsymbol{x}_i$ denotes covariates and $\delta_i = 0$ indicates censored time for subject i. We are interested in modelling the overall and the NSCLC progression-free survival, where the events of interest are *dead* and *progression*, respectively.

The Weibull distribution is a flexible model for survival data (Klein and Moeschberger, 2003) which considers that the survival time, T, follows a Weibull distribution of parameters α and λ, $\mathcal{W}(\alpha, \lambda)$. We consider this model in the context of accelerated failure-time models where covariates are introduced. Considering the logarithm of the survival time, it is possible to write

$$Z_i = \log(T_i) = \mu + \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma W_i, \tag{1}$$

where $\boldsymbol{x}_i$ is the vector of covariates of dimension p and W is the standard Gumbel distribution with density and survival functions, $f_W(w) = \exp(w - e^w)$, $S_W(w) = \exp(-e^w)$, respectively. The distribution for T_i results in a Weibull distribution with parameters $\alpha = 1/\sigma$ and $\lambda_i = \exp(-(\mu + \mathbf{x}_i'\boldsymbol{\beta})/\sigma)$. Note that the sign of parameters $\boldsymbol{\beta}$ indicates whether the covariates are related negatively or positively with the increase of the risk, with a negative coefficient indicating that the risk increases with the associated covariate.

In order to make inference about parameters in model (1) from a Bayesian perspective, we need to specify a prior distribution over parameters. In this first study we do not have previous information to use for the elicitation of this prior distribution, hence, we use a prior distribution reflecting minimum information, $\pi(\mu, \boldsymbol{\beta}, \sigma) \propto 1/\sigma$. This prior, that leads to a proper posterior distribution with $p + 2$ uncensored observations, has been proposed in Evans and Nigm (1980).

The posterior distribution can be written as:

$$\pi(\mu, \boldsymbol{\beta}, \sigma | \boldsymbol{z}, \boldsymbol{\delta}, \boldsymbol{X}) \propto \frac{1}{\sigma} \prod_{i=1}^{n} \left(\frac{1}{\sigma} f_W \left(\frac{z_i - (\mu + \boldsymbol{x}_i'\boldsymbol{\beta})}{\sigma} \right) \right)^{\delta_i} \left(S_W \left(\frac{t_i - (\mu + \boldsymbol{x}_i'\boldsymbol{\beta})}{\sigma} \right) \right)^{1-\delta_i},$$

where $\boldsymbol{X}$, is a design matrix composed by the p covariates observed in the n subjects in the sample. We have approximated the posterior distribution using Markov Chain Monte Carlo (MCMC) methods.

4 Prognostic Variables Selection

In order to select the covariates that are significantly related with the survival, we have used the Bayesian Information Criterion (BIC). The BIC was derived by Schwarz (1978) as a large sample approximation to twice the logarithm of the Bayes Factor, which quantifies the evidence for one hypothesized model against another (Kass and Raftery, 1995). In particular, we use the BIC proposed in Volinsky and Raftery (2000) for censored data. When different values of BIC are obtained for different models, the smaller the value of BIC the greater the evidence in favor of the model. Once we have sorted all the possible models with additive effects following the BIC, we have interpreted the relations between the covariates included in these models and the survival, excluding those models where the sign of the coefficients was not compatible with previous medical knowledge.

5 Results

Among the possible models for overall survival, we have selected the model with the smallest BIC, in which covariates are calcaemia and smoking habit. The fitted model has an increasing baseline hazard along time, with probability 0.86. Regarding the effect of calcaemia, β_{calc}, is clearly lower than 0. The effect of having a greater value of calcaemia significantly increases the risk. Similar comments apply to the effect of smoking.
Regarding NSCLC progression-free survival, we select the model of smallest BIC among the models compatible with medical knowledge. The final model has the following covariates: cea, albumin, body mass index and ldh.

6 Discussion

Instead of the Weibull regression model we could have used nonparametric models. However, although these are very popular in survival analysis, as stated in Ibrahim *et al.* (2001), parametric models play an important role in Bayesian survival analysis, since they offer straightforward inference even for small sample sizes.
The methodology used for model selection combines the BIC together with the expert knowledge, leading to an effective way of extracting the prognostic factors for overall and progression-free survival. We acknowledge that the small sample size can make problematic the use of BIC and on this pourpose expert knowledge has been relevant for model selection.

The results obtained are in the line of the expert knowledge and suggest that calcaemia and smoking habit are related to the overall survival of a patient. Regarding progression-free survival, we find that cea, albumin, body mass index and ldh are the most related variables. This result also agrees with the expert knowledge, as cea, albumin and ldh are related to the progression of the tumour, and body mass index is connected with the state of the patient that is receiving the treatment.
It is important to note that the aim of this work is to present an application of survival analysis and model selection and that the strength of eventual clinical conclusions need to be measured in the light of the available sample. It would be interesting to apply these methods to more data.

Acknowledgments: This research has been partially supported by the Ministerio de Ciencia e Innovación grant MTM2010-19528 and Fundación Mutua Madrileña grant AP75942010, Ministero dell'Istruzione, dell'Università e della Ricerca of Italy and the visiting professor program of the Regione Autonoma della Sardegna.

References

Evans, I.G., and Nigm, A. (1980). Bayesian prediction for two-parameter weibull lifetime models. *Comm. Statistics - Theory and Methods*, **9(6)**, 649-658.

Ibrahim, J.G., Chen, M.H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer.

Kass, R.E., and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Klein, J.P., and Moeschberger, M. (2003). *Survival Analysis. Techniques for Censored and Truncated Data*. Springer.

Sant M., Allemani C., Santaquilani M., Knijn A., Marchesi F. and Capocaccia R. (2009). EUROCARE-4. Survival of cancer patients diagnosed in 1995-1999. *Eur. J. Cancer* **45**, 931-91.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Volinsky, C.R., and Raftery, A. (2000). Bayesian information criterion for censored survival models. *Biometrics*, **56(1)**, 256-262.

On probabilities of avalanches triggered by alpine skiers. Models with random effects taking the stratified data into account.

Christian Pfeifer[1]

[1] Institut für Statistik, Universität Innsbruck, Universitätsstraße 15, A–6020 Innsbruck

Abstract: In this paper we take into account the stratified data structure of former avalanche models using mixed effect models.

Keywords: avalanche danger; random effects.

1 Introduction

In Austria, most fatal snow avalanche accidents are caused by skiers or snowboarders. 79 avalanche accidents (17 fatalities) were reported during the winter of 2001/02. 16 out of 17 these fatalities were caused by alpine skiers or snowboarders. By far the highest number of accidents took place in Tyrol (2001/02: 47 accidents/ 12 fatalities).
However, it is rather difficult to predict the risk (=probability) of avalanche events on a backcountry ski slope under given conditions. About 10 years ago, the mountain guide Werner Munter suggested a quantitative method in order to estimate the risk of avalanche events. Assuming that the variables

- danger levels from the local avalanche information service (low=1 to very high=5),
- incline of the slope (flat to steep),
- aspect of the slope (north, south) and
- skiers behaviour

have an influence on the risk, he calculated a quantity which he calls "remaining risk". On the base of this quantity, he developed a strategy for backcountry skiers whether to go or not to go on a skiing tour (stop if "remaining risk" is larger than 1, see [3]). But Munter's quantity cannot be understood as a probability of avalanche events. Moreover, there is no empirical evidence for his method because he does not take skiing incidents without avalanche accidents into account ([5]). At least, it is necessary to

include some information on frequencies of skiers on slopes under specific conditions.
At the 25^{th} International Workshop on Statistical Modelling (see [7] and [8]) we proposed a logistic regression model (no vs. on or more accidents as dependent variable for days i with avalanche reports from the Tyrolean avalanche information service), in order to estimate the probabilities $\mathbf{p}$ in question.

$$\text{logit}(\mathbf{p}) = \texttt{LWS} + \texttt{NEIG} + \texttt{EXPOS} + \texttt{WOENDE} + \texttt{TOURV}$$

Beside danger level *lws*, incline of slope *neig* and aspect of slope *expos* we took the qualitative variates skiing conditions *tourv* and day of the week *wotag* into consideration. There is some evidence that frequencies of skiers on slope strongly depend on weather and snow conditions and on the days of the week (weekend, working days). We used accident data and avalanche forecasts in Tyrol within the seasons 1999-2002 reported by the Tyrolean avalanche information service (497 days of observation). Based on this very simple model, we established a decision strategy for backcountry skiers based on empirical/statistical arguments.

2 Mixed models

In this paper we are considering some comments of the discussion at the 25^{th} IWSM meeting (see [8]). For example, we are going to take the stratified data structure of the avalanche data into account. We notice that several observations (one at each class of incline and aspect of the slope) are taken on the same day of observation. As a result of this, we introduce the variable *day* as a random effect (ν) considering at the same time the variables *lws*, *neig*, *expos*, *woende*, *tourv* as fixed effects (β).

1 In the case of the logistic model above, there are several `R`-packages (Bayesian and non-Bayesian) which allow us to calculate estimates of this random effect model. We are using the package `glmmML` written by G. Broström and H. Holmberg ([1]).

Table 1 shows the parameter estimates and standard errors for the logit model with (mixed model) and without random effects. Beside the fixed effects, σ denotes the estimated standard deviation of the random effect ν.

	Logit		Logit mixed	
	param	se	param	se
ICPT	-7.228	0.639	-8.548	0.871
LWS	0.912	0.178	1.087	0.242
NEIG	0.832	0.146	0.916	0.157
EXPOS	-0.578	0.215	-0.644	0.228
WOENDE	0.401	0.215	0.497	0.284
TOURV1	-0.123	0.290	-0.191	0.379
TOURV2	-0.936	0.381	-1.108	0.492
σ			1.328	0.216

2 If we consider counts of avalanche events instead of logit probabilities as dependent variable (see [6]), we have to employ ZIP models (or something which is similar to ZIP models): Avalanche accidents are rather rare events and thus we assume the counts to come from a mixture of a Bernoulli and Poisson distribution. In order to define the covariate effects on the observations we define link functions of the logit and the loglinear model in our case as follows:

$$\text{logit}(\mathbf{p}) = \mathbf{X}\beta + \nu \qquad \log(\lambda) = \tau(\mathbf{X}\beta + \nu),$$

which is denoted as ZIP(τ) model, using τ as a shape parameter (see [4] and [9]). In the case of random effects with regard to ZIP models, we are doing maximum likelihood estimation (`nlminb`) using numerical integration (Gauss-Hermite quadrature) for calculating the full likelihood ([2]).

Table 2 shows results (parameter estimates and standard errors and log-likelihood) for the Poisson, the ZIP(τ) and the mixed ZIP(τ) model.

	Poisson		ZIP(τ)		ZIP(τ) mixed	
	param	se	param	se	param	se
ICPT	-7.025	0.584	-5.426	1.278	-7.242	0.636
LWS	0.937	0.165	0.805	0.242	0.942	0.180
NEIG	0.795	0.136	0.678	0.193	0.820	0.144
EXPOS	-0.541	0.200	-0.464	0.203	-0.563	0.211
WOENDE	0.323	0.199	0.292	0.186	-0.367	0.215
TOURV1	-0.314	0.256	-0.248	0.245	-0.222	0.284
TOURV2	-1.090	0.343	-0.928	0.389	-1.025	0.376
σ					0.355	0.013
τ			0.302	0.363	0.728	0.150

3 Discussion

Comparing the results of the logit models in table 1, we notice that the effects are stronger in the mixed case. If we consider the results of the mixed ZIP model in table 2 we take notice that the coefficients are more similar to those of the logit model than to those of the ZIP model.
Considering the estimated probabilities (which are the quantities that are of interest for the 'avalanche community'), we notice that they are slightly smaller in the mixed logit case and more or less equal to the mixed ZIP case if we compare the quantities with the logit model in table 1.

References

Broström (2003). Generalized linear models with random intercepts; http://www.stat.umu.se/forskning/reports/glmmML.pdf

McCulloch C.E. Searle S.R. (2001): Generalized, Linear, and Mixed Models; Wiley, New York.

Munter W. (1997): 3x3 Lawinen; Pohl & Schellhammer, Garmisch-Partenkirchen.

Lambert D. (1992): Zero-Inflated Poisson regression, with an application to defects in manufacturing; Technometrics 34, 1-14.

Pfeifer C. Rothart V. (2002): Die Reduktionsmethode zur Beurteilung der Lawinengefahr für Schitourengeher aus statistischer Sicht; Jahrbuch der Kuratoriums für alpine Sicherheit 2002, Innsbruck.

Pfeifer C. Rothart V. (2004): On probabilities of avalanches triggered by alpine skiers. An application of models for counts with extra zeros; Proceedings International Workshop of Statistical Modelling 2004 Florence.

Pfeifer C. (2009): On probabilities of avalanches triggered by alpine skiers. An empirically driven decision strategy for backcountry skiers based on these probabilities; Natural Hazards 2009; 48/3: 425-438.

Pfeifer C. (2010): On probabilities of avalanches triggered by alpine skiers. An empirically driven decision strategy for backcountry skiers based on these probabilities; Presentation at the International Workshop of Statistical Modelling 2010 Glasgow.

Welsh, A.H. et al. (1996): Modelling the abundance of rare species: statistical models of counts with extra zeros; Ecological Modelling 88, 297-308.

Cancer incidence in kidney transplant recipients

Salvador Pita-Fernández[1], Teresa Seoane-Pillado[1],Francisco Valdés-Cañedo[2],Rocio Seijo-Bestilleiro[1], Sonia Pértega-Díaz[1], Constantino Fernández-Rivera[2], Angel Alonso-Hernández[2], Dolores Lorenzo-Aguiar[2], Beatriz López-Calviño[1], Andrés López-Muñiz[2]

[1] Clinical Epidemiology and Biostatistics Unit. A Coruña Hospital. As Xubias, 84. Hotel de Pacientes 7 Planta. 15006 A Coruña, Spain.
[2] Department of Nephrology. A Coruña Hospital. As Xubias, 84. Hotel de Pacientes 7 Planta. 15006 A Coruña, Spain.

Abstract: The Kaplan-Meier method is commonly used to estimate the incidence of an event over time. The use of models that take into account the presence of competing risks will allow more precise estimates in this context.The aims of this study are: (i) to establish the incidence of cancer in recipients of renal transplants performed in a hospital in A Coruña (Spain) during the period 1981–2008 compared to that experienced by the general population, ii) to demonstrate the importance of apporpiately estimating the cumulative incidence of an event of interest in the presence of competing risk events.
Analysis of cancer incidence rates was calculated using the indirect standardisation method. Kaplan-Meier and competing risk analysis were used to analyze the incidence of cancer during the follow-up.
Neoplasm incidence rates are higher after kidney transplantation compared with the general population. Kaplan-Meier methodology overestimates the incidence of cancer in kidney transplant recipients.

Keywords: Kidney transplant; Competing risks; Kaplan-Meier.

1 Background

The Kaplan-Meier method is commonly used to estimate the incidence of an event over time. It assumes independence between the event of interest and any competing event that precludes the event of interest to occur.
Often times, a patient may experience an event other than the one of interest which alters the probability of experiencing the event of interest. Such events are known as competing risk events. Any subject who does not experience the event of interest can be treated as censored. However, a patient experiencing a competing risk event is censored in an informative manner. Hence, the Kaplan-Meier estimation procedure may not be directly

applicable. In this setting, it would often be of interest to calculate the cumulative incidence of a specific event of interest. The use of models that take into account the presence of competing risks will allow more precise estimates in this context (Aalen (1978), Prentice y Kalbfleish (1980), Gray (1988), Fine y Gray (1999)).

Cardiovascular illnesses and neoplasms are the two main causes of death with normal function of the graft in the long-term follow-up of patients who have received kidney transplants (Morales (2006), Campistol (2009)). The presence of neoplasms is a major threat and cause of morbidity in kidney transplant patients.

According to data published in other countries, the accumulated incidence of neoplasms can reach 20% 10 years from the transplant (Buell et al. (2005)) and nearly 30% after 20 years (Chapman y Webster (2004), Chapman y Webster (2004b)). The rate of expected cancers compared to those which are observed varies in the different registers. On average, it is estimated that the incidence of cancer in patients who have received kidney transplants is 3 times higher than that for the general population. By localizations, this ratio can reach an incidence rate of between 8 and 14 times more for kidney cancer in transplant patients, and an incidence of between 65 and 92 times more of non-melanoma skin tumours (Chapman y Webster (2004), Chapman y Webster (2004b), Kasiske et al (2004), Birkeland et al. (1995), Jensen et al (1999), Lindelof et al. (2000), Birkeland et al. (2000)). In the largest study on the incidence rates of malignancies among first-time recipients of deceased or living donor kidney transplantation (n = 35765) the rates for most malignancies are higher after kidney transplantation compared with the general population (Kasiske et al (2004)). Similar results were observed in studies from five national tumour registries in Denmark (Birkeland et al. (2000)), Finland (Kyllonen et al. (2000)), Sweden (Adami et al. (2003)), Australia (Vajdic et al. (2006)), and Canada (Villeneuve et al. (2007)) with a total sample size of 31,977 transplant recipients.

The aims of this study are: (i) to establish the incidence of cancer in recipients of renal transplants performed in a hospital in A Coruña (Spain) during the period 1981–2008 compared to that experienced by the general population, ii) to demonstrate the importance of apporpiately estimating the cumulative incidence of an event of interest in the presence of competing risk events.

2 Methods

An observational prospective follow-up study with a retrospective component, carried out in the health district of A Coruña (northwest Spain) during the period 1981–2008. During that period 2059 kidney transplants were performed in the University Hospital Complex of A Coruña, which corresponded to 1794 patients. Patients with pretransplant neoplasms were

excluded from the analysis (n = 91). A follow-up study was designed in order to estimate cancer incidence after kidney transplantation. This sample size would make it possible to detect relative risks of ≥ 1.2 estimating an exposed proportion of 35% and a proportion of censured observations of 40%, with a security of 95% and a statistical power of 80%.

The methodology of this study was described previously (Pita-Fernández et al. (2009)).

Incident cancer is considered as new cases of cancer which occur after the transplant and which have anatomopathological confirmation. Their localization is classified according to the International Classification of Diseases-9 (ICD-9).

Analysis of cancer incidence rates was calculated using the indirect standardisation method. Estimates of age-adjusted cancer incidence rates in the general population of Spain are obtained from the Carlos III Health Institute, the National Epidemiology Centre of Spain's Ministry of Science and Technology. Crude first, second and third-year post-transplantation cancer incidence rates are calculated for male and female recipients. The number of observed cases of cancer at each site is calculated from data in the clinical records. The expected number of cancers is calculated from data supplied by the Carlos III Health Institute. For each tumour location we estimate the standardized incidence ratios (SIRs) of cancer, using sex-specific cancer incidence rates, by dividing the incidence rate for the transplant patients by the rate of the general population. The 95% confidence intervals of the SIRs and their associated p-values are calculated by assuming that the observed cancers follow a Poisson distribution.

Competing risk survival analysis methods are applied to estimate the cumulative incidence of developing cancer over time from kidney transplantation. This method allows for the fact that a patient may experience an event which is different from the event of interest. These events are known as competing risk events, and may preclude the onset of the event of interest, or may modify the probability of the onset of the event of interest. In particular, a transplanted patient may die or lose the graft without developing any kind of cancer. In a Kaplan-Meier estimation approach, these persons would be treated as censored and would be eliminated from the risk set. This could lead to misleading results, as it is based on the assumption that censoring is "non-informative", meaning that a censored patient has the same risk of developing cancer as those who have complete follow-up. This is not the case in patients who die before without developing cancer, as they are no longer at risk. Occurrence of cancer is the event of interest. Any other event, such as death of graft failure, are considered competing events. Estimates of cumulative incidence functions are calculated based on the two-step process developed by Kalbfleisch and Prentice (2002). In the first step, we calculate the Kaplan-Meier estimate of the overall survival from any event. In the second step, the conditional probabilities of developing cancer are calculated. The cumulative incidences are estimated

using these probabilities.
Statistical analysis are performed by using the R statistical package (version 2.9. 2009; The R Foundation for Statistical Computing) and EPIDAT statistical software (version 3.1, 2006; Dirección Xeral de Saude Pública, Organización Panamericana de la Salud).

3 Results

Mean age is 46.2(SD=14.3) years, 62.8% are males. One hundred twenty nine patients were diagnosed of cancer during the follow-up period, the more frequent locations were: skin(non-melanoma), kidney and non-Hodking lynfoma. Comparing the observed to expected cancer incidence in the Spanish population, using standardized incidence ratios, a significant increase is observed in the incidence of cancer in transplant patients in non-melanoma skin cancers (SIR=19.59), kidney (SIR=24.07), breast (SIR=3.45) and non-Hodgkin lymphoma (SIR=8.7). In the other locations (except bladder), a non statistically significant increase of the cancer incidence was detected.
Five years after kidney transplantion, 4.05% of the patients presented a neoplasm, 26.32% lost the allograft, 9.08% died and 60.55% were alive. Ten years after transplantion, these figures were: 8.03% of the patients presented a neoplasm, 33.99% lost the allograft, 15.10% died and 42.88% were alive, respectively (Figure 1).
Differences in the estimated cancer cumulative incidence using the Kaplan-Meier method and the competing risk analysis are shown in Figure 2. Using the Kaplan-Meier method, the neoplasm incidence at 5 and 10 years were 5.72% and 12.48%, respectively. These values overestimate the incidence of cancer in the follow-up estimated with the competing risk methodology (4.05% and 8.03%, respectively).

4 Conclusions

Neoplasm incidence rates are higher after kidney transplantation compared with the general population. Kaplan-Meier methodology overestimates the incidence of cancer in kidney transplant recipients.

References

Aalen, O.O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *Annals Statistics*, **6**, 534-545.

Adami J., Gabel H., Lindelof B., et al. (2003). Cancer risk following organ transplantation: a nationwide cohort study in Sweden. *Br J Cancer*, **89(7)**, 1221-1227

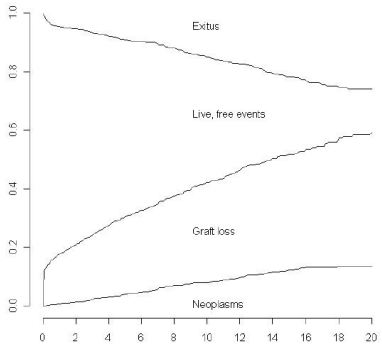


FIGURE 1. Evolution of kidney transplant recipients after transplantation: competing risk analysis.

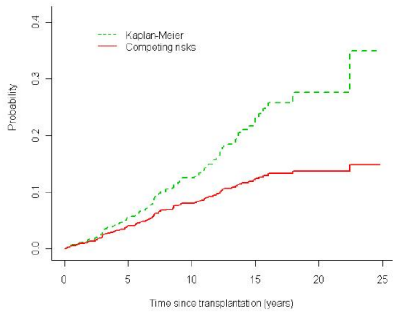


FIGURE 2. Cancer incidence in kidney transplant recipients after transplantation: Kaplan-Meier method vs. competing risk analysis.

Birkeland S.A., Storm H.H., Lamm L.U., et al. (1995). Cancer risk after renal transplantation in the Nordic countries, 1964 – 1986. *Int J Cancer*, **60(2)**, 183-189.

Birkeland S.A., Lokkegaard H., and Storm H.H. (2000). Cancer risk in patients on dialysis and after renal transplantation. *Lancet*, **355(9218)**, 1886-1887.

Buell J.F., Gross T.G., and Woodle E.S. (2005). Malignancy after transplantation. *Transplantation*, **80(2 Suppl)**, S254-264.

Campistol J.M. (2009). Minimizing the risk of posttransplant malignancy. *Transplantation*, **87(8 Suppl)**, S19-22.

Chapman J.R., and Webster A.C. (2004) *Cancer report. ANZ-DATA Registry 2004 Report*, Chapter 10, 99-103 `http://www.anzdata.org.au/anzdata/AnzdataReport/27thReport/files/Ch10Cancer.pdf`

Chapman J.R., and Webster A.C. (2004b). Cancer after renal transplantation: the next challenge. *Am J Transplant* **4(6)**, 841-842.

Fine J.P., and Gray R.J. (1999). A proportional hazards model for the subdistribution of a competing risk. *JASA*, **94**, 496-509.

Gray R.J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics* **16**, 1141-1154.

Jensen P., Hansen S., Moller B., et al. (1999). Skin cancer in kidney and heart transplant transplant recipients and different long-term immunosuppressive therapy regimens. *J Am Acad Dermato* **40(2 Pt 1)**, 177-186.

Kalbfleisch J.D., and Prentice R.L. (2002). *The statistical analysis of failure time data*. New York: Wiley, 163-188.

Kasiske B.L., Snyder J.J., Gilbertson D.T., and Wang C. (2004) Cancer after kidney transplantation in the United States. *Am J Transplant* **4(6)**, 905-913.

Kyllonen L., Salmela K., and Pukkala E. (2000). Cancer incidence in a kidneytransplanted population. *Transpl Int*, **13(Suppl 1)**, S394-398

Lindelof B., Sigurgeirsson B., Gabel H., and Stern R.S. (2000). Incidence of skin cancer in 5356 patients following organ transplantation. *Br J Dermatol* **143(3)**, 513-519.

Morales J.M. (2006). Neoplasias y trasplante *Nefrologia*, **26(2)**, 12-20.

Pita-Fernandez S., Valdes-Cañedo F., Pertega-Diaz S., et al. (2009). Cancer incidence in kidney transplant recipients: a study protocol. *BMC Cancer*, **9**, 294.

Prentice R., and Kalbfleish J. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley.

Vajdic C.M., McDonald S.P., McCredie M.R., et al. (2006). Cancer incidence before and after kidney transplantation. *JAMA*, **296(23)**, 2823-2831.

Villeneuve P.J., Schaubel D.E., Fenton S.S., et al. (2007). Cancer incidence among Canadian kidney transplant recipients. *Am J Transplant*, **7(4)**, 941-948.

Evaluating Change Detection in Data Streams

Gina-Maria Pomann[1], Tamraparni Dasu[2], Shankar Krishnan[2]

[1] gpomann@ncsu.edu, North Carolina State University, Raleigh, NC
[2] {tamr,krishnas}@research.att.com, AT&T Labs - Research, Florham Park, NJ

Abstract: Change detection algorithms for data streams typically return binary decisions of "Change" or "No Change". However, binary responses provide no additional information about the properties of an algorithm such as sensitivity to different types of changes, or stability with respect to small perturbations in the distribution. In this paper, we propose a general statistical framework, for the evaluation of change detection algorithms, based on an objective performance measure, *streaming power*. We model the change of distribution in data streams using a mixture model and vary the change to study the behavior of change detection algorithms. We demonstrate using simulated data examples.

Keywords: Change detection, data streams, hypothesis testing, nonparametric.

1 Evaluation of Change Detection Algorithms

Data streams are rapidly accumulating data sets that are used for real time decision making and thus pose computational challenges. Examples include telecommunications data, financial ticker streams, and network polling data. An important problem is determining distributional shifts within a data stream. Kifer et. al. (2004), Song et al.(2007), and Dasu et. al.(2009) are three popular change detection (*CD*) algorithms but their behavior is not well understood. No single algorithm can detect all types of changes with equal efficacy. Current evaluative methods are ad hoc, relying on relative benchmarking based on a handful of data sets, with no general framework of reference that reflects the true behavior of the algorithm.

The binary outcome from a single run does not provide any insight into the true behavior of a given *CD* algorithm. In this paper, we propose a rigorous, objective performance measure, *streaming power* (*SP*), to evaluate and identify desirable properties that an effective *CD* algorithm should have. In doing so, we provide the user with a framework that enables them to compare different algorithms and choose the one that best meets their needs.

Typically a *CD* algorithm $\mathcal{A}$ compares a reference distribution $F_0 \in \mathcal{F}$, where $\mathcal{F}$ is the space of distributions, with a test distribution, $F_1 \in \mathcal{F}$ and associates a binary response $I_{\mathcal{A}} \in \{0, 1\}$. While we only have access to the mapping $I_{\mathcal{A}} : \mathcal{F} \to \{0, 1\}$, in reality $\mathcal{A}$ maps $\mathcal{F}$ to $p \in [0, 1]$ where p is the probability of detecting change from F_0 to any $F_1 \in \mathcal{F}$.

1.1 Streaming Power

We adapt the notion of statistical power to build a framework for associating a probability of detecting change which we call *streaming power* (*SP*). This allows for the general comparison of *CD* algorithms for data streams by measuring the ability of an algorithm to discriminate between F_0 and F_1.

Consider a multi-dimensional data stream of observations, $\mathbf{X}_t = (x_1, ..., x_d)_t$ where d is the dimension and t indexes time. Assume that the *CD* algorithm under analysis uses the sliding window framework, as presented by Kifer et. al. (2004). A window refers to a contiguous segment of the data stream containing a specified number of data points n. The generating distribution of the data in each window W_t corresponds to some $F \in \mathcal{F}$, where t is the starting point of the window in the data stream. The distribution F of the data in W_t is compared to the distribution F_0, of the data in a reference window W_0. The size of the window is typically dependent on the data set. Initially, the window W_0 contains only samples from the reference distribution F_0. When the distribution changes, samples in the data stream are generated from a new distribution F_1. As the window slides over the data stream, we define the mixing proportion $\delta(t) \in [0, 1]$ to be the proportion of samples from F_1 in the current window, which we now denote $W_{\delta(t)}$.

We model the distribution in W_t as coming from the mixture distribution,

$$F_{\delta(t)} = (1 - \delta(t)) \cdot F_0 + \delta(t) \cdot F_1. \tag{1}$$

This is a natural model for the way change occurs in a data stream as illustrated in Figure 1. From here on we denote $\delta(t)$ as δ.

*The **streaming power** of an algorithm $\mathcal{A}$, with binary response $I_{\mathcal{A}}$, is the probability of detecting a change from F_0 to a F_1,*

$$S_{\mathcal{A}}(\delta) = P(I_{\mathcal{A}} = 1 | F_\delta), \tag{2}$$

where δ is the mixing proportion.

SP can be thought of as a temporal version of statistical power from the hypothesis testing context and is empirically estimated through commonly used simulation techniques, such as bootstrapping. The algorithm is used to compare N pairs of simulated windows $\{W'_0, W'_{\delta(t)}\}_{k=1}^{N}$, created by sampling with replacement from W_0 and $W_{\delta(t)}$. The resulting binary outcomes $\{I_{\mathcal{A}}^k\}_{k=1}^{N}$ are i.i.d. Bernoulli trials with probability $p_\delta := S_{\mathcal{A}}(\delta)$. The estimated *SP* is given by,

$$\hat{p_\delta} = \sum_{k=1}^{N} I_{\mathcal{A}}^k / N. \tag{3}$$

By the central limit theorem, $\hat{p_\delta} \sim AN(p_\delta, \frac{p_\delta(1-p_\delta)}{n})$. A high value of *SP* represents a strong ability to discriminate between the distributions. Figure 1 illustrates how $\hat{p_\delta}$ can be used to measure an algorithm's change detection ability. For a given algorithm and small values of δ, the *SP* should

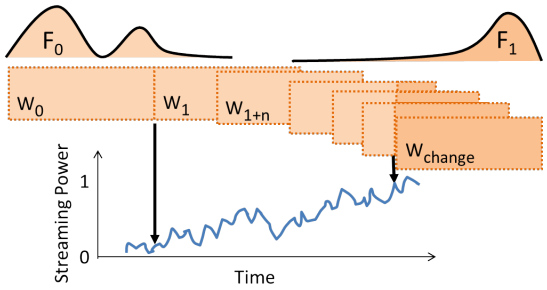


FIGURE 1. *SP* Framework: Initial reference window W_0 consists of samples from distribution, F_0 (lighter shade). As the stream advances, its distribution changes to F_1 (darker shade).

be low as shown by the first downward arrow in Figure 1. The estimated value of *SP* , $\hat{p_\delta}$ should increase with the mixing proportion δ for a good *CD* algorithm.

The ability of an algorithm to detect change varies by the type of change. Some algorithms focus on keeping down the false positive rate. In doing so, the algorithms allow for a broader interpretation of F_0, reducing the ability to discriminate when F_1 and F_0 are close. This, in turn, reduces its *SP* . On the other hand, some algorithms define F_0 too specifically and are sensitive to the slightest of perturbations, reducing their usefulness in any realistic setting.

In order to analyze the robustness of an algorithm $\mathcal{A}$, we define its ***sensitivity*** as

$$\eta_{\mathcal{A}}(\delta) := \frac{1}{\delta} \frac{d\, S_{\mathcal{A}}(\delta)}{d\, \delta}. \tag{4}$$

Intuitively, sensitivity measures how the *SP* of an algorithm increases in relation to its distance from the reference distribution. Ideally, we would like a *CD* algorithm to detect statistically significant change, but not small perturbations. That is, there should be no sudden increase in *SP*, especially when δ is very small. The example in Section 2 demonstrates this notion.

2 Simulation Results

In this section, we present analysis of three *CD* algorithms, the Rank tests (3 variants) of Kifer et. al. (2004) (only for 1D data), the KL test of Dasu et. al. (2009), and the Density test of Song et. al. (2007). To demonstrate the use of *SP* for their comparison, we implement the method in Section 1.1 on two sets of data. The *SP* and sensitivity of the two algorithms are computed as δ goes from 0 to 1. Figure 2(a) displays power curves of Rank (3 variants) (lines with markers), KL (solid line) and Density (dashed line) tests for the two parameter 1D-Gamma distribution. The X-axis represents the mixing proportion of F_0 ($\Gamma(0.5, 0.5)$) and F_1 ($\Gamma(0.5, 0.6)$). Figure 2(b) shows the

corresponding sensitivity curves. The sensitivity of each algorithm occurs at the maxima of this curve. In this example, the Rank methods yielded sensitivity values of 2.99, 2.87 and 3.3, while those of KL and Density were 4.08 and 4.16, respectively. Note that the Rank tests have lower sensitivity than the other two tests. Figure 2(c) shows power curves of KL (solid line) and Density (dashed line) tests where $F_0 = \mathcal{N}^3(0,1)$ (the 3D standard Gaussian) and $F_1 = (\mathcal{N}^2(0,1), \mathcal{N}(0.2,1))$. Figure 2(d) the corresponding sensitivity curves. The Density method becomes more sensitive (20.36), and hence less stable, in higher dimensions compared to the KL method (2.46) even though the contamination is only in one dimension.

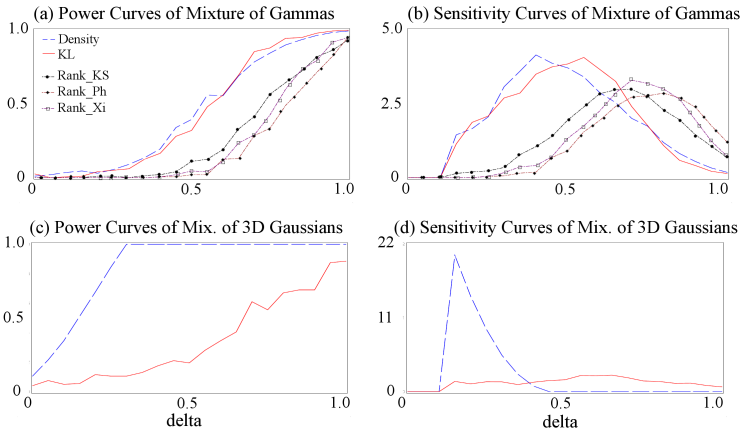


FIGURE 2. Simulation results

3 Conclusion

In this paper, we have proposed a novel framework for evaluating change detection algorithms for multidimensional data streams called *streaming power*. We model the change in data streams using a mixture model, and explore the sensitivity of algorithms by varying the amount of change. Sensitivity has implications for the robustness of a *CD* algorithm, which we will explore in future work.

References

T. Dasu, S. Krishnan, D. Lin, S. Venkatasubramanian, and K.Yi. (2009)., Change (Detection) you can believe in: Finding distributional shifts in data streams. In: *8th International Symposium on Intelligent Data Analysis*. 21-34.

D. Kifer, S. Ben-David, and J. Gehrke. (2004). Detecting changes in data streams. In: *Proceedingsof the 30th International Conference on Very Large Databases*. 180-191.

X. Song, M. Wu, C. Jermaine, and S. Ranka (2007). Statistical Change Detection for Multi-Dimensional Data. In *13th International Conf. on Knowledge Discovery and Data Mining*. 667-676.

Modelling the Timing of Marital Dissolution in Italy: censored quantile regression with additive terms

Mariano Porcu[1], Vito M. R. Muggeo[2], Vincenza Capursi[2]

[1] Dipartimento Ricerche Economiche e Sociali, Università di Cagliari, ITALY
email: mrporcu@unica.it
[2] Dipartimento Scienze Statistiche e Matematiche 'S. Vianelli', Università di Palermo, ITALY

Abstract: The analysis of marital dissolution in Italy represents a quite interesting and challenging topic from a substantive standpoint; in fact, despite of the decreasing number of marriages and the increasing number of divorces, the traditional family based on the marriage of heterosexual partners is still considered as a fundamental institution of the society. Here we present a censored quantile regression model with additive terms to investigate the determinants of the timing of marital dissolution on a large and substantial sample from a survey carried on in Italy.

Keywords: censored quantile regression; Timing of Marital Dissolution; Survival data; Smoothing

1 Introduction

It is commonly asserted that the family is the *fundamental institution* of the Italian society. The main consequence of this assertion is a widespread political support addressed to the upholding of the classical family built on the marriage of heterosexual partners. In this context the analysis of the possible determinants of the marital dissolution is a largely debated issue: the study of the factors that could affect the end of the marriage is prominent in the social research. The topic is also of great interest for the policy-makers as the marital dissolution affects some of the key features of the modern societies, such as economy, gender equality and especially fertility. Although the study of the time-to-separation can provide quite useful information and insights to evaluate trends and changes in the formation and dissolution of the marriage, relatively few studies take explicitly into account the time dimension, see Cavanagh and Huston (2008), Gottman and Levenson (2000).
Using a large sample surveyed by the Italian national statistics institute (ISTAT), we aim to model the time-to-marital dissolution in a regression

quantile framework. While the Cox model represents the most used framework to model survival data, censored quantile regression (CQR) offers a more flexible alternative by focusing the attention on narrow slices, lower or upper tails, of the conditional survival distribution of interest (Koenker, 2008).

2 The Data

The data considered in this paper come by from the sample survey on *Families and Social Subjects* (FSS), carried on in Italy by the official statistics institute at the end of 2003 on a sample of over 19,000 Italian families (nearly 50,000 individuals). The survey was addressed to collect broad information on the Italian households, such as the shapes, the network of kinship, the relations among partners, the permanence of young adults in the family, and the working life.

TABLE 1. Some descriptive statistics on data analysed.

	Males ($n = 4633$)		Females ($n = 5235$)	
Covariates	Separated	Non-Separ	Separated	Non-Separ
Area (*obs.*)				
North	335	1640	445	1763
Center	190	1164	223	1269
South	124	1180	151	1384
Education (*obs.*)				
1^{st} stage basic	75	591	63	684
2^{nd} stage basic	253	1529	268	1586
Upper secondary	232	1473	384	1713
Degree	89	391	104	433
Age at Marriage (*Years*)				
Mean (sd)	*26.3(5.7)*	*27.7 (5.8)*	*23.3(5.0)*	*22.2(5.7)*
Childless				
%	*33.4*	*12.8*	*32.1*	*12.9*
Work				
Yes at marr (%)	*80.1*	*86.7*	*51.2*	*45.0*
Yes at separ (%)	*88.0*	—	*64.7*	—

The data from the 2003 FSS survey here analyzed represent the most recent information available on the topic: the data from the last FSS survey carried out in 2010 are not yet available. We have omitted from the sample persons married before the 1970 when the divorce, understood as the 'total dissolution of marital status', was not allowed. Moreover, to keep away from any potential confounding interaction between sex, covariates and timing we have considered females and males independently (e.g., Schoen and Canudas-Romo, 2006). Table 1 summarizes some descriptive statistics for the sample.

3 Methods and Results

We aim to model the time-of-dissolution as a function of the following covariates in a CQR model: the categorical variables Education, Area, and Work at Marriage, and the numerical variables Age at Marriage, Number of Sons and Year of Birth. Until now, CQR has been discussed only with parametric linear terms, however for the aforementioned continuous covariates the linearity assumption is not tenable and more flexible alternatives are requested: we use B-spline bases with quadratic penalties on the coefficients to get smooth estimates of the nonlinear relationships. The additive CQR model with J nonparametric terms for the variables z and linear terms for the variables x, may be written as

$$Q_\tau(Y|x_i) = x_i^T \beta_\tau + \sum_j^J f_{\tau j}(z_{ij}), \tag{1}$$

where the subscript τ points the percentile of interest ($0 < \tau < 1$). Notice that, unlike the usual model for the conditional mean, here the covariate effect (parametric or nonparametric) depends on the percentile τ. The response Y measures the time span of their marriage up to the year of separation; we consider as uncensored the spouses (male or female) which stop living together regardless of the possibility of reconciliation; in fact in Italy only a slight proportion of separations ends with a reuniting of the couple (Castiglioni, 2008). We modify the iterative estimating algorithm described in Bottai and Zhang (2010) to include the additive (spline) terms in the linear predictor and to obtain parameter estimate of the additive CQR model (1). Although QR allows to model every quantile of the response conditional distribution, our analysis focuses on the lowest quantiles ($\tau \leq 0.10$). Indeed, early dissolutions (i.e. the left tail of the survival distribution) are of major interest in the present study since the first years of marriage are known to be crucial for fertility, children social development, changes in lifestyle and also for their influence on the probability of remarriage.

TABLE 2. Point estimates for the parameters of the linear terms in the four CQR models.

	Males		Females	
Linear Terms	$\tau = 0.05$	$\tau = 0.10$	$\tau = 0.05$	$\tau = 0.10$
EDUC (2 stage basic vs 1 stage basic)	-2.257	-1.666	-2.190	-3.971
EDUC (upper sec. vs. 1 stage basic)	-1.372	-1.138	-3.444	-5.668
EDUC (degree vs 1 stage basic)	-2.291	-2.474	-4.454	-6.503
Area (center vs. north)	0.197	1.081	0.770	1.116
Area (south vs. north)	2.319	2.730	1.317	1.486
Work at marr (yes vs. no)	1.500	2.157	0.469	0.130

For the four fitted additive CQR models (two quantiles 0.05 and 0.10 for males and females), Table 2 shows the point estimates for the parameters

of the linear terms and Figure 1 reports the fitted smooth effects of the three continuous covariates.

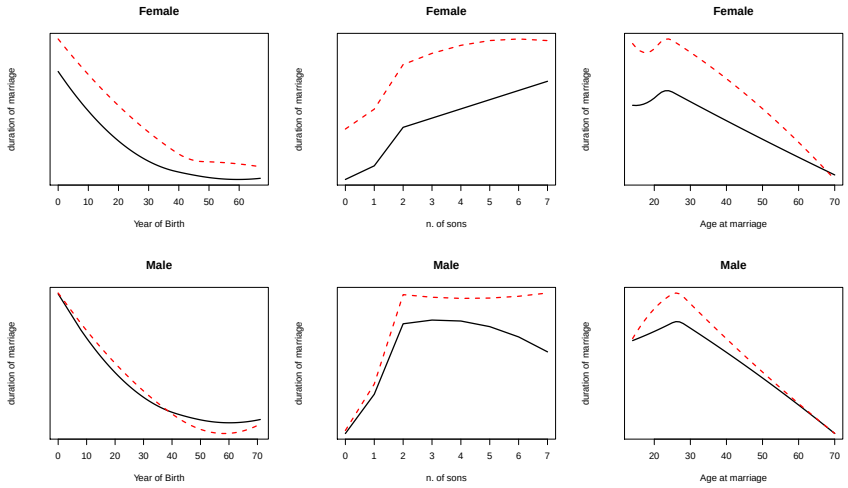


FIGURE 1. Fitted quantiles ($\tau = 0.05$, continuous line; $\tau = 0.10$ dashed line) for males and females.

In short, for the *early* marital dissolutions (i.e., the low percentiles 5% and 10%) we observe strong and somewhat expected effects of the 'area' and of the 'educational level' for both male and female groups. On the other hand, the effect of the 'working status' is somewhat different. The plots in Figure 1 emphasize the nonlinear effects of the continuous covariates, by highlighting a different role of the number of sons variable.

References

Bottai, M., Zhang, J. (2010) Laplace regression with censored data *Biometrical Journal*, **52**, 487-503.

Castiglioni, M., Dalla Zuanna, G. (2008) A Marriage-Cohort Analysis of Legal Separations in Italy. *Population*, **63**, 1, 173-193.

Cavanagh, S.E., Huston, A.C. (2008) The Timing of Family Instability and Children's Social Development. *Journal of Marriage and Family*, **70**, 1258-1269.

Gottman, J.M., Levenson, R.W. (2000) The Timing of Divorce: Predicting When a Couple Will Divorce Over a 14-Year Period. *Journal of Marriage and Family*, **62**, 737-745.

Koenker, R.W. (2008) Censored Quantile Regression Redux *Journal of Statistical Software*, **27**, 6 , 1-25.

Estimation of the density of the Antarctic Blue whales population using their sequences of sounds

Rocío Prieto González [12], María Cruz Valsero Blanco [1], Flore Samaran [3], Olivier Adam [2]

[1] Universidad de Valladolid, Departamento de Estadística e Investigación Operativa, 47005 Valladolid Spain. (e-mail: mcruz@eio.uva.es)
[2] Centre de Neurosciences de Paris Sud, CNPS-CNRS UMR8195 Université Paris 11, 91405 Orsay Cedex France. (e-mail: rocio.prieto-gonzalez@u-psud.fr) (e-mail: olivier.adam@u-psud.fr)
[3] Centre d'Études Biologiques de Chizé, CEBC-CNRS UPR1934, Villiers-en-Bois, France. (e-mail: flore.samaran@cebc.cnrs.fr)

Abstract: Method is developed for estimating the size or the density of cetacean populations using data from a fixed passive acoustic sensor. We must link the number of sounds detected with the number of individuals of the group. Nowadays research follows two ways: estimating the population density in a given area, and to estimate the recurrence the songs are produced. We introduce a Poisson process with two parameters, one spatial that deals with the density of whales in an area and the other one temporal which measures the intensity of the number of sequences of sounds detected by a whale in an interval of time. We suppose that every whale acts independent of each other. With these hypotheses, we develop a distribution spatio-temporal associated with the process and we use it to calculate the likelihood functions in order to find a maximum that would produce estimators of both parameters spatial and temporal. This distribution is used to fit the experimental data: The intervals of time between two consecutive sequences of calls. We specify sequences as we are using not each cue detected, but the whole sequence of song produced by an individual. This method is a preliminary study and it is potentially applicable other species.
Method is illustrated with a case study: To estimate the density of Antarctic blue Whales (BMi) population around the Crozet Islands, in the Austral Ocean, using the sounds detected from April 2003 to March 2004. Our data come from the records of only one fixed hydrophone. These records of the sound produced by Antarctic Blue Whales must be processed to produce a multiple count time-series.

Keywords: Population density; Poisson model; Passive acoustic; Cetaceans.

1 Introduction

Method of population assessment based on acoustic techniques is applied to the study of Antarctic Blue whales using a hydrophone located off the

Crozet Archipelago, in the southern Indian Ocean. An area where so few Blue whales have been seen that their density has not been previously estimated. This information is vital from the point of view of marine biology/ecology because the Blue whale is a specie endangered since 1967. It is estimated that its population has been reduced to 0.15% of the initial population and nowadays there are about 3000 individuals in all the oceans.
The object of this paper is to increase the repertoire of tools available for making species assessments. We present an initial Poisson process with two parameters. The first one is spatial, it measures the density of whales in an area and the latter is temporal that measures the intensity of sounds sequences detected by a whale in an interval of time. Also we suppose independence in the way, two whales sing. Then we develop a distribution spatio-temporal associated with the process, using the intervals of time between two consecutive sequences of calls. We use it to calculate the likelihood function in order to find a maximum that would produce estimators of the parameters for obtaining estimates of population stocks abundance. This paper presents a framework for estimating cetacean density from fixed passive acoustic detectors. It is general enough that it might be used under considerably different scenarios, with appropriate modifications that are also discussed.

2 Materials and methods

The data set used in this study was recorded from April 2003 to March 2004 at a station located in the South-western Indian Ocean (Crozet Islands - 46°51'S-51°53'E). The station is moored in the International Monitoring System (IMS) and support the Comprehensive Nuclear Test-Ban Treaty (CTBT). These hydrophone systems were designed to control nuclear tests in the ocean, but the recordings also contain some natural sounds produced by whales. So although the main objective was not register whales sounds, as the filter used is a low frequency one, the sounds of Blue whales were detected as a secondary product.
There were two arrays of three instruments each one located in the Northern and Southern coasts of Possession Island. The instruments were deployed on the seafloor at a depth between 1100 and 1500 meters in a triangular configuration (triad) with approximately 2 km spacing. The two arrays were located on opposite sides of the island and spaced 60 km apart. The hydrophones were suspended near the sound channel axis (SOFAR) at a depth of approximately 300 m. In this area of the ocean, the speed of propagation of sound waves is minimal, making that the ocean behaves at these depths as a wave guide. The hydrophones monitored sound continuously, 24 hours a day, 7 days a week.
The acoustic data from each hydrophone were analyzed to check for the presence of calls typically associated with Antarctic and Pygmy Blue whales.

An automatic detection method for both call types was designed (Samaran et al. 2008). The detector used a matched filter process that related the acoustic data with synthetic waveforms (templates) defined for both blue whale subspecies' calls (McDonals et al. 2006).
On this paper we analyzed the data of BMi calls (Antarctic Blue whales) of only one hydrophone located in the North of the island which covers a recording surface of 47123 m^2. We have data of one year. Although the hydrophone registered continuously, we have a gap of data due to technical incidents. We modify the data in order to obtain the time between two consecutive sequences of calls. We have used the R language to apply the initial model to these data and estimate the density of Antarctic blue whales in the area.

3 The model

Assessing the size of cetacean populations in the open ocean has traditionally relied on visual surveys alone. The addition of acoustic monitoring can complement these surveys. Nowadays we do not known to assign an individual acoustic signature in the case of whales and the only method for estimating the size of cetacean populations that has been adapted to the acoustic is Distance Sampling (Thomas et al. 2002). This method use a punctual estimation of the density adding a probability of detection based on the distribution of observed detection distances. It requires additional information like the probability of detecting cues and the rate at which animals produce it. To obtain estimation of this rate, (Marques et al. 2009) proposed to use specific tag including sensors attached on the back of the beaked whales. However, it is not possible to generalize this approach for different cases, especially when the individuals are not reachable. In this case visual observation was not possible and the knowledge of these parameters neither. So in this paper an initial model is presented to estimate Antarctic blue whale population based on near-continuous recording from a single hydrophone as a new alternative method.

Let $T(s)$: the observed time between two consecutive sequences of calls produced into the surface s

$N(t,s)$: be the number of calls in s in a interval of time t

$B(s)$: number of whales in the area s

$C(t)$: be the number of registered calls for whale in the time interval t

We model $B(s)$ and $C(t)$, unobserved, as a Poisson process of intensity λ and μ, so $B(s) \sim \mathcal{P}(\lambda s)$ and $C(t) \sim \mathcal{P}(\mu t)$.

The distribution function is given by $F_T(t) = 1 - P\Big(T(s) > t\Big)$., If the time between two detected sons is more than t, means that in the range t

have not had any song, then:

$$P\Big(T(s) > t\Big) = P\Big(N(t,s) = 0\Big) = \sum_{i=0}^{\infty} P\Big(N(t,s) = 0/B(s) = i\Big) P\Big(B(s) = i\Big)$$
$$= \sum_{i=0}^{\infty} P\Big(\big(C_0(t) + \cdots + C_i(t)\big) = 0\Big) P\Big(B(s) = i\Big)$$
$$= \sum_{i=0}^{\infty} e^{-\mu i t} e^{-\lambda s} \frac{(\lambda s)^i}{i!} = e^{-\lambda s} \sum_{i=0}^{\infty} \frac{(\lambda s\ e^{-\mu t})^i}{i!} = e^{-\lambda s}\ e^{\lambda s\ e^{-\mu t}}$$

Therefore its distribution and density function are:

$$F_T(t) == 1 - e^{\lambda s\,(e^{-\mu t} - 1)} \qquad f_T(t) = \mu\lambda s\, e^{-\lambda s} e^{-\mu t} e^{\lambda s\, e^{-\mu t}}$$

Suppose we have a sample of size m of the time between songs, so we have the log-likelihood function.

$$\ell_T(\lambda, \mu) = \lambda s \sum_{j=1}^{m} e^{-\mu t_j} - \mu \sum_{j=1}^{m} t_j + \ln\big(\mu\lambda s\big)^m - m\lambda s$$

4 Results

We have processed the data using R software.
First of all we have represented the number of calls observed. Through the model created we have estimated $\widehat{\lambda}_{EMV}$ and $\widehat{\mu}_{EMV}$, as the values that maximized the likelihood function. Our goal is to estimate the density of population of Antarctic Blue whales in the vicinity of the Crozet island, ie, the number of whales $B(s)$, in the area. Then we have calculated the confidence intervals of 90% y 95% of the number of Antarctic Blue whales present in the Archipelagos. We analyze if the data set shows some seasonality, (hypothesis supported by studies carried out earlier) but the lack of recording hours in our data make it no visible enough.
To sum up our model have estimate a number close to 5 individuals, which corresponds to the expectations of biologists.

References

T.A. Marques and L. Thomas and N. Ward and N. DiMarzio and P.L Tyack (2009). *Estimating cetacean population density using Fixed passive acoustic sensors: An example with Blainville s beaked whales.* Journal of the Acoustical Society of America, **125**, 1982-1994.

M.A. McDonald and S.L. Mesnick and J.A. Hildebrand (2006). *Biogeographic characterization of blue whale song worldwide: using song to identify populations.* Press, Oxford. JCRM, **8**, 55-65.

F. Samaran and O. Adam, and J.F. Motsch and C. Guinet (2008). *Definition of the Antarctic and Pygmy Blue whale call templates. Application to fast automatic detection.* Canadian Acoustic, **36**, 93-102.

H. M. Taylor, S. Karlin (1998). *An introduction to Stochastics Modeling.* Academic Press, third edition, **32**, 146-154.

Optimal DNA Pooling for the Detection of Single Nucleotide Polymorphisms

David M. Ramsey[1], Andreas Futschik[2]

[1] Department of Mathematics & Statistics, University of Limerick, Limerick, IRELAND, e-mail: david.ramsey@ul.ie
[2] Department of Statistics, University of Vienna

Abstract: We consider the optimal pooling of DNA to detect single nucleotide polymorphisms (SNPs), sites along the genome at which a population shows variation. The focus is on the detection of low frequency variants. Pooling individuals increases the probability that a rare variant appears in the sample. However, as the pool size increases, the mean number of reads from an individual decreases, making it harder to distinguish reads of a rare variant from errors. A hypothesis test for the detection of SNPs is defined. On the basis of this test, we determine the asymptotically optimal pool size given the parameters of the genome sequencer used, the number of lanes available and a specified significance level.

Keywords: genome sequencing; optimal pooling; single nucleotide polymorphisms.

1 Introduction

The genome consists of sequences made of 4 nucleotides (bases). At a majority of the sites in these sequences, each individual in a population has the same base. A site where there is variation is called a single nucleotide polymorphism (SNP). At such sites, in general, just two of the four bases appear. These variants are called alleles, the most common (rare) is termed the major allele (minor allele, respectively). We treat chromosomes, rather than members of a species, as individuals. However, our analysis can be generalized.

Since any reasonable test detects alleles of relatively large frequency with power close to 1, we concentrate on the detection of low frequency alleles. Following Futschik and Schlötterer (2010), one may use the following test: accept that there is a minor allele if in any lane the number of reads for a non-major allele exceeds a given threshold. We develop their work by specifying this threshold given the parameters of the sequencer and significance level required. An estimate of the power of this test is derived, which is used to find the optimal pool size for detecting low frequency alleles. For more on the practical issues involved in gene pooling see Kenny *et al.* (2010).

2 Description of the Problem and a Simplified Model

Genome sequencers read DNA from a pool (of m individuals) placed in a lane. Suppose we have k independent pools, i.e. the sample size is $n = km$. Consider a given site. Each lane gives a random number of reads for that site. If the same (large) amount of genetic material is taken from each individual, we may assume that the number of reads from an individual given that there are r reads in a lane has a binomial distribution with parameters r and $1/m$. Assume that each read is incorrect with a small probability ϵ, independently of other reads. Also, suppose that only two alleles are possible, the major allele and the putative minor allele.
Let $\mathbf{R} = (R_1, R_2, \ldots, R_k)$, where R_i is the total number of reads for that site in lane i. It is assumed that the R_i are i.i.d. from the Poisson(λ) distribution. In addition, suppose good estimates of λ and ϵ are available for the gene sequencer used.
The major allele is inferred to be the one with the largest number of reads in the whole sample. As we are interested in detecting low frequency alleles, we may assume that for reasonable sample sizes the major allele is correctly identified with probability 1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_k)$, where X_i is the number of reads of the putative minor allele in lane i.
Denote the minor allele frequency at a given locus by p. We wish to define an optimal pooling procedure (maximizing power) while controlling the type I error rate for a test of the following hypotheses.
H_0: The locus is not a SNP, i.e. $p = 0$.
H_A: $p = p_0$, where p_0 is some small positive value.

3 A Test for the Presence of a Minor Allele

Consider the test statistic $U = \max_{1 \le i \le k} X_i$, i.e. U is the maximum number of reads of a putative minor allele in a lane. Hence, under H_0, U is the maximum of independent observations from the Poisson($\lambda\epsilon$) distribution. The critical value for the test, u_k, is the smallest integer satisfying

$$P(U \le u_k | H_0) \ge 1-\alpha \Rightarrow P(X_i \le u_k | H_0)^k \ge 1-\alpha \Rightarrow P(X_i \le u_k | H_0) \ge \sqrt[k]{1-\alpha}.$$

Thus we can take the $\sqrt[k]{1-\alpha}$ quantile of the Poisson($\lambda\epsilon$) distribution as the critical value. We reject H_0 if and only if $U > u_k$. Note that this procedure takes into account the fact that we essentially have a multiple testing problem based on k test statistics $X_1, X_2, \ldots, X_k$. The critical value used in the test can be approximated using the Bonferroni procedure. However, this test does not take into account that such a procedure is repeated for each site. Hence, the value of α chosen should reflect this.
Under H_A, the number of minor alleles in the sample has a Bin(n, p_0) distribution. This can be approximated by the Poisson(np_0) distribution.

Result. *When there are b individuals with the minor allele, the distribution of the test statistic stochastically dominates the distribution of this statistic when one individual with the minor allele appears in each of b lanes.*
Let D denote the event that H_A is accepted given that it is true. Let the number of individuals with the minor allele in the sample be B and $\mu = E[B] = mkp_0$. We obtain

$$P[D] = \sum_{b=0}^{\infty} P[D|B=b]P(B=b) \geq \sum_{b=1}^{\infty} P[D|B=b]P(B=b).$$

Since $P[D|B=0] \leq \alpha$, we can treat the resulting bound as a good approximation of $P[D]$. For $b \geq 1$,

$$P[D|B=b] = P(U > u_c|B=b) = 1 - P(U \leq u_k|B=b) \geq 1 - P(V_1 \leq u_k)^b,$$

where V_1 is the number of correct reads from one individual. If p_0 is small enough to neglect the possibility of two individuals with the minor allele being in a pool, it follows that $P[D|B=b] \approx 1 - q_k^b$, where

$$q_k = \sum_{j=0}^{u_k} \frac{e^{-\lambda/m}(\lambda/m)^j}{j!}.$$

Hence,

$$P[D] \approx \sum_{b=1}^{\infty} \frac{e^{-\mu}\mu^b[1-q_k^b]}{b!} = 1 - e^{-\mu(1-q_k)}.$$

Since the exponent in this expression is linear in p_0, the asymptotically (as $p_0 \to 0$) optimal pool size is independent of p_0.

4 Results from Simulations

Simulations were carried out for each of the following models:

1. Mistakes from reading the major allele always resulted in observing the same allele (the minor allele, if one was present). Mistakes in reading the minor allele always resulted in observing the major allele.

2. Mistakes from reading an allele always resulted in observing the same allele (neither the major allele nor the minor allele, if one was present).

3. Mistakes from reading any allele gave the other three possibilities with equal probability.

It should be noted that Model 1 corresponds to the model described in Section 2. Under Models 2 and 3, more than two alleles can be observed at a site. In these cases, as before, the major allele is assumed to be the

allele with the largest number of reads in the whole sample. The putative minor allele is taken to be the non-major allele with the largest number of reads from a single lane. Note that it is possible to correctly reject H_0, but incorrectly infer which base is the minor allele. For such an error to occur, it is necessary for the number of errors in a lane to exceed both the threshold and the number of reads of the real minor allele. Hence, the probability of such an error is less than α. Tables 1-3 give results based on 10,000 simulations in each case. In each case $p = 0.01$ and $\alpha = 0.001$. It can be seen that the optimal pool size and empirical power are robust to deviations from the assumptions of the model.

TABLE 1. Optimal pool sizes (derived by simulation), theoretical and estimated power under Model 1. The power estimated by simulation is given in brackets.

	$k = 16$	$k = 40$	$k = 80$	$k = 120$
ϵ=0.01	4, 0.3752 (0.3864)	3, 0.6145 (0.6252)	3, 0.8514 (0.8489)	3, 0.9427 (0.9425)
ϵ=0.005	4, 0.3752 (0.3825)	4, 0.6915 (0.6973)	4, 0.9048 (0.9065)	4, 0.9706 (0.9728)
ϵ=0.002	6, 0.4628 (0.4733)	6, 0.7885 (0.7995)	7, 0.9553 (0.9583)	4, 0.9706 (0.9707)
ϵ=0.001	7, 0.4628 (0.4678)	7, 0.7885 (0.7946)	6, 0.9553 (0.9581)	6, 0.9905 (0.9917)

TABLE 2. Optimal pool sizes, estimated power and the probability of wrongly determining the minor allele (given in brackets) under Model 2.

	$k = 16$	$k = 40$	$k = 80$	$k = 120$
ϵ=0.01	4, 0.3696 (0.0006)	3, 0.6167 (0.0002)	3, 0.8504 (0.0000)	3, 0.9388 (0.0000)
ϵ=0.005	4, 0.3729 (0.0001)	4, 0.6948 (0.0001)	4, 0.9060 (0.0001)	4, 0.9728 (0.0000)
ϵ=0.002	6, 0.4723 (0.0001)	6, 0.7917 (0.0001)	6, 0.9588 (0.0001)	4, 0.9712 (0.0000)
ϵ=0.001	6, 0.4699 (0.0000)	6, 0.7883 (0.0000)	5, 0.9535 (0.0001)	6, 0.9907 (0.0000)

TABLE 3. Optimal pool sizes, estimated power and the probability of wrongly determining the minor allele (given in brackets) under Model 3.

	$k = 16$	$k = 40$	$k = 80$	$k = 120$
ϵ=0.01	4, 0.3752 (0.0000)	3, 0.6133 (0.0000)	3, 0.8549 (0.0000)	3, 0.9421 (0.0000)
ϵ=0.005	4, 0.3766 (0.0000)	4, 0.6993 (0.0000)	4, 0.9056 (0.0000)	4, 0.9703 (0.0000)
ϵ=0.002	6, 0.4694 (0.0001)	6, 0.7923 (0.0000)	7, 0.9575 (0.0000)	4, 0.9712 (0.0000)
ϵ=0.001	6, 0.4711 (0.0000)	6, 0.7942 (0.0000)	6, 0.9546 (0.0000)	6, 0.9922 (0.0000)

Acknowledgments: D. M. Ramsey is grateful for the support of Science Foundation Ireland under the BIO-SI project (no. 07MI012)

References

Futschik A. and Schlötterer C. (2010). Massively parallel sequencing of pooled DNA sample - the next generation of molecular markers *Genetics*, **186**, 207-218.

Kenny E. M., Cormican P., Gilks W. P., Gates A. S., O'Dushlaine C. T., Pinto C., Corvin A. P., Gill M., Morris D. W. (2010) Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Research* , doi: 10.1093/dnares/dsq029

Modelling seasonal patterns in longitudinal profiles with correlated circular random walks

Andrea Riebler[1], Leonhard Held[1] , Håvard Rue[2]

[1] Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland;
Email: andrea.riebler@ifspm.uzh.ch, leonhard.held@ifspm.uzh.ch

[2] Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway; Email: havard.rue@math.ntnu.no

Abstract: Seasonal patterns, as they occur in time series of infectious disease surveillance counts, are frequently modelled using a superposition of sine and cosine functions. However, in some cases this might be too simple. We propose the use of circular second order random walks instead and extend this approach to multivariate time series of counts. A correlated Gaussian Markov random field approach combines a uniform correlation matrix with a circular random walk to allow the seasonal pattern to be similar across regions, say, but not identical. Thus, spatially-varying disease onsets may be accounted for. The methodology is applied to weekly number of deaths from influenza and pneumonia in nine major regions of the USA.

Keywords: circular random walk; infections disease surveillance; INLA; Kronecker product; multivariate time series of counts.

1 Introduction

Time-series of infectious disease counts are marked by occasional outbreaks, but furthermore there are frequently seasonal variations, for instance harder strikes in winter than summer. To model seasonal variation a superposition of sine and cosine functions is often used, where the amplitudes can be described by a fixed coefficient or, to be more flexible, by smoothly time-varying coefficients, see for example Harvey and Koopman (1993), Eilers et al. (2008), Paul et al. (2008) or Fanshawe et al. (2008). However, in some cases this approach might be too simplistic and specific seasonal variations, for example sharp peaks around Christmas, might not be captured (Harvey and Koopman, 1993). Circular random walks (CRWs) are similar in spirit to periodic splines (see Harvey and Koopman, 1993) and represent a flexible alternative to adequately capture seasonal variations. In a multivariate setting, where different regions, say, show a similar seasonal pattern which is, however, likely to vary across regions, we propose the use of correlated CRWs. Analyses are performed using integrated nested

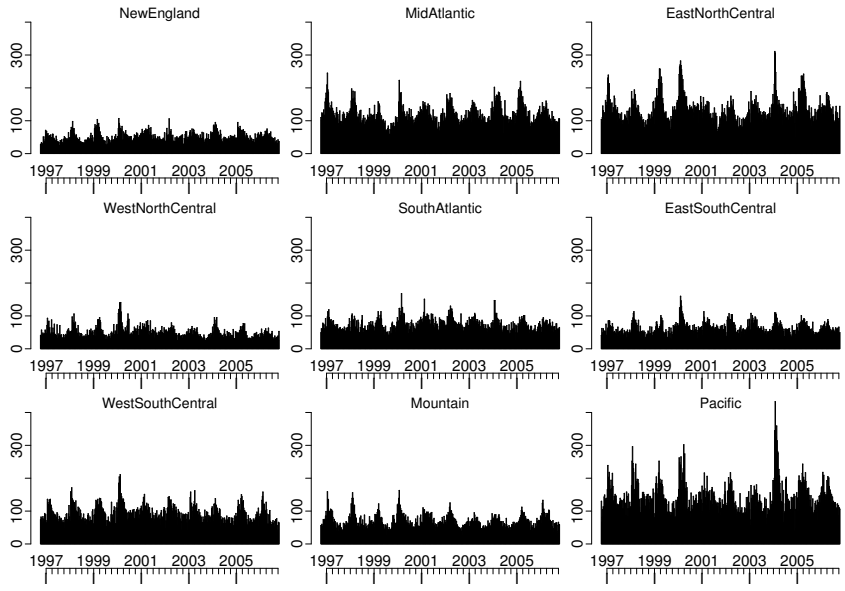


FIGURE 1. Weekly number of deaths from influenza and pneumonia in the USA from 40/1996 - 39/2006.

Laplace approximations (INLAs), see `www.r-inla.org`, which is a fast deterministic alternative to MCMC for latent Gaussian random field models (Rue et al., 2009). We apply the methodology to weekly number of deaths from influenza and pneumonia in the USA, previously analysed by Paul et al. (2008). Using the deviance information criterion (DIC) we compare the correlated approach with a model using independent CRWs for each region and a model assuming a common seasonal pattern across all regions.

2 Weekly data of influenza in nine regions of the USA

Weekly data on the number of deaths from influenza and pneumonia are provided for the weeks 40/1996 to 39/2006 in nine major geographic regions of the USA, see Figure 1. Region-specific population counts are not available for all years. Thus, we used the population counts derived from a census in the year 2000 in our analysis.

Let y_{tr} denote the number of deaths at time point t in region r, $r = 1, \ldots, R$, with $R = 9$. In our application, time is divided into weeks from 40/1996 to 39/2006, so that $t = 1, \ldots, 520$. We adopt a Poisson model with mean $n_r \lambda_{tr}$, where n_r denotes the population counts in region r (in the year

2000). To adequately model the seasonal pattern in a general and flexible way we use a CRW of second order (CRW2) for the 52 weeks. The precision matrix of a CRW2 is given by

$$\mathbf{R}^{CRW2} = \kappa \begin{pmatrix} 6 & -4 & 1 & 0 & \cdots & 0 & 1 & -4 \\ -4 & 6 & -4 & 1 & 0 & \cdots & 0 & 1 \\ 1 & -4 & 6 & -4 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -4 & 6 & -4 & 1 \\ 1 & 0 & \cdots & 0 & 1 & -4 & 6 & -4 \\ -4 & 1 & 0 & \cdots & 0 & 1 & -4 & 6 \end{pmatrix}, \quad (1)$$

with unknown precision parameter κ. As for all circulant matrices, only one column or row is sufficient to derive the whole structure matrix (Rue and Held, 2005, Section 2.6.1). To allow for similar but not equal seasonal patterns across the nine regions, we correlate the single CRW2s using the precision matrix $\mathbf{P} = \mathbf{C}^{-1} \otimes \mathbf{R}^{CRW2}$. Here, $\mathbf{C}^{-1}$ is the inverse of a 9×9 uniform correlation matrix $\mathbf{C} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$, where ρ denotes the unknown correlation parameter, $\mathbf{I}$ the identity matrix, $\mathbf{J}$ a matrix of ones, and $\mathbf{R}^{CRW2}$ is the precision matrix given in (1). In addition to seasonal variation, the disease incidence, as displayed in Figure 1, shows occasional outbreaks. To address such temporal dependence beyond seasonal variation, we additionally introduce an autoregressive process of order 1 (AR1) again coupled with a uniform correlation matrix. The linear predictor follows as:

$$\log(\lambda_{tr}) = \mu_r + \alpha_{tr} + \beta_{(t \bmod 52)r}, \quad (2)$$

where μ_r denotes the region-specific intercept, α_{tr} the outbreak-specific component modelled as a correlated AR1 and $\beta_{(t \bmod 52)r}$ the seasonal component modelled as a correlated CRW2.

All 5 hyperparameters (the seasonal precision, the correlation between the CRWs, precision and autoregressive parameter of the AR1 processes and correlation between the AR1s) are treated as unknown. For the unknown precision parameters we use gamma hyper-priors, namely a $Ga(1, 0.00005)$ for the precision κ of the correlated random walk, and a $Ga(0.1, 0.001)$ for the precision of the AR(1) process as proposed by Schrödle et al. (2011). For the Fisher's z-transformed autoregressive parameter we use a normal distribution with zero-mean and variance 0.2^{-1}, corresponding to a U-shaped prior. The same prior is used for the transformed correlation parameters between the CRWs and the AR1s. Here, the general Fisher's z-transformation (Fisher, 1958, page 219) is used, which ensures that the correlations only take values between $(-1/(R-1), 1)$, so that $\mathbf{C}$ is positive definite without imposing an additional constraint, see also Riebler et al. (2011).

TABLE 1. DIC for three different models using a CRW2 to model seasonal variation in the nine major regions of the USA.

	common CRW2	region-specific CRW2 uncorrelated	region-specific CRW2 correlated
DIC	36707	36716	36704

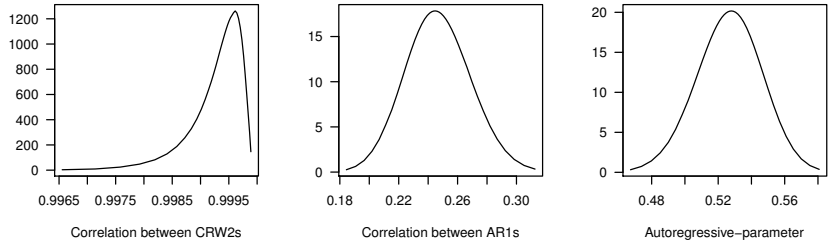


FIGURE 2. Approximated posterior marginals for the correlation parameters and the autoregressive parameter.

3 Results

We compared the model, which uses correlated CRW2s to model the seasonal pattern in the nine regions, with a model assuming independent CRW2s and a model assuming a common CRW2. The DIC values for all three models are shown in Table 1. The model assuming correlated region-specific CRW2s is classified as the best model for which Figure 2 shows the approximate posterior marginals for both correlation parameters and the autoregressive parameter. The correlation between the seasonal components is close to unity (0.999; 95% CI: [0.998, 1]). Figure 3 shows the seasonal pattern (mean within 95% CI) for NewEngland, and for the other regions the pair-wise differences of the estimated mean seasonal effects to NewEngland are shown. For all regions, the seasonal component is higher in the winter months and lower during the summer, but some small differences occur across regions. For example, in Mountain the peak in the winter months is higher, while the pattern in summer is lower compared to NewEngland. In SouthAtlantic it is the other way around.

Of note, the seasonal pattern is not completely smooth. The decreasing effect at the end of the year and the increasing effect at the beginning might be explained by a Christmas effect, where few cases are reported around Christmas but many after the holidays. Peaks throughout the year are not completely clear and need to be investigated in detail.

Turning to the correlated AR1 processes, we note that the estimated autoregressive parameter is 0.53 (95% CI: [0.49, 0.56]) and the correlation

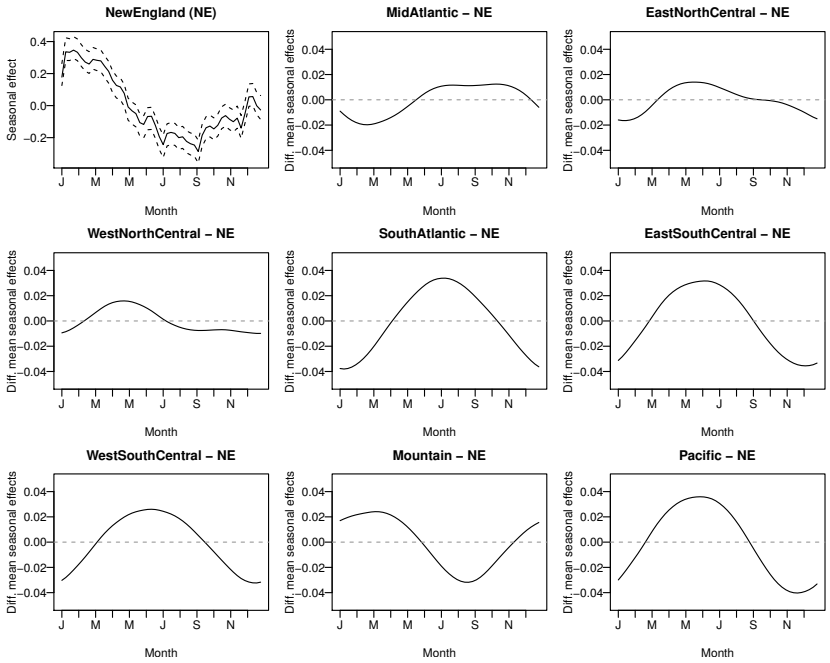


FIGURE 3. Estimated seasonal effects (mean and 95% CI) for NewEngland (NE) (top left). For the other regions the difference of the mean seasonal effects to NE is shown.

between the processes is estimated to be 0.25 (95% CI: [0.20, 0.29]) and thus also clearly different from zero.

4 Discussion and outlook

We proposed the use of correlated CRW2s for modelling seasonal variation in multivariate time series of counts. We applied the methodology to weekly numbers of deaths from influenza and pneumonia in nine major regions of the USA. Although, the correlation between the single seasonal trends was close to unity, this model was preferred compared to a model with one common seasonal component.

In certain aspects the CRW2 represents a quite flexible approach, as the seasonal pattern is not restricted in its functional form, so that also sharp peaks can be captured. However, it assumes that the temporal pattern repeats every 52 weeks, whereas ideally we would like to account for time-varying disease onsets.

The modulation model proposed by Eilers et al. (2008) is more flexible in this aspect. However, here the (co)sine function might be too simple in

certain applications. An unstructured non-parametric model as defined in Rue and Held (2005, page 122f) can also account for time-varying disease onsets. However, here the week indicators are treated exchangeable so that the seasonal pattern is not required to be smooth. Both the modulation model of Eilers et al. (2008) and the seasonal model of Rue and Held (2005, page 122f) can be implemented in INLA and could also be coupled across regions using a uniform correlation matrix. Currently, we are working on a comparison and if possible a combination of these models. Furthermore, we are exploring possibilities to include spatial correlation between the nine geographical regions of the USA.

Acknowledgments: This work received support from the Swiss National Science Foundation and the Research Council of Norway.

References

Eilers, P.H.C., Gampe, J., Marx, B.D. and Rau, R. (2008). Modulation models for seasonal time series and incidence tables. *Statistics in Medicine*, **27**, 3430–3441.

Fanshawe, T.R., Diggle, P.J., Rushton, S., Sanderson, R., et al. (2008). Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach *Environmetrics*, **19**, 549–566.

Harvey, A. and Koopman, S.J. (1993). Forecasting hourly electricity demand using time-varying splines. *Journal of the American Statistical Association*, **88**, 1228–1236.

Paul, M., Held, L. and Toschke, M.A. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, **27**, 6250–6267.

Riebler, A., Held, L. and Rue, H. (2011). Correlated multivariate age-period-cohort models. Technical report. University of Zurich. Available from: `http://www.biostat.uzh.ch/research/manuscripts/cmapc.pdf`.

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

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

Schrödle, B., Held, L. and Rue, H. (2011). Assessing the impact of network data of the spatio-temporal spread of infectious diseases. Technical report. University of Zurich. Available from: `http://www.biostat.uzh.ch/research/manuscripts/schroedle_etal_2011.pdf`.

Segmented smoothing with an L_0 penalty

Ralph C. A. Rippe [1], Paul H. C. Eilers[2]

[1] Institute of Psychology, Leiden University, The Netherlands
[2] Department of Biostatistics, Erasmus Medical Center, The Netherlands

Abstract: Copy number variations in tumor DNA show sudden jumps between constant segments. We propose a smoother with a roughness penalty on the number of jumps, implemented by an L_0 norm. A simple iterative weighting algorithm finds the solution.

Keywords: SNP; penalty; smoother.

1 Introduction

In normal (human) DNA, the autosomal chromosomes form pairs, so there are two very similar copies of each segment of a chromosome. In tumors aberrations can occur: some segments get lost and others occur three or more times. This is called copy number variation (CNV). A useful way to detect and quantify CNV is to use the (fluorescence) signals that are delivered by microarrays for genotyping of SNPs (single nucleotide polymorphisms). Each SNP has two alleles, and the signals are proportional to the number of alleles in a biological sample. For normal DNA the sum of the signals should be 2 (times an unknown scaling factor), in tumors we find segments where this sum is either smaller or larger.
Biologists and medical doctors are highly interested in these segments. Recently, quite some work was done (Morganella et al, 2010; Tsuang et al., 2010; Winchester et al., 2009) comparing available methods to determine the presence of CNVs. The common denominator is the suggestion to use multiple methods in conjunction. Here we present a new smoothing algorithm. Our main goal is better (clearer) visualization, by either signal smoothing or scatterplot smoothing, but preliminary tests show that it can also be used for actual CNV detection.
The observed signal can be quite noisy, so we have to smooth it to enhance segments. However, standard smoothers do not respect the sharp boundaries between segments that occur in these data. The smoother we start with, based on ideas from Whittaker (Eilers, 2003), uses as roughness penalty the sum of squared differences of adjacent fitted values. This is fine if we aim for a "rounded" smooth result, as is often the case. For the present application this is not what we want; instead we like to see sharp jumps between constant segments.

Eilers & de Menezes (2005) showed how replacing sums of squares by sums of absolute values (the L_1 norm) goes a long way towards our goal. They used linear programming for the computations. Here we further improve their approach by introducing the L_0 norm in the penalty, which is essentially a penalty on the number of jumps. Also we introduce a simple iterative weighting scheme to avoid linear programming.

2 Theory and application for CNV signals

In this section we introduce the model. We develop it in several steps and illustrate results directly, not in a separate section.
The data (for one chromosome) are m data pairs (x_i, y_i), where x_i gives the position of SNP i ($x_i < x_{i+1}$ for all i) and y_i is the copy number signal. We are going to compute a smooth series z. Our starting point is a variant of the Whittaker smoother (Eilers, 2003). It rounds off edges, which is fine in many applications, but not here. Therefore, Eilers & de Menezes (2005) replaced the sum of squares (the L_2 norm) by sums of absolute values (the L_1 norm). Their objective function is

$$S_1 = \sum_{i=1}^{m} |y_i - z_i| + \lambda \sum_{i=2}^{m} |z_i - z_{i-1}|.$$

As can be seen from the top panel of Figure 1, this goes in the right direction. The bold middle line shows the smoother with a λ chosen by eye. Segments become more clearly visible, although the jumps are not perfect. The shifted thin line above shows the effect of setting the smoothing parameter too large (misses major jumps) and the shifted line below shows the result for a too small λ: it shows too many small jumps.
Notice that the L_1 norm occurs not only in the penalty but also in the first term of S_1. The reason is that Eilers and de Menezes use linear programming to minimize S_1. The first term measures the quality of the fit to the data. The L_1 norm there implies median smoothing. This increases robustness, but decreases sensitivity to the data, relative to the L_2 norm. In practice robustness is hardly an issue in CNV studies.
We propose the following modification:

$$S_q = \sum_{i=1}^{m} (y_i - z_i)^2 + \lambda \sum_{i=2}^{m} |z_i - z_{i-1}|^q$$

where q is a number between 0 and 1. Actually we will concentrate on $q = 0$, the L_0 norm. Essentially this is a penalty on the number of non-zero difference between neighboring elements of z. As a result, we are no longer detecting relatively small deviations: only large(r) regions are picked up, as illustrated in the bottom panel in Figure 1. Again, the thin lines above

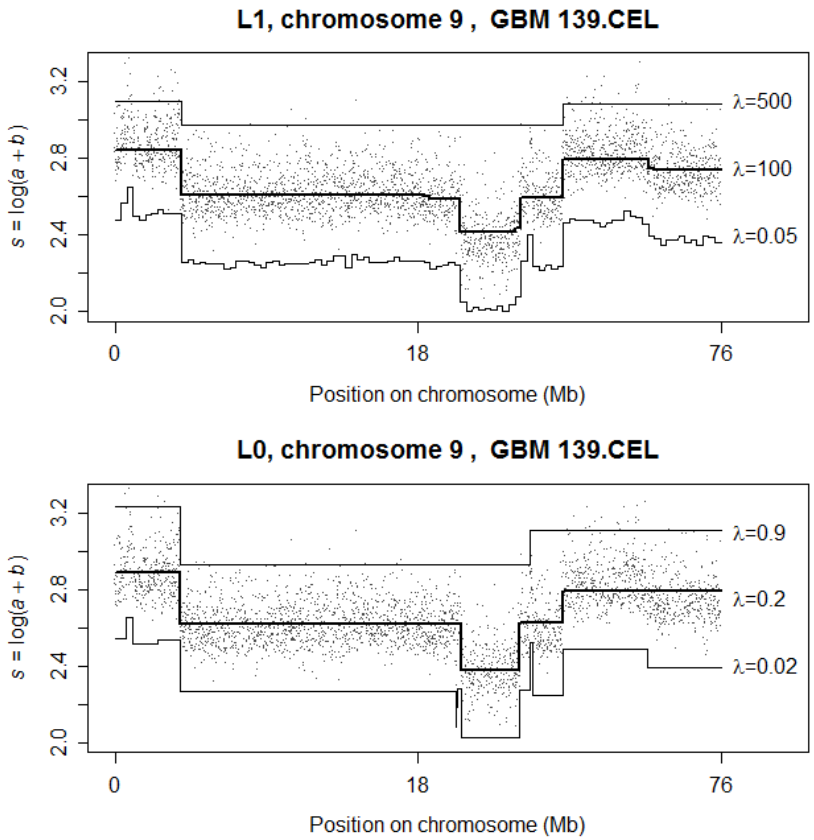


FIGURE 1. Illustration of smoothing with two different norms for the roughness penalty. Top: L_1 norm; bottom: L_0 norm, giving better segmentation. In both panels results are shown for three values of the smoothing parameter λ (chosen subjectively).

show results when λ is too large, while the lower lines illustrate a too small smoothing parameter.
It is easy to find the solution for the Whittaker smoother, using matrix-vector operations. If D is a matrix that forms first differences of z, $Dz = \Delta z$, the objective function can be written as $S_2 = ||y-z||^2 + \lambda||Dz||^2$, with an explicit solution that follows form the linear system $(I + \lambda D'D)\hat{z} = y$. The system is banded and thus very sparse, which can be exploited in native Matlab or in R, using the `spam` package. For large m this reduces computation time and memory use by orders of magnitude, compared to non-sparse matrix operations.
We propose a simple, but effective, algorithm to minimize S_q, using iterated weights in an adapted Whittaker smoother. Obviously $|a|^q = a^2|a|^{q-2}$, for any number a. If we do not know a itself, but an approximation $\tilde{a}$, then $|a|^q \approx a^2|\tilde{a}|^{q-2}$. Using this relation, we approximate $|z_i - z_{i-1}|^q$ by $v_i(z_i - z_{i-1})^2$, with $v_i = |\tilde{z}_i - \tilde{z}_{i-1}|^{q-2}$. If $V = \text{diag}(v)$, the system to be solved becomes $(I + \lambda D'VD)\hat{z} = y$. This gives a new approximation to the solution from which new weights are computed. These steps are iterated until convergence.
To improve numerical stability and reduce the number of iterations, we modify the weights somewhat: $v_i = 1/[(\tilde{z}_i - \tilde{z}_{i-1})^2 + \alpha^2]^{1-q/2}$, where α is a small number, of the order of 1/1000th the expected size of the jumps.

3 The L_0 norm in scatterplot smoothing

Another useful application of the L_0 norm can be found as an addition to the so-called scatterplot smoother (Eilers & Goeman, 2004). Here we visualize the ratio of the signals for the two alleles, $r = log(b/a)$. In normal, healthy, tissue we find three signal bands: one centered around 0, representing the heterozygous genotype AB, and two above and below 0, representing the homozygous AA and BB genotypes. In tumor tissue regions can occur where the middle band is missing: loss of the heterozygosity (LOH). It can get worse: one or both alleles can be missing. These changes usually occur in larger regions on a chromosome, so we want to smooth the scatterplot, while keeping the clear segmentation. This is not the case with the existing scatterplot smoother (Figure 2, top panel).
The scatterplot smoother first computes a two-dimensional histogram. Rows and columns are smoothed with a modified Whittaker smoother, in which both first and second order differences occur. This is done to guarantee that smoothed counts are always positive. The penalty is $\lambda^2 D_2'D_2 + 2\lambda D_1'D_1$. We keep this penalty for the signal (vertical) direction, but for the location direction we use a penalty based on the the L_0 norm of first differences: $\lambda D_1'VD_1$. Because we apply the smoothing to all rows of the histogram simultaneously, the weights in V have to be determined by a whole column. We found that $1/\tilde{v_j} = \sum_i (z_{ij} - zi, j-1)^2/m + \alpha^2$ works well. Results are shown in the bottom panel in Figure 2.

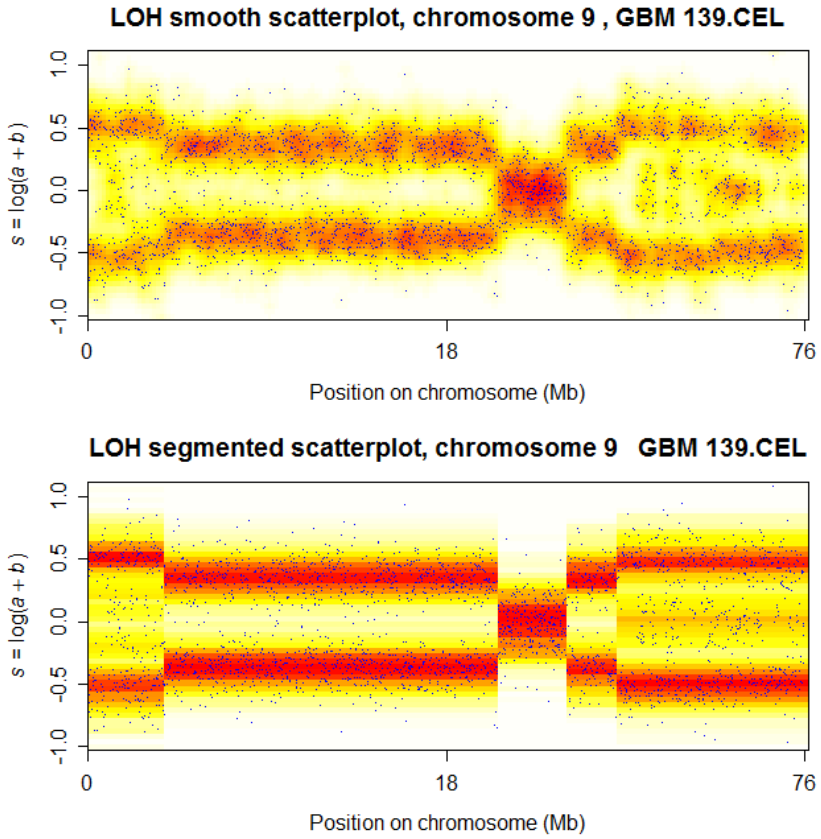


FIGURE 2. Illustration of scatterplot smoothing without (top) and with (bottom) built-in L_0 norm in horizontal direction.

4 Discussion

We have shown that using the L_0 norm in the difference penalty of a smoother works well to give segmented results, as is required by the biological application.
It is remarkable that this norm works so well, because the objective function is non-convex. In the first iteration we use the Whittaker smoother, which probably sets the scene well for the iterations that follow. Yet we cannot be sure that a solution is optimal: it might have reached a local minimum. At present we do not try to optimize the choice of λ, the smoothing parameter, in some objective way. The primary goal was to deliver an improved tool for data visualization and exploration. However, we will study performance as a CNV detection algorithm in simulation experiments. Then automatic smoothing will be needed. The linear equation system after the final iteration is suitable to compute the effective model dimension and hence allows for an AIC or a cross-validation measure.

References

Bralten L.B., Kloosterhof, N.K., Gravendeel, L.A., Sacchetti, A. et al. (2010). Integrated genomic profiling identifies candidate genes implicated in glioma-genesis and a novel LEO1-SLC12A1 fusion gene. *Genes, Chromosomes and Cancer*, **49(6)**, 509–517.

Eilers, P. H. C. (2003). A perfect smoother, *Analytical Chemistry*, **75**, 3299–3304.

Eilers, P.H.C. & de Menezes, R. (2005). Quantile smoothing of array CGH data, *Bioinformatics*, **21(7)**, 1146–1153.

Morganella et al. (2010). VEGA: Variational segmentation for copy number detection, *Bioinformatics*, **21(7)**, 1146–1153.

Tsuang et al. (2010). The effect of algorithms on copy number variant detection. *PLoS One*, **5(12)**, e14456.

Winchester L, Yau C, Ragoussis J (2009). Comparing CNV detection methods for SNP arrays, *Brief Funct Genomic Proteomic*, **8**, 353–366.

Testing for covariate effects in ROC-GAM regression models based on bootstrap methods

M. X. Rodríguez-Álvarez[1,4], J. Roca-Pardiñas[2] , C. Cadarso-Suárez[3,4]

[1] Unit of Clinical Epidemiology and Biostatistics, Complexo Hospitalario Universitario de Santiago de Compostela, Spain. mariajose.rodriguez.alvarez@usc.es.
[2] Dept. of Statistics and Operational Research. University of Vigo, Spain.
[3] Dept. of Statistics and Operations Research, University of Santiago de Compostela, Spain.
[4] Instituto de Investigación Sanitaria de Santiago (IDIS), Santiago de Compostela, Spain.

Abstract: In this work a bootstrap-based procedure for testing continuous covariate effect on the ROC-GAM regression model is proposed. The validity of the bootstrap-based tests is examined in a simulation study, and endocrine data are analysed with the aim of assessing the performance of the waist circumference (WC) in detecting patients having a higher risk of cardiovascular problems.

Keywords: ROC curve; generalized additive models; bootstrap

1 Introduction

The receiver operating characteristic (ROC) curve is the most widely used measure for evaluating the accuracy of continuous diagnostic tests. Recently, Rodríguez-Álvarez et al. (2009) have proposed a new flexible estimator for the conditional ROC curve, based on direct modelling (Alonzo and Pepe, 2002). In that approach, the effect of the covariates and false positive fraction on the ROC curve is modelled non-parametrically using generalised additive models (GAM) combined with local polynomial kernel smoothers (Fan and Gijbels, 1995). More precisely, given $\mathbf{X}$ a set of p continuous covariates, the following ROC-GAM regression model for the ROC curve is assumed:

$$ROC_{\mathbf{X}}(t) = g\left(\alpha + \sum_{k=1}^{p} f_k(X_k) + h_0(t)\right), t \in (0,1), \qquad (1)$$

where f_j and h_0 are smooth and unknown functions.
In this work we introduce bootstrap-based procedures to test for continuous covariate effect on the ROC-GAM regression model specified in (1).

Specifically, for each continuous covariate, X_r, in (1), our interest is focused on the null hypothesis

$$H_0^r:\ f_r(X_r)=0.$$

2 Testing for continuous covariate effect

The test for the null hypothesis

$$H_0^r: ROC_{\mathbf{X}}(t)=g\left(\alpha+\sum_{k=1}^{r-1}f_k(X_k)+\sum_{k=r+1}^{p}f_k(X_k)+h_0(t)\right) \qquad (2)$$

versus the general hypothesis

$$H_1^r: ROC_{\mathbf{X}}(t)=g\left(\alpha+\sum_{k=1}^{p}f_k(X_k)+h_0(t)\right)$$

is based on the estimate $\hat{f}_r$. For this purpose, L_1 and L_2 norms are considered yielding the following test statistics:

$$T^{||}=\sum_{j=1}^{n_D}\left|\hat{f}_r(x_{jr})\right|,\quad T^2=\sum_{j=1}^{n_D}\hat{f}_r(x_{jr})^2.$$

2.1 Bootstrap-based procedure

To approximate the distributions of $T^{||}$ and T^2 under the null hypothesis, a general bootstrap procedure is proposed:

Step 1. Estimate $\mu_{\bar{D}}$, $\sigma_{\bar{D}}$, and $S_{\bar{D}}$ from $\left\{\left(\mathbf{x}_i^{\bar{D}},y_i^{\bar{D}}\right)\right\}_{i=1}^{n_{\bar{D}}}$.

Step 2. Estimate the null ROC-GAM regression model (2) from $\left\{\left(\mathbf{x}_j^{D},y_j^{D}\right)\right\}_{j=1}^{n_D}$, and obtain the bootstrap pilot estimates $\widehat{ROC}^{0}_{\mathbf{x}_j^D}(t)$, $j=1,\ldots,n_D$.

Step 3. For $b=1,\ldots,B$, generate the bootstrap resamples $\left\{\left(\mathbf{x}_i^{\bar{D}},y_{i,b}^{\bar{D}*}\right)\right\}_{i=1}^{n_{\bar{D}}}$ and $\left\{\left(\mathbf{x}_j^{D},y_{j,b}^{D*}\right)\right\}_{j=1}^{n_D}$ where

$$y_{i,b}^{\bar{D}*}=\hat{\mu}_{\bar{D}}\left(\mathbf{x}_i^{\bar{D}}\right)+\hat{\sigma}_{\bar{D}}\left(\mathbf{x}_i^{\bar{D}}\right)\varepsilon_{i,b}^{\bar{D}*},$$

$$y_{j,b}^{D*}=\hat{\mu}_{\bar{D}}\left(\mathbf{x}_j^{D}\right)+\hat{\sigma}_{\bar{D}}\left(\mathbf{x}_j^{D}\right)\hat{S}_{\bar{D}}^{-1}\left(\left(\widehat{ROC}^{0}_{\mathbf{x}_j^D}\right)^{-1}\left(u_{j,b}^{*}\right)\right),$$

$\left\{\varepsilon_{i,b}^{\bar{D}*}\right\}_{i=1}^{n_{\bar{D}}}$ is an i.i.d. sample from $\hat{S}_{\bar{D}}$, $\left\{u_{j,b}^{*}\right\}_{j=1}^{n_D}$ is an i.i.d. sample from $U[0,1]$, and $\left(\widehat{ROC}^{0}_{\mathbf{x}_j^D}\right)^{-1}\left(u_{j,b}^{*}\right)=\inf\left\{t:\ \widehat{ROC}^{0}_{\mathbf{x}_j^D}(t)\geq u_{j,b}^{*}\right\}$.

Step 4. From $\left\{\left(\mathbf{x}_i^{\bar{D}}, y_{i,b}^{\bar{D}*}\right)\right\}_{i=1}^{n_{\bar{D}}}$ and $\left\{\left(\mathbf{x}_j^{D}, y_{j,b}^{D*}\right)\right\}_{j=1}^{n_{D}}$ obtain T_b.

Since the bootstrap resamples are constructed under the null hypothesis, this procedure approximates the distribution of T ($T^{||}$, T^2) under H_0. Consequently, the test rule based on T consists of rejecting the null hypothesis if $T > T_\alpha$ where T_α is the empirical (1-α)-percentile of the values of $T_1, \ldots, T_B$ obtained in Step 4.

3 Simulation study

Data were simulated from

$$Y_D = \sin(\pi X) + \sqrt{0.2 + 0.5\exp(X)} + \sqrt{0.2 + 0.5\exp(X)}\varepsilon_D,$$

$$Y_{\bar{D}} = \sin(\pi X) - a0.3X^3 + \sqrt{0.2 + 0.5\exp(X)}\varepsilon_{\bar{D}},$$

where a is a real constant, $X \sim U[-1, 1]$, and $\varepsilon_{\bar{D}}$, $\varepsilon_D \sim N(0, 1)$. With the above configurations, the corresponding covariate-specific ROC curves is

$$ROC_X(t) = \Phi\left(\frac{a0.3X^3 + \sqrt{0.2 + 0.5\exp(X)}}{\sqrt{0.2 + 0.5\exp(X)}} + \Phi^{-1}(t)\right).$$

It should be noted that $a = 0$ corresponds to the null hypothesis, and the more the constant a shifts towards zero, the greater the effect of the covariate on the ROC curve. The bootstrap procedure described above was applied using $B = 200$ bootstrap samples for determining the critical values of the tests. The type I error, as well as the power, have been calculated as the proportion of rejections of H_0 in 1000 runs. The results are shown in Table 1.

4 Application to endocrine data

We applied the proposed bootstrap-based tests to an endocrine study, with the aim of assessing the effect of age on the accuracy of the WC when predicting clusters of cardiovascular risk factors. The study was carried out with a random sample of Galician adult population (2945 subjects, 46.2% men; age range 18-85 years). Subjects having two or more cardiovascular disease risk factors (raised triglycerides, blood pressure and plasma glucose, and reduced HDL-cholesterol) were considered as diseased. The following ROC-GAM model was considered:

$$ROC_{(Age,Gender)}(t) = \Phi\left(\alpha_0 + \alpha_1 \mathbf{1}_{\{Gender=Men\}} + f(Age) + h_0(t)\right).$$

Figure 1 depicts the conditional areas under the ROC curve (AUCs) together with 95% pointwise bootstrap confidence bands. The resulting p-value of the proposed $T^{||}$ and T^2 tests (with $B = 200$) was lower than 0.001 in both cases.

TABLE 1. Estimated type I error ($a = 0$) and rejection probabilities under the alternative hypothesis ($a = 2$) of the $T^{||}$ and T^2 tests

			Level				
a	Sample size	Test	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**
0	**50**	$T^{\|\|}$	0.024	0.074	0.122	0.175	0.236
		T^2	0.024	0.074	0.117	0.174	0.241
	200	$T^{\|\|}$	0.025	0.077	0.120	0.180	0.226
		T^2	0.017	0.066	0.117	0.171	0.224
	500	$T^{\|\|}$	0.017	0.064	0.120	0.170	0.223
		T^2	0.024	0.068	0.120	0.174	0.218
2	**50**	$T^{\|\|}$	0.089	0.202	0.291	0.38	0.449
		T^2	0.103	0.201	0.286	0.365	0.423
	200	$T^{\|\|}$	0.411	0.605	0.706	0.791	0.833
		T^2	0.415	0.616	0.714	0.792	0.834
	500	$T^{\|\|}$	0.834	0.938	0.971	0.987	0.991
		T^2	0.891	0.955	0.978	0.990	0.993

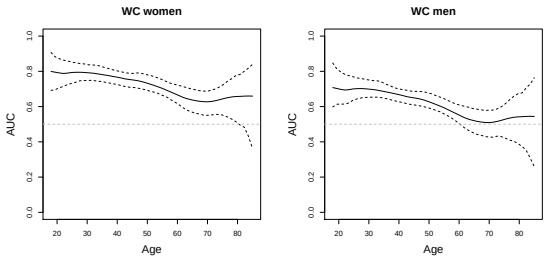


FIGURE 1. Estimated AUCs adjusted by age and gender with 95% pointwise bootstrap confidence bands for Women and Men.

References

Alonzo, T.A., and Pepe, M.S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, **3**, 421-432

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall: CRC.

Rodríguez-Álvarez, M.X., Roca-Pardiñas, J., and Cadarso-Suárez, C. (2009). A new flexible direct ROC regression model. Detection of cardiovascular risk factors by anthropometric measures. *Report 09/06. Discussion Papers in Statistics and Operation Research*. Universidade de Vigo.

D-Optimum designs in random effect logistic regression models

J. M. Rodríguez-Díaz[1], M. T. Santos-Martín[1], C. Tommasi[2]

[1] Department of Statistics, University of Salamanca, Spain,
e-mail: juanmrod@usal.es and maysam@usal.es
[2] Department of Economics, Business and Statistics, University of Milan

Abstract: In the context of nonlinear models, the analytical expression of the Fisher information matrix is essential to compute optimum designs. The Fisher information matrix of the random effect logistic regression model is proved to be equivalent to the information matrix of the linearized model, which depends on some integrals. *D*-optimum designs are computed for the univariate logistic regression model with Gaussian random effects. It is proved that *D*-optimum designs are invariant with respect to a scale transformation of the design region.

Keywords: binary regression model; information matrix; optimal design of experiments.

1 Introduction

The interest in finding optimum designs in the context of regression models with random effects is steadily increasing. See for instance, Holland-Letz et al.(2011) and Debusho and Haines (2011). Another setting where optimal designs have been extensively studied is the context of (fixed effect) binary regression models. Recently, Ouwens et al.(2006) have studied optimum designs for logistic models with random intercept. In this paper, D-optimum designs are derived for the logistic regression model where not only the intercept but all the coefficients are random.

2 Binary regression model and Fisher information matrix

Let Y be a binary response variable such that $P(Y = 1|\beta) = F(\mathbf{x}'\beta)$ where "$'$" denotes transposition, $\mathbf{x} = (1, x_1, \ldots, x_{k-1})'$ is a $k \times 1$ vector of experimental conditions which may be chosen in an experimental domain $\mathcal{X} \subseteq \mathbb{R}^k$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_{k-1})'$ is $k \times 1$ vector of random coefficients. In other words, for each experimental unit there is a vector of unobservable random coefficients $\beta \in \Omega_\beta \subseteq \mathbb{R}^k$ such that $\beta \sim \phi(\beta; \theta)$, where $\phi(\cdot)$ is a k-variate probability density function (pdf) which depends on some unknown

parameters $\theta \in \Theta \subseteq \mathbb{R}^m$, with $k \leq m$. It is well known that if β is a vector of constant unknown coefficients then the Fisher information matrix coincides with information matrix corresponding to the linearized binary regression model. This equivalence holds even when the coefficients are random variables.

In order to have n independent observations, $y_1, \ldots, y_n$, it is assumed that one observation per individual is taken, as in Graßhoff et al. (2009). With this assumption, the log-likelihood function is

$$\log L(\theta) = \sum_{i=1}^{n} \log \int [F(\mathbf{x}_i' \beta)]^{y_i} \, [1 - F(\mathbf{x}_i' \beta)]^{1-y_i} \, \phi(\beta; \theta) d\beta,$$

where from now on the integration is taken over $\Omega_\beta \subseteq \mathbb{R}^k$.

The Fisher information matrix of an exact design $\xi = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the $m \times m$ matrix $\mathcal{I}(\xi; \theta) = \{\mathcal{I}_{rs}(\xi; \theta)\}$, whose (r, s)-item $(r, s = 1, \ldots, m)$ is

$$\mathcal{I}_{rs}(\xi; \theta) = \sum_{i=1}^{n} \frac{\int F(\mathbf{x}_i' \beta) \, \frac{\partial}{\partial \theta_r} \phi(\beta; \theta) \, d\beta \cdot \int F(\mathbf{x}_i' \beta_i) \, \frac{\partial}{\partial \theta_s} \phi(\beta; \theta) \, d\beta}{\int F(\mathbf{x}_i' \beta) \, \phi(\beta; \theta) \, d\beta \cdot \left[1 - \int F(\mathbf{x}_i' \beta) \, \phi(\beta; \theta) \, d\beta\right]} \qquad (1)$$

where the last equality is obtained after some algebra, assuming the usual regularity conditions on the pdf $\phi(\cdot)$.

An alternative expression for the binary regression model is $Y_i = E[Y_i] + \varepsilon_i$, where $E[Y_i] = \int F(\mathbf{x}_i' \beta) \, \phi(\beta; \theta) \, d\beta$ and ε_i is such that $E[\varepsilon_i] = 0$ and

$$Var[\varepsilon_i] = Var[Y_i] = \int F(\mathbf{x}_i' \beta) \, \phi(\beta; \theta) \, d\beta \cdot [1 - \int F(\mathbf{x}_i' \beta) \, \phi(\beta; \theta) \, d\beta].$$

If this model is linearized at some nominal values of the parameters, then the information matrix for one observation is given by $\mathbf{g}(\mathbf{x}_i; \theta)\mathbf{g}(\mathbf{x}_i; \theta)'$ where $\mathbf{g}(\mathbf{x}_i; \theta)$ is the $m \times 1$ vector whose j-th item, $(j = 1, \ldots, m)$, is

$$g_j(\mathbf{x}_i; \theta) = \frac{\int F(\mathbf{x}_i' \beta) \frac{\partial}{\partial \theta_j} \phi(\beta; \theta) d\beta}{\sqrt{\int F(\mathbf{x}_i' \beta) \, \phi(\beta; \theta) \, d\beta \cdot [1 - \int F(\mathbf{x}_i' \beta) \, \phi(\beta; \theta) \, d\beta]}}.$$

From (1) it follows that $\mathcal{I}(\xi; \theta) = \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i; \theta)\mathbf{g}(\mathbf{x}_i; \theta)'$ and this proves the equivalence between the Fisher information matrix and the information matrix corresponding to the linearized model.

3 Univariate logistic regression model

In this section the previous results are applied to the case of univariate logistic regression model with Gaussian random coefficients. Thus, the success probability of the binary response Y is given by the following model $P(Y = 1|\beta) = F(\mathbf{x}' \beta) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$; $\beta \sim N_2(\mathbf{b}, \mathbf{V})$, where $x \in \mathcal{X} \subseteq \mathbb{R}$ is an experimental condition, $\beta = (\beta_0, \beta_1)'$ is the vector of random coefficients $(k = 2)$, which comes from a Normal distribution with mean vector $\mathbf{b} = (b_0, b_1)'$ (the so called fixed effects) and dispersion matrix

$\mathbf{V} = diag(v_0^2, v_1^2)$. If $\mathbf{v} = (v_0, v_1)'$, the unknown vector of model parameters is $\theta = (\mathbf{b}', \mathbf{v}')'$, thus $m = 4$.
For an approximate design ξ, the Fisher information matrix is proportional to $\mathcal{I}(\xi;\theta) = \int_{\mathcal{X}} \mathbf{g}(x;\theta)\mathbf{g}(x;\theta)' d(\xi)$ where $\mathbf{g}(x;\theta)$ is a 4×1 vector whose items are given by,

$$g_1(x;\theta) = \frac{I_1(x;\theta)}{v_0\sqrt{I_0(x;\theta)[1-I_0(x;\theta)]}}, g_2(x;\theta) = \frac{I_2(x;\theta)}{v_1\sqrt{I_0(x;\theta)[1-I_0(x;\theta)]}},$$

$$g_3(x;\theta) = \frac{I_3(x;\theta) - I_0(x;\theta)}{v_0\sqrt{I_0(x;\theta)[1-I_0(x;\theta)]}}, g_4(x;\theta) = \frac{I_4(x;\theta) - I_0(x;\theta)}{v_1\sqrt{I_0(x;\theta)[1-I_0(x;\theta)]}},$$

where integrations are taken over $\mathbb{R}^2$ and

$$\begin{aligned}
I_0(x;\theta) &= \int F(\mathbf{x}'\,\beta)\,\phi(\beta;\theta)\,d\beta = \frac{1}{2\pi}\int e^{h_1(\tilde{\beta};x;\theta)}\,d\tilde{\beta},\\
I_1(x;\theta) &= \frac{1}{2\pi}\int \tilde{\beta}_0\, e^{h_1(\tilde{\beta};x;\theta)}\,d\tilde{\beta}, \quad I_2(x;\theta) = \frac{1}{2\pi}\int \tilde{\beta}_1\, e^{h_1(\tilde{\beta};x;\theta)}\,d\tilde{\beta},\\
I_3(x;\theta) &= \frac{1}{2\pi}\int \tilde{\beta}_0^2\, e^{h_1(\tilde{\beta};x;\theta)}\,d\tilde{\beta}, \quad I_4(x;\theta) = \frac{1}{2\pi}\int \tilde{\beta}_1^2\, e^{h_1(\tilde{\beta};x;\theta)}\,d\tilde{\beta},\\
h_1(\tilde{\beta};x;\theta) &= b_0 + v_0\tilde{\beta}_0 + b_1 x + v_1\tilde{\beta}_1 x - \frac{1}{2}\tilde{\beta}_0^2 - \frac{1}{2}\tilde{\beta}_1^2\\
&- \log\left[1 + \exp\left(b_0 + v_0\tilde{\beta}_0 + b_1\, x + v_1\tilde{\beta}_1 x\right)\right],
\end{aligned}$$

$\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1)'$ with $\tilde{\beta}_0 = (\beta_0 - b_0)/v_0$ and $\tilde{\beta}_1 = (\beta_1 - b_1)/v_1$.
Thus, integrals $I_j(x;\theta)$, $j = 0, 1, \ldots, 4$, must be evaluated to compute the Fisher information matrix of the logistic regression model with Gaussian random effects.

4 *D*-optimum designs

Among all optimality criteria for precise estimation of the whole vector of parameters, the D-criterion is indeed the most popular. It is defined by the following criterion function $\Phi_D(\xi) = \begin{cases} -\log|\mathcal{I}(\xi;\theta)| & \text{if } |\mathcal{I}(\xi;\theta)| \neq 0 \\ \infty & \text{otherwise} \end{cases}$
A design which minimizes $\Phi_D(\xi)$ is a D-optimum design. The following theorem states that the D-optimum design is invariant to a scale transformation of the design region.
Theorem *Let ξ_D^* be a D-optimum design to estimate θ on $\mathcal{X}$ and let q be a positive constant, then a D-optimum design on the scaled design space $\mathcal{Z} = \{z = q\,x\,|x \in \mathcal{X}\}$, is $\eta_D^*(z) = \xi_D^*(x), \quad x \in \mathcal{X}$ and $z = q\,x$.*
Table 1 shows locally D-optimal designs on $\mathcal{X} = [0, 1]$ for the logistic random effect model and different nominal values of the parameters. Weights are not shown because most of designs are equally-weighted four-point designs. The six-point design has respective weights 0.246, 0.206, 0.151, 0.231,

TABLE 1. Support points of D-optimal designs on $\mathcal{X} = [0, 1]$, for different nominal values of the parameters. All the designs are equally-weighted, but the six-point one, whose weights are 0.246, 0.206, 0.151, 0.231, 0.042 and 0.124.

	$\mathbf{v} = (v_0, v_1)'$		
$\mathbf{b} = (b_0, b_1)'$	(0.02,0.03)'	(0.2,0.3)'	(2,3)'
(0.1,0.2)'	{0,0.276,0.723,1 }	{0,0.275,0.720,1}	{0,0.192,0.579,1}
(1,2)'	{0,0.226,0.648,1}	{0,0.227,0.647,1}	{0,0.248,0.649,1}
(10,20)'	{0,0.1,0.84,0.95}	{0,0.07,0.2,0.93,0.94,0.98}	{0,0.03,0.12,0.29}

0.042 and 0.124, respectively. The case $b_0 = 10, b_1 = 20$ is very difficult to solve from the computational point of view. After several trials (with different initial designs) D-optimal designs with the largest support point different from 1 are found, which is an unusual result. In addition, the 6-point design cannot be reduced, even when some support points are very close to each other. For instance, the joining of points 0.93 and 0.94 would produce a deep decreasing of the lower bound for the D-efficiency.

5 Conclusions

The Fisher information matrix of the random effect logistic regression model is proved to be equivalent to the information matrix of the linearized model. Besides, D-optimal designs are proved to be invariant with respect to a scale transformation of the design region, thus when the experimental domain changes according to a scale transformation, the optimal experimental conditions change in the same way. Finally D-optimal designs are computed for different values of the parameters.

References

Debusho, L.K., Haines, L.M. (2011). D- and V-optimal population designs for the quadratic regression model with a random intercept term. *Journal of Statistical Planning and Inference* **141**, 889-898.

Graßhoff, U., Holling, H., Schwabe, R. (2009). On Optimal Design for a Heteroscedastic Model Arising from Random Coefficients. In: *Proceedings 6th St. Petersburg Workshop on Simulation*, pp. 387-392.

Holland-Letz, T., Dette, H., Pepelyshev, A. (2011). A geometric characterization of optimal designs for regression models with correlated observations. *Journal of the Royal Statistical Society B* **73**, in press.

Ouwens, M.J.N.M., Frans, T.E.S., Martijn, B.P.F. (2006). A maximin criterion for the logistic random intercept model with covariates. *J. Statist. Plann. Inference* ,**136**, 962-981.

Adaptive Spectral Estimation for Nonstationary Time Series

Ori Rosen[1], Sally Wood[2], David Stoffer[3]

[1] Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas 79968, U.S.A., E-mail: ori@math.utep.edu,
[2] Melbourne Business School, University of Melbourne, Victoria, 3053, Australia
[3] Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, U.S.A.

Abstract: We propose methodology for analyzing possibly nonstationary time series by adaptively dividing the time series into an unknown but finite number of segments and estimating the corresponding local spectra by smoothing splines. The model is formulated in a Bayesian framework, and the estimation relies on reversible jump Markov chain Monte Carlo (RJMCMC) methods. For a given segmentation of the times series, the likelihood function is approximated via a product of local Whittle likelihoods. Thus, no parametric assumption is made about the process underlying the time series. The number and lengths of the segments are assumed unknown and may change from one MCMC iteration to another. The method is illustrated with simulated and real data.

Keywords: MCMC; Nonstationary Time Series; Whittle Likelihood.

1 Introduction

This paper proposes methodology for analyzing possibly nonstationary time series by adaptively dividing the time series into an unknown but finite number of segments and estimating the corresponding local spectra by smoothing splines. The model is formulated in a Bayesian framework, and the estimation relies on reversible jump Markov chain Monte Carlo (RJMCMC) methods. For a given segmentation of the times series, the likelihood function is approximated via a product of local Whittle likelihoods. Thus, no parametric assumption is made about the process underlying the time series. The number and lengths of the segments are assumed unknown and may change from one MCMC iteration to another.

The analysis of nonstationary time series is important in many fields. For example, in epilepsy research, understanding seizure initiation and its propagation is a critical task (Qin and Wang (2008)) which relies on analyzing EEG time series. Another example is weather research where global warming is of major concern. One time series which is often analyzed in this context is the Southern Oscillation Index which is an indicator of the

El Niño Southern Oscillation (ENSO) phenomenon. A related question is whether human-induced global warming has changed the structure of the ENSO time series (Timmermann et al. (1999)). In Section 5 we present some results for this time series.
Rosen, Wood and Stoffer (2009) estimate the log of the local spectrum using a Bayesian mixture of splines. The basic idea of this approach is to first partition the data into small sections. It is then assumed that the log spectral density of the evolutionary process in any given partition is a mixture of individual log spectra. A mixture of smoothing splines model with time varying mixing weights is used to estimate the evolutionary log spectrum. The mixture model is fit using MCMC techniques that yield estimates of the log spectra of the individual subsections. As described above, unlike Rosen et al. (2009), the current paper adaptively divides the time series into segments of variable lengths, rendering the mixture model unnecessary. In addition to more accurate estimation, this also leads to computational saving.

2 Spectral Estimation for Stationary Time Series

We first explain our approach to estimating the spectral density of a stationary process. Suppose that a stationary time series, $\{X_t\}$, has a bounded positive spectral density, $f(\nu)$, for $-1/2 < \nu \leq 1/2$. Given a realization, $x_1, \ldots, x_n$, the periodogram of the data at frequency ν (measured in cycles per unit time) is

$$I_n(\nu) = \frac{1}{n}\left|\sum_{t=1}^{n} x_t \exp(-2\pi i \nu t)\right|^2 .$$

Let $\nu_k = k/n$, for $k = 0, \ldots, n-1$, be the Fourier frequencies. Whittle (1957) showed that, under appropriate conditions, for large n, the likelihood of $\boldsymbol{x} = (x_1, \ldots, x_n)'$, given $\boldsymbol{f} = (f(\nu_0), \ldots, f(\nu_{n-1}))'$, can be approximated by

$$p(\boldsymbol{x} \mid f) = (2\pi)^{-n/2} \prod_{k=0}^{n-1} \exp\left\{-\frac{1}{2}\Big[\log f(\nu_k) + I_n(\nu_k)/f(\nu_k)\Big]\right\}. \quad (1)$$

Note that in (1), there are only $[n/2] + 1$ distinct observations since the spectral density and the periodogram are both even functions of ν. The notation $[n]$ means the largest integer less than or equal n. For ease of notation, in what follows, we assume that n is even. Letting $y_n(\nu_k) = \log I_n(\nu_k)$ and $g(\nu_k) = \log f(\nu_k)$, we place a smoothing spline prior on $g(\nu_k)$, (Wahba (1990)) and estimate it by MCMC methods.

3 Spectral Estimation for Locally Stationary Time Series

Consider a time series $\boldsymbol{x} = (x_1, \ldots, x_n)'$ with an unknown number of locally stationary segments. Before specifying the model, we introduce some notation. Let m be the unknown number of segments and $n_{j,m}$ be the number of observations in the jth segment of m locally stationary segments. We assume that $n_{j,m} \geq t_{\min}$, where $t_{\min}$ is taken to be large enough in order for the local Whittle likelihood to provide a good approximation to the true local likelihood. The location of the end of the jth segment is denoted by $\xi_{j,m}$, $j = 0, \ldots, m$, where $\xi_{0,m}$ and $\xi_{m,m}$ are $t = 0$ and $t = n$, respectively. Given a partition $\boldsymbol{\xi}_m = (\xi_{0,m}, \ldots, \xi_{m,m})'$ of the time series $\boldsymbol{x}$, the jth segment consists of the observations $\boldsymbol{x}_{j,m} = \{x_t : \xi_{j-1,m} + 1 \leq t \leq \xi_{j,m}\}$, $j = 1, \ldots, m$, with underlying spectral densities $f_{j,m}$ and periodograms $I_{n_{j,m}}$, evaluated at frequencies $\nu_{k_j} = k_j/n_{j,m}$, $0 \leq k_j \leq n_{j,m} - 1$. For a given partition $\boldsymbol{\xi}_m$, the approximate likelihood of the time series is thus

$$L(f_{1,m}, \ldots, f_{m,m} \mid \boldsymbol{x}, \boldsymbol{\xi}_m) = \prod_{j=1}^{m} (2\pi)^{-n_{j,m}/2} \prod_{k_j=0}^{n_{j,m}-1} \exp\Big\{-\tfrac{1}{2}\Big[\log f_{j,m}(\nu_{k_j}) + I_{n_{j,m}}(\nu_{k_j})/f_{j,m}(\nu_{k_j})\Big]\Big\}.$$

Prior distributions are placed on all the parameters, including the number of segments, m, and the partition, $\boldsymbol{\xi}_m$. The estimation is performed via reversible jump MCMC methods.

4 Simulations

To illustrate the methodology, we present results based on single realizations from a slowly-varying autoregressive process and a piecewise autoregressive process. The model is fitted to the data with a total of 10,000 iterations, 2000 of which are used as burn-in. The value of $t_{\min}$ is set to 40. Data were generated from each of the models

$$x_t = a_t x_{t-1} + \epsilon_t \text{ where } a_t = -0.5 + t/500 \text{ for } t = 1, \ldots, 500 \quad (2)$$

$$x_t = a_t x_{t-1} + \epsilon_t \text{ where } a_t = \begin{cases} -0.5 & \text{for } t \leq 250 \\ 0.5 & \text{for } t > 250 \end{cases} \quad (3)$$

and $\epsilon_t \sim N(0, 1)$.
Figure 4 displays the log spectrum as a function of frequency and time, corresponding to the realizations from models (2) and (3) (top and bottom, respectively.) As Figure 4 shows, our methodology may handle situations where a time series changes slowly, as well as cases where the change is abrupt.

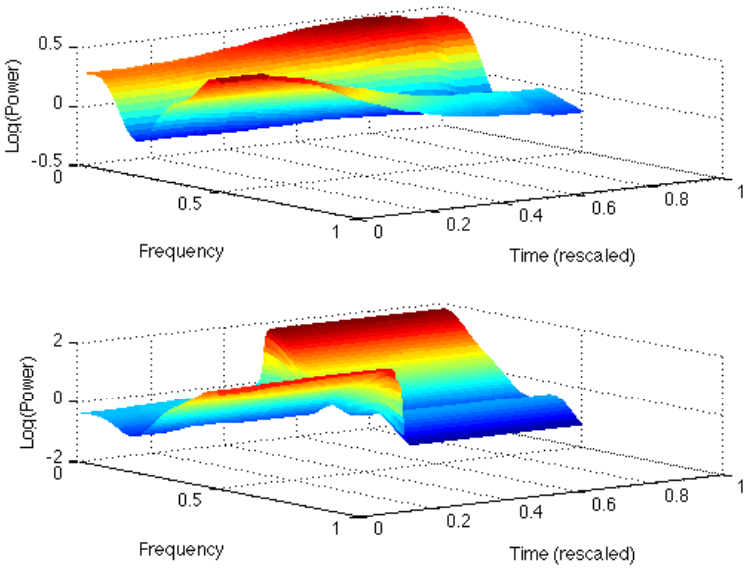


FIGURE 1. Log spectrum vs. time and frequency. The upper plot corresponds to the time series generated from (2); the lower plot correspond to the series generated from (3).

5 Analysis of the Southern Oscillation Index

The Southern Oscillation Index (SOI) is calculated from the monthly or seasonal fluctuations in the air pressure difference between Tahiti and Darwin. The SOI time series is presented in Figure 2. Sustained negative values of the SOI often indicate El Niño episodes. These are characterized by sustained warming of the central and eastern tropical Pacific Ocean, decrease in the strength of the Pacific Trade winds and a reduction in rainfall over eastern and northern Australia. Positive values of the SOI indicate La Niña episodes. These tend to be accompanied by stronger Pacific Trade winds and warmer sea temperatures to the north of Australia, cooling of the waters in the central and eastern tropical Pacific Ocean and an increased probability that eastern and northern Australia will be wetter than normal. As mentioned in Section 1, it is of interest to find out whether the structure of the SOI time series has changed as a result of human-induced global warming. Our analysis of the SOI time series shows that there is a probability of about 10% that a change has occurred in this time series. Figure 3 displays the estimated posterior cumulative distribution function

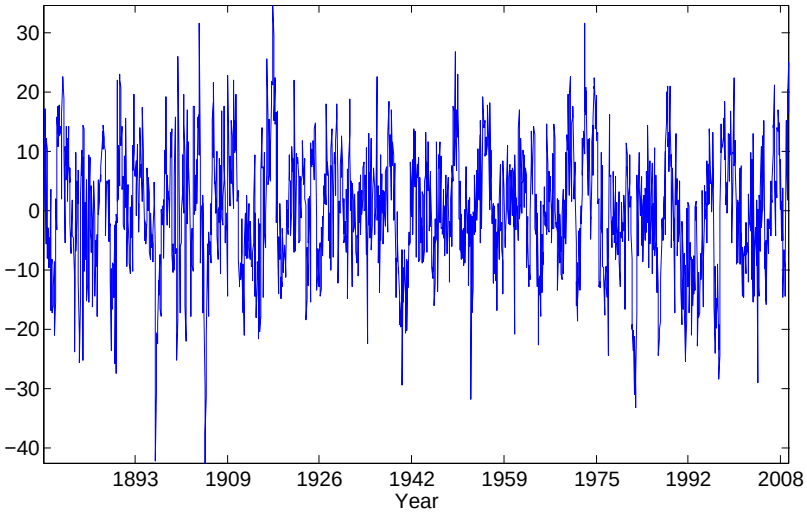


FIGURE 2. The Southern Oscillation Index

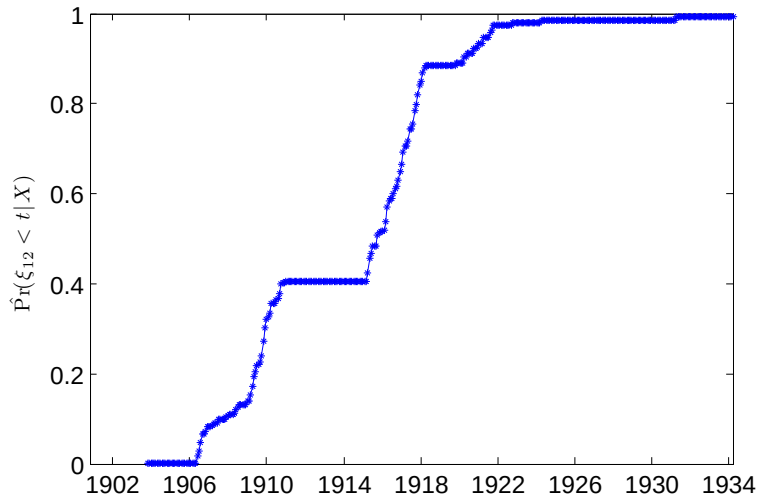


FIGURE 3. $\hat{\Pr}(\xi_{12} < t)$ as a function of time.

$\hat{\Pr}(\xi_{12} < t)$ as a function of time. It shows that in 10% of the time where the SOI time series is partitioned into two segments, the transition between these segments occurs slowly between 1906 and 1934. We have found no evidence for a more recent change.

References

Qin, L. and Wang, Y. (2008). Nonparametric spectral analysis with applications to seizure characterization using EEG time series. *The Annals of Applied Statistics*, **2**, 1432-1451.

Rosen, O., Wood, S. and Stoffer, D. (2009). Local spectral analysis via a Bayesian mixture of smoothing splines. *Journal of the American Statistical Association*, **104**, 249-262.

Timmermann, A., Oberhuber, J., Bacher, A., Esch, M., Latif, M. and Roeckner, E. (1999). Increased El Niño frequency in a climate model forced by future greenhouse warming. *Nature*, **398**, 694-697.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia.

Whittle, P. (1957). Curve and periodogram smoothing. *Journal of the Royal Statistical Society B*, **19**, 38-47.

Distributed lag models for hydrological data

A. M. Rushworth[1], A. W. Bowman[1], M. J. Brewer[2], S. J. Langan[3]

[1] School of Mathematics and Statistics, University of Glasgow
[2] Biomathematics and Statistics Scotland, Aberdeen
[3] James Hutton Institute, Aberdeen

Abstract: The distributed lag model (DLM), used prominently in air pollution studies, finds application wherever the effect of a covariate is delayed and distributed through time. We explore the use of a modified formulation of a DLM on flow data obtained from a Scottish mountain river, with particular emphasis on how changes in the relationship between environmental covariates and flow regimes can be captured.

Keywords: hydrology; non-parametric regression; P-splines; time series

1 Introduction

The relationship between river flow and precipitation and the models that attempt to capture it have long been of interest to environmental scientists. Such models find application in flood prediction, as part of larger river catchment-scale models and in attempting to simulate and understand the changing climatic scenarios that might be expected in the future. We wish to investigate the relationship between precipitation and flow rates observed on the River Dee. In particular, we would like to construct a model capable of capturing the state and sensitivity of the catchment, enabling identification of any changes in the strength of the relationship between flow rates and rainfall for which evidence is already emerging (Baggaley et al., 2009). We review some current methodology before proposing a model which we fit to hourly river discharge (m^3s^{-1}) and hourly precipitation (mm) observed on the River Dee between December 2006 and November 2007. The River Dee system is located in the North East of Scotland, covering an area of over $2000km^2$ and has an average annual discharge of $45m^3s^{-1}$

2 Background ideas

A wealth of time series models exist for modelling the relationship between two time dependent variables, most only allowing a limited amount of flexibility in model structure. We propose models based on B-spline basis func-

tions and difference penalties (Eilers and Marx, 1996) that allow particularly flexible specification of time-lagged dependence. The models are modified versions of the distributed lag model (DLM) most commonly found in the air pollution literature (Zanobetti et al., 2000; Welty et al., 2009; Gasparrini et al., 2010). The former propose models in which the temporal dependence between response and lagged covariate is assumed static, while Gasparrini et al. (2010) extend the DLM allowing nonlinearity through interaction with other covariates. A DLM is a natural choice for flow modelling, as flow can be considered as directly responsive to rainfall at lag s, and is sometimes expressed as the transfer function $f(t) = \int_0^\infty r(t)h(t-i)dt$ (Sherman, 1932) where h is some response function; the DLM framework allows h to be estimated from within a wide class of functions. In the current context we use a similar DLM specification to Gasparrini et al. (2010), but use time as the only effect modifier of the DLM curve.

3 Model

3.1 Specification

We set up a model for river discharge, f_t, in terms of preceding rainfall r_{t-i} where $1 \leq i \leq p$. We specify that the contribution each rainfall lag variable r_{t-i} makes to f_t is allowed to change smoothly through time, in turn determined by a regression on a set of B-spline basis functions $\overline{B}(t) = (B_1(t), \ldots, B_k(t))^T$. The model can be represented thus:

$$\begin{aligned} f_t &= \sum_{i=1}^{p} \beta_i(t) r_{t-i} + \epsilon_t = \sum_{i=1}^{p} \sum_{j=1}^{k} a_{ij} B_j(t) r_{t-i} + \epsilon_t \\ \text{and so} \quad \mathbf{f} &= \mathbf{X}\mathbf{a} + \epsilon \end{aligned}$$

where $\mathbf{a} = vec(\mathbf{A})$, $\mathbf{A}$ is the matrix of a_{ij}'s, ϵ_t is an iid error sequence and the i-th row of $\mathbf{X}$ is given by $\overline{B}(p+i) \otimes (r_{p+i-1}, r_{p+i-2}, \ldots, r_i)$ where $\otimes$ is the Kronecker product. We assume here that the smooth change in each β_i through time can be captured by the same basis set.

3.2 Penalties

We specify two penalties on $\mathbf{a}$, each of which is a weighted sum of squared differences of 'neighbouring' $a_{ij}s$. The first penalty, $\lambda_1 \mathbf{D}_1^T \mathbf{D}_1 \mathbf{a}$, penalises the way each rainfall lag variable r_{t-i} is allowed to influence f_t as t changes. If $\mathbf{P}_k$ is a quadratic difference matrix on k parameters (so that $\mathbf{P}_k \mathbf{a} = \sum_{i=1}^{k-2} (a_i - 2a_{i+1} + a_{i+2})^2$), then $\mathbf{D}_1 = \mathbf{P}_k \otimes \mathbf{I}_p$.
The second penalty term, $\lambda_2 \mathbf{D}_2^T \mathbf{D}_2 \mathbf{a}$, is for controlling differences between r_{t-i} and $r_{t-(i+1)}$ for $1 \leq i \leq p-1$ for any time t. This is a penalty on differences between neighbouring elements in the columns of $\mathbf{A}$, or equivalently, elements of $\mathbf{a}$ that are spaced k elements apart, so that $\mathbf{D}_2 = \mathbf{I}_k \otimes \mathbf{P}_p$.

we can then combine the two penalties so that the estimating equation is

$$\mathbf{a} = \left(\mathbf{X}^T\mathbf{X} + \lambda_1\mathbf{D}_1^T\mathbf{D}_1 + \lambda_2\mathbf{D}_2^T\mathbf{D}_2\right)^{-1}\mathbf{X}^T\mathbf{f}.$$

3.3 Results and comments

The optimal smoothing parameters λ_1 and λ_2 were chosen by minimising GCV. The mid-monthly fitted β_is for a model with $k = 100$, $p = 70$ are given in Figure 1 from which it is clear that the shape and strength of the temporal dependency between rainfall and flow varies greatly throughout the year.

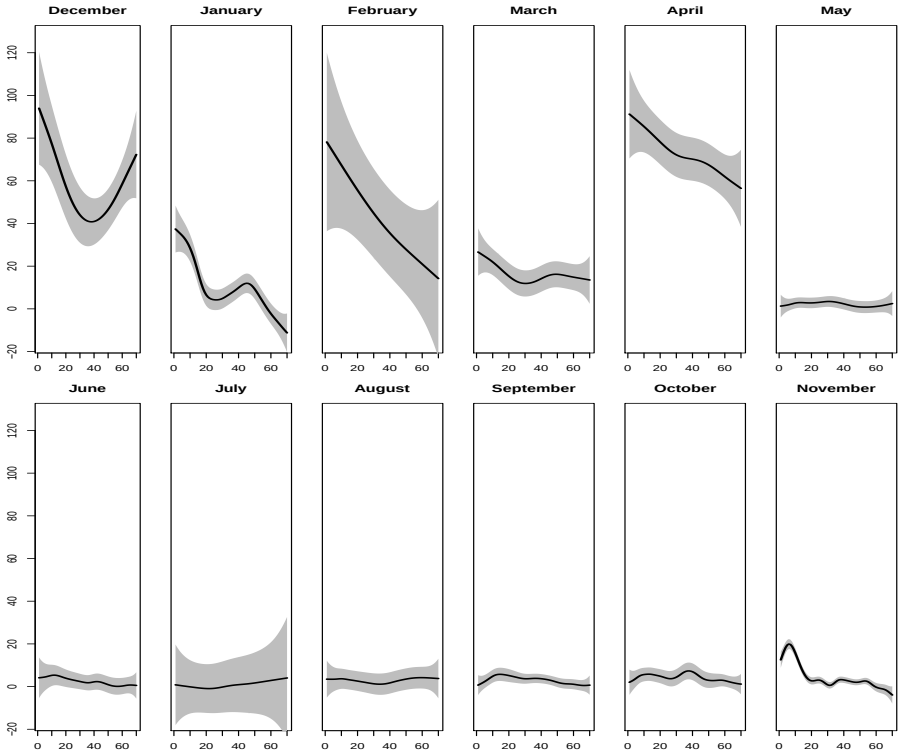


FIGURE 1. Estimated distributed lag curves at monthly intervals with 95% confidence regions

We note much higher parameter estimates in winter and spring months than during summer, probably an expression of nonlinearity between rainfall and flow levels occurring seasonally. Figure 2 shows the fitted flows with those observed for a spring period and and a late autumn period, from which it can be seen that the model performs best during periods of high and continuous rainfall.

It may be desirable to interpret the curves of Figure 1 as discrete estimates of the function h described in section 2, which combined with rainfall lag variables represent a discrete transfer function of rainfall through time. Some of the estimated curves suggest unrealistic relationships if interpreted as the function h, for example during April where rainfall even 70 hours ago appears to heavily influence flows occurring in the present. It is likely that these estimates represent the influence of unobserved, spatially and temporally correlated rainfall, presenting itself in the coefficient estimates as anomalously high values. Adjustments are suggested for improving model performance during periods of decreased rainfall and for periods when snow and ice accumulation and ablation have an impact.

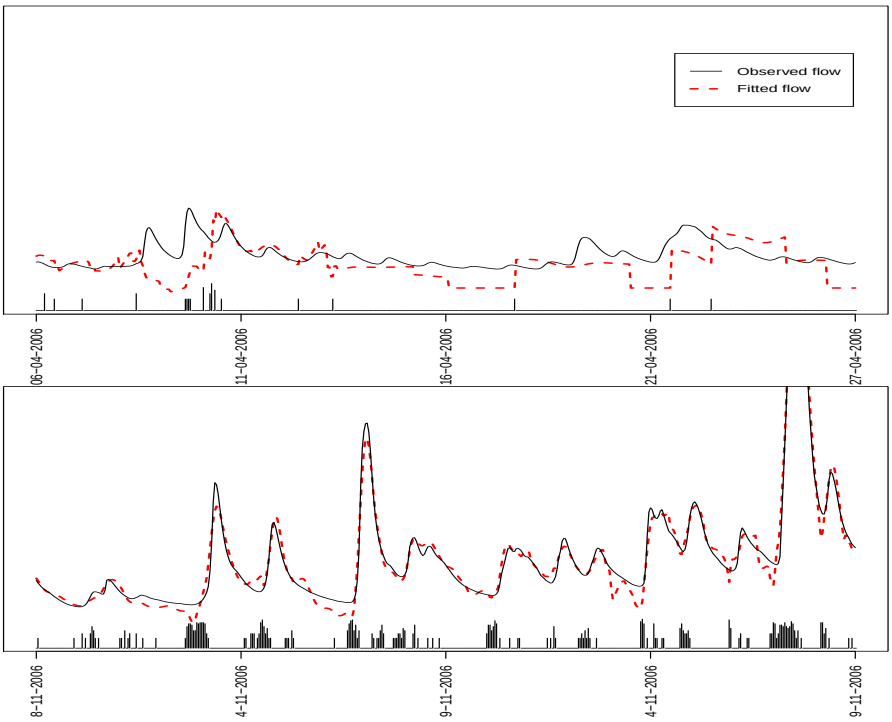


FIGURE 2. Observed and estimated flow values with 95% confidence regions

3.4 Future work

Our analysis was limited by a lack of hourly rainfall data and was unable to discover how much variation in dependence exists between years. By applying similar models to daily data, we hope to discover whether longer term changes can be identified in the flow response to rainfall.

Acknowledgments: Thanks to the Scottish Environmental Protection Agency (SEPA) for providing the river discharge data, the British Atmospheric Data Centre for providing precipitation and temperature data and to Nikki Baggaley for help and advice with the River Dee data

References

Baggaley, N.J., Langan, S.J., Futter, M.N., Potts, J.M., and Dunn, S.M. (2009). Long-term trends in hydro-climatology of a major Scottish mountain river. *Science of the Total Environment*. **407**(16), 4633–4641.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*. **11**(2), 89-102 .

Gasparrini, A., Armstrong, B. and Kenward, M.G. (2010). Distributed lag non-linear models. *Statistics in medicine*. **29**, 2224–2234

Sherman, L.K. (1932). Streamflow from rainfall by the unit-graph method *Engineering News Record*. **108**(14), 501–505

Welty, L.J., Peng, R.D., Zeger, SL., Dominici, F. (2009). Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. *Biometrics*. **65**(1), 282–291.

Zanobetti, A. and Wand, M.P., and Schwartz, J. and Ryan, L.M. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*. **1**(3), 279.

Exact and approximate inferences for nonlinear mixed-effects heavy-tailed models

Cibele M. Russo[1], Victor Hugo Lachos[2], Reiko Aoki[1] , Gilberto A. Paula[3]

[1] Departamento de Matemática Aplicada e Estatística, ICMC, Universidade de São Paulo, Caixa Postal 668, CEP 13560-970, São Carlos, SP, Brazil, e-mail: cibele@icmc.usp.br and reiko@icmc.usp.br

[2] Departamento de Estatística, IMECC, Universidade Estadual de Campinas, CEP 13083-859, Campinas, Brazil, e-mail: hlachos@ime.unicamp.br

[3] Departamento de Estatística, IME, Universidade de São Paulo, Caixa Postal 66281 (Ag. Cidade de São Paulo), CEP 05311-970, São Paulo, SP, Brazil, e-mail: giapaula@ime.usp.br

Abstract: Nonlinear mixed-effects models provide an useful alternative to model nonlinear correlated data. The most usual assumption is that the random effects and errors jointly follow a normal distribution, which may not be adequate in cases of outliers or heavy-tailed data. In this work, we suppose a multivariate scale mixture of normal distributions for the random effects and errors. Aiming to obtain an efficient estimation procedure, we compare two estimation methods, an exact method via Monte Carlo EM and an approximate method, which approaches the problem to linear case. A real data set illustrates the modelling and a simulation study is performed to compare the methodologies.

Keywords: nonlinear mixed-effects models; approximate inference; Monte Carlo EM.

1 Introduction

The interest for nonlinear mixed-effects models (NLMEMs) comes from different application areas as pharmacokinetics longitudinal data or growth curves, for instance. In the literature, the most common assumption for the distribution of the errors and the random effects in NLMEMs is the normality (see, for instance, Wu, 2004), which may not be the most appropriate choice in cases of heavy-tailed data or in the presence of outlying observations. Nonlinear elliptical mixed-effects models were discussed by Russo et al. (2009), where the random effects were included linearly to the model. Recent discussions on heavy-tailed nonlinear mixed-effects models can be found in Meza et al. (2010) and Lachos et al. (2011). We assume that the joint distribution of the random effects and errors belongs to the class of scale mixture of normal (SMN) distributions, which covers important families as the multivariate Student-t (MSt), the multivariate slash (MSl), the

multivariate contaminated normal (MCN), among others. Considering this model we present two methods of estimation, a Monte Carlo EM method which is referred to as an "exact" method, since it is based on the exact likelihood, and an approximate method based on iterative approximations to a linear mixed-effects model.

2 The model

Suppose that $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$ is a vector of observed continuous multivariate responses in which $\mathbf{y}_i$ denotes an $(n_i \times 1)$ vector containing the observations for the experimental unit i, $i = 1, \ldots, n$, such that

$$\begin{array}{l} \mathbf{y}_i = \mathbf{g}(\boldsymbol{\phi}_i, \mathbf{X}_i) + \boldsymbol{\epsilon}_i, \ i = 1, \ldots, n, \\ \boldsymbol{\phi}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \end{array} \tag{1}$$

in which $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i})^\top$ is a matrix of explanatory variables for the ith unit, $\mathbf{b}_i$ is a $(q \times 1)$ vector of random effects, $\boldsymbol{\epsilon}_i$ is a $(n_i \times 1)$ vector of random errors for $i = 1, \ldots, n$, $\boldsymbol{\beta}$ is a $(p \times 1)$ location vector and $\mathbf{A}_i$ and $\mathbf{B}_i$, with dimensions $(p \times p)$ and $(p \times q)$ respectively, are full rank matrices of known constants. In this work, we will assume that

$$\begin{pmatrix} \boldsymbol{\epsilon}_i \\ \mathbf{b}_i \end{pmatrix} \overset{ind.}{\sim} \mathrm{SMN}_{n_i+q}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}; \mathrm{H}(u_i)\right), \tag{2}$$

where $\mathbf{D}$ and $\boldsymbol{\Sigma}_i$ are positive-definite dispersion matrices. For simplicity, we assume that $\mathbf{D} = \mathbf{D}(\boldsymbol{\tau})$ is a diagonal matrix and denote its elements by $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q)^\top$. Matrix $\boldsymbol{\Sigma}_i$ with dimension $(n_i \times n_i)$ is typically dependent upon i through its dimension, and is initially assumed to be of the form $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}_{n_i}$ for $i = 1, \ldots, n$ and $\sigma > 0$ a scalar. Since $\mathbf{A}_i$, $\mathbf{B}_i$ and $\mathbf{X}_i$ are known matrices, we will simplify the notation by writing $\mathbf{g}(\boldsymbol{\beta}, \mathbf{b}_i)$ to represent $\mathbf{g}(\boldsymbol{\phi}_i, \mathbf{X}_i) = \mathbf{g}(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \mathbf{X}_i)$. The quantity $\mathrm{H} = \mathrm{H}(u, \boldsymbol{\nu})$ is the cdf generator that determines the specific SMN model that was assumed. The obtention of maximum likelihood estimates for the parameters are drawn by using two methods, a Monte Carlo EM (MCEM) algorithm and an approximate method. MCEM method is called exact, although it entails a simulation step, because the model is maintained as originally proposed, whereas the approximate method brings the problem to a linear solution. A similar approach was discussed by Wu (2004). To select the model, the IC_Q criterion is used (see Ibrahim et al., 2008). The same criterion may be used to choose the parameters from the scale mixture of normal distributions, which are assumed to be fixed.

For the MCEM algorithm, $\mathbf{b}_i$ and the scale factor U_i is considered as missing data, so that the "complete data" is given by $\{(\mathbf{y}_i, \mathbf{b}_i, u_i), i = 1, \ldots, n\}$. For individual i, let $\{(\mathbf{b}_i^{(1)}, u_i^{(1)}), \ldots, (\mathbf{b}_i^{(m_i)}, u_i^{(m_i)})\}$ denote a random sample of size m_i generated from $[u_i, \mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}^{(t)}]$ then the E step at the $(t+1)$th

EM iteration can be written as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \propto \sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{m_i}\left[-\frac{n_i}{2}\log\sigma^2 - \frac{\kappa^{-1}(u_i^{(j)})}{2\sigma^2}||\mathbf{y}_i - \mathbf{g}(\boldsymbol{\beta}, \mathbf{b}_i^{(j)})||^2\right] + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{m_i}\left[-\frac{1}{2}\log|\mathbf{D}| - \frac{\kappa^{-1}(u_i^{(j)})}{2}\mathbf{b}_i^{(j)\top}\mathbf{D}^{-1}\mathbf{b}_i^{(j)}\right],$$

and the M step can be obtained straightforwardly.
For the approximate method, first-order Taylor expansion of g_{ij} around the current parameter estimate $\widehat{\boldsymbol{\beta}}$ and the random effect estimate $\widehat{\mathbf{b}}_i$, and the problem is reduced to the iterative solution of the LME response model

$$\widetilde{\mathbf{y}}_i = \mathbf{W}_i\boldsymbol{\beta} + \mathbf{T}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \tag{3}$$

where $\widetilde{\mathbf{y}}_i = \mathbf{y}_i - \mathbf{g}_i(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i) + \mathbf{W}_i\widehat{\boldsymbol{\beta}} + \mathbf{T}_i\widehat{\mathbf{b}}_i$ and the elements of $\mathbf{W}_i$ and $\mathbf{T}_i$ are related to the derivatives of the entries of $\mathbf{g}_i$ with respect to $\boldsymbol{\beta}$ and $\mathbf{b}_i$, respectively.

3 Numerical illustration

Considering the growth soybean data set analysed by Pinheiro & Bates (2000, chap. 6), a possible nonlinear mixed effects model could be written in the form

$$y_{ij} = \frac{\beta_1 + b_i}{1 + \exp\{-[x_{ij} - \beta_2]/\beta_3\}} + \epsilon_{ij},\ j = 1, \ldots, n_i,\ i = 1 \ldots, n, \tag{4}$$

where n_i assumes the values 8, 9 or 10 depending on the value of $i \in \{1, \ldots, n = 48\}$. The measurements of leaf weights were taken within approximately weekly intervals after planting, over three years, 1988, 1989 and 1990, and two genotypes, P (*plant introduction*) and F (*forrest*). The observed value y_{ij} represents the jth mean weight (in g) of leafs from a soybean plant in the ith plot, after t days of being planted, where for each of the 6 year-genotype combination there were 8 plots. In this case, β_1, β_2 and β_3 represent the asymptotic leaf weight, the time at which the leaf reaches half of its asymptotic weight and the time elapsed between the leaf reaching half and $1/(1 + e^{-1})$ of its asymptotic weight, respectively. The maximum likelihood estimates of the parameters obtained by the exact and the approximate method are given in Tables 1 and 2. We can observe that the fixed effects parameters estimates are close, but the scale elements estimates are distant when the two different methodologies are applied. Simulation studies showed that the approximate method is more efficient and quite reasonable in this case.

Table 1: Maximum likelihood estimates of the parameters with standard errors (S. E.) obtained by the MCEM method.

	normal		Student-t (3)		slash (3)	
	Estimate	(S. E.)	Estimate	(S. E.)	Estimate	(S. E.)
β_1	19.196	(0.48)	19.657	(0.41)	19.134	(0.43)
β_2	55.541	(0.47)	56.487	(0.48)	55.611	(0.43)
β_3	8.902	(0.29)	9.078	(0.25)	8.913	(0.26)
σ^2	1.744	(0.13)	0.944	(0.13)	0.93	(0.09)
τ_1	15.089	(3.33)	13.266	(3.60)	10.726	(2.44)
IC_Q	1653.595		1397.845		1378.642	

Table 2: Maximum likelihood estimates of the parameters with standard errors (S. E.) obtained by the approximate method.

	normal		Student-t (3)		slash (3)	
	Estimate	(S. E.)	Estimate	(S. E.)	Estimate	(S. E.)
β_1	18.939	(0.26)	19.439	(0.21)	18.828	(0.23)
β_2	55.277	(0.39)	56.306	(0.31)	55.393	(0.35)
β_3	8.7651	(0.26)	8.9903	(0.20)	8.8053	(0.23)
σ^2	1.7478	(0.12)	0.91984	(0.06)	0.93157	(0.07)
τ_1	14.7437	(3.01)	12.9969	(2.65)	10.5225	(2.15)
IC_Q	1656.199		1384.248		1379.955	

Acknowledgments: This research is supported by FAPESP and CNPq, Brazil.

References

Ibrahim, J. G., Zhu, H. and Tang, N. (2008). Model Selection Criteria for Missing-Data Problems Using the EM Algorithm. *Journal of the American Statistical Association* 103, 1648–1658.

Lachos, V.H., Bandyopadhyay D. and Dey D. K. (2011). Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. To appear in *Biometrics*.

Meza, C., Osorio, F, De la Cruz, R. (2010). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing* DOI 10.1007/s11222-010-9212-1.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*, Springer.

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

Wu, L. (2004). Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. *Journal of the American Statistical Association* 99 467, 700–709.

Hyper-g Priors for Generalised Additive Model Selection

Daniel Sabanés Bové[1], Leonhard Held[1] , Göran Kauermann[2]

[1] Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, CH-8001 Zurich, Switzerland, Email: daniel.sabanesbove@ifspm.uzh.ch, leonhard.held@ifspm.uzh.ch

[2] Centre for Statistics, Department of Economics and Business Administration, University Bielefeld, Postfach 300131, D-33501 Bielefeld, Germany, Email: gkauermann@uni-bielefeld.de

Abstract: We propose an automatic Bayesian approach to the selection of covariates and penalised splines transformations thereof in generalised additive models. Specification of a hyper-g prior for the model parameters and a multiplicity-correction prior for the models themselves is crucial for this task. We introduce the methodology in the normal model and illustrate it with an application to diabetes data. Extension to non-normal exponential families is finally discussed.

Keywords: Penalised splines; Bayesian variable selection; Shrinkage.

1 Introduction

Suppose we have p metrical covariates $x_1, \ldots, x_p$ and use the additive model $y = \beta_0 + \sum_{j=1}^{p} m_j(x_j) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. When x_j is included non-linearly in the model, we assume

$$m_j(x_j) = x_j \beta_j + \boldsymbol{Z}_j(x_j)^T \boldsymbol{u}_j$$

where $\boldsymbol{Z}_j(x_j)$ is the $K \times 1$ spline basis vector at position x_j and $\boldsymbol{u}_j \sim N(\boldsymbol{0}, \sigma^2 \rho_j \boldsymbol{I})$ is the corresponding coefficients vector. In order to combine n observations, we stack these to the $n \times 1$ vector $\boldsymbol{x}_j$ and the $n \times K$ basis matrix $\boldsymbol{Z}_j$, both modified to be zero-centred and orthogonal to each other. We then translate the variance parameter ρ_j into the corresponding degree of freedom (Aerts, Claeskens and Wand, 2002, section 2.2)

$$d_j(\rho_j) = \mathrm{tr}\{(\boldsymbol{Z}_j^T \boldsymbol{Z}_j + \rho_j^{-1} \boldsymbol{I})^{-1} \boldsymbol{Z}_j^T \boldsymbol{Z}_j\} + 1 \in (1, K+1). \tag{1}$$

A larger ρ_j (or a larger d_j) leads to a weaker penalty on the non-linear component of the function m_j. If x_j is excluded from or linearly included in the model we have $m_j(x_j) \equiv 0$ or $m_j(x_j) = x_j \beta_j$ and set $d_j = 0$ or $d_j = 1$, respectively. Thus, the function m_j is exactly defined by d_j, which we may restrict to a finite set of values, say $d_j \in \{0, 1\} \cup \{2, 3, \ldots, K\}$.

As default prior for the parameters β_0, $\boldsymbol{\beta} = (\beta_j : d_j \geq 1)$ and σ^2 in a given model specified via $\boldsymbol{d} = (d_1, \ldots, d_p)$,

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2 \sim \mathrm{N}(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\, \sigma^2 \boldsymbol{I}) \tag{2}$$

with $\boldsymbol{X} = (\boldsymbol{x}_j : d_j \geq 1)$, $\boldsymbol{Z} = (\boldsymbol{Z}_j : d_j > 1)$ and $\boldsymbol{u} = (\boldsymbol{u}_j^T : d_j > 1)^T$, we propose the hyper-$g$ prior (Liang *et al.*, 2008) described in Section 2. For the models we propose a multiplicity-correction prior in Section 3. The methodology is applied to diabetes data in Section 4 and extended to generalised additive models in Section 5.

2 Hyper-g Priors for Additive Models

Integrating out the spline coefficients vector $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{D})$, where $\boldsymbol{D} = \mathrm{diag}\{\rho_j \boldsymbol{I} : d_j > 1\}$, from the conditional model (2) yields the marginal model

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}\left(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta},\, \sigma^2 \boldsymbol{V}\right) \tag{3}$$

with $\boldsymbol{V} = \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T$ having Cholesky decomposition $\boldsymbol{V} = \boldsymbol{R}^T\boldsymbol{R}$. The transformed response vector $\tilde{\boldsymbol{y}} = \boldsymbol{R}^{-T}\boldsymbol{y}$ follows a linear model with similarly transformed design matrix $\tilde{\boldsymbol{X}}$ and diagonal covariance matrix $\sigma^2\boldsymbol{I}$. It turns out that we can use the hyper-g prior (Liang *et al.*, 2008) for this transformed model, *i. e.* a locally uniform prior $p(\beta_0) \propto 1$ on the intercept, Jeffreys' prior $p(\sigma^2) \propto (\sigma^2)^{-1}$ on the variance and the g-prior (Zellner, 1986)

$$\boldsymbol{\beta} \,|\, g, \sigma^2 \sim \mathrm{N}\left(\mathbf{0},\, g\sigma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\right) \tag{4}$$

on the coefficients are combined with a uniform prior on the shrinkage coefficient $g/(1+g)$. Note that $\sigma^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}$ is the Fisher information matrix of $\boldsymbol{\beta}$ in the marginal model (3). The hyper-g prior leads to a closed form of the marginal likelihood, which we can compute on the original response scale via the change of variables formula:

$$p(\boldsymbol{y} \,|\, \boldsymbol{d}) \propto \left\|\tilde{\boldsymbol{y}} - \bar{\tilde{\boldsymbol{y}}}\right\|^{-(n-1)} (l_{\boldsymbol{d}} + 2)^{-1}\, {}_2F_1\left(\frac{n-1}{2}; 1; \frac{l_{\boldsymbol{d}}+4}{2}; \tilde{R}^2\right) |\boldsymbol{R}|^{-1},$$

where $l_{\boldsymbol{d}}$ is the dimension of $\boldsymbol{\beta}$, ${}_2F_1$ is the Gaussian hypergeometric function and $\tilde{R}^2$ is the classical coefficient of determination in model (3).

3 Model Prior

We propose a prior $p(\boldsymbol{d})$ on the model space which explicitly corrects for the multiplicity of testing inherent in the simultaneous analysis of many covariates (see Scott and Berger, 2010): *A priori*, the number of covariates included in the model ($l_{\boldsymbol{d}}$) is uniformly distributed on $\{0, 1, \ldots, p\}$. Then the number of non-linearly included covariates ($s_{\boldsymbol{d}}$) is uniformly distributed on $\{0, 1, \ldots, l_{\boldsymbol{d}}\}$. The respective choice of the $l_{\boldsymbol{d}}$ and $s_{\boldsymbol{d}}$ covariates

TABLE 1. Marginal posterior probabilities (x_1: age, x_2: systolic blood pressure, x_3: cholesterol/HLD ratio, x_4: BMI, x_5: waist/hip ratio, x_6: gender).

	x_1	x_2	x_3	x_4	x_5	x_6
not included ($d_j = 0$)	0.00	0.65	0.00	0.14	0.50	0.65
linear ($d_j = 1$)	0.71	0.33	0.93	0.81	0.48	0.35
non-linear ($d_j > 1$)	0.29	0.03	0.07	0.05	0.02	—

is uniformly distributed on all possible configurations. Finally, the degrees of freedom of the non-linearly modelled covariates are independent and uniformly distributed on $\{2, 3, \ldots, K\}$. Altogether, this gives

$$1/p(\boldsymbol{d}) = \binom{p}{l_{\boldsymbol{d}}}(p+1)\binom{l_{\boldsymbol{d}}}{s_{\boldsymbol{d}}}(l_{\boldsymbol{d}}+1)(K-1)^{s_{\boldsymbol{d}}}$$

and leads to marginal prior probabilities $\Pr(d_j = 0) = 1/2$, $\Pr(d_j = 1) = \Pr(d_j > 1) = 1/4$.

4 Application

We illustrate our modelling approach with the diabetes data from Harrell (2001). We study the association of (the negative reciprocal of) glycosolated haemoglobin of $n = 377$ study participants with the continuous covariates age (in years), systolic blood pressure (in mmHg), cholesterol/HDL ratio, body mass index (BMI, in kg/m^2) and waist/hip ratio as well as the binary covariate gender. As the computational complexity is quadratic in the spline basis dimension K, we want to use splines with few quantile-based knots. Therefore, we choose cubic O'Sullivan splines (Wand and Ormerod, 2008). Here, we get basis matrices $\boldsymbol{Z}_j$ with $K = 9$ columns from 7 knots. The exhaustive evaluation of the posterior model probabilities $p(\boldsymbol{d} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{y} \,|\, \boldsymbol{d})p(\boldsymbol{d})$ of all $(K+1)^5 \cdot 2 = 200\,000$ models takes only 585 seconds due to an efficient `C++` implementation which is available in an `R`-package from the first author. In Table 1 the marginal posterior probabilities for linear and non-linear inclusion of the six covariates are shown. There is strong evidence for linear inclusion of cholesterol/HDL ratio and BMI, while the posterior probability for inclusion of systolic blood pressure or gender is only 35%. There is overwhelming evidence for (non-linear) inclusion of age, and the posterior odds for (linear) inclusion of waist/hip ratio are around 1. The *maximum a posteriori* model includes age, cholesterol/HDL ratio and BMI all linearly. Note that these are the covariates which have inclusion probabilities larger than 50%, thus defining the set of median probability models (Barbieri and Berger, 2004) $\boldsymbol{d}$ with $d_1, d_3, d_4 \geq 1$ and $d_2 = d_5 = d_6 = 0$. Figure 1 shows the estimated covariate effects from the resulting model average. While the age effect is slightly non-linear (with 38% probability in the median probability models), both other covariates have essentially linear effect estimates.

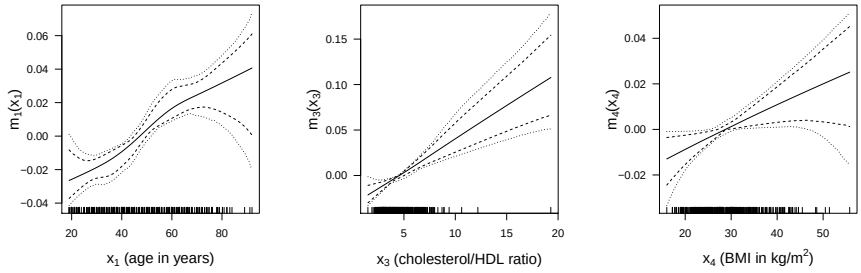


FIGURE 1. Estimated covariate effects in the median probability model average, based on 10 000 samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals as well as positions of data points (ticks above x-axes) are shown.

5 Extension to Generalised Additive Models

Now we assume more generally that the covariate effects $m_j(x_j)$ enter additively into the linear predictor $\eta = \beta_0 + \sum_{j=1}^{p} m_j(x_j)$ of an exponential family distribution with canonical parameter θ, mean $\mathrm{E}(y) = h(\eta) = db(\theta)/d\theta$ and variance $\mathrm{Var}(y) = \phi/w \cdot v(\mu) = \phi/w \cdot d^2b(\theta)/d\theta^2$ (see McCullagh and Nelder, 1989). We restrict our attention to non-normal distributions with fixed dispersion ϕ (as $\phi = 1$ for the Bernoulli and Poisson distribution) and known weight w. For n observations, the linear predictor vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ is $\boldsymbol{\eta} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$, where $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$, and the likelihood is

$$p(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \propto \exp\left\{ \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} \right\}. \tag{5}$$

A reasonable generalisation of (1) is (see Ruppert, Wand and Carroll, 2009, section 11.4)

$$d_j(\rho_j) = \mathrm{tr}\{(\boldsymbol{Z}_j^T \widehat{\boldsymbol{W}} \boldsymbol{Z}_j + \rho_j^{-1}\boldsymbol{I})^{-1} \boldsymbol{Z}_j^T \widehat{\boldsymbol{W}} \boldsymbol{Z}_j\} + 1, \tag{6}$$

which uses a fixed weight matrix $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\mathbf{1}\widehat{\beta}_0)$ for all models, where $\boldsymbol{W}(\boldsymbol{\eta}) = \mathrm{diag}\{(dh(\eta_i)/d\eta)^2 v(h(\eta_i))^{-1}\phi^{-1}w_i\}_{i=1}^n$ is the usual generalised linear model (GLM) weight matrix and $\widehat{\beta}_0$ is the intercept estimate from the null model. Therefore, we now arrange $\mathbf{1}$, $\boldsymbol{x}_j$ and the columns of $\boldsymbol{Z}_j$ to be orthogonal with respect to the inner product in terms of $\widehat{\boldsymbol{W}}$, so that (6) correctly captures the degrees of freedom associated with the non-linear part of m_j.

In order to derive a generalised g-prior for $\boldsymbol{\beta}$, we will use the iterative weighted least squares (IWLS) approximation to (5) to come back to a normal model and then derive the resulting g-prior (4). So let

$$\boldsymbol{z}_0 = \boldsymbol{\eta}_0 + \mathrm{diag}\{dh(\boldsymbol{\eta}_0)/d\boldsymbol{\eta}\}^{-1}(\boldsymbol{y} - h(\boldsymbol{\eta}_0))$$

be the adjusted response vector resulting from a first-order approximation to $h^{-1}(\boldsymbol{y})$ around $h(\boldsymbol{\eta}_0)$, such that

$$\boldsymbol{z}_0 \,|\, \beta_0, \boldsymbol{\beta}, \boldsymbol{u} \sim \mathrm{N}\big(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\, \boldsymbol{W}(\boldsymbol{\eta}_0)^{-1}\big)$$

is the working normal model. This can be rewritten to

$$\tilde{\boldsymbol{z}}_0 \,|\, \beta_0, \boldsymbol{\beta}, \boldsymbol{u} \sim \mathrm{N}(\tilde{\mathbf{1}}\beta_0 + \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{u},\, \boldsymbol{I}) \tag{7}$$

by setting $\tilde{\boldsymbol{z}}_0 = \boldsymbol{W}(\boldsymbol{\eta}_0)^{1/2}\boldsymbol{z}_0$ *etc.* Since (7) is analogous to (2), our proposal for a generalised g-prior is

$$\boldsymbol{\beta} \,|\, g \sim \mathrm{N}(\mathbf{0}, g\boldsymbol{J}^{-1}), \tag{8}$$

where $\boldsymbol{J}$ is the Fisher information for $\boldsymbol{\beta}$ in (7) with $\boldsymbol{\eta}_0 = \mathbf{0}$:

$$\begin{aligned}\boldsymbol{J} &= \tilde{\boldsymbol{X}}^T(\boldsymbol{I} + \tilde{\boldsymbol{Z}}\boldsymbol{D}\tilde{\boldsymbol{Z}}^T)^{-1}\tilde{\boldsymbol{X}} \\ &= \boldsymbol{X}^T\boldsymbol{W}_0^{1/2}(\boldsymbol{I} + \boldsymbol{W}_0^{1/2}\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{W}_0^{1/2})^{-1}\boldsymbol{W}_0^{1/2}\boldsymbol{X},\end{aligned}$$

abbreviating $\boldsymbol{W}_0 = \boldsymbol{W}(\mathbf{0})$. Note that this prior directly generalises the prior proposed by Sabanés Bové and Held (2011) for GLMs, to which it reduces when there are no spline effects in the model.
The generalised hyper-g prior then consists of the improper prior $p(\beta_0) \propto 1$ on the intercept β_0, the g-prior (8) on the linear effects vector $\boldsymbol{\beta}$, the penalty prior $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$ on the spline coefficients vector $\boldsymbol{u}$ and some proper hyper-prior $p(g)$ on the hyper-parameter g in the g-prior. For the implementation of posterior inference we can easily extend the approach of Sabanés Bové and Held (2011, section 3). Let $\boldsymbol{X}_a = (\mathbf{1}, \boldsymbol{X}, \boldsymbol{Z})$ and $\boldsymbol{\beta}_a = (\beta_0, \boldsymbol{\beta}^T, \boldsymbol{u}^T)^T$, such that $\boldsymbol{\eta} = \boldsymbol{X}_a\boldsymbol{\beta}_a$. The prior for $\boldsymbol{\beta}_a$ conditional on g has Gaussian form with mean zero and singular precision $\boldsymbol{R}_a = \mathrm{diag}\{0, g^{-1}\boldsymbol{J}(\mathbf{0}), \boldsymbol{D}^{-1}\}$. Thus, the Laplace approximation of $p(\boldsymbol{y} \,|\, g, \boldsymbol{d})$, which is based on a Gaussian approximation to the conditional posterior $p(\boldsymbol{\beta}_a \,|\, \boldsymbol{y}, g)$, can be obtained by the Bayesian IWLS algorithm (West, 1985). Afterwards, the marginal likelihood

$$p(\boldsymbol{y} \,|\, \boldsymbol{d}) = \int_0^\infty p(\boldsymbol{y} \,|\, g, \boldsymbol{d})p(g)\, dg,$$

can be approximated by numerical integration of the Laplace approximation $\tilde{p}(\boldsymbol{y} \,|\, g, \boldsymbol{d})$. Note that this strategy of integrated Laplace approximations was proposed more generally by Rue, Martino and Chopin (2009). Finally, for sampling from the posterior of $\boldsymbol{\beta}_a$ and g in a specific model $\boldsymbol{d}$ we can use a tuning-free Metropolis-Hastings algorithm.

References

Aerts, M., Claeskens, G., and Wand, M.P. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, **103**, 455-470.

Barbieri, M.M., and Berger, J.O. (2004). Optimal predictive model selection. *Annals of Statistics*, **32**, 870-897.

Harrell, Jr., F.E. (2001). *Regression Modeling Strategies*. New York: Springer.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., and Berger, J.O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, **103**, 410-423.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

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

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

Sabanés Bové, D., and Held, L. (2011). Hyper-g Priors for Generalized Linear Models. *Bayesian Analysis*, **6**, forthcoming article. URL: `http://ba.stat.cmu.edu/abstracts/Sabanes.php`

Scott, J.G., and Berger, J.O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, **38**, 2587-2619.

Wand, M.P., and Ormerod, J.T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, **50**, 179-198.

West, M. (1985). Generalized linear models: scale parameters, outlier accommodation and prior distributions. In: *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*, 531-558. Amsterdam: North-Holland.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233-243. Amsterdam: North-Holland.

Optimal time scaling for plant growth analysis

Sabine K. Schnabel[1,3], Paul H. C. Eilers[1,2] , Fred A. van Eeuwijk[1,3]

[1] Biometris, Wageningen UR, Postbus 100, 6700 AC Wageningen, The Netherlands; sabine.schnabel@wur.nl (communicating author)
[2] Erasmus MC, Department of Biostatistics, Postbus 2040, 3000 CA Rotterdam, The Netherlands
[3] Center for Biosystems Genomics, Postbus 98, 6700 AB Wageningen, The Netherlands

Abstract: In field trials the development of plants is regularly scored on a visual scale. Plots of the data show strongly curved relationships with time. We investigate optimal scaling of the time axis in order to get linear curves and apply it to decay data of potato plants.

Keywords: Optimal scaling; time axis; linear model

1 Introduction

In plant research field experiments are a common instrument to study the behavior of a plant population under different environmental conditions. However, there are many aspects that contribute to the uncertainty of the data. On the one hand some phenotypic traits for plant development are only registered on an ordinal scale by qualitatively judging the level of development. It is not clear how this translates into a numeric scale used in data analysis. The distances between two subsequent levels of the observation are often not known. On the other hand another data problem might result from the different environmental conditions during a field experiment. Weather cannot be kept constant. Temperature and exposure to daylight are the main factors driving plant development. For better comparison between field experiments in different environments it is necessary to have an uniform and adapted time scale that can capture these differences. In this analysis we will focus on transforming the time axis.

2 Data and objective

During a field experiment in Finland in 2004, haulm senescence of 200 potato varieties was recorded at 11 days (Zaban et al., 2006) on a discrete scale from “green plant” (1), “upper leaves with first signs of yellowing” (2)

etc. to "dead plant" (7). Figure (1.2) shows examples of three varieties. The horizontal axis is not calendar time, but it is beta-thermal time (PBTT) (Yin et al., 1995). This is a scale, developed by plant physiologists, in which the history of daylength and temperature is integrated over the growing season.
To summarize the senescence data for each variety, we would like to fit a simple curve to them, so that only a few clearly interpretable curve characteristics can be carried on to a genetic analysis. The simplest curve is a straight line. We assume that PBTT is a first step in the direction of a linear relationship between ϕ, i.e. transformed PBTT (indicated by τ), and y, the observed scores. The same transformation of PBTT is to be used for all varieties of the population.
We do not consider transformation of the response scale — nor of both scales simultaneously — but we will return to this issue in the Discussion.

3 Theory

Let there be m time points and n varieties. The senescence scores are collected in a matrix $Y = [y_{ij}], i = 1, \ldots, m, j = 1, \ldots, n$. For simplicity we assume Y to be complete. If that is not the case, an appropriate indicator vector can be introduced easily.
Our goal is to find a vector ϕ with m elements, such that ϕ is the optimal transformed PBTT τ. Optimal means that we get the best possible linear correlation between ϕ and each column of Y. This leads to the following objective function to be minimized:

$$S = \sum_j \sum_i (y_{ij} - \alpha_j - \beta_j \phi_i)^2$$

Given ϕ, we are looking for the least squares regression line for each variety. This is an ill-posed problem, because any arbitrary shifting and scaling of ϕ can be compensated by inverse scaling of β_j and shifting of α_j. In order to find the unique solution we want ϕ to be standardized, i.e. $\sum_i \phi_i = 0$ and $\sum_i \phi_i^2 = m$. This is an arbitrary constraint, and it is only used for fitting the model. Afterwards a linear transformation to a more meaningful scale can be applied (see Figure 1 for an example, where we have made minimum and maximum of ϕ equal to those of τ).
An intuitive algorithm repeats the following steps which start with an approximate solution $\tilde{\phi}$:

1. Estimate (new) α_j and β_j by linear regression of column j of Y on $\tilde{\phi}$ for each variety j.

2. Improve $\tilde{\phi}$ by linear regression of $y_{ij} - \alpha_j$ on β_j for each time point i.

In our experience this works well and convergence is obtained in few iterations. As starting values for $\tilde{\phi}$ we take the integers from 1 to m and standardize them.

The actual PBTT times, τ, do not occur in the estimation only their index i does. This is a consequence of the fact that all varieties have been scored on the same days. If this would not be the case, a second matrix $T = [\tau_{ij}]$ would give the actual observation times. Instead of a vector ϕ we would have to estimate a continuous function $f(\tau)$ that gives the transformation at every point in time.
A possible approach is to use B-splines for this purpose: $f(\tau) = \sum_k B_k(\tau)\gamma_k$. The second step of the algorithm above would then involve fitting the B-splines, scaled by b_j, to $y_{ij} - \alpha_j$.

4 Application

Optimal time scaling to linearity seems to be suitable for the present data. The results for the transformation of the scale as well as the linear fit before and after transformation are presented in Figure 1. The model is parsimonious and the estimated coefficients can be directly related to the development process. The slope b_j describes the speed of senescence for variety j. An important other characteristic deduced from the results is the halfway point of the senescence process, i.e. the transformed time at which the score halfway between 1 and 7 is reached.

5 Discussion and outlook

Optimal scaling is a standard tool in psychometric research and practice. To our knowledge it has not been used in plant research yet. As shown we got interesting — but also somewhat worrying — results: linearity is much better on the real time scale (just counting the days after planting) than on PBTT. The transformation we found was almost the inverse of the one from real time to PBTT.
Similar ideas can be used to transform only the senescence scores in relation to time. Technically it also possible to transform both time and scores simultaneously. However, interpretation of the results is unclear. Any monotone transformation of time combines with the corresponding inverse transform of the scores. There is a fundamental identification problem, for which we have no solution yet.
We did not show it in the examples, but a number of varieties show early saturation at the highest possible score. This leads to an S-shaped curve. No time transformation can accommodate that. Our next step will be to introduce a standard S-shaped curve –like the logistic function– with $a_j + b_j\phi_i$ as its argument. This is similar to the link function in generalized linear models.
A still more ambitious effort will be to investigate time transformations including cumulative solar radiation and temperature as explanatory variables.

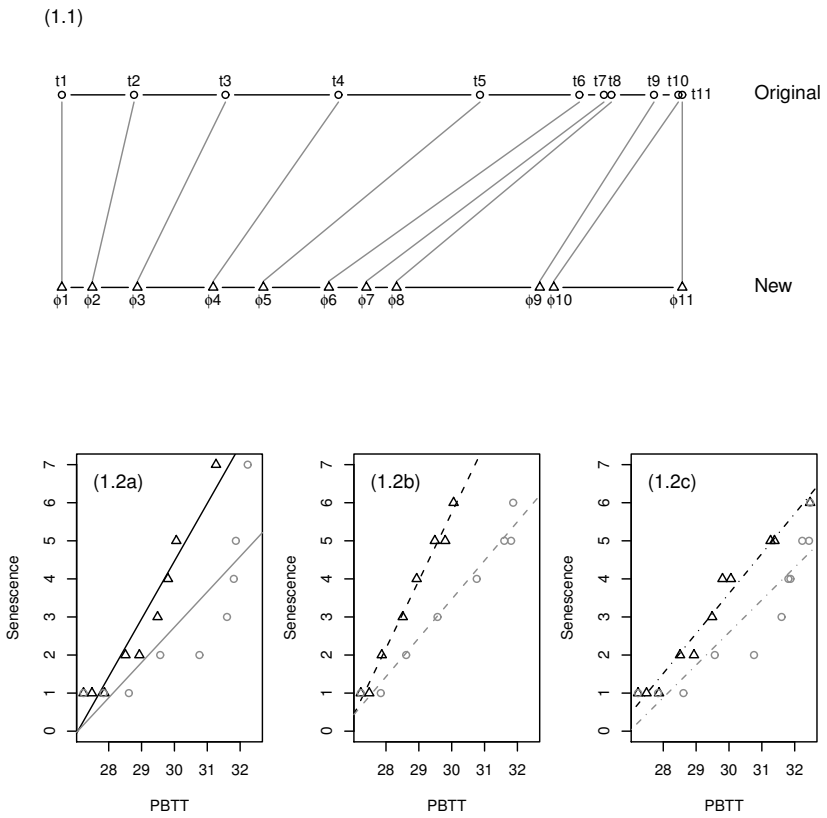


FIGURE 1. (1.1) Original scale versus new scale retransformed to original range. (1.2a-c) Original data (grey) and data on transformed time (black). Linear fit before (grey) and after transformation (black) for three selected potato varieties.

References

Yin X., Kropff M.J., McLaren G., and Visperas R.M. (1995) A nonlinear model for crop development as a function of temperature. *Agricultural and Forest Meteorology*, **77**, 1- 16.

Zaban A., Veteläinen M., Celis-Gamboa C., van Berloo R., Häggman H., and Visser R.G.F. (2006) Physiological and genetic aspects of a diploid potato population in the Netherlands and Northern Finland. *Maat-aloustieteen Päivät*, 1-7.

Introducing a Model to Determine True Counts via the Conway-Maxwell-Poisson Distribution

Kimberly F. Sellers[1]

[1] 306 St. Mary's Hall; Department of Mathematics and Statistics; Georgetown University; Washington, DC 20057; kfs7@georgetown.edu

Abstract: Under-reporting of Poisson counts has been widely studied, given concerns of respondents either forgetting or purposefully neglecting to provide accurate count information. One can likewise argue, however, that a respondent may overestimate a count for some reason (e.g. counting over a larger reference period than of interest). Thus, we need a means by which to determine a "true count" based on recorded data that presumably contains dispersion caused by the count bias associated with misreporting (either via under- or over-reporting) information. We consider a model based on the Conway-Maxwell-Poisson distribution that incorporates such general dispersion to serve as a flexible alternative for modeling report bias and determining true count estimation.

Keywords: Conway-Maxwell-Poisson distribution; report bias; under-reporting.

1 Introduction

"Underreported count data are generated when only a fraction of the actual events of interest (e.g. purchases) are reported" (Fader and Hardie, 2000). The problem of underreporting is pervasive in any arena involving data collection. Many surveying agencies contend with survey respondents who forget or intentionally neglect to report a true event count. This has a detrimental effect in that associated inferences are based on inaccurate information. Because this is such a significant problem in the statistics arena, there exists several proposed methodological approaches to adjust under-reported counts to obtain true estimated count information. The works of Winkelmann and Zimmermann (1995) and Fader and Hardie (2000), for example, both model under-reporting via a Poisson model for the true counts, and a Binomial distribution to represent the conditional distribution of recorded counts given knowledge of the true count. This formulation seems quite natural for describing under-reporting in that the success probability associated with the Binomial distribution denotes an individual's reporting rate, where one views the associated reporting as a sum of Bernoulli trials over the true response count. While the details of their works differ beyond this point, both develop a marginal distribution for the reported counts

that is overdispersed. Still focusing on under-reported counts, Neubauer and Djuras (2009) instead model the reported counts directly using a generalized Poisson distribution, recognizing that the mean-variance relationship can result in over- or under-dispersion depending on whether or not underlying parameters are presumed random.

One can likewise argue, however, that a respondent may overestimate a count for some reason (e.g. counting over a larger reference period than of interest). This introduces the broader question of misreporting, i.e. under- or over-reporting of data. Li et al. (2003) address this by assuming the true count to have a negative binomial distribution (and thus be overdispersed), while the observed count is (conditionally) Poisson distributed (conditional on the true count). Pararai et al. (2010) meanwhile use an approach similar to Li et al. (2003) where the reported counts are still assumed to be (conditionally) Poisson distributed, however the authors instead model the true counts via the generalized Poisson distribution.

We consider the broader problem of misreporting in a manner that borrows strength from both ideologies. Maintaining the belief that the true count information can be modeled as a Poisson distribution, we assume that the reported count information demonstrates its reporting bias (be it under- or over-counting) through its associated dispersion in the distribution. Thus, we need a flexible model distribution that can describe the misreporting in order to determine a more accurate assessment or "true count" based on the recorded information. Below, we introduce the Conway-Maxwell-Poisson (COM-Poisson) distribution and the sum of COM-Poisson (sCOM-Poisson) distributions as motivators for establishing this model. The broader consideration of misreporting, in turn, will produce a marginal distribution that exhibits over- or under-dispersion.

2 The COM-Poisson distribution

The Conway-Maxwell-Poisson (COM-Poisson) distribution is a general count distribution that relaxes the equidispersion assumption of the Poisson distribution. The COM-Poisson probability mass function (pmf) is

$$P(Y_i = y_i) = \frac{\lambda_i^{y_i}}{(y_i!)^\nu Z(\lambda_i, \nu)}, \quad \nu \geq 0, \quad y_i = 0, 1, 2, \ldots, \quad i = 1, \ldots, n \quad (1)$$

where $\lambda_i = E(Y_i^\nu)$, and $Z(\lambda_i, \nu) = \sum_{s=0}^{\infty} \frac{\lambda_i^s}{(s!)^\nu}$ is the normalizing constant. This distribution includes the Poisson ($\nu = 1$), geometric ($\nu = 0, \lambda_i < 1$), and Bernoulli $\left(\nu \to \infty \text{ with probability } \frac{\lambda_i}{1+\lambda_i}\right)$ distributions as special cases. Statistical properties are discussed in Shmueli et al. (2005).

2.1 Sum of COM-Poisson random variables

The sum distribution of Conway-Maxwell-Poissons (sCOM-Poisson(λ, ν, n)) is likewise a general count distribution whose pmf is

$$P(\mathcal{Y} = y) = \frac{\lambda^y}{(y!)^\nu Z^n(\lambda, \nu)} \sum_{\substack{a_1,\ldots,a_n=0 \\ a_1+\ldots+a_n=y}}^{y} \binom{y}{a_1, \cdots, a_{n-1}, a_n}^\nu, y = 0, 1, 2, \ldots,$$

for $\mathcal{Y} = \sum_{i=1}^n Y_i$, where $Y_i \sim$ COM-Poisson(λ, ν) are independent and identically distributed, $Z^n(\lambda, \nu)$ is the nth power of $Z(\lambda, \nu)$, and $\binom{y}{a_1, a_2, \cdots, a_n}$is a multinomial coefficient. The sCOM-Poisson distribution encompasses the Poisson$(n\lambda)$ distribution (for $\nu = 1$), negative binomial$(n, 1-\lambda)$ distribution (for $\nu = 0$ and $\lambda < 1$), and Binomial$\left(n, \frac{\lambda}{\lambda+1}\right)$ distribution (as $\nu \to \infty$) as special cases. Further, for $n = 1$, the sCOM-Poisson$(\lambda, \nu, n = 1)$ is simply the COM-Poisson(λ, ν) distribution. This distribution thus allows us to generalize the ideas of Winkelmann and Zimmermann (1995) and Fader and Hardie (2000), who use a Binomial distribution to model the reported counts in the case of under-reporting, as well as the analogous notion that a negative binomial distribution can model reported counts if over-reporting is assumed.

3 The model and its properties

Let $N_i = n_i$ and $Y_i = y_i$ equal the true and reported counts respectively for individual i, and assume that N_i conditional on covariates $\mathbf{x}_i$ is Poisson distributed with mean $E(N_i \mid \mathbf{x}_i) = \kappa_i = \exp(\mathbf{x}_i'\beta)$ and

- $Y_i \mid N_i > 0 \sim$ sCOM-Poisson$(\lambda_i = \exp(\mathbf{x}_i'\gamma), \nu, n_i)$, and
- $Y_i \mid N_i = 0 \sim$ COM-Poisson$(\mu_i = \exp(\mathbf{x}_i'\delta), \nu_0)$.

Accordingly, the marginal distribution for the recorded count is

$$\begin{aligned} P(Y_i = y_i) &= \frac{\lambda_i^{y_i} e^{-\kappa_i}}{(y_i!)^\nu} \sum_{n_i=1}^{\infty} \left\{ \frac{\kappa_i^{n_i}}{n_i! Z^{n_i}(\lambda_i, \nu)} \sum_{\substack{a_1,\ldots,a_{n_i}=0 \\ a_1+\ldots+a_{n_i}=y_i}}^{y_i} \binom{y_i}{a_1, \ldots, a_{n_i}}^\nu \right\} \\ &\quad + \frac{\mu_i^{y_i} e^{-\kappa_i}}{(y_i!)^{\nu_0} Z(\mu_i, \nu_0)}, \end{aligned} \tag{2}$$

where $\kappa_i = \exp(\mathbf{x}_i'\beta)$, $\lambda_i = \exp(\mathbf{x}_i'\gamma)$, and $\mu_i = \exp(\mathbf{x}_i'\delta)$; we denote Equation (2) as $f(y_i; \beta, \gamma, \delta, \nu, \nu_0)$. This distribution has an associated mean,

$$E(Y_i) = \mu_i e^{-\kappa_i} \left(\frac{\partial \log Z(\mu_i, \nu_0)}{\partial \log \mu_i} \right) + \kappa_i \left(\frac{\partial \log Z(\lambda_i, \nu)}{\partial \log \lambda_i} \right), \tag{3}$$

and variance,

$$\begin{aligned} V(Y_i) &= \mu_i e^{-\kappa_i}\left(\frac{\partial^2 \log Z(\mu_i,\nu_0)}{\partial\mu_i \partial \log \mu_i}\right) + \kappa_i \left(\frac{\partial^2 \log Z(\lambda_i,\nu)}{\partial(\log\lambda_i)^2}\right) \\ &\quad + e^{-\kappa_i}(1-\mu_i^2 e^{-\kappa_i})\left(\frac{\partial \log Z(\mu_i,\nu_0)}{\partial \log \mu_i}\right)^2 \\ &\quad + \kappa_i \left(\frac{\partial \log Z(\lambda_i,\nu)}{\partial \log \lambda_i}\right)\left[\left(\frac{\partial \log Z(\lambda_i,\nu)}{\partial \log \lambda_i}\right) - 2\mu_i e^{-\kappa_i}\left(\frac{\partial \log Z(\mu_i,\nu_0)}{\partial \log \mu_i}\right)\right]. \end{aligned} \tag{4}$$

3.1 Parameter estimation

With the model established, we will estimate the parameters, $\beta, \gamma, \delta, \nu, \nu_0$, via maximum likelihood estimation. Considering the log-likelihood equation, $\log L(\beta,\gamma,\delta,\nu,\nu_0;\mathbf{y}) = \sum_{i=1}^{n} \log f(y_i;\beta,\gamma,\delta,\nu,\nu_0)$, where $f(y_i;\beta,\gamma,\delta,\nu,\nu_0)$ is provided in Equation (2), we use a bounded nonlinear optimization tool to determine the maximum likelihood estimates, $\hat{\beta},\hat{\gamma},\hat{\delta},\hat{\nu},\hat{\nu}_0$.

4 Motivating example

The Bureau of Labor Statistics (BLS) collects data from the Survey of Occupational Injuries and Illnesses (SOII), which is a federal and state program where employers annually report the number of workplace injuries and illnesses in a calendar year. The information (including average annual employment size; total recordable cases; and total cases with days away from work, job transfer, or restriction) is collected from employers' respective Occupational Safety and Health Administration (OSHA) logs and transferred to their SOII survey, along with other relevant survey information requested by SOII. As a result, SOII provides the most complete information regarding injuries and illness throughout the country, as well as at the state level.

While SOII represents a comprehensive compilation of injury and illness data across different industries, the BLS is concerned that the SOII undercounts the number of workplace illnesses and injuries. While the amount of believed discrepancy varies, Ruser (2008) notes that the causes of the underreporting in the SOII dataset include the failure to count illnesses that have a latency period, injuries and illnesses incurred by out-of-scope workers, injuries and illnesses that are included in other data systems (e.g. workers' compensation), and injuries and/or illnesses that are not included anywhere. Some of these factors will always exist, given the construct of the SOII survey; nonetheless, the BLS considers various methods to resolve these concerns in order to provide accurate information.

We will work to apply the model described in Section 3 to capture and describe the level of misreporting that exists in the SOII dataset, and any relationship associated with the misreporting. Through the model estimates, we will then be able to describe the amount of count data misreporting. Further (time permitting), we will consider and compare this approach with alternative methods cited to adjust for misreporting.

5 Discussion

This work introduces a model to estimate true count information stemming from a survey that contains misreported count data. Here, we assume that the true count distribution is represented via a Poisson distribution. To allow for added flexibility, however, one may likewise consider a COM-Poisson distribution to represent the true distribution of counts, arguing that the true count distribution contains some underlying form of data dispersion.

Acknowledgments: This research is supported in part by the ASA/NSF/BLS Research Fellowship Program. The views expressed here are those of the author and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

References

Fader, P. and Hardie, B. (2000). A note on modeling underreporting Poisson counts. *Journal of Applied Statistics*, **27**, 953-964.

Li, T., Trivedi, P. K., and Guo, J. (2003). Modeling response bias in count: a structural approach with an application to the National Crime Victimization Survey data. *Sociological Methods and Research*, **31**, 514-544.

Neubauer, G. and Djuras, G. (2009). A beta-Poisson model for underreporting. In: *Proceedings of the 24th International Workshop on Statistical Modelling*. 255-260, Ithaca, NY.

Pararai, M., Famoye, F., and Lee, C. (2010). Generalized Poisson-Poisson mixture model for misreported counts with an application to smoking data. *Journal of Data Science*, **8**, 607-617.

Ruser, J.W. (2008). Examining evidence on whether BLS undercounts workplace injuries and illnesses. *Monthly Labor Review*, 20-32.

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

Winkelmann, R. and Zimmermann, K. F. (1995). Recent developments in count data modelling: Theory and application. *Journal of Economic Surveys*, **9**, 1-24.

Fast genome-wide association analysis in longitudinal studies

Karolina Sikorska[1], Patrick Groenen[2], Fernando Rivadeneira[3], Paul Eilers[1] , Emmanuel Lesaffre[1,4]

[1] Department of Biostatistics, Erasmus Medical Centre, Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands (k.sikorska@erasmusmc.nl, p.eilers@erasmusmc.nl, e.lesaffre@erasmusmc.nl)
[2] Econometric Institute, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (groenen@ese.eur.nl)
[3] Departments of Internal Medicine and Epidemiology, Erasmus Medical Centre, P.O. Box 2040, 3000 CA Rotterdam Rotterdam, The Netherlands (f.rivadeneira@erasmusmc.nl)
[4] L-Biostat, Catholic University of Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium

Abstract: We propose computational fast approaches for GWA analysis in case of longitudinal studies. We explored their performance with a simulation study and a practical example

Keywords: genome-wide association; longitudinal design; linear mixed model.

1 Introduction

The gross of genome-wide association (GWA) studies have focused on associations with cross-sectionally measured phenotypes. However it may be desirable to identify genetic variations that are associated with the longitudinal development of a trait over time. This requires a longitudinal study in order to characterize within-individual changes of the considered phenotype. A popular approach to analyze repeated measures is given by the linear mixed model (LMM). Unfortunately this technique can be computationally demanding when many subjects are involved and becomes prohibitively time consuming when it has to be executed a large number of times. The Generalized Estimating Equations (GEE) approach is an alternative which is computationally less demanding. However this method requires the specification of a so called working correlation matrix and does not protect against a missing at random (MAR) process. In addition fitting a few millions of models using the GEE approach can still require a too long computational time. We investigated three fast alternatives: 1) the slope as outcome approach, 2) the two-step and 3) the conditional two-step approach. Those three methods as well as the GEE approach with various working correlation structures were explored as possible alternatives of the

linear mixed model. We evaluated their accuracy and necessary computational time.
Rivadeneira et al. (2009) identified several genetic variations associated with cross-sectional bone mineral density (BMD) values using data coming from the Rotterdam Study (Hofman, A. et al. 2009). In this prospective population-based cohort study the BMD from 4987 individuals was measured at baseline and then after 2 and 6 years. However not all 3 measurements were available for every individual. Of interest is to identify SNPs associated with the evolution of BMD over time. Computational time for fitting 500K LMMs (for 500K SNPs) was estimated as more than 120 hours.

2 Statistical approaches

Let Y_{ij} be a continuous variable measured for individual i (belonging to SNP group S_i) at time t_{ij} ($i = 1, ..., N$, $j = 1, ..., k$). We are interested in the effect of each of the SNPs on the evolution over time of the response Y_{ij}. We considered the following approaches. The **LMM** of interest is given by:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + \beta_3 t_{ij} S_i + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \tag{1}$$

where $b_{0i} \sim \mathcal{N}(0, \sigma_0^2)$, $b_{1i} \sim \mathcal{N}(0, \sigma_1^2)$ ($corr(b_{0i}, b_{1i}) = \rho$) represent subject-specific intercept and slope respectively and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the residual term. Additionally, $b_1, ..., b_N, \epsilon_1, ..., \epsilon_N$ are assumed to be mutually independent. The model is fitted using a full likelihood approach. Furthermore the marginal model

$$E(Y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + \beta_3 t_{ij} S_i \tag{2}$$

is fitted using the GEE approach based on quasi-likelihood, which makes the computation time shorter. We are mainly interested in testing if there is a statistically significant effect of SNP (S_i) on the evolution of Y_{ij}. This is verified by testing $H_0 : \beta_3 = 0$ resulting in the corresponding p-value. We explored three alternative methods that could provide approximately the same p-value in a much shorter time. Those methods split the analysis into steps in order to avoid fitting the full model (1) repeatedly for the different SNPs. The **slope as outcome method** is based on two-stage formulation of LMM. In the first stage the slope $\beta_{1i}^{\triangle}$ per individual is estimated according to the model:

$$Y_{ij} = \beta_{0i}^{\triangle} + \beta_{1i}^{\triangle} t_{ij} + \epsilon_{ij}^{\triangle}. \tag{3}$$

In the second stage the estimated $\hat{\beta}_{1i}^{\triangle}$'s are regressed on S_i using ordinary least squares method of estimation On the other hand in the first step of the **two-step approach** all terms containing S_i are omitted from model (1), so the model becomes:

$$Y_{ij} = \beta_0^* + \beta_1^* t_{ij} + b_{0i}^* + b_{1i}^* t_{ij} + \epsilon_{ij}^* \tag{4}$$

In the second step we regress BLUPS of $\hat{b}^*_{1i}$ on S_i with a simple linear regression model:

$$\hat{b}^*_{1i} = \beta^{**}_0 + \beta^{**}_1 S_i + \epsilon^{**}_i \tag{5}$$

If Y_i denotes a vector of k_i measurements taken on the i-th individual, the LMM can be reformulated as

$$Y_i = X^{(1)}_i \beta^{(1)} + X^{(2)}_i \beta^{(2)} + Z^{(1)}_i b_{i0} + Z^{(2)}_i b_{i1} + \epsilon_i, \tag{6}$$

where $X^{(1)}_i$ and $X^{(2)}_i$ are the matrices of time stationary and time-varying covariates respectively. As given in Verbeke et al. (2001) the original data satisfying (6) can be transformed into

$$y^*_i \equiv A^T_i y_i = A^T_i X^{(2)}_i \beta^{(2)} + A^T_i Z^{(2)}_i b^{(2)}_i + A^T_i \epsilon_i = X^*_i \beta^{(2)} + Z^*_i b^{(2)}_i + \epsilon^*_i, \tag{7}$$

where A_i is $k_i \times (k_i - 1)$ matrix A_i such as $A^T_i \mathbf{1_{n_i}} = 0$. All the fixed cross-sectional effects as well as random intercepts have been vanished from the model (6). **The conditional two-step approach** is performed in the same way as the previously explained two-step, but on the data satisfying (7).

TABLE 1. Simulated balanced data. Approximate system time for 500K models (CPU 2.99 GHz, 3.21 GB of RAM).

Method	System time
linear mixed model	$\approx$ 130h
GEE-unstructured	$\approx$ 107h
GEE-exchangeable	$\approx$ 52h
GEE-fixed	$\approx$ 17h
GEE-independence	$\approx$ 24h
slope as outcome	$\approx$ 45min
two-step	$\approx$ 45min
conditional two-step	$\approx$ 45min

3 Simulation study and application to BMD data

To assess precision and computational time of proposed methods in comparison to full mixed model approach we conducted a simulation study for balanced and unbalanced scenarios (MCAR and MAR dropout). For each scenario we generated 200 data sets for 2000 individuals according to model (1). The p-values for the SNP*time interaction term obtained from the LMM were compared to the corresponding p-values from the GEE approach (with various choices of the working correlation matrix) and the three fast alternatives. The resulting plots for balanced scenario and MCAR

dropout (similar to MAR dropout) are given in Figure 1. Estimated computational times for the GWA scan (500K SNPs) using each of the method are given in Table 1. The simulation study showed that the conditional two-step approach is the most accurate method in the presence of missing responses.

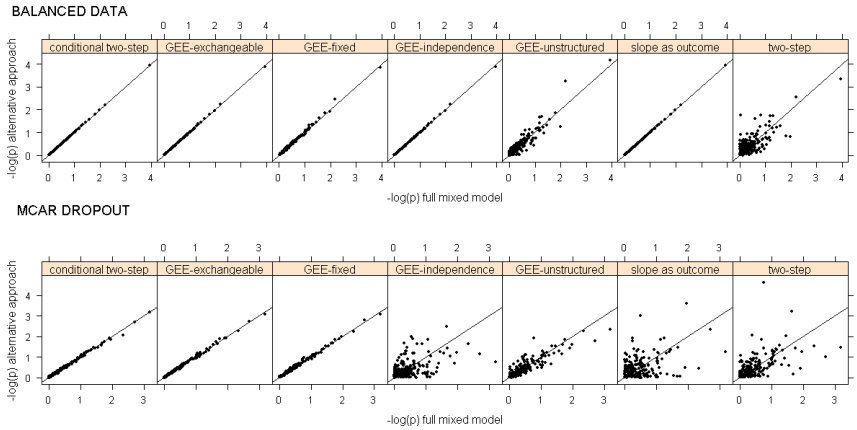


FIGURE 1. Simulation study for 2 different scenarios

We conducted GWA analysis of BMD data for $\approx$ 500K SNPs using the conditional two-step approach. None of the SNPs reached genomewide significance level ($p < 5 \times 10^{-8}$).

References

Hofman, A., Breteler, M., van Duijn, C., Janssen, H. et al. (2009). The Rotterdam Study: 2010 objectives and design update. *European Journal of Epidemiology*, **24(9)**, 553-572.

Rivadeneira, F. et al. (2009). Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies *Nature Genetics*, **41**, 1199-1206.

Verbeke, G., Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag New York, Inc.

Verbeke, G., Spiessens, B., Lesaffre, E. (2001). Conditional Linear Mixed Models *The American Statistician*, **Vol. 55, No. 1**, 25-34.

Analysis of Gene Duplication Data

Sarabdeep Singh[1] , S. Huzurbazar[1]

[1] Department of Statistics, University of Wyoming, Laramie, WY 82071, USA.
email:lata@uwyo.edu (Huzurbazar)

Keywords: evolutionary bioinformatics, measurement error.

1 Problem: Gene duplications and losses

Gene duplication is one of the most dominant driving forces behind genome evolution and the primary source for gene and protein functional evolution. Based on current theory, the most likely fate of any gene duplication or birth is subsequent gene loss or death, where duplicate genes may simply become silenced or nonfunctionalized by deleterious mutations. The other fates of duplicate genes, neofunctionalization, subfunctionalization and dosage compensation, are not of primary importance for this paper. Rates of gene birth and death are not well understood, and the main focus here is the estimation of these rates, which are key for understanding genome evolution as well as protein function evolution.
Past analyses of such gene duplication and loss data (Lynch and Conery, 2000 and Hughes and Liberles, 2007) have several shortcomings. In the sections that follow, we describe the data and how they are obtained, the main shortcomings of previous analyses and our approach for addressing them. We present results which have been obtained to date, and the next steps in our data modelling and analysis.

2 Data

To investigate the rates of gene birth and death, the primary data consist of estimates of time since genes duplicated, commonly denoted by dS. The dS estimates are of synonymous substitutions per synonymous sites within the DNA sequence of a species; these are silent mutations in the DNA sequences which do not alter the codon. To obtain these estimates, a long data generation process is undertaken. For a given species, the initial dataset of nucleotide or DNA sequences and their corresponding amino acid sequences are obtained from whole genome sequencing. The sequences are then filtered through different bioinformatics software packages. First, the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) is used to identify duplicate pairs within a genome. Next, the Multiple Sequence Comparison by Log-Expectation (MUSCLE) (Edger, 2004) package aligns

the sequences, and finally, the Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang, 1997) software allows for estimation of dS between any two sequence pairs. Within each software, a variety of decisions are made, all of which can affect the final dS data to some extent. One of our goals is to investigate the effect of these decisions, with the view that measurement error is accumulated at each step. It should also be noted that for various reasons having to do with the underlying biological processes, the final dS values are usually truncated at $dS = 0.3$, which corresponds to 30 million years ago.

The dataset on time since duplication analyzed here is for the *Oikopleura dioica* species. *Oikopleura dioica*, a type of sea squirt, is an organism of much interest in evolutinary genetics due to its rapidly evolving genome. Its genome was sequenced in early 2010, and the final dS data analyzed here were generated by our collaborator, Dr. David Liberles. Some statistical analyses of this data are part of our paper (Denoeud et al., 2010). Histograms for this and other similar datasets, not presented here due to space considerations, show that distribution of the dS values is extremely skewed, a large proportion of dS values are estimated to be zero, and there are various 'bumps' in the histograms indicating the presence of heterogeneity in the distributions.

Previous analyses of such dS data have several shortcomings. First, they used the frequency data as presented in the histograms and modelled the resulting counts with a Poisson distribution, whose mean varied either as an exponential (Lynch and Conery, 2000) or a Weibull (Hughes and Liberles, 2007). The original data and its truncation were ignored, and measurement error was not acknowledged. In our analysis, our first step is to fit a mixture distribution to the original dS values accounting for the heterogeneity and truncation, as well as accounting for the zero values in the data. Our next step is to begin to account for some measurement error by incorporating the likelihood-based standard errors for the dS values as produced by PAML.

3 Model and Data Analysis

3.1 Addressing truncation and heterogeneity with original data

For modeling dS on a continuous scale, we explored various candidate survival distributions. As the support of most survival distributions does not include zero, we used a mixture of a discrete component at $dS = 0$, combined with a continuous component for $dS > 0$. The final model is a 3 component finite mixture of the discrete component, with a mixture of 2 truncated Weibull distributions. The discrete component is modelled by defining a Bernoulli variable $Z(ds) \sim Ber(w_1)$ at $dS = 0$ and heterogeneity in $dS > 0$ is modelled with a mixture of 2 Weibull distributions, each truncated at 0.3; resulting in the following mixture distribution for dS,

$$f_{dS}(ds|w,\kappa,\lambda) = w_1^{z(ds)}\left[(1-w_1)(w_2 g_{dS}(ds|\kappa_1,\lambda_1) + (1-w_2)g_{dS}(ds|\kappa_2,\lambda_2))\right]^{1-z(ds)} \tag{1}$$

where w_2 is the probability corresponding to the mixture of two Weibull distributions. In eq(1) $g_{dS}(ds|\kappa_i, \lambda i)$ is a truncated Weibull distribution given by,

$$g_{dS}(ds|\kappa_i, \lambda_i) = \frac{\kappa_i ds^{\kappa_i - 1} e^{(\lambda_i - e^{\lambda_i} ds^{\kappa_i})}}{1 - e^{-e^{\lambda_i} 0.3^{\kappa_i}}}$$

for $0 < dS < 0.3$, where $\kappa_i > 0$ is the shape parameter and $\lambda_i > 0$ is the scale parameter, for $i = 1, 2$. For $\theta_2 = (w_1, w_2, \kappa_1, \lambda_1, \kappa_2, \lambda_2)$, the likelihood function based on eq(1), combined with the following priors, $P(w_i) \sim Beta(1, 1)$, $P(\kappa_i) \sim Gamma(1, 0.001)$, and $P(\lambda_i) \sim Gamma(1, 0.001)$ for $i = 1, 2$ was used to obtain posteriors for components of θ.
In simulating the posterior distributions, all the parameters were updated using a Metropolis Hastings random walk algorithm with a truncated normal proposal density. The MCMC algorithm was run with 3 independent chains for 2 million iterations, discarding the first 50,000 as burn-in. Trace plots showed good mixing, and standard convergence diagnostics indicated convergence for all the chains. Posterior modes and intervals for the parameters for the *Oikopleura dioca* data are displayed in Table 1.

3.2 Incorporating Measurement Error

The estimated dS values obtained from PAML are MLEs, and estimates of their standard errors are available. A plot of the SE(dS) versus dS values shows a curvilinear relationship where the scatter in the SE values increases with increasing dS values. We modelled this relationship using a non-linear model of the form $SE(dS) = \beta_0 + \beta_1(dS)^{\beta_2} + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$ and using WinBUGS obtained posterior distributions for the parameters $\beta_0, \beta_1, \beta_2$. The posteriors were then parameterized using Gamma distributions, $\beta_0 \sim Gamma(30, 14553)$, $\beta_1 \sim Gamma(3745, 28969)$, $\beta_2 \sim Gamma(2040, 2768)$, which were used as informative priors for the measurement error component of our data model. The resulting hierarchical model for dS is as follows. We denote Y as the observed data (time since genes were duplicated), X as the data generating mechanism (mixture model) defined in eq(1), and $\theta = (\theta_1, \theta_2)$ as the parameters. Then our data model has the following components: 1. Observed data (measurement error) model: $[Y|X, \theta_1] \sim N_T(X, \beta_0 + \beta_1(dS)^{\beta_2})$, truncated at $[0, 0.3]$ where $[\theta_1] = [\beta_0, \beta_1, \beta_2]$., 2. Data generating mechanism model: $[X|\theta_2] \sim$ Mixture distribution, from eq(1)., 3. Parameter model (priors): $[\theta] = [\theta_1]\,[\theta_2]$, with priors as stated above.
The results are sumarized in Table 1 for the *Oikopleura Dioca* data. Accounting for measurement error changes yields narrower posterior intervals but also yields greater separation of the two Weibull distributions. The results presented here only use the measurement error model for the $dS > 0$ data. Work in progress includes incorporating PAML SE information for $dS = 0$, and future work will include a biological process model.

TABLE 1. Posterior Distribution Summaries: with & without measurement error

	Without ME		With ME	
	Mode	(5th, 95th)	Mode	(5th, 95th)
w_1	0.067	(0.05,0.08)	0.067	(0.056,0.08
w_2	0.58	(0.38, 0.70)	0.49	(0.43,0.55)
κ_1	0.75	(0.65,0.83)	0.58	(0.54,0.63)
λ_1	0.37	(0.043 0.98)	0.31	(0.26,0.37)
κ_2	1.37	(1.24,1.46)	1.92	(1.86,1.97)
λ_2	2.6	(2.29,2.63)	3.33	(3.25,3.35)

Acknowledgments: Special Thanks to David A. Liberles and Anke Konrad for weekly discussions. The second author acknowledges the support of funding through NSF-DBI 0743374.

References

Altschul, S.F., Gish,W., Miller,W., Myers, E.W., Lipman,D.J. (1990). "Basic local alignment search tool". *J Mol Biol 215 (3): 403-410*

Denoeud et al., (2010). "Plasticity of Animal Genome Architecture Unmasked by Rapid Evolution of a Pelagic Tunicate". *Science, 330: 1381-1385.*

Edgar R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics 5:113.*

Hughes, T. and Liberles, D.A. (2007). The Pattern of Evolution of Smaller-Scale Gene Duplicates in Mammalian Genomes in More Consistent with Neo- than Subfunctionalism, *J. Mol. Evol. 65:574-588.*

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10:325–337.*

Lynch, M. and Conery, J.S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science, 290: 1151-1155.*

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences 13: 555-556*

Flexible Modelling of Functional Data using Continuous Wavelet Dictionaries

Leen Slaets [1], Gerda Claeskens [1], Maarten Jansen [2]

[1] ORSTAT and Leuven Statistics Research Centre, Katholieke Universiteit Leuven (KUL), Naamsestraat 69, 3000 Leuven - Belgium. leen.slaets@econ.kuleuven.be
[2] Departments of Mathematics and Computer Science, Université libre de Bruxelles (ULB), Boulevard du Triomphe CP213, 1050 Brussels - Belgium.

Abstract: A random effects model for functional data based on continuous wavelet expansions is proposed. It incorporates phase variation without the use of warping functions. Both coarse-scale features and fine-scale information are modelled parsimoniously, yet flexible. The regularity of the estimated function can be controlled, creating a joint framework for Bayesian estimation of smooth as well as spiky and possibly sparse functional data.

Keywords: Functional Data; Continuous Wavelet Dictionaries; Amplitude variation; Phase variation, Sparse Data, Random Effects.

1 Introduction

While functional data have been around for a long time, the availability of methodology recognizing their functional nature and corresponding features has blossomed more recently. For an overview see Ramsey and Silverman (2006). Samples $y_{nj} = y_n(t_j)$ are often encountered when observing a process over a certain time interval (at discrete time points t_j, $j = 1, \ldots, T_n$) for several subjects or instances $n = 1, \ldots, N$. A key element of the functional data framework is the recognition of phase variation (variation in timing of features) as a source of variability in the data, in addition to amplitude variation (variation in amplitude of features). A monotone increasing function transforming the time-axis, called a warping function, is typically used to take phase variation into account, prior to or joint with the analysis of the amplitude. These warping functions behave differently than the actual curves in the sample, complicating a combined analysis and a proper understanding of the total variation as a mixture of the two. With clustering in mind, Liu and Yang (2009) circumvented the warping function by representing the curves as B-splines with randomly shifted basis functions. Along that line we introduce a model which incorporates phase variation in a natural and intuitive way, by avoiding the use of warping functions, while still offering a good and controllable degree of complexity and flexibility. By building a model around wavelet transformations, we

use the location and scale notion of wavelet functions to model phase variation. The wavelet coefficients represent amplitude. Wavelets have already greatly shown their efficiency for the representation of single functions, and it are exactly those strengths that we aim to generalize towards samples of curves. Our methodology differs from that of Morris and Carroll (2006), in that they generalize a classic mixed effects model towards functional data. The discrete wavelet transformation is used to fit their proposed model. Our goal is to use wavelet functions for a direct modelling of the data, not to fit general functional mixed effects models. An additional advantage of using wavelets is that by choosing an appropriate wavelet many types of data can be analyzed, ranging from smooth processes to spiky spectra. The proposed model serves as a basis for a variety of applications, such as (graphical) exploration and representation, clustering and regression with functional responses.

2 Modelling Functional Data by means of Continuous Wavelet Dictionaries

The proposed model is built around a scaling function ϕ and a wavelet function ψ, the latter often forms an orthonormal basis ψ_{jk}, $j, k \in \mathbb{Z}$, by shifting and rescaling the mother wavelet ψ, subject to the dyadic constraints: $\psi_{jk}(t) = 2^{j/2}\psi\left(2^j t - k\right)$. A downside of obtaining orthonormality, is the fact that the functions need to be observed on an equidistant grid of time points. Therefore continuous wavelet transformations, using an overcomplete set of wavelet functions with arbitrary locations and scales, continue to gain popularity. In a functional setting, an overcomplete wavelet dictionary can represent the sample of curves in the following way:

$$\begin{aligned} y_n(t_j) = \quad & \sum_{m=1}^{M} c_{n,m}\sqrt{a_{n,m}}\phi\left(a_{n,m}(t_{n,j} - b_{n,m})\right) \\ & + \sum_{k=M+1}^{M+K} c_{n,k}\sqrt{a_{n,k}}\psi\left(a_{n,k}(t_{n,j} - b_{n,k})\right) + e_{n,j}, \end{aligned} \tag{1}$$

with random scales $a_{n,m}$,$a_{n,k}$, random shifts $b_{n,m}$, $b_{n,k}$, random amplitudes $c_{n,m}$,$c_{n,k}$ and independent random errors $e_{n,j}$. Also $a_{n,k} \geq a_{n,m}$, $\forall m = 1, \ldots, M, k = M+1, \ldots, K$. Denote

$$\begin{aligned} \mathbf{a}_{n,M} &= (a_{n,1}, a_{n,2}, \ldots, a_{n,M}), \quad \mathbf{a}_{n,K} = (a_{n,M+1}, a_{n,M+2}, \ldots, a_{n,M+K}), \\ \mathbf{b}_{n,M} &= (b_{n,1}, b_{n,2}, \ldots, b_{n,M}), \quad \mathbf{b}_{n,K} = (b_{n,M+1}, b_{n,M+2}, \ldots, b_{n,M+K}), \\ \mathbf{c}_{n,M} &= (c_{n,1}, c_{n,2}, \ldots, c_{n,M}), \quad \mathbf{c}_{n,K} = (c_{n,M+1}, c_{n,M+2}, \ldots, c_{n,M+K}), \end{aligned}$$

with the following random effects distributions:

$$\begin{aligned} (\mathbf{a}_{n,M}, \mathbf{b}_{n,M}, \mathbf{c}_{n,M}) &\sim \mathcal{N}_K(\boldsymbol{\mu}_M, \Sigma_M), \quad \text{for } n = 1, \ldots, N \\ (\mathbf{a}_{n,K}, \mathbf{b}_{n,K}, \mathbf{c}_{n,K}) &\sim \mathcal{N}_K(\boldsymbol{\mu}_K, \Sigma_K), \quad \text{for } n = 1, \ldots, N \end{aligned}$$

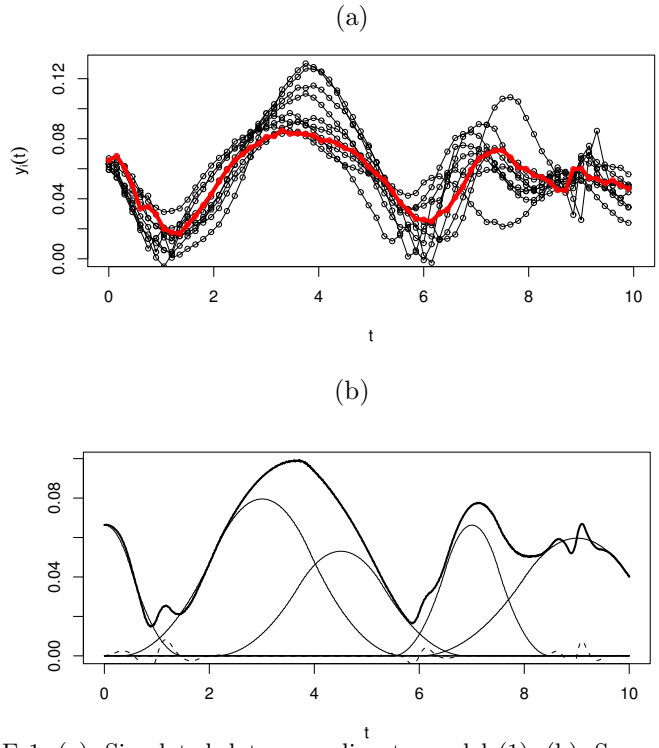


FIGURE 1. (a): Simulated data according to model (1). (b): Sum of scale and wavelet functions corresponding to $\boldsymbol{\mu}_M$ and $\boldsymbol{\mu}_K$ (bold line), separate scale functions (normal lines) and wavelet functions (dashed lines).

$$e_{n,j} \sim \mathcal{N}(0, \sigma^2), \qquad \text{for } n = 1, \ldots, N \text{ and } j = 1, \ldots, T_n,$$

with $\boldsymbol{\mu}_K = (\boldsymbol{\alpha}_K, \boldsymbol{\beta}_K, \boldsymbol{\gamma}_K) = (\alpha_1, \alpha_2, \ldots, \alpha_K, \beta_1, \beta_2, \ldots, \beta_K, \gamma_1, \gamma_2, \ldots, \gamma_K)$ and likewise for M. The index K (and M) refers to the dimensionality of the vector which depends on the number of wavelet functions K (or scale functions M) in expansion (1). While M is a fixed constant, K is a parameter in the model. Figure 1 shows a simulated data example corresponding to model (1) with spline wavelets. It illustrates the intuition behind the model by means of a sample of 10 curves, with $M = 5$ scaling functions, $K = 3$ wavelet functions and a positive correlation between the location of the first scale function and the amplitude of the last scale function. The scale functions corresponding to $\boldsymbol{\mu}_M$ represent the main features in a homogeneous functional data sample (as shown in Figure 1 (b)). The random effects $a_{n,m}$, $b_{n,m}$, $c_{n,m}$ allow for curve-specific deviations in respectively scale, location and amplitude from these average features, while maintaining parsimoniousness. Phase and amplitude variation are thus being modelled in an intuitive way, by means of random scale, location (both

representing phase) and amplitude of the scale functions. The covariance matrix Σ_M explains how the random effects corresponding to a certain feature relate to others. The bold grey line in 1 (a) corresponds to an observation with a late occurrence of the first scaling function and which has a relatively high amplitude of the last scaling function, illustrating the positive correlation. These kind of patterns are often impossible to detect by eye or by more simple methods. In this model there is no need for a fixed or equispaced grid of time points, as continuous wavelets are being used and information is borrowed within and across curves by means of the random wavelet functions. This makes the method suitable for the analysis of sparse data as well. For a single curve y ($N = 1$), model (1) fits the framework introduced in Abramovich et al (1999). They established conditions on the model parameters under which the smoothness of the expansion can be controlled. In the Bayesian framework, Chu et al (2009) do so by an appropriate choice of priors on the model parameters. For the estimation they use a reversible jump Markov chain Monte Carlo algorithm to improve computational efficiency. The ideas in both papers are used here. In case the data are heterogeneous, the model can be used for a clustering procedure following a k-centers type algorithm. The model can also be extended by incorporating additional covariates, giving rise to a regression model with functional responses. In summary, we create a framework to analyze many different types of functional data (smooth, spiky, sparse), while still being flexible and easy to understand, estimate and use.

References

Abramovich, F., Sapatinas, T. and Silverman, B.W. (2000). Stochastic Expansions in an Overcomplete Wavelet Dictionary. *Probability Theory and Related Fields*, **117**, p.133-144.

Chu, J.-H., Clyde, M.A. and Liang, F. (2009). Bayesian Function Estimation using Continuous Wavelet Dictionaries. *Statistica Sinica*, **19** , p.1419-1438.

Liu, X. and Yang, M.C.K. (2009). *Simultaneous Curve Registration and Clustering for Functional Data*, Computational Statistics and Data Analysis, **53** , p.1361-1376.

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

Ramsay, J.O. and Silverman, B.W. (2006). *Functional Data Analysis*. New York: Spinger.

Boundary identification in 3D images

Joanna Smith[1], Adrian Bowman[1]

[1] Department of Statistics, University of Glasgow, G12 8QQ
j.smith@stats.gla.ac.uk

Keywords: Shape analysis; principal curves; breast reconstruction; surface representation.

1 Introduction

3D images captured by stereo-photogrammetry sometimes contain areas of the surface that are not required for analysis, therefore it can be useful to find a means to extract only the area which contains the feature of interest. Our data consists of a set of such images from 44 women, who have all undergone a unilateral mastectomy and reconstruction procedure. It is of interest to determine whether the asymmetry between the two breasts is more severe than would normally be seen in the wider population. However, the images all include varying amounts of chest wall as well as the breast tissue itself, so it is necessary to find where the boundary of each breast lies in order to extract it from the rest of the image and analyse it independently.

2 Examining Surface Curvature

A first step in doing so is to examine the curvature of the surfaces, by calculating the principal curvature scores at each point. It was found that the boundary of the breast is characterised by high values of minimum curvature, due to the concavity of the surface where the breast tissue meets the chest wall. The aim is to use this information to try to extract the breasts from the remainder of the surface, by means of a radial line algorithm.

3 Calculating the boundary

3.1 Fitting the principal curve

Although our surface is a 2-dimensional manifold, it is possible to simplify the problem and work in lower dimensions. We can do so by taking transects in various directions across the surface to give us a series of plane curves - curves which are contained in a two-dimensional plane. It is then

a far simpler calculation to analyse the curvature of these plane curves, as opposed to the curvature of the surface.
To find these curves, the first step was to calculate a local axes system at *prom*, the most prominent point on each breast surface, and use this to find the strip of points which lie in a certain given direction (i.e. points lying along the x-axis). We could then examine the curvature of this set of points by fitting a principal curve to them. Principal curves are defined by Hastie and Stuetzle (1989) as smooth one-dimensional curves that pass through the middle of a data set. They minimise the orthogonal distance to all points subject to certain smoothing constraints, and are self-consistent, meaning that each point on the curve is the average of all the points that project there.
Due to the way in which our strip of points is found, there is very little variation in one direction (as all points were within a very small threshold distance on the y-axis). This means that we can simplify the calculation by considering the points in two dimensions only and fit a principal curve using the x and z coordinates alone.

3.2 Calculating the Curvature

We can now calculate the curvature at each point on this curve in order to determine where the boundary may lie. As the principal curve is parameterised by arc length s, the curvature $\kappa(s)$ can be calculated using the standard formula

$$\kappa(s) = \frac{x'y'' - x''y'}{(x'^2 + y'^2)^{3/2}}, \tag{1}$$

where $x' = \frac{dx(s)}{ds}$ and $y' = \frac{dy(s)}{ds}$. The calculated curvature scores for our strip of points can be seen in Figure 1. As we are interested in finding areas where the surface is concave and has a high minimum curvature, we wish to look for peaks in the curvature function.
As can be seen in Figure 1(a), it is possible for there to be more than one peak in the curvature. This occurs when there are other concave areas on the strip, for example if there are ridges or dents on the breast surface. Once the peaks in curvature have been found, the corresponding three-dimensional points can be taken to be candidate points for where the boundary may lie, as shown in Figure 1(b).
This process can then be repeated to find strips in different directions, and subsequently the potential boundary points lying on them, simply by rotating the local-axes system. Repeating this process through 2π radians gives us a set of candidate points all around the perimeter of the breast.

3.3 Point Selection

As mentioned previously, there is not always just one possible point found on each strip. Therefore it is necessary to find a way in which to select

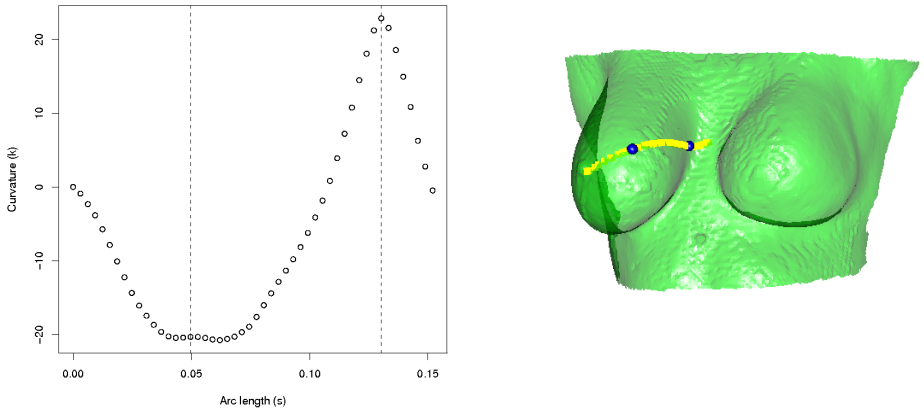


FIGURE 1. (a) Plot of curvature scores against arc length, with peaks in curvature shown by dashed lines. (b) The corresponding points on the surface, along with the strip of points to which the curve was fitted.

the correct points to use. We have several criteria to help us assess which points should be included and excluded from our boundary. Firstly, we can exclude any points which lie too close to *prom*. We can also use the knowledge that the boundary will lie reasonably close to the chest wall, and therefore remove points that lie too far away from this plane.
At this stage it is possible that we will have a combination of strips with multiple points and strips with single points, as well as having some strips with no candidate points at all (if the only candidate found was too close to *prom*, for example). We now need to produce a boundary from these sets of points. To do so, a principal curve was fitted to the single points only and this curve was used to assess which of the points should be included from the other strips (by assessing which points most retained the smoothness of the curve). The curve was also interpolated to find where the boundary should lie in missing positions, where necessary. An example of the boundaries found for two of our images can be seen in Figure 2.

4 Creating a surface model

Once these boundaries have been found, it is desirable to create a set of corresponding points across all breasts in order to make them more easily comparable. We wish to have the same number of points which are in corresponding positions across all surfaces, and this was done using the radial lines that had been calculated for finding the boundary. As there were a variety of shapes and sizes of breasts, it was thought that taking points at proportionally equal distances along these lines would give us a set

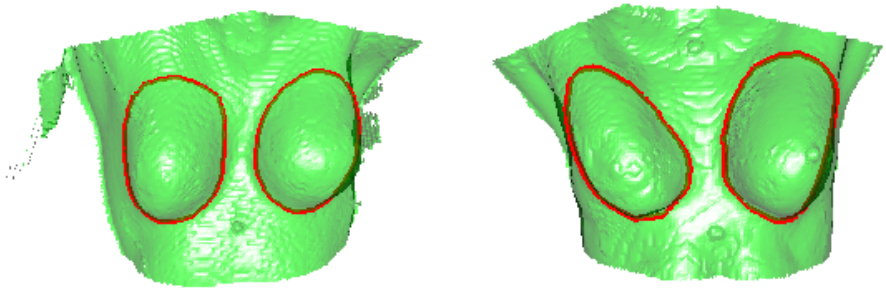


FIGURE 2. The calculated boundaries for two of our patients

of correspondingly placed points on all breasts. As all lines converge at the most prominent point, it was necessary to select points more sparsely at this end of the strip to achieve a more regular spacing. This was done by using a sequence constructed from equally spaced quantiles of the exponential distribution, in order to ensure an increasing distance between points as you progressed along the strip. A smoothing spline was then used to fit a smooth curve to the observed points, and this allowed us to predict the coordinates at these set distances along the curve. An example can be seen in Figure 3. The process was repeated across all strips in order to obtain a representation of the entire breast surface, an example of which can be seen in Figure 4.

5 Discussion

The algorithm can detect the boundary very well in many cases and works particularly well in larger chests where the curvature around the edge of the breast is very strongly defined. However, there are several cases which are more problematic. In smaller breasted women the surface is much flatter and the curvature information sometimes simply isn't there. This can lead to a lack of fixed points and a greater amount of predictions, which aren't necessarily reliable when based on a small number of points. Due to the fact that the shapes of the breasts are so varied, it is difficult to build a fully automated system for detecting the boundary. However, by manually adjusting various thresholds where necessary we were able to produce a set of boundaries which we feel are acceptable for all patients and capture the outline of the breasts well.

As our surface representations consist of a smaller number of points than the original surface the ease of analysis is improved. The corresponding nature of the points is also useful. For example, we can treat these points

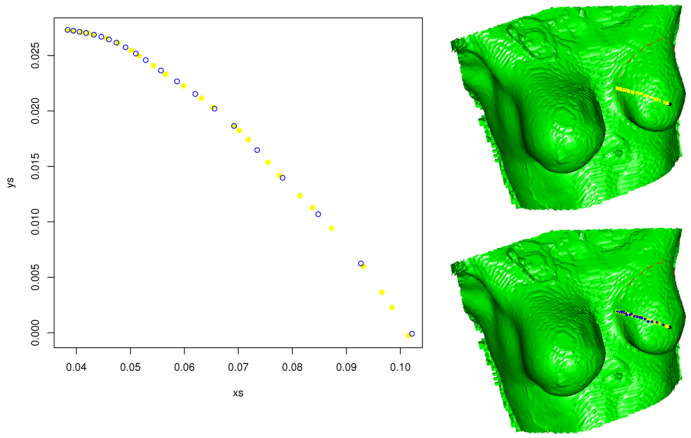


FIGURE 3. The strip of points in a particular direction (grey) and the selected points at standardised distances along the curve (black), in both a 2-dimensional representation (a) and on the original 3-dimensional image (b). (c) The representative surface points selected for our patient.

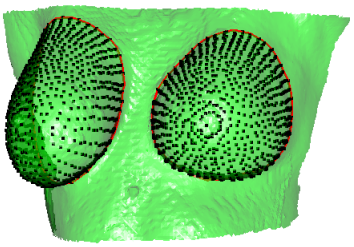


FIGURE 4. An example of the set of points taken to be representative of the breast surfaces.

as landmarks (points on each object which match between and within populations) and go on to investigate the asymmetry of the surfaces using our existing landmark methods.

Acknowledgments: Many thanks to Dr. Helga Henseler, Prof. Ashraf Ayoub and Dr. Balvinder Khambay from the Glasgow Dental Hospital for the collection and provision of data.

References

Goldfeather, J. and Interrante, V (2004). A novel cubic-order algorithm for approximating principal direction vectors. *ACM Transactions on Graphics 23*, **(1)**, 45-63.

Hastie, T. and Stuetzle, W. (1989). Principal Curves. *Journal of the American Statistical Association 84* **(406)**, 502-516.

Miller, J. (2009). *Shape Curvature Analysis Using Curvature*. Ph. D. thesis, University of Glasgow.

Spivak, M. (1979). *Differential Geometry*, Volume 2. Publish or Perish, Inc.

Confidence intervals for geoadditive expectile regression models

Fabian Sobotka[1], Thomas Kneib[1], Göran Kauermann[2]

[1] Department of Mathematics, Carl von Ossietzky University Oldenburg, 26111 Oldenburg, Germany, {fabian.sobotka, thomas.kneib}@uni-oldenburg.de
[2] Department of Economics and Business Administration, University Bielefeld, 33501 Bielefeld, Germany, gkauermann@wiwi.uni-bielefeld.de

Abstract: While a simple mean regression attempts to describe the expectation of a response as a function of the covariates, the results of a quantile or expectile regression offer a much broader view. In principle, a dense set of expectiles or quantiles allows for an analysis of the complete conditional distribution of the response. This can lead to new insight into the dependency between the response and its covariates. In our work, we allow for additive regression models with nonlinear as well as spatial effects. Further, we aim to construct pointwise confidence intervals for each fitted expectile. These shall return a clue about the precision of the estimated expectile curve and therefore into the amount of information that can be drawn from the expectiles. The methodological results are then applied to data about childhood malnutrition in India.

Keywords: Expectiles; Geoadditive Regression; Confidence Intervals; Bootstrap; P-splines.

1 Introduction

Quantile regression has emerged into one of the standard tools for regression analysis that enables a proper assessment of the complete conditional distribution of responses even in the presence of heteroscedastic errors. Quantile regression estimates are obtained by minimising an asymmetrically weighted sum of absolute deviations from the regression line, a decision theoretic formulation of the estimation problem that avoids a full specification of the error term distribution (Koenker, 2005). Recent advances in mean regression have concentrated on making the regression structure more flexible by including nonlinear effects of continuous covariates, random effects or spatial effects. These extensions often rely on penalised least squares or penalised likelihood estimation with quadratic penalties and may therefore be difficult to combine with the linear programming approaches often considered in quantile regression. As a consequence, geoadditive expectile regression based on minimising an asymmetrically weighted sum of squared residuals was introduced. Different estimation procedures are available including least asymmetrically weighted squares, boosting (Sobotka

and Kneib, 2010) or restricted expectile regression (Schnabel and Eilers, 2010). We propose to investigate these point estimators by constructing pointwise confidence intervals to each expectile regression curve. For the construction of the confidence intervals we use a nonparametric bootstrap or the asymptotic normality of the regression coefficients.

2 LAWS

The results of a quantile regression can be acquired by minimising the asymmetrically weighted sum of the absolute residuals and in analogy an expectile regression is computed from the least asymmetrically weighted squares (LAWS) of the residuals. LAWS minimises

$$S = \sum_{i=1}^{n} w_\tau(y_i)(y_i - \mu_i(\tau))^2 \tag{1}$$

with weights

$$w_\tau(y_i) = \begin{cases} \tau & \text{if } y_i > \mu_i(\tau) \\ 1-\tau & \text{if } y_i < \mu_i(\tau) \end{cases} \tag{2}$$

where y_i is a continuous response and $\mu_i(\tau)$ is the estimated expectile for different values of the asymmetry parameter $\tau \in (0,1)$. Hence the computation of expectile regression is much easier, since it avoids the non-differentiable absolute value criterion, but expectiles lack the intuitive interpretation of quantiles. While the quantile of a random variable Z immediately depicts the amount probability that lies below it, the τ-expectile $\mu(\tau)$ can only be defined imlicitely:

$$\tau = \frac{\int_{-\infty}^{\mu(\tau)} |z-\mu(\tau)| f_Z(z)dz}{\int_{-\infty}^{\infty} |z-\mu(\tau)| f_Z(z)dz} = \frac{G(\mu(\tau)) - \mu(\tau)F(\mu(\tau))}{2(G(\mu(\tau)) - \mu(\tau)F(\mu(\tau))) + (\mu(\tau) - \mu(0.5))}$$

where $G(m) = \int_{-\infty}^{m} z f_Z(z)dz$ and $G(\infty) = \mu(0.5)$ is the expectation of Z. On the other hand and in addition to the computational advantages, one can build additive models that contain different kinds of effects. We portray these effects by design matrices $B^{(j)}$ and assign a vector of regression coefficients β_j to each effect. We can then create the following additive expectile regression model:

$$\mu(\tau) = 1\beta_0 + B^{(1)}\beta_1 + \ldots + B^{(r)}\beta_r + \varepsilon_\tau.$$

For continuous univariate covariates, smooth expectile curves can be fitted using penalised splines (see Schnabel and Eilers, 2009, 2010). Additionally the model can include spatial effects based on either Markov random fields or tensor product splines (see Sobotka and Kneib, 2010). The smoothing can be induced by a quadratic penalty on the regression coefficients:

$$pen(\beta_{j,\tau}) = \lambda_j \beta'_{j,\tau} K_j \beta_{j,\tau}$$

with adaptable smoothing parameter λ and penalty matrix K.

3 Asymptotics

For the resulting estimated regression coefficients of the LAWS method (1) we derived asymptotic normality:

$$\hat{\beta}_\tau \overset{a}{\sim} N(\beta_\tau^0, \mathrm{Cov}(\hat{\beta}_\tau))$$

with

$$\mathrm{Cov}(\hat{\beta}_\tau) = (B'WB + K)^{-1}B'W^2\mathrm{diag}(y_i - B_i\hat{\beta}_\tau)^2B(B'WB + K)^{-1}$$

where $W = \mathrm{diag}(w_\tau(y_1), \ldots, w_\tau(y_n))$ and $B = (1, B^{(1)}, \ldots, B^{(r)})$.
From mean regression we already know that without further assumptions for the distribution of the residuals we have

$$\mathrm{Var}\left\{\left(y_i - B_i\hat{\beta}_\tau^0\right)\right\} = \mathrm{Var}\left\{\left(y_i - B_i'\beta_\tau^0\right)\right\}(1 - h_{ii}) \tag{3}$$

with h_{ii} being the ith diagonal element of the hat matrix H. For expectile regression we obtain a generalised hat matrix H^w with

$$h_{ii}^w = w_{i,\tau}^0 B_i' \left(\sum_{j=1}^n w_{j,\tau}^0 B_j B_j' + K\right)^{-1} B_i \tag{4}$$

that we can use to improve our estimation of $\mathrm{Cov}(\hat{\beta}_\tau)$.
Now we can compute the estimated variance of the unknown true expectile μ_τ for covariates x_i, $i = 1, \ldots, n$ as

$$\mathrm{Cov}(\hat{\mu}_\tau) = B\mathrm{Cov}(\hat{\beta}_\tau)B'.$$

With the knowledge of the asymptotic normality we can derive a confidence interval for the true expectile at covariate value x_i

$$\left[\hat{\mu}_{\tau,i} \pm z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{Var}(\hat{\mu}_{\tau,i})}\right]$$

with $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$ the $1 - \frac{\alpha}{2}$-quantile of the standard normal distribution.
We compare this method to the results of bootstrap percentile intervals, which in general need a large number of bootstrap replications to be accurate.

4 Simulation

After introducing this new method for expectile regression confidence intervals, its merits and disadvantages are investigated in terms of a simulation study. The method is compared with the results of bootstrap percentile

intervals as a reference method. The data structures considered in the simulation study are linear on the one hand and additive nonlinear on the other in order to simulate different data scenarios. Using different error distributions we generate data situations with several properties like high probability for outliers or heteroscedasticity.
From the results we can see that the confidence level will not be met for extreme asymmetries ($\tau \to 0, \tau \to 1$). Especially for growing sample sizes the asymptotic intervals will deliver a smooth result and increase in precision, but all that depends on the available data. For example, at the edge of the covariate support with less observations available, the width of the intervals increases strongly.
In general we can observe nice attributes for confidence intervals like a decreasing width for larger samples. The performance of the asymptotic confidence intervals can outperform the numerical alternative in many scenarios. Therefore we will from here on refrain from using the computationally far more challenging bootstrap.

5 Application: Childhood Malnutrition in India

Finally we apply our methods to a data set from the MEASURE Demographic and Health Surveys (DHS) that provides national studies on health and population development. In our case we use data on childhood malnutrition in India from the year 2001. We attempt to model a malnutrition score in a geoadditive specification that allows for an analysis of parametric as well as nonlinear effects of the age and BMI of the mother and the spatial distribution of malnutrition simultaneously. The regression predictor can then be calculated as

$$\begin{aligned}\eta_\tau \;=\;& x\beta_\tau + f_{\tau,1}(\text{age of child}) + f_{\tau,2}(\text{duration of breastfeeding}) \\ & + f_{\tau,3}(\text{BMI of mother}) + f_{\tau,4}(\text{age of mother}) \\ & + f_{\tau,5}(\text{education years of mother}) + f_{\tau,6}(\text{education years of partner}) \\ & + f_{\tau,\text{spat}}(\text{district}).\end{aligned} \tag{5}$$

For each effect we are now also able to provide confidence intervals as shown in Figure 1. For the nonlinear effects we can mainly observe homoscedasticity while there are larger differences for parametric effects. We can also observe the influence of the observation density on the confidence interval width. Further Figure 2 indicates a positive or negative effect of the districts of India on the nutritional status for several expectiles.
The analyses are done using our R–package "expectreg" (Sobotka, Schnabel, Schulze Waltrup, 2011).

Acknowledgments: Special Thanks to Sabine Schnabel, Paul Eilers and Linda Schulze Waltrup for our cooperation on expectiles. Financial support from the German Research Foundation (DFG) grant KN 922/4-1 is gratefully acknowledged.

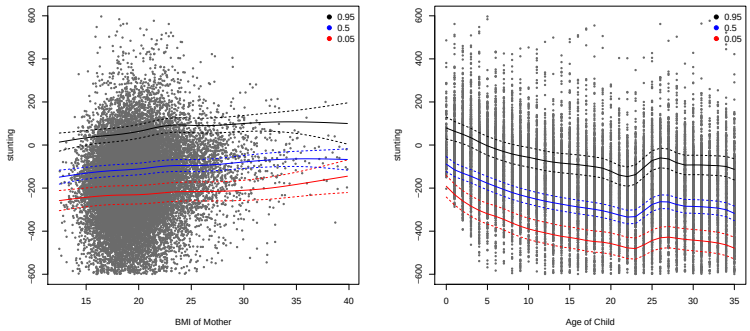


FIGURE 1. The two figures depict the estimated nonlinear effect for BMI of the mother and the age of the child. A 0.95 confidence interval for each expectile is marked with dashed lines.

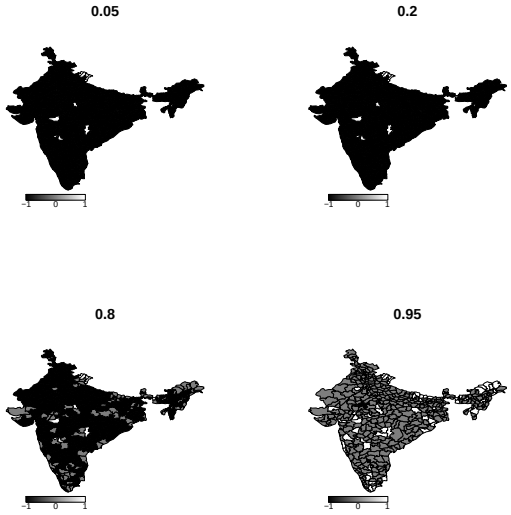


FIGURE 2. The figures show if the influence of a district is positive (white), negative (black) or insignificant for the 0.05, 0.20, 0.80 and 0.95-expectiles.

References

Koenker, R. (2005). *Quantile Regression.* New York: Cambridge University Press.

Schnabel, S. and P. Eilers (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis* **53**, 4168-4177.

Schnabel, S. and P. Eilers (2010). A location scale model for non-crossing expectile curves. *Working Paper*.

Sobotka, F. and T. Kneib (2010). Geoadditive expectile regression. *Computational Statistics and Data Analysis*, doi: 10.1016/j.csda.2010.11.015.

Sobotka, F., S. Schnabel and L. Schulze Waltrup (2011). *expectreg: Expectile and Quantile Regression*. R package version 0.21.

Measuring Efficiency of Trial Designs with Unreplicated or Partially Replicated Test Lines

Katia Stefanova[1]

[1] Department of Agriculture and Food WA, Perth WA 6151, Australia

Abstract: In this paper efficiency measure for optimal design is presented and illustrated on the example of unreplicated field trials laid out on plots with spatial errors defined by uniformity trials. A simulation study is conducted by randomizing the allocation of genotypes to the plots of four uniformity trials.

Keywords: Efficiency Factor; Linear Mixed Models; Unreplicated Trials Design.

1 Introduction

Unreplicated trials in early generation plant breeding enable breeders to select material from a large number of genotypes with limited quantities of seed. The greater the number of genotypes grown, the greater will be the probability of detecting superior plants. Traditionally, a small number of control varieties are planted in a systematic pattern across the trial and test lines randomly allocated to the intervening plots. Kempton (1984) gives an excellent historical account of such unreplicated trials. Optimal design, particularly the spatial arrangement of control plots and their number forms the central question addressed in this paper.

Unreplicated trial designs fall into three categories. Firstly, the most widely used designs, with systematic control plot arrangements in a diagonal or knight's move pattern. Secondly, the "augmented designs" introduced by Federer (1961) and later developed and improved (Lin & Poushinsky, 1983; Reynolds & Crossa, 2001; Williams & John, 2003). The experimental area is split into blocks and control varieties randomly arranged in each block in accordance with some incomplete or complete block design. The third type of design is the "partially replicated design", where some proportion of the test lines is replicated (Cullis *et al.*, 2006).

The analysis of the unreplicated trial has evolved and now a linear mixed model is most commonly used. In this paper our simulated data are analysed using spatial mixed models with REML estimation of variance components (Gilmour *et al.*, 1997 and Stefanova *et al.*, 2009). Also we have chosen to compare designs using an "efficiency factor", which is the standard error of the comparison between test line and control as a ratio of that standard error in a completely randomized trial.

2 Simulation and modelling

2.1 Description of the uniformity trials and choice of designs

Uniformity trials data has been used in an extensive simulation study to compare trial designs. At each of 4 locations in Western Australia (Katanning, Merredin, Newdegate and Wongan Hills), a rectangle of 60m×37.5m was marked out on a field with 300 plots in 12 columns and 25 rows. At each plot single wheat variety was grown and harvested and it's yield was individually recorded. Spatial mixed models were fitted to each data set. The analyses were performed using ASREML R (Butler *et al.*, 2007). A total of 34 different trial designs were used in the simulation study, representing designs from the three categories described in the Introduction.

2.2 The simulation process and fitting of mixed models

Let $\mathbf{u}_t$=$\{u_1, u_2 ... u_{n_t}\}$ denote true, known fixed effects for n_t test varieties and $\mathbf{u}_c = \{u_{c1}, u_{c2} ... u_{cn_c}\}$ denote the effects for n_c controls. In the simulation exercise, $\mathbf{u}_t$ and $\mathbf{u}_c$ are fixed and unchanged. The field experiment which we emulate has 300 plots and consequently $n_t + rn_c = 300$ where the controls are each replicated r times. The parameters n_t , n_c and r change according to the particular design being investigated.
Let $\mathbf{z} = \{z_1, z_2 ... z_{300}\}$ denote the known plot yields from a uniformity trial. Simulated yields are obtained by randomizing the test and control variety labels to the plots in accordance with the design rule in question. If the i^{th} plot is allocated to test variety m , the simulated yield for that plot is $y_i = z_i + u_m$, and likewise for all plots, thus generating $\mathbf{y} = \{y_1, y_2 ..., y_{300}\}$. We fit a mixed model to $\mathbf{y}$ and hence calculate estimates of all test line effects $\widehat{u_t}$ and control variety effects $\widehat{u_c}$. For a given design, given uniformity trial and given true variety effects $\mathbf{u} = \{\mathbf{u}_t, \mathbf{u}_c\}$, one randomization of plot labels results in one data vector of yields $\mathbf{y}$ and fitting a specified linear mixed model to this data gives one set of parameter estimates $\widehat{\mathbf{u}} = \{\widehat{u_t}, \widehat{u_c}\}$. An entire simulation procedure entails repeating this cycle of randomization and estimation many times. The values of $\mathbf{u}$ are defined as Normal deviates from a population with mean 0 and variance σ_g^2. The controls comprising u_{c50}, u_{c70} and u_{c90} are defined such that 50%, 70% and 90% respectively of the $\mathbf{u_t}$ values are less than the control values.

3 Results and Discussion

As explained earlier, each model fit to a particular data set is summarized by one efficiency factor, κ . We might interpret κ as the standard error (SE) of the contrast between a test line mean and a control mean in the design under study, as a percentage of that standard error with completely randomized plots. More specifically,

$$\kappa = 100(SE(\widehat{u_t} - \widehat{u_c})/\sqrt{\sigma^2(1 + 1/r)}) \quad (1)$$

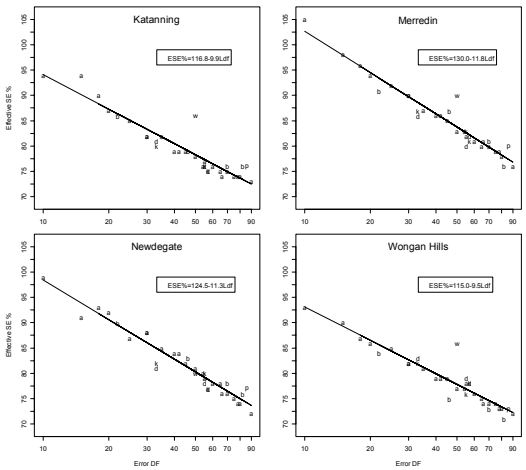


FIGURE 1. Efficiency factor κ graphed against Error DF. The symbols denote the type of design, a=augmented, d=diagonal, k=knight's move, b=partial rep, w=WA p=LP.

where σ^2 is the plot error variance and r is the number of replications of the control. The adjustment for r ensures that when comparing designs on the basis of their κ's such comparisons are unconfounded with changed values of r. In Figure 1 κ is graphed against log(Error DF); a linear model fits this data very well. The results show that by increasing the Error DF in any design, we reduce κ. However, increasing Error DF comes at the cost of reducing the number of test lines that can be accommodated in an experiment of fixed size. We look at this question in two ways, namely at maximizing expected genetic gain and assessing the overall probability of making successful selections. The dual problem is firstly to have sufficient test lines in a trial to ensure a reasonable probability that at least some superior lines are present in that trial (the inclusion probability P_p) and secondly to have a trial design which has sufficient Error DF to ensure that there is a reasonable probability that the superior lines will be correctly identified (the selection probability P_s). Use P_{ps} to denote the joint probability of a superior line being included and selected. To illustrate the calculations, define

$$P_s(u_i) = prob(\widehat{u}_i > \widehat{u}_c | u_i, u_c) = \Phi((u_i - u_c)/\sqrt{\sigma^2(1 + 1/r)}), \qquad (2)$$

$$P_{ps} = \int_{u^*\sigma_g}^{\infty} P_s(x)\, \phi(x\sigma_g)\sigma_g dx, \qquad (3)$$

and

$$Pr(x\,or\,more\,successful\,selections) = 1 - \sum_{i=0}^{x-1} \frac{n_t!}{i!(n_t-i)!} P_{ps}^{i}(1-P_{ps})^{n_t-i} \tag{4}$$

where $\Phi(.)$ denotes the cumulative Normal distribution function and $\phi(.)$ the density function. In these formulae u_c denotes the genetic effect of that control used to identify apparently superior lines based on their phenotype, whilst u^* denotes the genetic effect in the infinite population above which a test line is genuinely superior. These effects, u_c and u^* need not be equal. The critical factor in defining a good design for unreplicated trial, is the number of Error DF. By calculating expected genetic gain, we have determined a reasonable compromise to be about 50 Error DF, greater than 50 leads to a decline in expected genetic gain. The most important finding has been the marked superiority of the partially replicated designs. Finally, the benefit of analysing unreplicated trials by fitting a spatial mixed model has been shown to be very substantial, where fitting variety as random or fixed depends on the aim of the analysis.

Acknowledgments: The study was done in collaboration with Emeritus Professor G. P. Y. Clarke and some of the results are presented here.

References

Butler, D., Cullis, B.R., Gilmour, A.R., Thompson, R., and Gogel, B.J. (2007). *ASREML R Reference Manual. Release 2.0.* QLD DPIF & NSW DPI.

Cullis, B.R., Smith, A., and Coombes, N.E. (2006). On the design of early generation variety trials with correlated data. *J Agr Biol Environ Stat*, **11**, 381-393.

Gilmour, A.R., Cullis, B.R., and Verbyla, A.P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *J Agr Biol Environ Stat*, **2**, 269–273.

Kempton, R.A. (1984). The design and analysis of unreplicated trials. *Vortage fur Pflanzenzuschtung*, **7**, 219-242.

Stefanova, K.T., Smith, A.B., and Cullis, B.R. (2009). Enhanced diagnostics for the spatial analysis of field trials. *J Agr Biol Environ Stat*, **14**, 1-19.

A Markov switching model for vine copulas

Jakob Stöber [1], Claudia Czado[1]

[1] Lehrstuhl für Mathematische Statistik, Technische Universität München, Parkring 13, 85748 Garching-Hochbrück, Germany. corresponding email: stoeber@ma.tum.de

Abstract: Regular vine (R-vine) copulas, which are entirely constructed from bivariate copulas as building blocks, constitute a flexible class of high dimensional dependency models. In this paper we introduce a Markov switching R-vine copula model, combining the flexibility of general R-vine copulas with the possibility for dependence structures to change over time. Bayesian parameter estimation in this context is discussed and we apply the newly proposed model to examine the dependence of exchange rates during times of crisis.

Keywords: Bayesian estimation; Markov switching; regular vines; copulas.

1 Introduction

The recent financial crisis demonstrated that there are two key features of financial time series that have not been adequately addressed in risk modeling: extremal dependencies and sudden changes in behavior.
While recent developments in the area of multivariate dependence modeling tend towards flexible copula structures which are able to cover extremal dependence properties, the second feature is more difficult to deal with. As risk modeling is always based on experiences of the past, mathematical models cannot take new types of behavior, which have never been observed before, into consideration. While we cannot predict new types of behavior we can use Markov switching (MS) models to account for changes to more extreme types of behavior during times of crisis.
This paper presents an MS model for regular vine (R-vine) copulas and a Bayesian procedure for estimating its parameters to address the aforementioned problems. Our contribution is twofold: First, we extend the Bayesian estimation procedure of Min and Czado (2010) for Student-t copulas on drawable (D-)vines, which constitute a subclass of R-vines, to general R-vines and arbitrary bivariate copulas. Furthermore, we combine it with a Bayesian estimation procedure for the underlying MS model as it has been developed by Kim and Nelson (1998).

2 Markov switching regular vine copulas

In the following, we briefly recall both components of our model.

2.1 Regular Vines

One of the most promising structures for multivariate modeling is the hierarchical R-vine structure which has first been used by Joe (1996) and been formally introduced by Bedford and Cooke (2001). They define a regular vine $\mathcal{V}$ on d variables as a sequence of connected trees (undirected, acyclic graphs) $T_1, \ldots, T_{d-1}$, with nodes N_i and edges E_i, $1 \leq i \leq d-1$, which satisfy the following properties:

1. T_1 is a tree with nodes $N_1 = \{1, \ldots, d\}$ and a set of edges E_1.
2. For $i \geq 2$, T_i is a tree with nodes $N_i = E_{i-1}$ and edges E_i.
3. If two nodes in T_{i+1} are joined by an edge, the corresponding, edges in T_i must share a common node.

A five-dimensional R-vine is shown in Figure 1. The notation we employ throughout our paper follows Czado (2010).

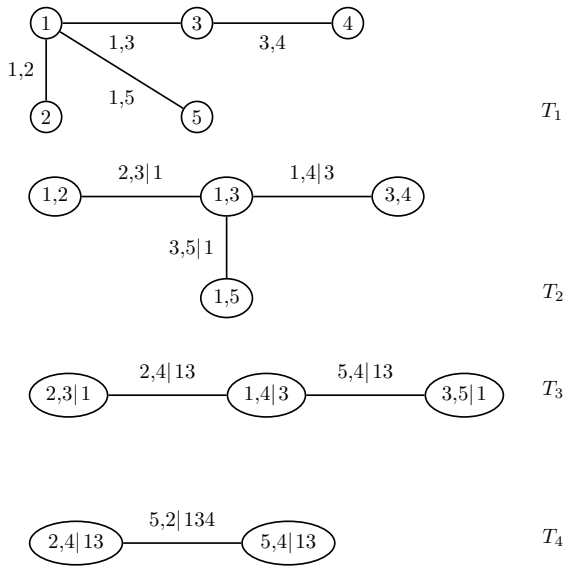


FIGURE 1. An R-vine tree sequence in five dimensions with edge indices.

Since Aas et al. (2009) considered vine copulas in an inferential context, there is increasing interest in developing estimation methods and their applications. To build up a statistical model on the graph theoretic object of a (d-dimensional) R-vine, a bivariate copula is associated to each edge of the vine. In particular, the bivariate copula $C_{j(e),k(e)|D(e)}$ corresponding to edge $e \hat{=} j(e), k(e)|D(e)$ is the copula corresponding to the conditional bivariate distribution of $X_{j(e)}$ and $X_{k(e)}$ given $\mathbf{X}_{D(e)} = \mathbf{x}_{D(e)}$. This construction uniquely determines the density of the joint distribution of a d-dimensional

random vector, whose conditional copulas correspond to the copulas on the vine. A detailed treatment of R-vine copulas can be found in Kurowicka and Cooke (2006) and parameter estimation is considered in Dißmann (2010). We will denote the density of an R-vine copula corresponding to a vine $\mathbf{V}$ with set of copulas $\mathbf{B}$ and parameters $\boldsymbol{\theta}$ as $c(.|\mathbf{V}, \mathbf{B}, \boldsymbol{\theta})$.

2.2 Markov switching model

An MS model is a nonlinear specification, in which different states of the world or the economy affect the development of a time series. Assuming the densities $c(.|\mathbf{V}_k, \mathbf{B}_k, \boldsymbol{\theta}_k)$ of n R-vine copulas to be given, we want to combine them in an MS model such that at each point in time the present regime determines the dependence structure.
For this, let (S_t) be a Markov chain with states $\{1, \ldots, n\}$. For simplicity, we assume it to be of first order such that it can be completely characterized by its transition matrix $P(S_t = i|S_{t-1} = j) = P_{i,j}$. In a general setting, the probabilities in this matrix may change over time and depend on other internal or external variables of the model.
Given the Markov chain (S_t), the simplest MS vine copula model for a time series $(\mathbf{u}_t)$ is characterized by the conditional densities

$$c(\mathbf{u}_t|S_t, (\mathbf{V}_k, \mathbf{B}_k, \boldsymbol{\theta}_k)_{k=1,\ldots,n}) = \sum_{k=1}^{n} 1_{\{S_t=k\}} \cdot c(\mathbf{u}_t|\mathbf{V}_k, \mathbf{B}_k, \boldsymbol{\theta}_k).$$

Here, the specification $\mathbf{V}_k$ and $\mathbf{B}_k$ is assumed to be given while $\boldsymbol{\theta}_k$ needs to be estimated.

2.3 Estimation

For computational reasons, parameter estimation for copula structures is usually performed in a two-step approach as described by Joe and Xu (1996). Marginal structures are fitted first and posterior mode or posterior mean estimates for their parameters are chosen to convert the residuals to uniform data by applying the probability integral transform. As our main interest lies in developing methods for describing and estimating the copula structure we will assume for the remainder of this paper that estimation uncertainties for the marginal models can be neglected when examining the copula for the uniform data.
Estimation of the parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ and P in the MS vine copula model is performed by a Gibbs sampler with three main steps:

1. Sample P conditional on the states S_t.
2. Sample the states S_t conditional on $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ and P.
3. Sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ conditional on the states S_t.

For Steps 1. and 2. we apply the procedure of Kim and Nelsen (1998). Step 3. is done by our extension of the algorithm in Min and Czado (2010). As the set of parameters is augmented by the hidden state variables for the purpose of estimation, we also obtain posterior estimates for the probability that the economy is in state k at time t.

3 Application

The dataset we use consists of 9 exchange rates against the US dollar, namely Euro, British pound, Canadian dollar, Australian dollar, Brazilian real, Japanese yen, Chinese yuan, Swiss franc and Indian rupee. We consider 1007 daily observations from July 22, 2005 to July 17, 2009. An extensive analysis of the dataset including the fitting of marginal structures and the conversion to copula data on the unit interval has been performed by Czado et al. (2010).

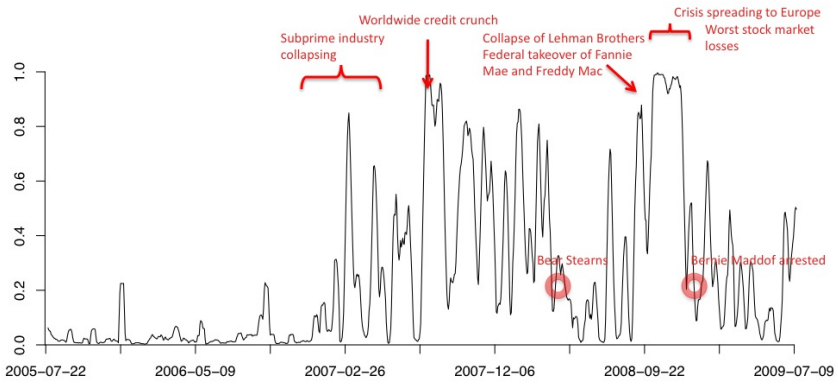


FIGURE 2. Smoothed probability that the state variable in the first model indicates the crisis regime. The annotations highlight important events during the financial crisis.

For our application, we assume two regimes to be present: one describing "normal" dependencies and one describing the dependencies in times of crisis. We consider two different sets of R-vine structures $\mathbf{V}$. In the first case, we assume that the same copula structure is present in addition to the same vine structure ($\mathbf{V}_1 = \ldots = \mathbf{V}_n, \mathbf{B}_1 = \ldots = \mathbf{B}_n$) in the whole dataset and that only the parameters do vary over time following the underlying Markov chain. For the second case, we do also allow for the R-vine structure $\mathbf{V}$ and copula structure $\mathbf{B}$ to change during times of crisis.
Using methods described by Dißmann (2010), we select an appropriate R-vine structure covering the average dependence in the long run. Conducting a rolling window analysis, we further select an R-vine structure describing the dependence during peak times of the financial crisis. In this second case, where we have a different "crisis" structure, the bivariate copulas in the "normal" regime are assumed to be Gaussian. For the bivariate copulas in the "crisis" regime we select rotated Gumbel copulas. All R-vine structures are truncated after the second tree, i.e. we assume that

the copulas corresponding to bivariate conditional marginal distributions conditioning on more than one variable are independence copulas. Given this setup, the estimation procedures outlined in Section 2.3 can be applied. The probabilities for the hidden state variable to indicate the presence of the crisis regime are plotted in Figures 2 and 3. They illustrate our finding that the times where the crisis regime is predominant correspond to events of high impact during the financial crisis.

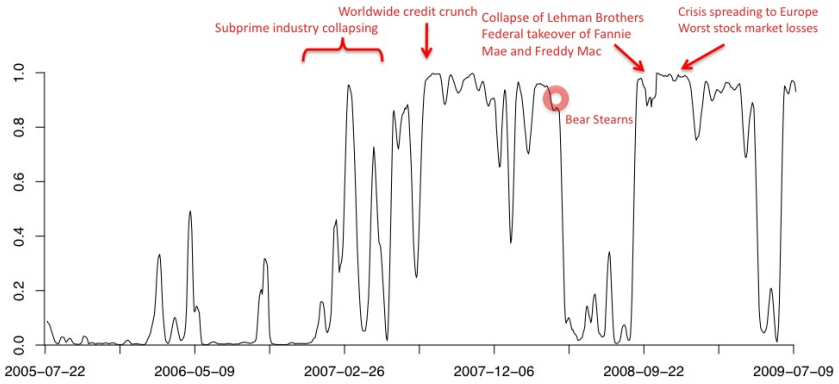


FIGURE 3. Smoothed probability that the state variable in the second model indicates the crisis regime. Again, annotations are made to highlight important events during the financial crisis.

Comparing Figure 2 to Figure 3 it shows that MS-models which allow for a change of the R-vine and copula structure are more successful in detecting a financial crisis regime than models in which regimes differ only by their parameter values.

4 Conclusion

The model presented in this paper enables us to flexibly analyze dependence structures varying over time. Applying it to exchange rates, we discover that their dependence was similar during all peaks of the recent financial crisis. This supports the use of MS vine copula models to account for behavioral changes in these times.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics* **44** (2), 182-198.

Bedford, T., and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence* **32**, 245-268.

Czado, C. (2010) Pair copula constructions of multivariate copulas. In *Copula Theory and Its Applications*, Jaworski, P., Durante, F., Härdle, W., and Rychlik, T. (Eds.). Springer, Berlin.

Czado, C., Schepsmeier, U., and Min, A. (2010). Maximum likelihood estimation of mixed C-vines with application to exchange rates. To appear in *Statistical Modeling.*

Dißmann, J. (2010). Statistical inference for regular vines and application. *Diploma thesis, Technische Universität München, Germany.*

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. In *Distributions with Fixed Marginals and Related Topics*, Rüschendorf, L., Schweizer, B. and Taylor, M. D. (Eds.)

Joe, H. and Xu, J. J. (1996). The estimation method of inference functions of margins for multivariate models *Technical Report 166, Department of Statistics, University of British Columbia*

Kim, C.-J., and Nelson, C. R. (1998). Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. *The Review of Economics and Statistics* **80** (2), 188-201.

Kurowicka, D., and Cooke, R. M. (2006). *Uncertainty Analysis with High Dimensional Dependence Modeling.* Wiley, Chichester.

Min, A., and Czado, C. (2010). Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics* **8** (4), 511-546.

Bayesian residual analysis in Poisson regression models.

James Sweeney[1] , John Haslett[1]

[1] Dept of Statistics, School of Computer Science & Statistics, Trinity College Dublin, Ireland.

Abstract: In recent times, the use of Bayesian methods has become more widespread in regression problems where complex noisy data is a frequent occurrence. In using a Bayesian approach however, data and model criticism methods generally do not take the form of those of classical residual analysis. In this paper we seek to bring together both approaches when considering some Poisson count data. We address the problem in a Bayesian manner using Gaussian random effect terms to model potential overdispersion of the Poisson counts. The posterior random effects are used as a "surrogate" for classical residuals to aid in outlier detection and model criticism. We apply the proposed approach to the palaeoclimate dataset of Huntley et al, and use some exploratory tools from classical residual analysis to gain an extra insight into underlying model dynamics from the posterior random effect terms.

Keywords: Bayesian inference; residual analysis; Gaussian random effect; Gaussian approximation

1 Introduction

In this paper we propose a novel approach for the Bayesian analysis of residuals in settings where the response variable is non-Gaussian and discrete. In the interest of brevity however, we will constrain our discussion to the specific scenario where the response variable consists of (assumed) Poisson distributed counts. Our aim is to create an automatic Bayesian approach for outlier detection and model criticism in studies where the data is non-Gaussian in nature and we have no recourse to conventional Gaussian residual theory.

Chaloner & Brant (1988) proposed a simple Bayesian approach for the detection of outliers in Gaussian linear regression models. A priori, a model residual, ϵ_i, is distributed $N(0, \sigma^2)$. A posteriori, the residual is considered "outlying" if the posterior distribution of a given ϵ_i is located far from zero. The posterior probability of such an event is $Pr_i = Pr(|\epsilon_i| > k\sigma|Y)$ and values of $Pr_i > 2\Phi(-k)$ may be regarded as suspicious; k is available from standard Gaussian residual theory and values of $k = 1.96$, $2\Phi(-1.96) = .05$ are typically used.

Albert and Chib (1995) provided a method for Bayesian residual analysis in the presence of binary counts. Given the binary regression model $E(y_i) = p_i = F(x_i\beta)$, a model residual is specified as $r_i = y_i - p_i$; r_i is a continuous valued random variable under the Bayesian framework. Each posterior r_i has support on the interval $(y_i - 1, y_i)$ and outlier detection involves the identification of residuals which tend towards the "extremes" of their respective support region. Souza (2010) promoted the inclusion of random effect terms in binary regression models for outlier detection purposes. The prior distribution for each random effect term is specified as a two-component scale mixture of normals; $\gamma_i|k_i \sim N(0, [(1-k_i)\sigma^2 + ck_i\sigma^2]), c > 1$ and $k_i|\pi \sim Bern(\pi)$; outlier detection involves the identification of observations with large $p_{k_i} = Pr(k_i = 1)$.
Our proposed approach is similarly built upon the use of Gaussian random effect terms to accommodate potential overdispersion of the observed count data. We propose to utilise these posterior random effects, defined on a continuous scale, as a "surrogate" for classical residuals. We approximate each posterior random effect as a weighed mixture of Gaussians; this provides access to standard Gaussian residual theory and thus an automatic and efficient tool for outlier detection. We also aim to use the posterior random effects in a "classical fashion" as a tool for model criticism. We make use of diagnostic tools such as Q-Q plots and use simple plots of the posterior random effects to provide extra insight into underlying model dynamics.
The motivating application for our work is the palaeoclimate dataset of Huntley et al (1993). Conventional Bayesian investigation of each available datum would require the visual analysis of simulated posterior probability distributions for 7742×28 observations ($\approx 200,000$ in total) which is infeasible with regard to time constraints.

2 Bayesian outlier detection and exploratory residual analysis.

In this section we present a step by step description of our proposed approach in the context of a simple Poisson linear regression problem. Potential overdispersion of the Poisson counts is accounted for by the addition of (mean zero) Gaussian random effect components to the linear predictor. In the following let $Y = \{y_1, .., y_n\}$ represent the observed data, $X = \{x_1, ..., x_n\}$ the observed predictor variables and β some regression coefficients. We specify a mean zero normal random effects model as the prior for each random effect term, $u_i \sim N(0, \sigma^2)$, where σ^2 is a model hyperparameter representing the global variance of the random effect terms.

$$y_i|w_i \sim Poisson(e^{w_i}) \quad (1)$$

$$w_i = x_i\beta + u_i \quad (2)$$

$$u_i \sim N(0, \sigma^2) \quad (3)$$

We adopt a Bayesian hierarchical model for the inference task; if the model under consideration is more complex than the presented example, the addition of extra stages to the model hierarchy is simple and easy to implement.

$$\begin{aligned} \pi(U, \beta, \sigma^2|Y) &\propto \pi(Y|\beta, U, \sigma^2)\pi(U, \beta|\sigma^2)\pi(\sigma^2) &(4)\\ &= \prod_{i=1}^{n} \pi(y_i|u_i, \beta, \sigma^2)\pi(U, \beta|\sigma^2)\pi(\sigma^2) &(5)\end{aligned}$$

The use of random effect terms for capturing extra Poisson variation can be computationally burdensome; a posterior random effect must be inferred for each datum. Furthermore, the joint posterior distribution in (4) is not known in closed form as it is not possible to analytically integrate out the latent u_i's. As numerical algorithms are infeasible given even moderate amounts of data, simulation based algorithms such as MCMC can be used to provide samples from the posterior distributions of interest. However, we must then address the usual issues regarding convergence and mixing of the Markov chains. To sidestep these issues, we make use of some recent advances in approximate Bayesian computation, namely the INLA algorithm of Rue (2009). For $W = X\beta + U$:

$$\pi(\sigma^2|Y) \approx \left.\frac{\pi(Y, W, \sigma^2)}{\tilde{\pi}_G(W|\sigma^2, Y)}\right|_{W=W^*(\sigma^2)} \quad (6)$$

$\tilde{\pi}_G(W|\sigma^2, Y)$ is the Gaussian approximation to the full conditional of W and $W(\sigma^2)$ is the mode of the full conditional for W for a given value of σ^2. In most cases such an approximation is nearly exact. If the posterior for σ^2 is computed on an (arbitrarily fine) discrete grid, probability weights can then be calculated which enable representation of the posterior distribution for each random effect, $\pi(u_i|Y)$, as a weighed mixture of Gaussians.

$$\pi(u_i|Y) = \sum_k \tilde{\pi}_G(u_i|\sigma_k^2, Y) \times \pi(\sigma_k^2|Y) \times \triangle_k \quad (7)$$

$\tilde{\pi}_G(u_i|\sigma_k^2, Y)$ is available from $\tilde{\pi}_G(w_i|\sigma_k^2, Y)$ and $\triangle_k$ are area weights which ensure the posterior probability distributions for each random effect sum to one. The posterior random effects in (7) can be used visually, to identify patterns within the data or alternatively to pinpoint suspicious observations. Consider the problem of outlier detection: suppose we say the i^{th} observation is an outlier if $Pr_i = Pr(|u_i| > k\sigma|Y)$ is sufficiently "large".

$$\begin{aligned} Pr(|u_i| > k\sigma|Y) &= \int Pr(|u_i| > k\sigma)\pi(\sigma^2|Y)d\sigma^2 &(8)\\ &= E_{\sigma^2}[1 - Pr(u_i(\sigma) < k\sigma|Y) + \end{aligned}$$

$$Pr(-u_i(\sigma) < -k\sigma|Y)] \tag{9}$$

The probability that a given u_i is "outlying" is efficiently calculated from the Gaussian approximation to the full conditional; $Pr(u_i < k\sigma|Y) = \Phi(k\sigma, \mu_i, \tau_i^2)$ where $\pi(u_i|\sigma^2, Y) \approx N(\mu_i(\sigma^2), \tau^2(\sigma^2))$ and k is available from standard Gaussian residual theory. This probability can be compared to $2\Phi(-k)$ to identify "suspicious" observations.
The posterior random effects are used as a surrogate for classical residuals to aid with outlier detection. However, they may additionally be used in an exploratory fashion (as in classical residual analysis) for model criticism purposes. In the following sections we illustrate how the examination of mean posterior random effects ($E(U|Y)$) may help identify patterns within the data masked by the discrete nature of the response variable or even provide an informative insight into underlying model dynamics.

2.1 Toy Problem.

We create a simple toy problem to highlight one of the model criticism features of our proposed approach. Data is simulated from the the Poisson regression model $y_i \sim Pois(e^{w_i}), w_i = 2 + 2x_i + u_i$, where u_i is simulated from (a) $u_i \sim N(0, 1)$ and (b) $u_i \sim \Gamma(1, 1)$. We proceed to infer model parameters as presented in section 2. Our interest lies in determining if examination of the posterior random effects indicates possible misspecification of the model. In figure (1b) we see that our incorrect specification of the random effects as a-priori Gaussian is detected.

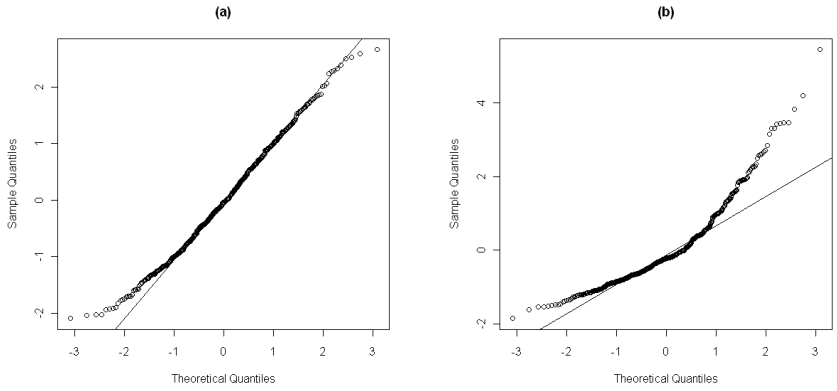


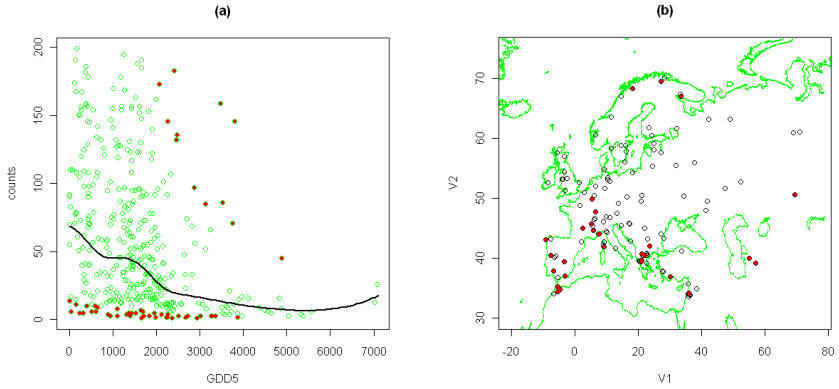
FIGURE 1. Q-Q plot of $E(U|Y)$ where (a) $u_i \sim N(0, 1)$ & (b) $u_i \sim \Gamma(1, 1)$.

3 Palaeoclimate application.

We apply our proposed approach to a subset (400 counts) of the data of the Alnus (common alder) taxon considering only one climate predictor

variable, GDD5. We model the latent response surface of the pollen counts as a smooth nonparameteric function (f) of the GDD5 predictor variables (see Haslett(2006) for further details) and account for overdispersion in the pollen counts via Gaussian random effect terms as previously. Our interest lies in detecting potentially "outlying" observations and validating our a-priori distributional assumptions for the random effects. We specify a vague $\Gamma(1, .001)$ prior for each model hyperparameter.

$$y_i|w_i \sim Poisson(e^{w_i}) \quad (10)$$
$$w_i = f(x_i) + u_i \quad (11)$$
$$u_i \sim N(0, \sigma^2) \quad (12)$$



FIGUR E 2. (a) Pollen counts vs GDD5 ($E(f)$ ($-$), counts ($\circ$), outliers ($\bullet$)) & (b) Observed counts ($\circ$) and suggested outliers ($\bullet$) on a map of Europe.

4 Results & Conclusions

The examination of the posterior random effect terms proved to be extremely efficient at identifying potential outliers in the Alnus data subset. In figure (3), we plot a Q-Q plot of the mean posterior random effects ($E(U|Y)$); this allows us to confirm the a-priori belief that the random effects are approximately Gaussian. Figure (2(b)) illustrates how the use of the posterior random effects as an exploratory tool can provide additional insight into underlying model dynamics. Count observations and outliers are plotted on a map of Europe and we observe that there appear to be a large number of outliers around the Mediterranean region. Further investigation reveals that each of the outliers are recorded at high altitudes where there is more moisture available to the Alnus plant and hence more pollen

is produced than would generally be expected at this particular world latitude.

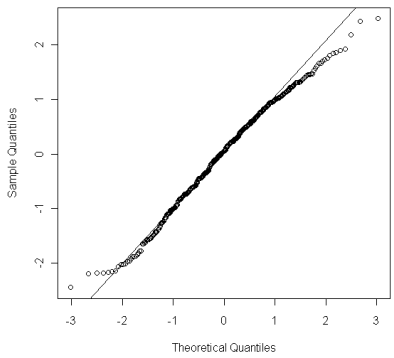


FIGURE 3. Q-Q plot mean posterior random effects ($E(U|Y)$).

References

Albert, J. and Chib, S. (1995). Bayesian residual analysis for binary response regression models, *Biometrika*, **82**, 747 - 759.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis, *Biometrika*, **75**, 651-9.

Haslett, J., Bhattacharya, S., Whiley, M., Salter-Townshend, M., Wilson, S. (2006). Bayesian Palaeoclimate Reconstruction *J. R. Statist. Soc. A.*, **169, Part 3**, 1-36.

Huntley, B (1993). The use of climate response surfaces to reconstruct palaeoclimate from quartenary pollen and plant macro-fossil data. *Philosophical Transactions of the Royal Society of London, Series B - Biological Sciences*, **341**, 215-233.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B*, **71**, 319-392.

Souza, A.D.P. and Migon, H. S. (2010). Bayesian outlier analysis in binary regression. *Journal of Applied Statistics*, **Vol 37, No. 8**, 1355-1368.

Prediction for an observation in a new cluster for Multilevel Logistic Regression considering k random coefficients

Karin Ayumi Tamura [1], Viviana Giampaoli[2]

[1] Departamento de Estatística, IME-USP, Rua do Matão, 1010, Cidade Universitária, São Paulo, Brazil, karinat@ime.usp.br
[2] Departamento de Estatística, IME-USP, Rua do Matão, 1010, Cidade Universitária, São Paulo, Brazil, vivig@ime.usp.br

Abstract: This article addresses the problem of predicting the outcome variable for an observation in a new group, using the multilevel logistic regression model (MLRM) with k random effects. We fitted the MLRM considering the random intercept and one random slope, nonetheless the method can be implemented for k random effects. With this objective, we used two estimation methods in the multilevel models: Penalized Quasi-Likelihood (PQL) and Laplace Approximation. The prediction multilevel approach was applied to a set of data related to nutritional status of children.

Keywords: multilevel logistic model; outcome prediction; random slopes.

1 Introduction

This paper presents a novel approach to predict the outcome variable of an observation in a new cluster using the Multilevel Generalized Linear Model (MGLM) considering k random effects. In this case, it is not possible to predict the outcome using the traditional method, because the estimates of the random effects there are not available for the new groups. Extending the method presented in Tamura and Giampaoli (2010), our proposal is to predict the outcome using the MLRM with k random effects.

2 Prediction

Let a MGLM with 2 levels and y_{ij} be the outcome for the j-th observation in the i-th cluster, in which $i = 1, \ldots, q$ and $j = 1, \ldots, n_i$. Let the density function be defined by

$$f(y_{ij}|\boldsymbol{\alpha}_i) = \exp\left[\left(\frac{a_{ij}}{\phi}\right)(y_{ij}\gamma_{ij} - b(\gamma_{ij})) + c\left(y_{ij}, \left(\frac{a_{ij}}{\phi}\right)\right)\right], \qquad (1)$$

where a_{ij} is a weight, $b(\cdot)$ and $c(\cdot,\cdot)$ are known functions associated with the exponential family and ϕ is the dispersion parameter. γ_{ij} is associated

with $\mu_{ij} = E(y_{ij}|\boldsymbol{\alpha})$, which in turn is associated with a linear function through a link function $g(\cdot)$. Besides, $g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^t\boldsymbol{\beta} + \mathbf{z}_{ij}^t\boldsymbol{\alpha}_i$, where $\mathbf{x}_{ij}$ is a vector of known independent variables $(p \times 1)$ associated with $\boldsymbol{\beta}$, $\boldsymbol{\beta}$ is a vector of fixed effects $(p \times 1)$, $\mathbf{z}_{ij}$ is a vector of known independent variables associated with $\boldsymbol{\alpha_i}$ $(k \times 1)$ and $\boldsymbol{\alpha}_i$ is a vector of the random effects $(k \times 1)$ of the i-th cluster, $i = 1, \ldots, q$. The vector $\mathbf{z}_{ij}^t$ is defined by $\mathbf{z}_{ij}^t$=(1, z_{1ij}, z_{2ij}, ..., $z_{(k-1)ij}$). In particular, under the so called canonical link, $\gamma_{ij} = \eta_{ij}$. Furthermore, $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q$ are i.i.d. with $\boldsymbol{\alpha}_i \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$, in which $\boldsymbol{\Sigma}$ is the unknown covariance matrix of the random effects.

2.1 Prediction Function for k random effects

Consider the prediction problem as $\tilde{\varsigma} = \varsigma(\boldsymbol{\beta}, \boldsymbol{\alpha}_S)$, in which S is a subset of $\{1, \ldots, q\}$ and $\boldsymbol{\alpha}_S = (\boldsymbol{\alpha}_{i \in S})$. Let $y_S = (y_i)_{i \in S}$, where $y_i = (y_{ij})_{1 \leq j \leq n_i}$. Under the above model, in sense of minimum MSPE (Mean Squared Prediction Error), the BP (Best Predictor) of ς for $S = \{i\}$ is given by $\tilde{\varsigma}_i = E(\varsigma_i|y_i) = E(\varsigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha}_S|y_S))$.
Tamura and Giampaoli (2010), based on Jiang and Lahiri (2006), presented a method to predict the outcome variable of an observation in a new cluster, considering only the random intercept. This method was extended to k random effects in this work. In order to predict the outcome variable for a new cluster, we considered the distribution of $\boldsymbol{\alpha}_i$, where $\boldsymbol{\alpha}_i \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$, by the assumption of the model. To predict $\boldsymbol{\alpha}_i$, for $i \notin S$, as we do not know its value, we used the multivariate linear transformation $\boldsymbol{\alpha_i}$=$\boldsymbol{\Sigma}^{1/2}\boldsymbol{\xi}$, with $\boldsymbol{\xi} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I})$. Furthermore, we denoted $\boldsymbol{\alpha_i} = \boldsymbol{u}$ to explicitly indicate the prediction as follows

$$\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{pmatrix} = \begin{pmatrix} \xi_1 v_{11} + \xi_2 v_{12} + \cdots + \xi_k v_{1k} \\ \xi_1 v_{21} + \xi_2 v_{22} + \cdots + \xi_k v_{2k} \\ \vdots \\ \xi_1 v_{k1} + \xi_2 v_{k2} + \cdots + \xi_k v_{kk} \end{pmatrix}.$$

Thus, the BP prediction function, based on (1), can be written as

$$\tilde{\varsigma}_{ij}(\boldsymbol{\beta}, \xi) = \frac{\int_{\xi_1} \cdots \int_{\xi k} (\varsigma(\boldsymbol{\beta}, (\xi_1, \ldots, \xi_k)) \cdot \exp(\phi^{-1} S_i(\boldsymbol{\beta}, (\xi_1, \ldots, \xi_k)))) f(\xi_1, \ldots, \xi_k) d\xi_1 \ldots d\xi_k}{\int_{\xi_1} \cdots \int_{\xi_k} \exp(\phi^{-1} S_i(\boldsymbol{\beta}, (\xi_1, \ldots, \xi_k))) f(\xi_1, \ldots, \xi_k) d\xi_1 \ldots d\xi_k}, \tag{2}$$

where $f(\xi_1, \ldots, \xi_k) = f(\xi_1) \cdots f(\xi_k)$ with $f(\xi_m)$ defined by the univariate standard normal density, with $m = 1, \ldots, k$, and $S_i(\cdot) = \sum_{j=1}^{n_i} a_{ij}(y_{ij}\gamma_{ij} - b(\gamma_{ij}))$.

2.2 Prediction: Multilevel Logistic Regression Model

Consider the MLRM with the binary response y_{ij}. This model can be defined by $g(\mu_{ij}) = \text{logit}(P(y_{ij} = 1|\boldsymbol{\alpha}_i)) = \text{logit}(p_{ij}) = \mathbf{x}_{ij}^t\boldsymbol{\beta} + \mathbf{z}_{ij}^t\boldsymbol{\alpha}_i$. The Binomial distribution belongs to the exponential family and based on (1),

$b(\gamma_{ij}) = \log(1 + \exp(\gamma_{ij}))$ and $\phi = n_i$. As logit function is a canonical link, then $\gamma_{ij} = \eta_{ij}$. This paper particularized the BP function presented in (2) for MLRM. Lets define $\tilde{\varsigma}(\boldsymbol{\beta}, \boldsymbol{\alpha}_i) = \tilde{p}_{ij}$. Since $\tilde{\varsigma}$ is usually unknown, we replace $\tilde{\varsigma}$ by an estimator $\hat{\varsigma}$. Then, $\hat{\varsigma} = \varsigma(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}\xi) = \hat{p}_{ij}$ is denominated EBP (Empirical Best Predictor) and is given by

$$\frac{\exp(\mathbf{x}_{ij}^t\hat{\boldsymbol{\beta}}) \int_{\xi_1} \dots \int_{\xi_k} \frac{\exp((y_{i.}+1)\mathbf{z}_{ij}^t\hat{\boldsymbol{u}})}{1+\exp(\mathbf{x}_{ij}^t\hat{\boldsymbol{\beta}}+\mathbf{z}_{ij}^t\hat{\boldsymbol{u}})} \cdot \prod_{l=1}^{n_i} \frac{1}{1+\exp(\mathbf{x}_{il}^t\hat{\boldsymbol{\beta}}+\mathbf{z}_{il}^t\hat{\boldsymbol{u}})} f(\xi_1, \dots, \xi_k) d\xi_1 \dots d\xi_k}{\int_{\xi_1} \dots \int_{\xi_k} \exp(y_{i.}\mathbf{z}_{ij}^t\hat{\boldsymbol{u}}) \cdot \prod_{l=1}^{n_i} \frac{1}{1+\exp(\mathbf{x}_{il}^t\hat{\boldsymbol{\beta}}+\mathbf{z}_{il}^t\hat{\boldsymbol{u}})} f(\xi_1, \dots, \xi_k) d\xi_1 \dots d\xi_k},$$

where $\mathbf{z}_{ij}^t\hat{\boldsymbol{u}} = (\xi_1\hat{v}_{11} + \ldots + \xi_k\hat{v}_{1k}) + \ldots + (\xi_1\hat{v}_{k1} + \ldots + \xi_k\hat{v}_{kk})z_{(k-1)ij}^t$. Note that, since we do not know $y_{i.}$ and a_{ij}, we assumed that $y_{i.} = n_{i.}/2$ and $\sum_{j=1}^{n_i} a_{ij} = n_{i.}$.

3 Application

In order to evaluate the nutritional condition of the newborn children, we considered a longitudinal data including 241 newborn males. The information of each child was observed 2, 4, 6, 9, 12, 15 and 18 months after birth. HAZ-score is a nutritional classification based on the height of the children, which was collected at each observation time. The outcome of the problem is the HAZ-score, classified two categories: 1 - heavy unnourished and 0 - otherwise.

To illustrate the procedure, we considered a random sample of 50% of the clusters (120 children) in the training data set. The remaining groups, 121 children, were considered in the validation data set. The training data set was used to fit the models. The validation data set was assessed to predict the outcome, based on the parameters estimate of the model provided by the training data set. The independent variable associated with the fixed effect was ln(weight of the newborn child). In the multilevel model, we considered as independent variable associated with the random slope: the deviation from the child's weight expected for the period. This variable was calculated as $z_{ij} = \text{weight}_{ij} - (\overline{w}_{.j} - \sigma_{.j})$, in which $\overline{w}_{.j} = \sum_{i=1}^{q} \text{weight}_{ij}/q$, $\sigma_{.j}^2 = \sum_{i=1}^{q}(\text{weight}_{ij} - \overline{w}_{.j})^2/q$, where i indexes the children in the j-th observation months.

In Table 1, we can observe the parameters estimate of each model. Analyzing the fixed effects, in all models, the intercept and the slope were significant. Comparing the estimation methods in multilevel model, one possible reason for the differences between the parameters estimate can be explained by the fact that the PQL produces biased estimates (see, Cole et al., 2003), although the PQL is more efficient than Laplace regarding the estimate of the standard error. The values presented in Table 2 were based on the cut off which maximized the $KS = |Sensibility - (1 - Specificity)|$. In both data sets, the KS measure outperformed in the multilevel models in comparison to the traditional logistic regression.

TABLE 1. Usual and multilevel models fitted with PQL and Laplace methods.

	Estimate	SE	P-value	SE/Estimate
Usual Estimation				
Intercept	7.011	0.736	<0.00	10.50%
Ln(Weight of the newborn child)	-6.939	0.663	<0.00	-9.55%
PQL Method				
Intercept	9.877	2.624	<0.00	26.57%
Ln(Weight of the newborn child)	-10.671	2.314	<0.00	-21.68%
Standard Deviation of the Random Effects (σ_1, σ_2)	(4.36, 2.97)			
Intraclass correlation	(-0.82)			
Laplace Method				
Intercept	6.830	2.441	0.005	35.40%
Ln(Weight of the newborn child)	-8.418	2.142	<0.00	-25.44%
Standard Deviation of the Random Effects (σ_1, σ_2)	(4.50, 2.17)			
Intraclass correlation	(-0.92)			

TABLE 2. KS for the usual and multilevel models.

Performance Measures	Usual Model	PQL Method	Laplace Method
KS for Training data set	44.3	83.2	81.0
KS for Validation data set	43.5	49.5	57.9

4 Conclusions

The main advantage of the proposed methodology is the possibility to predict the outcome variable in MLRM with k random effects. For a future work, it is interesting to apply the EBP methodology in others distributions belonging to exponential family.

Acknowledgments: This work received partial financial support from FAPESP and CNPq.

References

Cole, D. J., Morgan, B. J. T. and Ridout, M. S. (2003). Generalized linear mixed models for strawberry inflorescence data. *Statistical Modelling*, **3** , 273-290.

Jiang, J., and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1-96.

Tamura, K. A., and Giampaoli, V. (2010). Prediction in multilevel logistic regression.*Communications in Statistics - Simulation and Computation*, **39(6)**, 1063-1076.

Multivariate regression smoothing through the "falling net"

James Taylor[1] , Jochen Einbeck[1]

[1] Department of Mathematical Sciences, University of Durham, Durham, DH1 3LE, UK.

Abstract: We consider multivariate regression smoothing through a conditional mean shift procedure. By computing local conditional means iteratively over a set or grid of target points, at each iteration a "net" is formed which gently drifts towards the data cloud, until it settles at the conditional modes of the response distribution. The method is edge-preserving and allows for multi–valued response.

Keywords: Conditional density; modal regression; smoothing

1 Methodology

Given d-variate covariates $X_i = (X_{i1}, ..., X_{id})^T$ and scalar response values Y_i where $i = 1, ..., n$, we find the regression surface via the conditional modes of Y given $X = x$. These are determined by the conditional density function, $f(y|x)$, which can be estimated through

$$\hat{f}(y|x) = \frac{\hat{f}(x,y)}{\hat{f}(x)} = \frac{\sum_{i=1}^{n} G\left(\frac{Y_i - y}{b}\right) \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_j}{h_j}\right)}{b \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_j}{h_j}\right)}, \qquad (1)$$

where G and K are univariate (e.g. Gaussian) kernels, and the subscript j denotes the $j-$th component of the corresponding vector. The values b and h_j are bandwidth parameters to be selected. At each x there may be more than one conditional mode since $\hat{f}(y|x)$ can have several maxima. By setting $\frac{\partial \hat{f}(y|x)}{\partial y} = 0$, one obtains a conditional mode y_m (argument x omitted for ease of notation) as the solution to the estimation equation $y_m = \mu(y_m)$, with

$$\mu(y_m) = \frac{\sum_{i=1}^{n} G\left(\frac{Y_i - y_m}{b}\right) \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_j}{h_j}\right) Y_i}{\sum_{i=1}^{n} G\left(\frac{Y_i - y_m}{b}\right) \prod_{j=1}^{d} K\left(\frac{X_{ij} - x_j}{h_j}\right)}. \qquad (2)$$

Since this cannot be solved analytically, we solve it iteratively using the result by Cheng (1995) that, starting from any $y_0 \in \mathbb{R}$, the mean shift

procedure $y_{\ell+1} = \mu(y_\ell)$ converges to a nearby conditional mode. In order to detect more than one mode for each x it is necessary to specify more than one starting point for the mean shift, typically two. For bivariate predictors, if y_0 is (for all x) set greater than all Y_i, the simultaneous iterative execution of the mean shift resembles visually a net falling onto the data and forming a surface. Of course, if y_0 is below rather than above all Y_i, we would talk about a "rising" net. We emphasize that the techniques proposed in this section do neither require the estimation of any density function, nor the solution of any optimization problem (such as least squares) at any stage; all computational work is carried out by the mean shift.

2 Examples

Figure 1 (left) shows data from a wheat yield trial, where latitude and longitude serve as covariates (the data are part of R package **nlme**, Pinheiro et al. (2008)). Figure 1 (right) provides the surface formed after 30 iterations of the mean shift process on the dataset. Here $h_1 = 3.18$, $h_2 = 3.18$ and $b = 5.61$ after using the bandwidth selection methods described in Section 4.

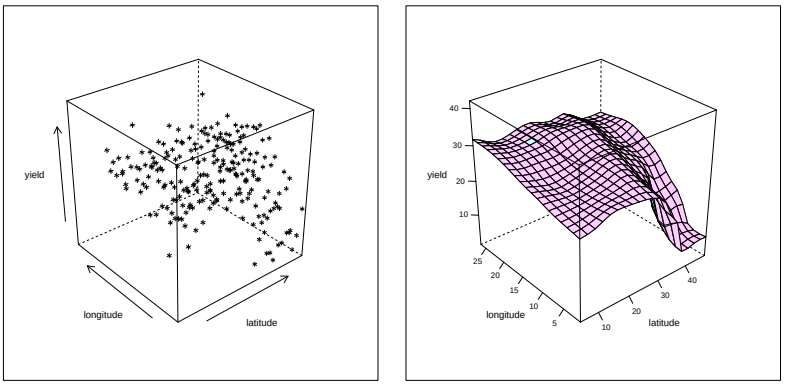


FIGURE 1. The procedure applied to the wheat yield dataset.

Figure 2 illustrates the characteristics of this smoothing technique through simulated data sets of size $n = 200$. Data set A is simulated from the univariate function $y = \sin(0.2x_1) + \cos(x_2)$ and subjected to Gaussian error with standard deviation 0.05. Data set B has a partially bimodal response, which splits for $x_1 \geq 0.5$ into two branches. For $x_1 < 0.5$ the response is simulated from the univariate function $y = 1.5 + 3x_1$ with Gaussian error of standard deviation 0.4. For $x_1 \geq 0.5$, the upper plane is centered at $y = 3$ and the lower plane at $y = 1$; the error standard deviation is 0.2 each. One observes from Figure 2 how the estimated surfaces develop after

different numbers of iterations, ℓ, with starting points positioned *above* (upper estimated surface) and *below* (lower estimated surface) all responses. The right hand column of Figure 2 demonstrates clearly that the procedure is edge-preserving, and able to identify multiple branches when the underlying conditional distribution is multimodal, where other regression techniques could not successfully describe it.

3 Relevance of a mode

When there exist more than one mode of the conditional response distribution for a given x, it is interesting to evaluate the relevance of the different modes. To estimate the probability associated with a conditional mode, one integrates numerically over the part of the estimated conditional density which forms that modal peak. The search for the minimum and the integration can be done simultaneously by descending in small steps from the modes and increasing the integral until either the boundary or the next dip separating the modes is reached (Einbeck and Tutz, 2006). For the simulated data from the right hand column of Figure 2, Figure 3 (left) shows a surface of probabilities, calculated as described, showing the probability of data being present in the mode captured by the "falling net". Figure 3 (right) shows the same for the "rising net." For this data set, the plots show a probability of 1 for about half of all values of x; this is expected since the response is unimodal for these x.

4 Bandwidth selection

In the case of multivariate predictors, the problem of bandwidth selection is more challenging than in the univariate case, since values must be selected for all the h_j as well as for b. For the selection of bandwidth b, one can resort to univariate conditional density bandwidth selectors, such as *cde.bandwidths* in the package **hdrcde**, Hyndman (2010), since this bandwidth does not directly depend on d. Performing this for each covariate separately and then taking the mean as b is effective here. Given b, the h_j are successfully selected by adapting Bashtannyk and Hyndman's (2001) univariate *regression-based bandwidth selector* for use with multivariate covariates, as the authors themselves suggest doing. Therefore we standardize the covariates and search for an optimal $h = h_1 = ... = h_d$. The extended regression-based bandwidth selector minimizes the penalized average squared prediction error $Q(h)$ with respect to h, for a fixed b, where

$$Q(h) = \frac{\Delta}{n}\sum_{k=1}^{N}\sum_{i=1}^{n}\left\{\frac{1}{b}G\left(\frac{Y_i - y'_k}{b}\right) - \hat{f}(y'_k|X_i)\right\}^2 \times p\left(\frac{(K(0))^d}{\sum_{l=1}^{n}\prod_{j=1}^{d}K\left(\frac{X_{ij}-X_{lj}}{h}\right)}\right) \quad (3)$$

where $\{y'_1, ..., y'_N\}$ are equally spaced over the sample space Y with $y'_{i+1} - y'_i = \Delta$ and where $p(u) = (1-u)^{-2}$ is a penalty function. This $p(u)$ is identical to that used in generalized cross-validation, but differs from the one used typically in the univariate case for this technique, since this was found to perform badly in the multivariate setting. Once h has been found, it is unstandardized and the modal regression is then carried out with unstandardized covariates and bandwidths. Following this procedure for the wheat yield data gives the bandwidths stated in Section 2.

5 Discussion

This work constitutes essentially a multivariate extension of the multimodal regression technique introduced in the context of traffic data modelling in Einbeck and Tutz (2006). The problem of bandwidth selection has been addressed by appropriately extending bandwidth selectors which were developed for conditional density estimation with univariate predictors by Bashtannyk and Hyndman (2001).

Attractive features of the technique are the computational simplicity, the edge-preserving property, and the visual appeal. Moreover, the method is able to deal with multi–valued response, though it should be admitted that data of this type are relatively rare, and that multiple modes in the response distribution may be an indicator that important covariates have been omitted from the model. Nevertheless, the presented approach may still serve to detect and visualize situations of this type.

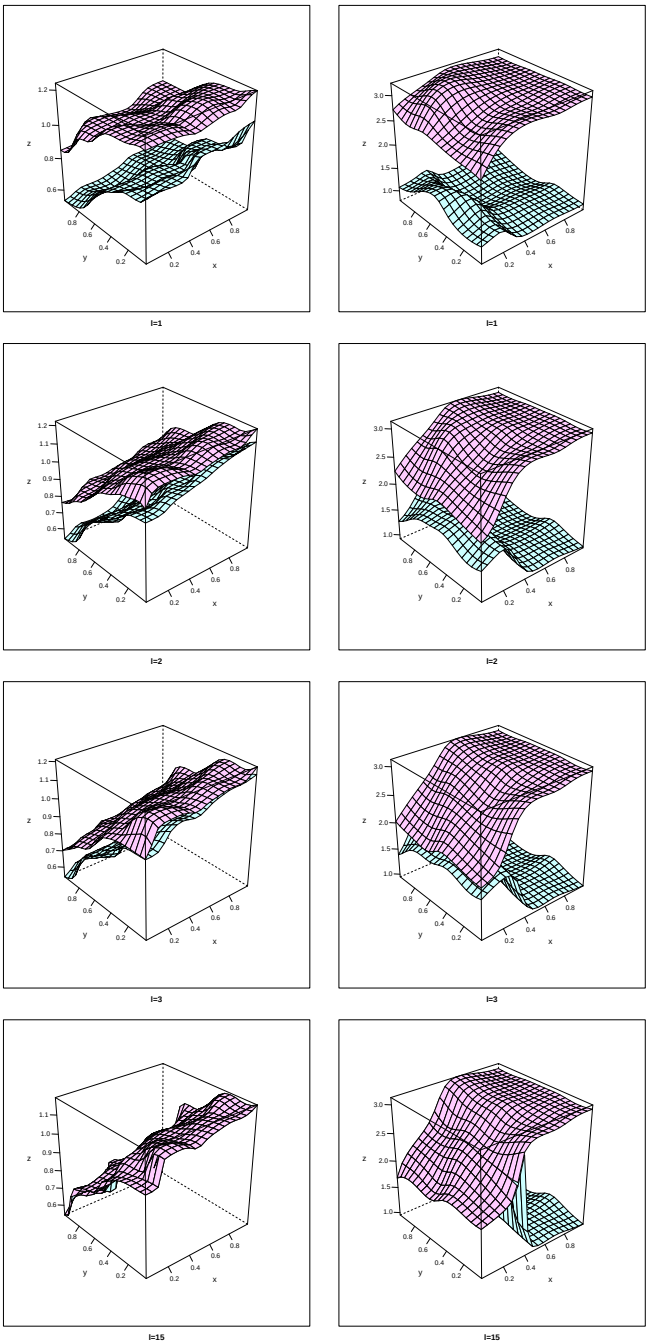


FIGURE 2. The left column displays the surfaces for simulation A, for $\ell = 1, 2, 3, 15$ (from top to bottom). The right column shows the same for simulation B.

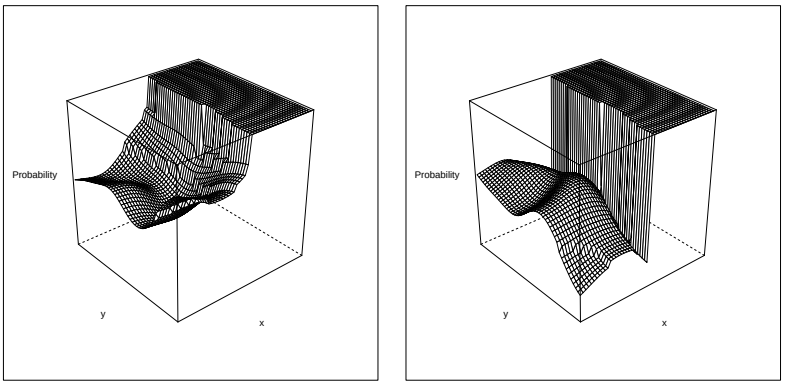


FIGURE 3. Left: Bivariate probability plot for the "falling net" (left) and the "rising net" (right); each for the fitted surface from Figure 2 (bottom right). Note that the orientation is rotated in order to allow for a better view of the probability surfaces.

References

Bashtannyk, D. and Hyndman, R. (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics and Data Analysis*, **36**, 279–298.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Machine Intell.*, **17**, 790–799.

Einbeck, J. and Tutz, G. (2006). Modelling Beyond Regression Functions: An Application of Multimodal Regression to Speed-flow Data. *Applied Statistics*, **55**, 461–475.

Hyndman, R. (2010). **hdrcde**: Highest density regions and conditional density estimation. R package version 2.15.

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D. and the R Dev. Core Team (2011). **nlme**: Linear and Nonlinear Mixed Effects Models. R package version 3.1-98.

Robust model selection in additive penalized regression splines models

K. Tharmaratnam[1], G. Claeskens [1]

[1] ORSTAT and Leuven Statistics Research Center, Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium

Abstract: This paper studies model selection strategies for semiparametric additive models fit with penalized regression splines. This estimation method is attractive because of its link to mixed models. We work specifically with outlier robust versions. In the context of mixed models there exist two different forms of AIC. The marginal AIC (MAIC) is used for selecting covariates in the model, and is based on the marginal likelihood. The conditional AIC (CAIC) is based on the conditional likelihood given the random effects, and is used for estimating smoothing parameters as well as for selecting covariates in the parametric part of the model. Our proposal leads to robust versions of the MAIC and CAIC that are based on S-estimators. Simulated data and real data examples are used to illustrate the effectiveness of the proposed method.

Keywords: Additive model; Linear mixed model; Penalized regression spline; S-estimator.

1 Introduction

Additive penalized regression spline models have found many applications in the last few years. Variable selection in these models is challenging. We need to select variables in the nonparametric component as well as identify significant variables in the parametric component. An AIC based on the marginal likelihood is generally used in linear mixed models (marginal AIC) and returned by standard statistical software. The paper by Vaida and Blanchard (2005) focuses on model selection for linear mixed models using the conditional Akaike information criterion and shows that the classical AIC is not appropriate for conditional inference to take into account both the fixed and random parts of linear mixed models. Liang et al (2008) propose the corrected conditional AIC that accounts for the estimation of the variance parameters. Greven and Kneib (2010) study the theoretical properties of both the marginal and the conditional AIC. All of these variable selection procedures need special care in the presence of outliers in the data. The main purpose of this paper is to study robust versions of the marginal and conditional AIC in linear mixed models obtained from additive semiparametric regression splines.

2 Methodology

Consider the semiparametric additive regression model

$$Y_i = \sum_{j=1}^{p} \beta_j X_{ji} + \sum_{j=p+1}^{q} m_j(X_{ji}) + \varepsilon_i, \qquad i = 1, \cdots, n. \tag{1}$$

where the functions m_j are smooth, but not specified. Model (1) is written using matrix notation as $y = X\gamma + Zu + \varepsilon$. where y is a $(n \times 1)$ response vector, γ and u are, respectively, fixed and random effects, X and Z are design matrices for, respectively, fixed and random effects; ε is the error term. We assume $u \sim N_{qK}(0, G)$, $\varepsilon \sim N_n(0, R)$ where $G = \sigma_u^2 I_{qK}$ and $R = \sigma^2 I_n$. We define $\lambda = \sigma^2/\sigma_u^2$. Estimation of the parameters γ and prediction of u leads to minimize the penalized least squares criterion

$$\parallel Y - X\gamma - Zu \parallel^2 + \lambda u^t D u.$$

Copt and Victoria-Feser (2006) compute the S-estimators of the marginal model $Y \sim N(X\gamma, V)$, where $V = (Z^t G Z + R)$. Instead, we consider the conditional model $Y|u \sim N(X\gamma + Zu, R)$. Assume the use of normal random effects. We consider the joint density of y and u, $f(y, u) = f(y|u)\, f(u)$. The constraint of the S-estimators corresponding to the conditional model is,

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\sqrt{(Y_i - X_i\gamma - Z_i u)^t R^{-1} (Y_i - X_i\gamma - Z_i u)}\right) = b_0. \tag{2}$$

The lagrangian L_{joint} is as follows,

$$\mathrm{L}_{joint} = \log|R| + c\left[\frac{1}{n}\sum_{i=1}^{n} \rho(d_i) - b_0\right] - \frac{1}{2}\log|G| - \frac{1}{2} u^t G^{-1} u + K, \tag{3}$$

where $d_i = \sqrt{(Y_i - X_i\gamma - Z_i u)^t R^{-1} (Y_i - X_i\gamma - Z_i u)}$ and c, b_0 and K are constants. Robust estimators $\widehat{\gamma}$, $\hat{u}$, $\hat{R}$ and $\hat{G}$ are computed by minimizing (3) and solving (2).

2.1 Marginal and Conditional AIC in additive models

The marginal AIC comes from the marginal model $Y \sim N(X\gamma, V)$,

$$\mathrm{MAIC} = -2\, \log\{f(y|\hat{\gamma}, \hat{V})\} + 2\,(p + v + 1),$$

where $f(y|\hat{\gamma}, \hat{V})$ is the maximized marginal likelihood and $(p+v+1)$ is the number of parameters in the marginal model. The corrected conditional AIC comes from the conditional model $Y|u \sim N(X\gamma + Zu, R)$,

$$\mathrm{CCAIC} = -2\, \log\{f(y|\hat{\gamma}, \hat{u}, \hat{R})\} + 2\,(\phi_0 + 1),$$

where $f(y|\hat{\gamma}, \hat{u}, \hat{R})$ is the maximized conditional likelihood and ϕ_0 represents the bias corrected form of the effective degrees of freedom as proposed in Greven and Kneib (2010) for unknown variance components.

2.2 AIC based on S-estimators in additive models

Based on the results in Tharmaratnam and Claeskens (2010), we can write the form of the AIC based on S-estimators in additive models as,

$$\text{AIC.S} = 2\,\log|\widehat{\Sigma}_s| + 2\,\text{penalty}$$

where $\widehat{\Sigma}_s$ is the covariance matrix of the S-estimator. A robust version of the marginal AIC is

$$\text{MAIC.S} = 2\,\log|\widehat{V}_s| + 2\,(p+v+1),$$

where $\widehat{V}_s$ is the S-estimator of V from the marginal model $Y \sim N(X\gamma, V)$, p is the number of columns of X and $v+1$ is the number of variance components. A robust version of the corrected conditional AIC is

$$\text{CCAIC.S} = 2\,\log|\widehat{R}_s| + 2\,\phi_s,$$

where $\widehat{R}_s$ is the S-estimator of R from the conditional model $Y \sim N(X\gamma + Zu, R)$, ϕ_s is the trace of $(\partial\hat{y}/\partial y)$ with all variance components unknown, $\hat{y} = X\hat{\gamma}_s + Z\hat{u}_s$, with $\hat{\gamma}_s$ and $\hat{u}_s$ the S-estimators from the conditional model.

3 Results and Conclusion

We conduct a simulation study to select a best model using marginal and conditional AIC based on ML- and S-estimators.

3.1 Simulation results

We consider the non-linear function $m_1(x) = 1 + x + 2d(0.3 - x)^2$, the covariate values $x \sim U(0,1)$. In the model, d is a constant and increasing values of d correspond to the increased non-linearity. We generate 11 different models corresponding to $d = (0, 5, 10, \ldots, 50)$. The model is linear in x when $d = 0$. In the case of no outliers, the error terms $\epsilon \sim N(0,1)$.
For each value of the constant d, for each simulated data set, we use the MAIC, CCAIC, MAIC.S and CCAIC.S to decide on either the linear model or the more complex model. To assess the performance of the marginal and the conditional AIC for distinguishing between linear and non-linear models, we compute the frequency of selecting the nonlinear model for each d value. We use 1000 simulated data sets for both cases with (a) no outliers and (b) 20% outliers on the error terms, generated from a $N(100, 0.5^2)$ distribution for the sample size $n = 100$.
From Figure 1 we observe that the CCAIC selects a larger proportion of nonlinear models than the MAIC (which is the true model when $d \neq 0$). This holds for both maximum likelihood estimators and S-estimators. In these penalized spline models, the random effects correspond to the spline coefficients. The CCAIC is better suited to decide on the inclusion of random effects (i.e. nonlinear effects in this setup) than the MAIC. The results do not change much for different values of d.

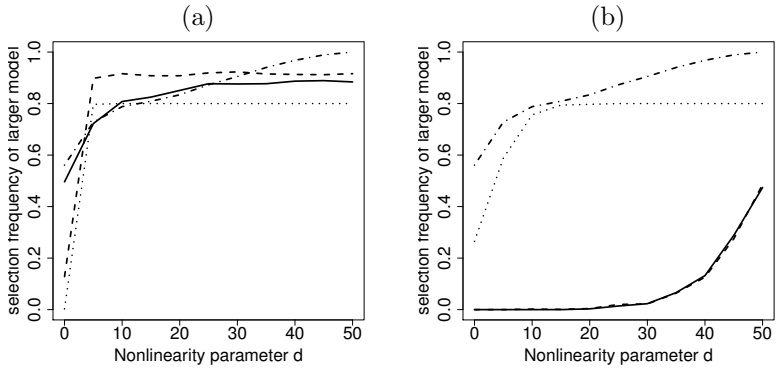


FIGURE 1. Proportion of selected larger models from MAIC (solid line), CCAIC (dashed line), MAIC.S (dotted line) and CCAIC.S (dot-dashed line) with mean function $m_1(x)$ (a) no outliers, (b) 20% of outliers in the data.

3.2 Conclusion

Robust versions of marginal and conditional AIC are needed for the data with outliers on the response variable in additive penalized regression splines models. The CCAIC selects higher proportions of more complex models than the MAIC, with or without outliers in the data.

References

Copt, S., and Victoria-Feser, M. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association*, **101(473)**, 292-300.

Greven, S., and Kneib, T. (2010). On the behavior of marginal and conditional Akaike information criteria in linear mixed models. *Biometrika*, **97(4)**, 773-789.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, **95(3)**, 773-778.

Tharmaratnam, K., and Claeskens, G. (2010). A comparison of robust versions of the AIC based on M, S and MM-estimators. *Statistics*, in press.

Vaida, F., and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351-370.

Statistical modeling of geographic risks for very low birth weights near Texas superfund sites

James A. Thompson[1]

[1] Texas Veterinary Medical Center, College Station, TX USA

Abstract: The first step for investigating potential geographic disease clusters has commonly been to use registry health data to assess the statistical probability of an elevated risk within some arbitrary boundary. Most investigations fail to demonstrate a statistically significant risk and the investigations are often ended. Recent advances in the statistical modeling of geographic risks offer potential to increase the sensitivity of this initial assessment. We propose statistical modeling of an exceedance probability based on geographic coordinates as an alternative to evaluating arbitrary locations. The current study mapped risks for very low birth weights using an Intrinsic Conditional Autoregressive (ICAR) model. The data were adapted to apply to a 20 by 20 grid of pixels centered at each of the 57 listed and recently de-listed federal superfund sites in Texas. A Geographical Information System (GIS) evaluation of the risk estimates and the exceedance probabilities was performed. The study identified several locations of high risk for births with very low birth weight.

Keywords: VLBW; disease mapping; statistical model.

1 Introduction

1.1 Very Low Birth Weight

Fetuses are known to be especially susceptible to environmental toxins and investigations of potential geographic clusters of adverse birth disorders including very low birth weights (VLBW; birth weight < 1500 g) are common (Maisonet et al., 2004). The National Center for Environmental Health (NCEH) defines a cluster as a greater-than-expected number of cases that occurs within a group of people in a geographic area over a defined period of time. The Centers for Disease Control and Prevention (CDC) provide guidelines for investigation of potential clusters but, in the United States, the investigation usually falls to the state health department (Kingsley et al., 2007). Commonly the state health agency examines its health registry data and performs statistical testing comparing incidence rates among arbitrary geographic areas (Wheeler, 2007). Most investigations of potential clusters fail to demonstrate statistical significance at a small p-value. Several developments in the statistical modeling of geographic risk provide

potential to increase the sensitivity of this initial evaluation. Collectively, these advances address multiple aspects of statistical precision, enable flexible cluster shapes and sizes and enable the direct modeling of the exceedance probability.

1.2 Modeling needs

The size and shape of potential clusters need to be flexible (Wheeler, 2007). Investigators typically use chi-square tests of differences in expected and observed case counts in census or political units. This approach does not consider if areas with significantly more cases than expected are close together or scattered across a larger map. Furthermore, when a point source is suspected, there should be a gradient of risk away from the source. Multiple point sources should demonstrate a more complex risk surface with multiple peaks and valleys. It is now possible to directly model the posterior probability that the standardized morbidity rate (SMR) estimate is greater than one (Richardson et al., 2004). This is often called the exceedance probability. This parameter is affected by both the magnitude and the precision of the SMR. We propose a statistical modeling of the exceedance probability, based on the risk at geographic coordinates. The objective of this study was to apply and evaluate small-scale Intrinsic Conditional Autoregressive (ICAR) modeling of VLBW near 57 listed and recently de-listed federal superfund sites in Texas.

2 Data

2.1 Registry health data

A database was created to evaluate the geographic risks according to the mother's living location at the time of childbirth as determined from birth certificate data. Briefly, the creation of the database involved the retrieval of all Texas birth records from January 1, 1990 to December 31, 2002 from the Texas Department of State Health Services (TDSHS). Geocoding was performed by the TDSHS, based on street addresses, and was 87 percent complete. The latitude and longitude of birth locations were projected into Universal Transverse Mercator 1983 (UTM83), Zone 14 units. All births were located within a pixel using the mother's address at birth and then using GIS analysis. All birth weights less than 1500 g were classified as very low birth weights.

2.2 Superfund sites

The identities and locations of the 57 listed and recently delisted superfund sites were first identified using latitude and longitude given on the Texas Site Status Summaries on the EPA Program Region 6 Superfund website (`http://www.epa.gov/earth1r6/6sf/6sf-tx.htm`). The superfund sites

were then visually identified on satellite imagery and the apparent centroid was used as the location. The latitude and longitude of toxic site locations were projected into Universal Transverse Mercator 1983 (UTM83), Zone 14 units.

3 Modeling

3.1 Intrinsic conditional autoregressive (ICAR)

We specified a convolution prior for the area-specific random effects using the BYM model which partitions the overall random effect for each area into the sum of a structured or spatial component plus an unstructured or non-spatial component (Besag et al., 1991). The original BYM model applied to continuous data that could be assumed to be normally distribution. In disease mapping studies, this has been adapted to incorporate normally distributed spatially correlated random effects into Poisson models for disease counts. This adaptation usually models the SMR, which is defined as the number of cases / number of expected cases. We used the UTM83 coordinates with a distance scale of 1 km between coordinates and modelled the 20 x 20 grid surrounding each of the superfund sites. Each x,y coordinate then represented the centroid of a 1 km x 1 km location that we call a pixel. For each of the 400 pixels, the number of cases was counted and the expected number of cases was the sum of individual expectations accounting for individual risk factors. All models employed Bayesian inference, with vague or flexible prior beliefs and an MCMC implementation. The MCMC implementation was performed by use of WinBUGS version 3.1.1 (Spiegelhalter et al., 2003). The initial 5,000 iterations were discarded to allow for convergence and every hundredth of the following 1,000,000 iterations were sampled for the posterior distribution. Observing convergence of two chains with widely different initial values checked convergence to the posterior distribution.

3.2 DIC evaluation

To evaluate the spatial variance component, the full model was compared to a reduced model using the Deviance Information Criterion (Spiegelhalter et al., 2002). For the reduced model, spatial covariance was removed from the model and random covariance retained. The DIC was modified to use half the variance of the Deviance (var(D)/2) as the measure of model complexity (Gelman et al., 2004).

3.3 GIS evaluation

When the DIC comparison supported the spatial modeling (i.e., the DIC for the convolution model was lower than the DIC for the random model), a GIS was used to evaluate SMR The parameterization used for GIS evaluation was the posterior probability that the spatial SMR estimate was

greater than one and were taken directly from the full posterior distributions. We chose to report maps that identified at least 1 pixel with exceedance probability greater than 0.95 and we also identified the number of maps that had adjacent pixels with exceedance probability greater than 0.95.

4 Main results

The areas around 57 federal superfund sites in Texas were evaluated for incidence of very low birth weights. The incidence data were fit better by a convolution model than the random model in 53 percent of the locations (30/57). Spatial covariance should be considered dependent upon the existence of one or more spatially oriented causes. The search for local cancer clusters, based on local spatial covariance, is considered both more useful and also more specifically addresses public concerns The current study deals with local spatial covariance by restricting the modeling of spatial covariance to locations very near the putative toxic source and not statewide. Of these superfund site locations 30 percent (17/57) had one or more pixel, each representing a square kilometer, with an exceedance probability of greater than 95 percent. Many of the locations had single or multiple but isolated high-risk pixel locations. When the map of risks shows isolated pixels with high exceedance probabilities, it indicates that the random component of the convolution model is important. For locations with this finding, the possibility that there is a spatial pattern within a 1 km x 1 km location exists and our study did not evaluate this possibility. Our objective was to increase the sensitivity of the initial evaluation of registry data. The current study was successful in identifying high-risk locations. Lower exceedance probabilities will also generate considerable interest and could be used to justify further investigation. Twelve percent of the locations (7/57) had at least two neighboring pixels with exceedance probabilities of greater than 95 percent. Evaluation of these maps was very subjective and each of these maps will be presented. It is difficult to gain causal inference from these maps because there very often are multiple known point-sources near the cluster and that the exact size and shape of the cluster varies with the choice of the center of the mapping location. Furthermore, the issue of confounding will be very difficult to deal with, especially with small numbers of cases.

5 Conclusions

The study identified several locations of high risk for births with very low birth weight. These findings support a more vigilant approach of risk assessment before investigators dismiss public concerns regarding a potential disease cluster.

Acknowledgments: This research was funded by the National Institutes of Health

References

Besag J., York J.C., Mollie A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion) *Annals of the Institute of Statistical Mathematics*, **43**, 1.

Gelman A., Carlin J.B., Stern H.S., Savitz D.A. (2004). ıtBayesian Data Analysis. Boca Raton: Chapman and Hall/CRC.

Kingsley B.S., Schmeichel K.L., Rubin C.H. (2007). An update on cancer cluster activities at the Centers for Disease Control and Prevention *Environmental Health Perspectives* , **115**, 165.

Maisonet M., Correa A., Misra D., Jaakkola J.J.K. (2004). A review of the literature on the effects of ambient air pollution on fetal growth. *Environmental Research*, **95**, 106.

Richardson S., Thomson A., Best N., Elliott P. (2007). Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives* , **112**, 1016.

Spiegelhalter D.J., Best N.G., Carlin B.P., van der Linde A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* , **64**, 1.

Spiegelhalter D., Thomas A., Best N., Lunn D. (2003). ıtWinBUGS User Manual: Version 1.4.. Cambridge: MRC Biostatistics Unit.

Wheeler D. (2007). A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996 - 2003. *International Journal of Health Geographics* , **6**, 13.

Spatio-temporal risk smoothing and forecasting with P-splines

M. D. Ugarte[1], T. Goicoa[1], J. Etxeberria[1,2], A. F. Militino[1]

[1] Department of Statistics and O. R., Universidad Pública de Navarra
[2] CIBER in Epidemiology and Public Health

Abstract:
To study the evolution of geographical patterns of mortality or incidence risks is very useful for epidemiologists and public health researchers to have a full picture of the disease in space and time. However, official figures about mortality/incidence up to a certain year are only available after a few years later. This is why predicting mortality risks/incidence for future years is a necessary task. These predictions are very valuable for health agencies to plan and coordinate public health programs and clinical services. In this work, a P-spline spatio-temporal model is considered for smoothing and forecasting space-time risks. The mean squared error of the forecast values will be also derived. Results are illustrated with prostate cancer data in Spain from 1975 to 2008. Forecast values will be provided up to 2011.

Keywords: prostate cancer; mortality risk prediction; mean squared error.

1 Introduction

Epidemiologists are interested in studying the spatio-temporal distribution of mortality/incidence risks to find some clues about the disease etiology. It is also of great interest to make risk predictions for future years because official figures about mortality/incidence up to a certain year are only available after a few years later, and this information is extremely valuable to plan and coordinate public health programs, clinical services, and management strategies. There are some work on prediction in disease mapping. Clements et al. (2005) consider generalized additive models for cancer rate predictions and compare them with Bayesian age-period-cohort models, but they do not consider the spatial dimension. Malvezzi et al. (2011) simply use joint point regression models to forecast cancer rates in the European Union and they do not include the spatial component either.
In this work, a spatio-temporal P-spline model is considered to model the spatio-temporal distribution of risks and to forecast risk values for future years. This model incorporates spatio-temporal interactions. P-spline models have been used in the literature for forecasting purposes. For example, in the context of insurance and pensions industry, Currie et al. (2004)

propose a method for fitting and forecasting simultaneously when the coefficients are estimated using the expressions dependent on the B-splines basis. Ugarte et al. (2009) consider a semiparametric longitudinal model to smooth and forecast dwelling prices in several neighborhoods (small areas) of a Spanish city. They also derive the prediction MSE of both fitted and forecast values, as well as estimators of those quantities.
Very recently Ugarte et al. (2010) use a spatio-temporal P-spline model incorporating space-time interactions to smooth risks. These authors exploit the mixed model representation of the P-spline model and use the well known penalized quasi-likelihood technique (PQL)(Breslow and Clayton, 1993) for model estimation. The authors also provide the mean squared error of the log risk predictor. In this work, this model is considered, and attention is focussed on predicting risks for future years. Using the mixed model reformulation of the P-spline model, forecasting will be carried out by extending the B-spline basis in the time dimension.
The methodology will be used to analyze male prostate mortality cancer data for 50 Spanish provinces (excluding Ceuta and Melilla) in the period 1975-2008. Risks predictions for future years will be also provided.

2 Spatio-temporal P-spline model

In this section, a spatio-temporal P-spline model (Ugarte et al., 2010) is described. This model depends on a B-spline basis and on a penalty matrix to control function wiggliness.
Let us consider n adjacent regions labelled $i = 1, \ldots, n$, and T time periods denoted by $t = 1, \ldots, T$. Then, conditional on the relative risks r_{it}, the number of deaths from a rare disease in each area and time period, C_{it}, is assumed to be Poisson distributed with mean $\mu_{it} = e_{it} r_{it}$, that is

$$C_{it}|r_{it} \sim Poisson(\mu_{it} = e_{it} r_{it}), \quad \log \mu_{it} = \log e_{it} + \log r_{it}. \tag{1}$$

Then, the $\log r_{it}$ is modeled as

$$u_{it} = \log r_{it} = f(x_{1i}, x_{2i}, t) = \theta_1 B_1(x_{1i}, x_{2i}, t) + \ldots + \theta_K B_K(x_{1i}, x_{2i}, t),$$

where x_{1i} and x_{2i} are the coordinates of the centroid of the ith small area (longitude and latitude respectively), t is the time, f is a smooth function to be estimated using P-splines with B-spline bases, θ_k are the model coefficients, and B_k are the elements of the B-spline basis defined as $\mathbf{B} = \mathbf{B}_3 \otimes \mathbf{B}_s$. Here, $\mathbf{B}_3$, is the marginal B-spline basis for time and $\mathbf{B}_s$ is the "row-wise" Kronecker product of the marginal B-spline bases for latitude and longitud (see Eilers et al., 2006).
To control function wiggliness, the P-spline approach places penalties on the coefficients . If $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_3$ are penalties for the marginal basis the following penalty is considered

$$\mathbf{P} = \lambda_1 I_{c_3} \otimes I_{c_2} \otimes \mathbf{P}_1 + \lambda_2 I_{c_3} \otimes \mathbf{P}_2 \otimes I_{c_1} + \lambda_3 \mathbf{P}_3 \otimes I_{c_2} \otimes I_{c_1} \tag{2}$$

where λ_i and c_i, $i = 1, 2, 3$ are smoothing parameters and the number of columns of the marginal B-spline bases respectively, $\mathbf{P}_i = \mathbf{D}_i'\mathbf{D}_i$, and the matrices $\mathbf{D}_i$ are difference matrices to impose smoothing over adjacent coefficients.
One of the most interesting aspects of the P-spline models is that they can be reformulated as linear mixed models (in our case, as generalized linear mixed models) using a one-to-one (orthogonal) transformation. Hence, the P-spline model can be represented as

$$\mathbf{u} = \mathbf{B}\theta = \mathbf{X}\beta + \mathbf{Z}\alpha, \quad \alpha \sim N(\mathbf{0}, \mathbf{F}^{-1}), \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are the fixed and random effects matrices, and $\mathbf{F}^{-1}$ is the diagonal covariance matrix for the random effects α derived from the mixed model representation of the P-spline model (see Ugarte et al., 2010 for more details).

3 Illustration

To illustrate the methodology, Figure 1 displays relative risks estimates (1975-2008) and predictions (2009-2011) for six selected Spanish provinces, together with 95% confidence bands. A decreasing trend in mortality can be observed.

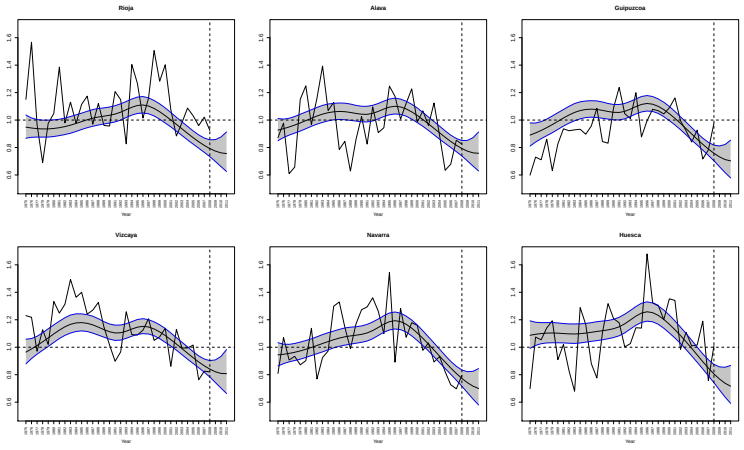


FIGURE 1. Smoothed prostate cancer mortality risks estimations and predictions with 95% confidence bands.

Acknowledgments: This research has been supported by the Spanish Ministry of Science and Innovation (MTM 2008-03085/MTM). The authors

would like to thank to Marina Pollán from the National Epidemiology Center (area of Environmental Epidemiology and Cancer) for providing the data.

References

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

Clements M.S., Armstrong B.K. and Moolgavkar S.H. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics*, **6**, 576-589.

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

Eilers P.H.C., Currie I.D., and Durbán M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, **50**, 61-76.

Malvezzi M., Arfé A., Bertuccio P., Levi F., La Vecchia C. and Negri E. (2011). European cancer mortality predictions for the year 2011. *Annals of Oncology*, **22**, 947-956.

Ugarte M.D., Goicoa T., Militino A.F., and Durbán M. (2009). Spline smoothing in small area trend estimation and forecasting. *Computational Statistics and Data Analysis*, **53**, 3616-3629.

Ugarte, M.D., Militino, A.F., and Goicoa, T. (2010). Spatio-temporal modelling of mortality risks using penalized splines. *Environmetrics*, **21**, 270-289.

Bioassays models with natural mortality and random effects

Mariana Ragassi Urbano[1], John Hinde[2], Clarice Garcia Borges Demétrio[1]

[1] Departamento de Ciências Exatas, ESALQ/USP, Piracicaba, São Paulo, Brazil, mrurbano@esalq.usp.br, clarice@esalq.usp.br
[2] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland, john.hinde@nuigalway.ie

Abstract: In fitting dose-response models to entomological data it is often necessary to take account of natural mortality and/or overdispersion. The standard approach to handle natural mortality is to use Abbott's formula. Standard overdispersion models include beta-binomial models, logistic-normal, and discrete mixtures. Here we consider combining these two aspects with extensions that allow for the modelling of the natural mortality and overdispersion. Two models are developed: one including a random effect in the linear predictor and other including a random effect in the natural mortality. We consider the application of these models to data from an experiment on the use of a virus (*PhopGV*) for the biological control of worm larvae (*Phthorimaea operculella*) in potatoes. Using the models with random effects, we obtained a better fit than that provided by the standard model.

Keywords: Bioassay; Natural mortality; Overdispersion; Random effects.

1 Introduction

Models for binary and binomial response grew out of the needs of a type of experimental investigation known as bioassay. In a typical bioassay, different concentrations of a chemical compound are applied to batches of experimental subjects and the number of subjects in each batch that respond to the chemical is then recorded. These values are regarded as observations on a binomial response variable. Some experiments in entomology exhibit evidence that responses can occur even at zero dose; here the response of interest is death and this phenomenon is referred to as *natural mortality*. Also the variation of the data may be greater than that predicted by the model, commonly described as *overdispersion*.

1.1 Natural mortality and overdispersion

Among the available methods for the analysis of data with natural mortality, only a few can also handle overdispersion. According to Collet (2002),

this additional variation can be attributed to relevant explanatory variables that have not been adequately measured or controlled. This situation can be modeled by the inclusion of a random effect and so a mixed model can be used in modelling overdispersion. We modified the usual model, first proposed by Abbott (1925), and considered two other models: one with a random effect in the linear predictor of the dose levels, and another in which the natural mortality was taken to be random.

1.2 Description of the dataset

The application here is to an experiment in which potatoes (*Solanum tuberosum L.*) were each infected with $m_{ij} = 30$ larvae of *Phthorimaea operculella*, and then, D different concentrations of a virus ($PhopGV$) were applied to samples of n_i potatoes (observations are indexed by $i = 1, \ldots, D$ and $j = 1, \ldots, n_i$). There was also a control sample (no virus, $i = 0$) with $n_0 = 9$ potatoes. The experiment was conducted at 18^oC, and after 60 days the numbers of dead larvae y_{ij} were counted.

2 Methodology

In modeling the observed proportions y_{ij}/m_{ij}, the y_{ij} can be assumed to have a $B(m_{ij}, \pi^*_{ij})$ distribution, where π^*_{ij} the probability of response depends on the natural mortality and the dose-response relationship.
A model for π^*_{ij} (Morgan, 1992) is therefore

$$\pi^*_{ij} = \omega_{ij} + (1 - \omega_{ij})\pi_{ij}, \quad j = 1, ..., n_i \quad \text{and} \quad i = 0, ..., D \tag{1}$$

where π_{ij} is given by the tolerance distribution (normal, logistic or extreme value), and ω_{ij} is the natural response probability. In general, we can model π_{ij} and ω_{ij} as function of covariates and parameters defining three different models:

Model (a) - Standard model

$$\log\left(\frac{\omega_{ij}}{1-\omega_{ij}}\right) = \boldsymbol{\gamma'}\mathbf{u}_{ij} \quad \text{and} \quad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{\beta'}\mathbf{x}_{ij}.$$

Model (b) - Random effect in the linear predictor of the dose levels

$$\log\left(\frac{\omega_{ij}}{1-\omega_{ij}}\right) = \boldsymbol{\gamma'}\mathbf{u}_{ij} \quad \text{and} \quad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{\beta'}\mathbf{x}_{ij} + \sigma z_i.$$

Model (c) - Random effect in natural mortality

$$\log\left(\frac{\omega_{ij}}{1-\omega_{ij}}\right) = \boldsymbol{\gamma'}\mathbf{u}_{ij} + \tau v_i \quad \text{and} \quad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{\beta'}\mathbf{x}_{ij},$$

where z_i and v_i are random effects with standard normal distribution.
If were possible to label the subjects who responded due to the applied dose as y_{ijd} and those who responded naturally as y_{ijc} then the total number of dead at dose d_{ij} would be

$$y_{ij} = y_{ijc} + y_{ijd}.$$

In the control group the number of larvae that died of a total of m_{0j} that did not receive the virus is $y_{0j} = y_{0jc}$.
The likelihood of Models (a), (b) and (c) is given by

$$\begin{aligned} L(\omega_{ij}, \pi_{ij}; y_{ij}) &\propto \prod_{i=1}^{D}\prod_{j=1}^{n_i} [(1-\omega_{ij})(1-\pi_{ij})]^{m_{ij}-y_{ij}} \, [(1-\omega_{ij})\pi_{ij}]^{y_{ijd}} \omega_{ij}^{y_{ijc}} \\ &\times \prod_{j=1}^{n_i} \omega_{ij}^{y_{0j}} (1-\omega_{ij})^{m_{0j}-y_{oj}}. \end{aligned} \tag{2}$$

The log-likelihood of (2) is given by

$$\begin{aligned} l(\omega_{ij}, \pi_{ij}; y_{ij}) &\propto \sum_{i=1}^{D}\sum_{j=1}^{n_i} \{(m_{ij}-y_{ij}) \log[(1-\pi_{ij})] + y_{ijd} \log(\pi_{ij}) \\ &+ (m_{ij}-y_{ij}) \log(1-\omega_{ij}) + y_{ijd} \log(1-\omega_{ij}) \\ &+ y_{ijc} \log(\omega_{ij})\} \\ &+ \sum_{j=1}^{n_0} [y_{0j} \log(\omega_{ij}) + (m_{0j}-y_{0j}) \log(1-\omega_{ij})] \\ &= l(\pi_{ij}; y_{ij}) + l(\omega_{ij}; y_{ij}). \end{aligned} \tag{3}$$

This log-likelihood is easy to maximize, because $l(\pi_{ij}; y_{ij}) + l(\omega_{ij}; y_{ij})$ can be maximized separately. The approach used to estimate the parameters was the EM algorithm (Dempster et al., 1977), as also used in bioassays with natural mortality by Hasselblad (1980). With the EM algorithm, the complete log-likelihood (3) is maximized iteratively by alternating between estimating y_{ijc} by its expectation under the current estimates of π_{ij} and ω_{ij} (E step) and then, with the y_{ijc}'s fixed at their expected values from the E step, maximizing $l(\omega_{ij}, \pi_{ij}; y_{ij})$ (M-step).
Let $\boldsymbol{\psi}$=$(\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma)$ be the combined parameter vector. The likelihood of Model (b) is given by

$$L(\boldsymbol{\psi}; \mathbf{y}) = \prod_{i=0}^{D} \left\{ \int_{-\infty}^{+\infty} \left[\prod_{j=1}^{n_i} P(y_{ij}|\boldsymbol{\psi}) \right] \phi(z_i) dz_i \right\}. \tag{4}$$

The integral in the likelihood (4) does not have a closed form except for Y normal, and so for other response models it is approximated by a Gaussian quadrature: the integral is replaced over the normal Z_i by the finite sum

over K Gaussian quadrature mass points z_k with masses α_k (Aitkin et al. 2009). The likelihood is then

$$L(\boldsymbol{\psi};\mathbf{y}) \quad = \quad \prod_{i=0}^{D}\left\{\sum_{k=1}^{K}\left[\prod_{j=1}^{n_i} P(y_{ij}|\boldsymbol{\psi})\right]\alpha_k\right\},$$

where $P(y_{ij}|\boldsymbol{\psi}) = \binom{m_{ij}}{y_{ij}}(\pi^*_{ij})^{y_{ij}}(1-\pi^*_{ij})^{m_{ij}-y_{ij}}$.
The likelihood is thus (approximately) the likelihood of a finite mixture of exponential families density with known mixture proportions α_k at know mass-points z_k, thus z_k becomes another observable variable in the regression, with regression coefficient σ. The log-likelihood is $l(\boldsymbol{\psi};\mathbf{y}) = \sum_{i=0}^{D}\log\left(\sum_{k=1}^{K}\alpha_k\rho_{ik}\right)$, with $\rho_{ik} = \prod_{j=1}^{n_i} P(y_{ij}|\boldsymbol{\psi})$.
Then

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=0}^{D}\frac{\sum_{k=1}^{K}\alpha_k\rho_{ik}\frac{\partial \log\rho_{ik}}{\partial\boldsymbol{\beta}}}{\sum_{k=1}^{K}\alpha_k\rho_{ik}} = \sum_{i=0}^{D}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}\mathbf{s}_{ijk}(\boldsymbol{\beta}),$$

where w_{ik} is the posterior probability that observation y_{ij} comes from component k,

$$w_{ik} = \frac{\alpha_k\rho_{ik}}{\sum_{l=1}^{K}\alpha_k\rho_{il}}$$

and $\mathbf{s}_{ijk}(\boldsymbol{\beta})$ is the $\boldsymbol{\beta}$-component of the score function for observation (ij) in component k,

$$\mathbf{s}_{ijk}(\boldsymbol{\beta}) = \frac{(y_{ij}-\mu_{ijk})\mathbf{x}_{ij}}{\left(\frac{m_i\mu-\mu^2}{m_i}\right)g'_{ijk}}.$$

Following Anderson and Hinde (1988), the estimate of σ can be found by regarding Z_i as an additional covariate and σ as an extra parameter in the linear predictor. Estimation proceeds by fitting a weighted generalized linear model using w_{ik} as additional weights. These weights are functions of Z_i, Y_i, σ and $\boldsymbol{\beta}$ and must themselves be estimated iteratively.
The steps of the EM algorithm for models (a) (b) and (c) are the following:

$\mathbf{E}-\mathbf{Step}$: Estimate $\mathrm{E}(y_{ijc}|y_{ij})$ under the current $\pi^{(k)}_{ij}$ and $\omega^{(k)}_{ij}$

$$y^{(k)}_{ijc} = \mathrm{E}(y_{ijc}|y_{ij},\pi^{(k)}_{ij},\omega^{(k)}_{ij}) = \left\{ \quad \frac{\omega_{ij}y_{ij}}{\omega_{ij}+(1-\omega_{ij})\pi_{ij}} \qquad \text{for } i=1,...,D;\right.$$

$\mathbf{M}-\mathbf{Step}$ for π_{ij}: Find $\pi^{(k+1)}_{ij}$ by maximizing $l(\pi_{ij};y^{(k)}_{ijc}|y_{ij})$: $\pi^{(k+1)}_{ij}$ can be found from a binomial logistic regression of the responses $y^{(k)}_{ijd} = y_{ij} - y^{(k)}_{ijc}$ with binomial denominator $m_{ij} - y^{(k)}_{ijc}$ and different design matrix for models (a), (b) and (c):

Model (a) - design matrix $\mathbf{X}$;

Model (b) - with weights w_{ik} for a design matrix $\mathbf{X}$ augmented by a vector $\mathbf{z}$ of the k Gaussian quadrature points;

Model (c) - design matrix $\mathbf{X}$;

$\mathbf{M-Step}$ for ω_{ij}: Find $\omega_{ij}^{(k+1)}$ by maximizing $l(\omega_{ij}; y_{ijc}^{(k)}|y_{ij})$: using a binomial logistic regression of the responses y_{0j} and $y_{ijc}^{(k)}$ with binomial denominators m_{0j} and m_{ij} and different design matrix for models (a), (b) and (c):

Model (a) - design matrix $\mathbf{U}$;

Model (b) - design matrix $\mathbf{U}$;

Model (c) - with weights w_{ik} for a design matrix $\mathbf{U}$ augmented by a vector $\mathbf{z}$ of the k Gaussian quadrature points.

In models (b) and (c) 20 quadrature points were used and the procedures were implemented in the R package.

3 Main Results and Conclusions

We included in the standard model for natural mortality random effects, with the aim to provide a better fit when the dataset exhibits overdispersion. We concluded that data from biological assays that present natural mortality and overdispersion can be more realistically modelled when a random effect is included to account for variability in the larvae that received the same dose. Table 1 presents the fit statistics (-2 Log Likelihood, AIC, and BIC) for models (a), (b) and (c). For these three statistics, the

TABLE 1. Fit Statistics: -2 Log Likelihood, AIC, and BIC for models (a), (b) and (c)

Fit Statistics	Model (a)	Model (b)	Model (c)
-2 Log Likelihood	394.30	355.20	356.30
AIC	400.30	363.20	364.30
BIC	399.47	362.09	363.19

smaller the value the better is the fit. Can conclude that the model with random effect in the linear predictor of the dose levels provides a better fit than the model without the random effect in the linear predictor .
In Figure 1 is the plot of model (b) with equation given by

$$\hat{\pi}_{ij} = 0.32 + (0.68)\frac{\exp[-7.61 + 1.66\log_{10}(d_{ij}) + 0.74z_j]}{1 + \exp[-7.61 + 1.66\log_{10}(d_{ij}) + 0.74z_j]}.$$

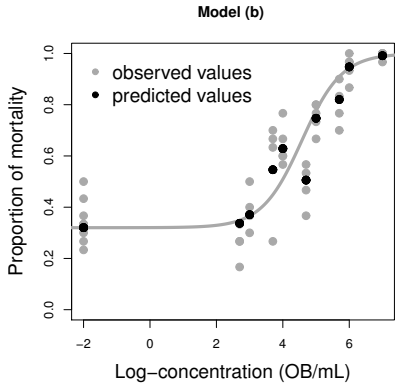


FIGURE 1. Proportion of mortality, fitted curve and predicted values for Model (b)

Acknowledgments: This work was supported by CAPES - Proc n^o 4942/10-8 and Science Foundation Ireland award 07/MI/012.

References

Abbott, W.S. (1925). A method of computing the effectiveness of an insecticide. *Journal of Economic Entomology*, **18**.

Aitkin, M., Francis, B., Hinde, J. (2009). *Statistical Modelling in R*. Oxford University Press, Oxford.

Anderson, D.A., Hinde, J. (1988). Random effects in generalized linear models and the EM algorithm. *Communications in Statistics - Theory and Methods*, **17**.

Collet, D. (2002). *Modelling Binary Data*. Chapman and Hall/CRC.

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

Hasselblad, V., Stead, A.G., Creason, J.P. (1980). Multiple probit analysis with a nonzero background. *Biometrics*, **36**.

Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*. Chapman and Hall/CRC.

R (2011). Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0. URL `http://www.R-project.org`.

A study to compare HGLM and GAMLSS in mixed linear models

Olga Cecilia Usuga Manco[1], Freddy Hernandez Barajas[1], Viviana Giampaoli[1]

[1] ousuga@gmail.com, fhernanb@gmail.com, vivigi08@gmail.com, Mathematics and Statistics Institute, Sao Paulo University

Abstract: This work presents a study to compare the goodness of fit of linear mixed models for some families of distributions through hierarchical generalised linear model (HGLM) and generalized linear additive model for location, scale and shape (GAMLSS). Simulations were used and the measure of goodness of fit considered was the average of mean squared error (MSE) . According to the simulations results it was found that two models showed similar results.

Keywords: Generalized linear additive model for location, shape and scale; Hierarchical generalized linear model; Linear mixed model.

1 Hierarchical generalized linear model

Lee and Nelder (1996) originally defined HGLM as follows:

1. Conditional on random effects u, the responses $\mathbf{y}$ follow a generalized linear model (GLM) family, satisfying

$$E(\mathbf{y}|u) = \mu, Var(\mathbf{y}|u) = \phi V(u).$$

The linear predictor takes the form

$$\eta = g(\mu) = \mathbf{X}\beta + \mathbf{Z}\upsilon$$

where $\upsilon = \upsilon(u)$ for some monotone function $\upsilon()$, are the random effects and β are the fixed effects.

2. The random component u follows a distribution conjugate to a GLM family of distributions with parameters λ.

2 Generalized linear additive model for location, shape and scale

The generalized linear additive models for location, shape and scale (GAMLSS) introduced by Rigby and Stasinopoulos (2005) are defined as follows:

Let $\mathbf{y}^T = (y_1, \ldots, y_n)$ be the vector of the response variable observations. Also, for $k = 1, \ldots, 4$ let $g_k(\cdot)$ be a known monotonic link function relating the parameters $(\mu_i, \sigma_i, \nu_i, \tau_i)$ to explanatory variables and random effects through an additive model given by

$$g_k(\theta_k) = \eta_k = \mathbf{X}_k\beta_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\gamma_{jk}$$

where θ_k and η_k are vectors of length n, $\mathbf{X}_k$ is a known design matrix of order $n \times J_k^{'}$, β_k is a vector of parameters of length $J_k^{'}$, $\mathbf{Z}_{jk}$ is a fixed known $n \times q_{jk}$ design matrix and γ_{jk} is a q_{jk}-dimensional random variable with independent normal distributions $\gamma_{jk} \sim N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^{-1})$, where $\mathbf{G}_{jk}^{-1}$ is the inverse of $q_{jk} \times q_{jk}$ symmetric matrix $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\lambda_{jk})$, which may depend on a vector of hyperparameters λ_{jk}.

3 Results and discussion

3.1 Poisson-Gamma model

The model structure is based on an example presented by Ronnegard et. al (2011).Let y_{ij} be the jth response variable on the ith group ($i = 1, \ldots, n$, $j = 1, \ldots, m$), where y_{ij} follows a distribution Poisson and $g(\mu_i) = \log(\mu_i)$ is the link function relating the parameter μ with the explanatory variable and random effect through

$$\eta = \log(\mu_i) = \mathbf{X}_i\beta + \mathbf{Z}_i\upsilon \tag{1}$$

where $\mathbf{X}_i$ is an $m \times 2$ design matrix for the fixed component, β is the vector of unknown parameters, $\mathbf{Z}_i$ is an $m \times n$ design matrix for the random component and υ is an n random vector, where $\mathbf{v} = \log(\mathbf{u})$ and $u \sim gama(1/\lambda, \lambda)$. The first column of $\mathbf{X}_i$ is represented by $1s$ and the second column is represented by the random sample of the Poisson distribution with mean 2. The values for the parameters considered were $\lambda \in \{0.01, 0.05, 0.1, 1.0\}$, $m = 5, 10, 15, 20, 25$ observations by group, $n = 5, 10, 15$ groups, and $\beta = (-1.5, 0.5)$ fixed.
For each combination of the parameters was generated the matrix $\mathbf{X}_i$, the matrix $\mathbf{Z}_i$ and the mean of $\mathbf{y}_i$ through the equation (1), then was estimated the vector of fixed effects $\hat{\beta}$ through the HGLM and the GAMLSS. To evaluate the performance of the two models was calculate the multivariate Mean Square Error (MSE) de $\hat{\beta}$ given by

$$MSE(\hat{\beta}) = tr(\Sigma(\hat{\beta})) + (\hat{\beta} - \beta)^T(\hat{\beta} - \beta) \tag{2}$$

where tr represents the trace of the covariance matrix Σ of $\hat{\beta}$. This procedure was repeated 10000 times.

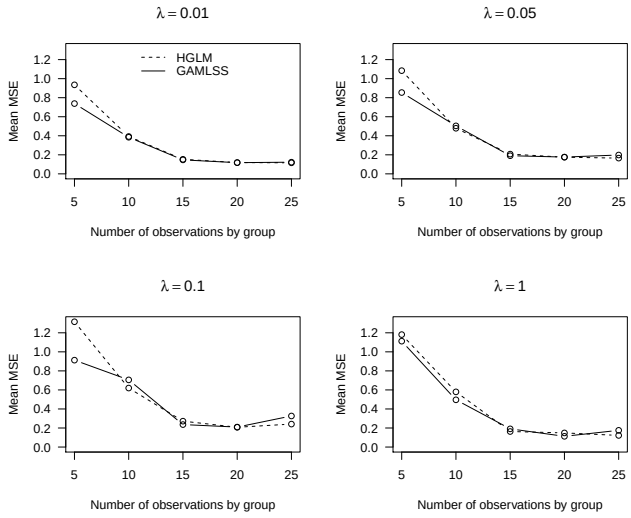


FIGURE 1. Mean MSE for the estimators obtained for the Poisson-gamma model with 5 groups

In Figure 1, the solid lines correspond to the mean of MSE of the model GAMLSS and the dashed ones to the mean of MSE of the model HGLM with $n = 5$. As seen on the graph, the two models presents similar mean MSE and the model HGLM has mean MSE larger than GAMLSS when $n = 5$. Were also performed simulations with number of groups of $n = 10, 15$ and the results showed the same pattern.

3.2 Poisson-normal model

In this model the response variable follow the Poisson distribution and the link function is given in equation (1), where υ is an n random vector with $\mathbf{v} = \mathbf{u}$ and $u \sim Normal(0, \sigma_u^2)$.
The values for the parameters considered were: $\sigma_u^2 \in \{0.1, 0.5, 1.0, 2.0\}$, $m = 5, 10, 15, 20, 25$ observations by group, $n = 5, 10, 15$ groups and $\beta = (-1.5, 0.5)$ fixed.
For each combination of the parameters was generated the matrix $\mathbf{X}_i$, the matrix $\mathbf{Z}_i$ and the mean of $\mathbf{y}_i$ through the equation (1), then was estimated the vector of fixed effects $\hat{\beta}$ through the HGLM and the GAMLSS. To evaluate the performance of the two models was calculate the multivariate MSE de $\hat{\beta}$ given by equation (2). This procedure was repeated 10000 times.
In Figure 2, the solid lines correspond to the mean of MSE of the model GAMLSS and the dashed ones to the mean of MSE of the model HGLM

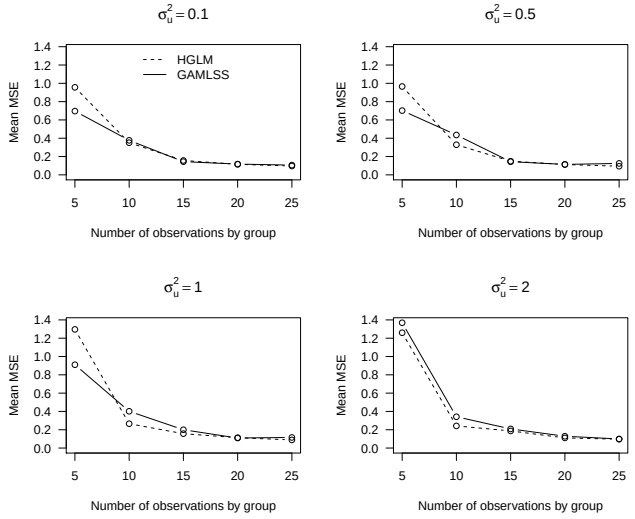


FIGURE 2. Mean MSE for the estimators obtained for the Poisson-normal model with 5 groups

with $n = 5$. Were also performed simulations with number of groups of $n = 10, 15$. As seen on the graph, the mean MSE decreases as the number of observations by group increase. Also, the model HGLM has mean MSE larger than GAMLSS when the number of observations by group is small.

4 Conclusions

In the Poisson-gamma and the Poisson-normal model was obtained similar results on the mean MSE for the models HGLM and GAMLSS. The mean MSE decreases as the number of observations by group increases.

References

Lee, Y. and Nelder, J. (1996). Hierarchical Generalised Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**,4, 619-678.

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

Ronnegard, L., Shen, X. and Alam, M. (2011). The hglm package, R Foundation for Statistical Computing . ISBN 3-900051-07-0.

A latent-class semi-parametric change point model for cognitive ability in older age

Ardo van den Hout, Graciela Muniz, Fiona E Matthews

[1] MRC Biostatistics Unit, Institute of Public Health, Robinson Way, CB2 0SR Cambridge, UK. E-mail: ardo.vandenhout@mrc-bsu.cam.ac.uk

Abstract: A random-effects change point model is formulated to describe cognitive decline in the older population in the years before death. Cognitive ability is measured using the sum score of the Mini-Mental State Examination with integer range 0-30. For the conditional distribution of the response variable the binomial and the beta-binomial distributions are used. To acknowledge the possibility that not everyone in the population experiences a change in cognition, two latent class are distinguished, namely one with change and one without. Estimation is by marginal maximum likelihood where a parametric population distribution for the random change point is combined with a non-parametric mixing distribution for other random effects. The approach is illustrated using data from a longitudinal study of ageing in Sweden.

Keywords: beta-binomial distribution, cognitive decline, mini-mental state examination, non-parametric maximum likelihood

1 Introduction

The scale of a cognitive test is often discrete. A typical example is the Mini-Mental State Examination (MMSE, Folstein et al. 1975) which has an integer scoring. The MMSE is a questionnaire for screening cognitive impairment and has items on, for instance, language, orientation, and memory. Scores for each of the questions are added up to obtain a final sum score ranging from 0 to 30.

We are interested in cognitive decline in the years before death. In case there is a change in the rate of decline, we estimate how many years before death this change takes place. Longitudinal MMSE data are available from the Swedish OCTO-twin study (McClearn et al. 1997). All the death times of the 656 individuals in our data are available and this allows years-to-death as the time scale in our model.

To acknowledge the discrete nature of MMSE data, we propose a change point model using discrete probability distributions for the response. Dependencies between the repeated measurements of an individual are dealt with by using random effects. A common assumption for the distribution of random effects is that these effects are multivariate-normally distributed.

Our model relaxes this assumption partly by using a non-parametric distribution for some of the random effects. To take into account that not everyone in the population experiences a change in cognition, two latent classes (one with change and one without) are distinguished in the model.

2 Methods

2.1 Model

Given response variable Y, predictor η, link function h, and time t as explanatory variable, the location is given by $E[Y|t] = h(\eta)$ and $\eta = f(t, \beta, \tau)$, where $\beta = (\beta_0, \beta_1, \beta_2)$ is the vector with the regression coefficients, and τ is the change point. The broken-stick change point model is defined by

$$\eta = f(t, \beta, \tau) = \begin{cases} \beta_0 + \beta_1 t & t < \tau \\ \beta_0 + \beta_1 \tau + \beta_2 (t - \tau) & t \geq \tau, \end{cases} \quad (1)$$

where the change is sudden and the function has no derivative at τ. The model is readily extended to a random-effects model by assuming that regression coefficients and the change point are individual-specific and follow a population distribution. A parametric distribution is, for example, $(\beta_{i0}, \beta_{i1}, \beta_{i2}, \tau_i) \sim MVN(\mu, \Sigma)$ for individual i, where MVN denotes the multivariate normal distribution with unknown mean μ and covariance matrix Σ.

For the conditional distribution of the response, we discuss two discrete distributions. The first is the well-known binomial distribution with the logit link $\pi = h(\eta) = \exp(\eta)/(1 + \exp(\eta))$. This distribution is denoted by $Y|\ t \sim B(\pi, n)$, where n is the maximum score ($n = 30$ for the MMSE).

The second is the beta-binomial distribution which is a mixture of two distributions. Assume, firstly, that $Y|\ t \sim B(\pi, n)$, and, secondly, that π has a beta distribution with parameters $\nu_1, \nu_2 > 0$. Then the marginal probability distribution function for Y is given by

$$P(Y = y|n, \nu_1, \nu_2) = \binom{n}{y} \frac{B(\nu_1 + y, n + \nu_2 - y)}{B(\nu_1, \nu_2)},$$

where $B(\nu_1, \nu_2)$ is the beta function. Given definitions $\mu = \nu_1/(\nu_1 + \nu_2)$ and $\phi = 1/(\nu_1 + \nu_2)$, the beta-binomial has $E[Y|t] = n\mu$ and $\mathrm{Var}[Y|t] = n\mu(1 - \mu)[1 + (n - 1)\phi/(1 + \phi)]$.

2.2 Semi-parametric maximum likelihood

Longitudinal data are given by $y = (y_1, ..., y_N)$, where N is the number of individuals in the sample. For each individual i, we have $y_i = (y_{i1}, ..., y_{in_i})$, where n_i is the number of observations for individual i. We assume conditional independence in the sense that $p(y|\beta, \tau) = \prod_{i=1}^{N} p(y_i|\beta_i, \tau_i)$, where $\tau = (\tau_1, ..., \tau_N)$, $\beta = (\beta_1, ..., \beta_N)$, and $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})$.

For a random-effects model with regard to parameter β, a marginal likelihood can be formulated by integrating out β_i using a parametric multivariate population distribution. As an alternative, we use non-parametric maximum likelihood (NPML) estimation where the distribution for β_i is a discrete distribution on a finite number K of mass points z_k, with masses π_k. The number of components K, the mass points and the masses are unknown and are estimated by maximum likelihood (Aitkin 1999). The likelihood conditional on τ is given by

$$p(y|\tau, \pi, z, K) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k p(y_i|\tau_i, z_k).$$

For the change point, we assume a parametric random-effects structure to allow for heterogeneity across individuals. The likelihood is thus given by

$$p_1(y|\pi, z, K, \tau_0, \sigma) = \prod_{i=1}^{N} \int \sum_{k=1}^{K} \pi_k p(y_i|\tau_i, z_k) p(\tau_i|\tau_0, \sigma) d\tau_i, \tag{2}$$

where τ_0 and σ are parameters for the distribution of the random change point τ_i. The distribution of τ_i is a truncated normal distribution (trN) with upper bound U equal to zero (death time) and lower bound L equal to a fixed number of years before death. That is, $\tau_i \sim trN(\text{mean} = \tau_0, \text{sd} = \sigma, \text{lower} = L, \text{upper} = U)$.
For the mixture model with a class in which there is change over time, and a class in which there is no change over time, assume that the parameter vectors are given by Θ_1 and Θ_2 respectively. Then the likelihood of the mixture model is given by

$$L(\theta, \Theta_1, \Theta_2) = \prod_{i=1}^{N} \theta p_1(y_i|\Theta_1) + (1-\theta) p_2(y_i|\Theta_2), \tag{3}$$

where $0 < \theta < 1$ is the mixture proportion such that θ is the probability to be in the class with the change, and $p_1(y_i|\Theta_1)$ is given by (2). For the stable class, and corresponding $p_2(y_i|\Theta_2)$, we assume an intercept-only logistic regression model, where the intercept is a random effect with K^* components for the NPML.

3 Analysis

For one-class NPML models with $K = 4$ and linear predictors, good results are obtained for the model with the beta-binomial distribution for the response. This is a GAMLSS model (Rigby and Stasinopoulos 2005) and can be fitted using the R package gamlss. With random intercept and slope for t, and a fixed-effect for t^2, we get -2Loglik = 10456 and AIC = 10482. For comparison, we also fitted this model with the normal distribution for the MMSE *ceteris paribus*. Although the conditional distribution of the

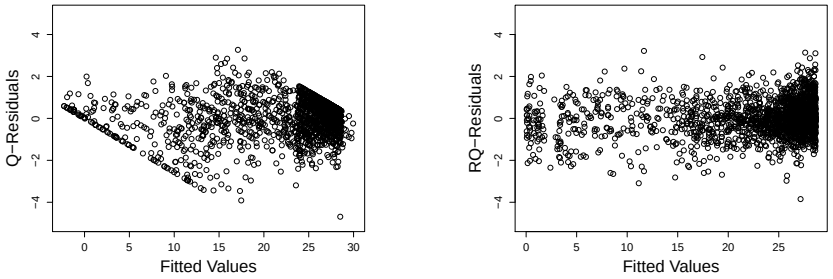


FIGURE 1. Quantile residuals for GAMLSS. Left panel for model with normal distribution, right panel for beta-binomial model.

MMSE is clearly not normal, this distribution is often chosen in linear mixed models for longitudinal MMSE data, see, e.g., Laukka et al. (2006). Quantile residuals (Dunn and Smythe 1996) for both models are depicted in Figure 1. Randomisation is applied for the residuals in case the fitted distribution is discrete (*ibid.*). Clearly, using the normal distribution for the MMSE is not wise: there is a strong dependence between fitted values and residuals. The AIC for the normal model is 12244.

Yet another alternative is using the binomial distribution in the GAMLSS model. This choice yields an AIC of 11851. From this we infer that the beta-binomial is the best choice to describe the change in the MMSE and we will use this distribution in what follows.

The results from fitting models in gamlss are used to determine starting values for the maximisation of the likelihood for the change point models. This maximisation is undertaken by using the multi-purpose optimiser optim in R.

In the latent-class model (3), the non-linear broken-stick predictor is used for the change class, and a linear predictor is used for the stable class. The distribution for the response is the beta-binomial for the change class, and the binomial for the stable class. We use $K = 4$ NPML components for β in (1), and $K^* = 2$ components for the intercept-only model for the stable class. The truncation of the normal distribution for the random change point is at -12 years and at zero. This model has -2Loglik = 10180, and AIC = 10224.

The mean and the variance of the truncated normal for the change point (and standard errors) are estimated as -5.07 (0.62) and 2.40 (0.24), respectively. This means that if there is a change in the cognitive decline, then in expectation this change will take place 5 years before death. The probability to be in the change class is estimated to be 0.63 (0.03). The left panel in Figure 2 depicts the marginal means for the four components in the change class. Parameters estimates can be found in Table 1.

Model validation was undertaken by looking at quantile residuals. First, class membership was estimated, and individual change points were esti-

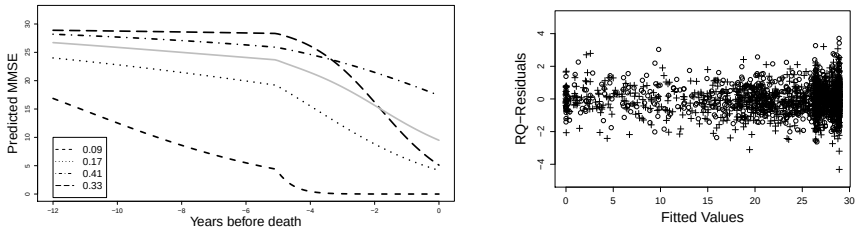


FIGURE 2. Marginal means for the $K = 4$ components in the change class (left panel, with masses in the legend and grey line for the overall mean trend. Quantile residuals for the broken-stick model (right panel, with + for change class).

mated for those allocated to the change class. Secondly, given this information, quantile residuals were assessed, see right panel of Figure 2.

TABLE 1. Parameters for the latent-class broken-stick model with $K = 4$ and $K^* = 2$. Standard errors in parentheses.

Latent-class mixture proportion				θ	0.63	(0.03)		
Change point model for change class								
	Mass points							
β_0	-3.24	(0.56)	-0.02	(0.58)	1.18	(0.54)	2.52	(0.45)
β_1	-0.29	(0.05)	-0.12	(0.05)	-0.13	(0.06)	-0.06	(0.06)
β_2	-2.26	(1.39)	-0.47	(0.09)	-0.30	(0.05)	-0.87	(0.14)
	Masses							
	0.09	(0.03)	0.17	(0.08)	0.41	(0.11)	0.33	(0.14)
μ	-5.07	(0.62)		σ	2.40	(0.24)		
ϕ	0.05	(0.01)						
Linear model for stable class								
	Mass points							
α	2.00	(0.09)	3.28	(0.08)				
	Masses							
	0.38	(0.06)	0.62	(0.06)				

4 Conclusion

A change point model with a discrete distribution for the response describes the OCTO data better than models with linear predictors. A latent class framework further improves statistical inference as it explicitly takes into

account that not all individuals in the data experience a change in cognition in the years before death.
In the analysis we have used the AIC to select between models. In general, model selection should be undertaken with care. Large sample properties of the likelihood ratio test statistic are violated in mixture models, see, e.g., Aitkin et al. (2009). In addition, a model with a linear predictor is not nested within a model with a (non-linear) change point predictor. Given the large differences in the reported AICs and the model validation in the current analysis, we are confident that the change point model in our data analysis outperforms the other models.
Currently we are looking into extensions of the model using smooth change point models and/or an increased number of NPML components.

Acknowledgments: The authors would like to thank Professor Boo Johanson, Department of Psychology, Göteborg University, Sweden, for the OCTO data.

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models, *Biometrics*, **55**, 117–128.

Aitkin, M., Darnell, R.E., Francis, B.J., and Hinde, J.P. (2009). *Statistical Modelling in R*, Oxford: Clarendon Press.

Dunn, P.K., and Smyth, G.K. (1996). Randomised quantile residuals, *Journal of Computational and Graphical Statistics*, **5**, 236–244.

Folstein, M.F., Folstein, S.E., and McHugh P.R. (1975). Mini-Mental State: A practical method for grading the state of patients for the clinician, *J. Psychiatr. Res.*, **12**, 189–198.

Laukka, E.J., MacDonald, S.W.S., and Bäckman, L. (2006). Contrasting cognitive trajectories of impending death and preclinical dementia in the very old, *Neurology*, **66**, 833–838.

McClearn, G.E., Johansson, B., Berg, S., Pedersen, N.L., Ahern, F., Petrill, S.A., and Plomin, R. (1997). Substantial genetic influence on cognitive abilities in twins 80 or more years old, *Science*, **276**, 1560–1563.

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

Measuring the Brier score for frailty models

R. Van Oirbeek [1], E. Lesaffre [1,2]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Katholieke Universiteit Leuven, Kapucijnenvoer 35, Blok D, bus 7001, B3000 Leuven, Belgium, and Universiteit Hasselt, Belgium; *robin.vanoirbeek@med.kuleuven.be*

[2] Department of Biostatistics, Erasmus Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands

Abstract: We apply the estimation technique of Graf *et al.* (1999) and Gerds and Schumacher (2006) in estimating the Brier score for frailty models. We exploid the conditional and marginal model formulations of a frailty model, resulting in two different definitions of the Brier score: the 'conditional Brier score' which measures the joint predictive contribution of the covariates and clustering effects, and the 'marginal Brier score' which measures the predictive ability of the covariate effects only. These two measures are computed using a Bayesian approach and both measures are applied to a dental data set.

Keywords: Brier score; frailty models; clustered data, survival.

1 Introduction

The Brier score of the survival curve is a well-developed measure of the predictive ability of a survival model. An estimator has been formulated that does not assume a certain class of survival models (Graf *et al.*, 1999) and that is shown to be consistent under independent censoring (Gerds and Schumacher, 2006). An adaptation to clustered data however, as developed for the concordance probability (Van Oirbeek and Lesaffre, 2010), is currently lacking. Therefore, we will adapt the Brier score to the frailty model and apply it to our motivating data set.

2 Motivating Data Set

Factors that influence amalgam restoration longevity were investigated in a clinical study (Kreulen *et al.*, 1998). 183 patients were recruited during the period 1977-1978 implying in total 1429 amalgam restorations. Patients were followed up for maximally 16 years and 189 amalgam restorations were replaced leading to a censoring percentage of 86.8 %. The clustered data structure is unbalanced, with 4, 8 and 12 restorations seen in 41, 97 and 35 patients, respectively. The primary covariates were 4 cavity wall treatments and the alloy of the amalgam. Of interest is the predictive ability of these covariates and how the predictive ability of the model changes when clustering effects are considered on top of covariate effects.

3 Brier score for frailty models

In the presence of clustering, frailty models can be used as a modeling tool. A frailty model accounts for clustering by introducing a frailty term w_q for each cluster q. The frailty term w_q is the realization of a positive random variable sampled from the frailty distribution $f(w|\boldsymbol{\zeta})$.
Two model formulations can be proposed for a frailty model: a conditional frailty model that explicitly contains the frailty terms $\boldsymbol{w}$ in its formulation and a marginal frailty model that is constructed by integrating out $\boldsymbol{w}$ from the conditional frailty model (Duchateau and Janssen, 2008). For both model formulations, two different types of survival curves can be constructed, i.e. the conditional survival curve $S_C(t|\boldsymbol{X}, \boldsymbol{w})$ and the marginal survival curve $S_M(t|\boldsymbol{X})$ with covariates $\boldsymbol{X}$. The Brier score $BS(t)$ consists of comparing the event status of a given subject to its predicted survival probability at a given time point t. Two different versions of $BS(t)$ can be defined for a frailty survival model, i.e. the conditional Brier score $BS_C(t)$:

$$BS_C(t) = E_{X,T,W}\{I(T > t) - S_C(t|\boldsymbol{X}, \boldsymbol{w})\}^2. \tag{1}$$

which compares for each subject the event status with the conditional survival curve and the marginal Brier score $BS_M(t)$:

$$BS_M(t) = E_{X,T,W}\{I(T > t) - S_M(t|\boldsymbol{X})\}^2. \tag{2}$$

which compares the event status with the marginal survival curve. Therefore, $BS_C(t)$ evaluates the predictive effect of the frailty model by explicitly correcting for both covariate and clustering effects, while $BS_M(t)$ evaluates the predictive effect of the covariate effects only. As such, the comparison of $BS_M(t)$ with $BS_C(t)$ expresses the added (predictive) value of considering clustering effects on top of covariate effects. In case that the model is correctly specified, it can be shown that $BS_M(t) \geq BS_C(t)$ for each time point t. By comparing $BS_M(t)$ and $BS_C(t)$ with the Brier score of a non-informative model $BS_0(t)$ such as the Kaplan-Meier model, a measure of explained variance can be constructed (Graf *et al.*, 1999).

4 Estimating $BS_C(t)$ and $BS_M(t)$ for frailty models

For univariate data, an uniformly consistent estimator has been provided for the Brier score (Graf *et al.*, 1999), even when censoring and failure times are only conditionally independent given the covariates (Gerds and Schumacher, 2006). Consider t_i as the observed failure time of subjects i such that $t_i = \min(T_i, T_{c,i})$ corresponds to the observed failure time with $T_{c,i}$ the right censoring time and T_i the true failure time. If $t_i = T_i$ then t_i is a true failure time and the censoring indicator δ_i equals to 1. If $t_i = T_{c,i}$ then t_i is a censoring time and the censoring indicator δ_i equals to 0. For a sample $i = 1, \ldots, n$, Gerds and Schumacher propose to estimate the Brier score as:

$$\widehat{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\{I(t_i > t) - \hat{S}(t|\boldsymbol{X})\}^2\, \omega(t, \hat{G}(t|\boldsymbol{X}_i)) \tag{3}$$

with $\hat{G}(t|\boldsymbol{X}_i)$ as the estimator of the conditional survival function of the censoring times $G(T_c > t|\boldsymbol{X}_i)$. The estimated weights ω correspond to:

$$\omega(t, \hat{G}(t|\boldsymbol{X}_i)) = \frac{I\{t_i \leq t\}\, \delta_i}{\hat{G}(t_i - |\boldsymbol{X}_i)} + \frac{I\{t_i > t\}}{\hat{G}(t|\boldsymbol{X}_i)}. \tag{4}$$

with t_i- is the time point just before t_i. $BS_C(t)$ and $BS_M(t)$ are estimated by plugging in the conditional $\hat{S}_C(t|\boldsymbol{X}, \boldsymbol{w})$ and marginal survival curves $\hat{S}_M(t|\boldsymbol{X})$ respectively in (3). Note that $\widehat{BS}(t)$ is a consistent estimator of the Brier score only when $\hat{G}(t|\boldsymbol{X})$ is a consistent estimator of $G(t|\boldsymbol{X})$.
We estimate $BS_C(t)$ and $BS_M(t)$ with a Bayesian approach and point estimates are obtained by taking the posterior median of the posterior distribution of each separate Brier score. In our simulation study, this approach resulted in well-performing point estimates. Since the point estimate of the Kaplan-Meier model is consistent in the presence of clustering, this model is used to calculate a point estimate of $BS_0(t)$.

5 Application

A Bayesian proportional hazards (PH) gamma frailty model with a gamma independent increments baseline hazard function was fitted in WinBUGS to the motivating data set. All the primary covariates were included in the model and the PH assumption was found to be acceptable for each. Convergence of the Markov chain was attained quickly, i.e. within the first 5000 iterations, such that a posterior sample of size 15000 was generated, discarding the first 5000 samples as burn-in. The $BS_0(t)$, $BS_C(t)$ and $BS_M(t)$ estimates are shown in Figure 1.
We see that for each time point $\widehat{BS}_0(t) \geq \widehat{BS}_M(t)$ implying a satisfactory performance of the marginal model, since the inclusion of the primary covariates reduces the prediction error at all times. The behaviour of $\widehat{BS}_C(t)$ is more peculiar, since at some time points, it performs worse than the null and the marginal model. This indicates that the frailty distribution might be misspecified since $BS_M(t) \geq BS_C(t)$ when the model is correctly specified.

6 Conclusion

The conditional and marginal Brier score allow to investigate the predictive ability of covariates and clustering effects separately. Moreover, the effect of certain model assumptions on the predictive ability of the frailty model can be quantified.

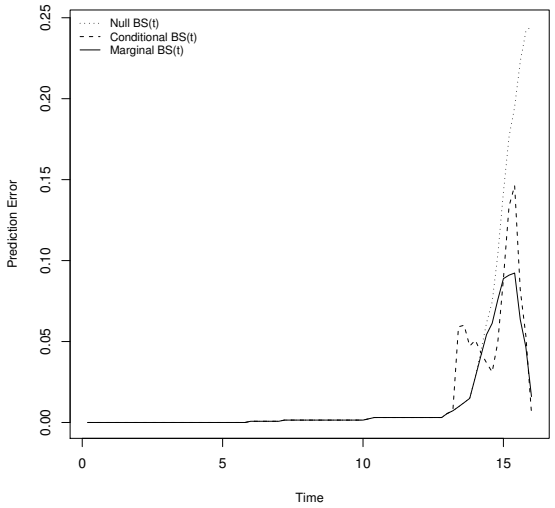


FIGURE 1. $BS_0(t)$, $BS_M(t)$ and $BS_C(t)$ estimates for the amalgam data set.

References

Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, **18**, 2529-2545.

Gerds, T.A., Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, **48**(6), 1029-1040.

Van Oirbeek, R., Lesaffre, E. (2010). An application of Harrell's C-index to PH frailty models. *Statistics in Medicine*, **29**(30), 3160-3171.

Duchateau, L., Janssen, P. (2008). *The Frailty Model*. Springer Science + Business Media: New York.

Kreulen, C.M., Tobi, H., Gruythuysen, R.J.M., van Amerongen, W.E., Borgmeijer, P.J. (1998) Replacement risk of amalgam treatment modalities: 15-year results. *Journal of Dentistry*, **26**, 627-632.

A Dipole Model for MEG Data

M. Ventrucci[1], A. W. Bowman [1], C. Miller (nee Ferguson)[1], J. Gross [2], K. Ghosh [1]

[1] School of Mathematics and Statistics, University of Glasgow, G12 8QW
[2] Institute of Neuroscience and Psychology, University of Glasgow, G12 8QW

Abstract: Nonlinear models have been fitted to MEG data in order to improve understanding of brain activity prior to exposure to a stimulus. Such models can be used to characterise brain activity using only a few parameters in order to study within and between subject variability. This knowledge will help to inform estimation of temporal and spatial locations of brain activation in response to a particular stimulus.

Keywords: MEG; dipole; smoothing

1 Introduction

Magnetoencephalography (MEG) measures the electromagnetic activity in the human brain by recording the magnetic fields outside the head. Data are acquired by sensitive devices (SQUIDS) embedded in a helmet placed over the head. The high temporal resolution of MEG is optimal for studying the transient magnetic fields associated with the highly dynamic processes of brain activations. Typically experiments consist of recording brain signals, under some experimental conditions or stimuli, for multiple trials on each subject. These MEG trials are then averaged to enhance signal identification and reduce noise. Ventrucci et al. (2010) highlights the benefits of using smoothing techniques and related statistical inference to estimate a smooth signal in single trials.
In general, exposure to a stimulus will result in a brain activation which can be characterised by a dipole (an area of the brain where MEG signals from the sensors will display high amplitude in two adjacent brain locations that are out of phase, see Figure 1). This dipole pattern can also have a temporal dimension (an oscillation with a given frequency). It is of particular interest to identify the spatial and temporal locations of a dipole effect associated to the stimulus. A dampened version of this dipole pattern is also evident in brain activity prior to a stimulus. This is a result of a 10Hz α band frequency that is believed to be continuously present (see, for example, Van Dijk et al. 2008; Schnitzler and Gross 2005). It is of interest in this paper to investigate statistical models for brain activity in the pre-stimulus period to estimate the location of a dipole and fit a dipole model. This will enable characterisation of the signal in terms of only a few parameters,

which can be used to study variability across trials and also estimate the temporal and spatial location of the post-stimulus dipole.

1.1 The Data

MEG data have been collected from replicates of an experiment conducted on 19 subjects to study event-related neural response. Each subject undertook 135 trials of the experiment and the MEG signal was recorded on $S = 248$ sensors and $T = 256$ time points which span a time window of one second with onset stimulus occurring at 500 msecs. For this paper, the analysis will focus on only one subject from this dataset in the pre-stimulus period in order to study brain activity prior to a stimulus.

2 Methods

2.1 A Fixed Location Dipole Model

In order to obtain information about the brain location at which a dipole occurs, Model 1 is fitted to each sensor ($s = 1, \ldots, S$), producing estimates of the amplitude and phase of the signal measured over time ($t = 1, \ldots, T$).

$$y_t = \alpha \cos(2\pi(t - \beta)/f^{-1}) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{1}$$

where α is the amplitude, t is the time point in the pre-stimulus period, f is fixed equal to 10 (the period f^{-1} is 0.1 for a 10Hz signal). The estimates of the parameters from this model can be used to identify the sensors with the largest amplitude that are out of phase, denoted s_1 and s_2. These two sensors indicate the brain location of the two opposite poles of the dipole. This knowledge can then be used to fit the fixed location dipole Model 2, to all of the sensors measurements in the pre-stimulus period.

$$y_{ts} = (\alpha_1 w_1 - \alpha_1 e^{\alpha_2} w_2)\cos(2\pi(t - \beta)/f^{-1}) + \epsilon \quad \epsilon \sim N(0, \sigma^2) \tag{2}$$

where for $i = \{1, 2\}$, $w_i = \exp(-0.5 d^2_{s,s_i}/h^2)$, d_{s,s_i} is the geodesic distance (see Ventrucci et al. 2010 for further details) between sensor s_i and sensor s, and h determines the weights assigned to the neighbouring sensor measurements, controlling the spatial extent of the fitted dipole.

2.2 A Varying Location Dipole Model

In order to estimate the location and orientation of a dipole in any trial, Model 2 has been extended to Model 3. This model removes the need for the preliminary search performed by Model 1.

$$y_{ts} = (\alpha w_1 - \alpha \alpha_0 w_2)\cos(2\pi(t - \beta)/f^{-1}) + \epsilon \quad \epsilon \sim N(0, \sigma^2) \tag{3}$$

where:

$$w_1 = \exp\left(-0.5\sqrt{(x_s - (x_d + \delta\cos\theta))^2 + (y_s - (y_d + \delta\sin\theta))^2}/h^2\right)$$

$$w_2 = \exp\left(-0.5\sqrt{(x_s - (x_d - \delta\cos\theta))^2 + (y_s - (y_d - \delta\sin\theta))^2}/h^2\right)$$

where (x_s, y_s) are the coordinates of sensor s. As in Model 2, w_1 and w_2 are weights which define a smooth surface representing the spatial pattern of the dipole. This surface is then assumed to have a temporal oscillating pattern governed by the cosine component. Model 3 enables estimation of the following parameters related to the dipole: location coordinates (x_d, y_d), orientation θ, oscillating frequency f, phase β, amplitudes α and α_0, spatial extent h, and the distance δ between the two poles. A standard optimization algorithm available in R (`optim()`) is used to fit the dipole spatiotemporal surface to the MEG data.

3 Results

Model 1 and Model 2 have been fitted to MEG data observed from a single trial of the experiment in a given subject, whereas the varying dipole location Model 3 has been fitted to each trial.

The brain map in Figure 1 (left panel) displays the amplitude values estimated by Model (1) over space. The circles identify the two marked sensors s_1 (white) and s_2 (black) which respectively have the highest positive and lowest negative amplitude at $t = 0$, and are out of phase. The raw data time series at s_1 and s_2 are also illustrated (top right panel), together with the signal fitted by Model 1 (bottom right panel). The out-of-phase oscillation of s_1 and s_2 provides evidence of a back-front dipole effect, which means that these two sensors identify the brain location which should be used to fit the fixed location dipole model.

Model 2 has been fitted using the two marked sensors s_1 and s_2 by a standard Newton-based iterative algorithm. The results from fitting this model are displayed in Figure 2 where the raw signals for each marked sensor s_1 and s_2 (plus a few sensors placed in their neighborhood) are displayed along with the fitted values for each of these sensors. It can be seen that for the marked sensors and their surrounding area the dipole model is quite a good representation of the underlying signal.

In order to explore variability of the dipole location and orientation across trials, Model 3 has been fitted to each trial. Figure 3 (left panel) illustrates the spatial dipole pattern estimated by Model 3 using the same data analyzed in Figures 1 and 2. As expected, the dipole location estimate (white circle) is approximately midway between the marked sensors s_1 and s_2 of

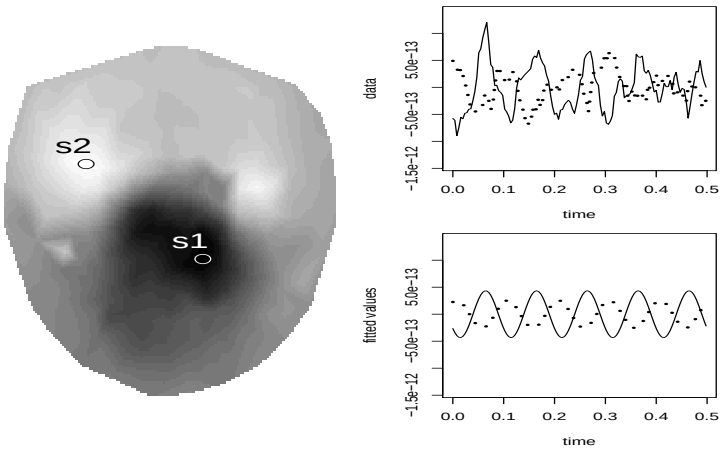


FIGURE 1. Results from Model 1 (single sensor sinusoid model). On the left the amplitude map highlighting s_1 and s_2 as the possible poles of a dipole effect. On the right the raw data (top) and fitted values (bottom) at sensors s_1 (dotted lines) and s_2 (solid lines) over the pre-stimulus time window. The spatiotemporal pattern of the fitted values taken at s_1 and s_2 is an example of a spatiotemporal dipole.

Figure 1, and the orientation estimate is also consistent with the back-front dipole. The white line at the dipole location depicts the orientation of the electrical current responsible for the dipole effect. More interestingly, the varying dipole location model has been fitted to all of the trials. Estimated locations and orientations for all trials are displayed in the central and right brain maps, the latter displaying the back-front oriented dipoles and the former the right-left. Results suggest that in most of the trials a pre-stimulus dipole occurs in the central lobe of the brain regardless of its orientation.

4 Current and Future Work

Current work includes extending Model 3 to model dipole effects occurring in the post-stimulus period. The dipole model will be developed to enable the frequency and amplitude of the signal to change over time and to allow the location and orientation of the dipole to vary across trials. It is of interest to develop a modeling framework able to characterise the signal using only a few parameters. The variability of these parameters can then be studied across trials and subjects to help inform the development of a mixed effects spatio-temporal model for the MEG data.

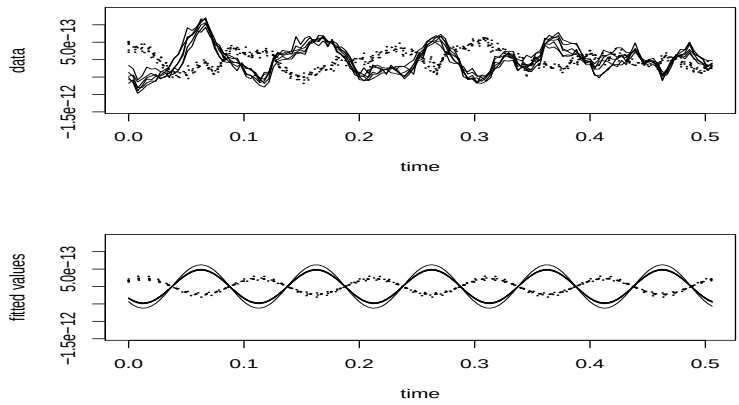


FIGURE 2. Results from Model 2 (fixed location dipole model). The raw data (top) and the fitted values (bottom) for sensors surrounding s_1 and s_2.

References

Van Dijk, H., Schoffelen, J.M., Oostenveld, R., Jensen, O. (2008). Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. *The Journal of Neuroscience* **28(8)**, 1816-1823.

Ventrucci, M., Ferguson, C., Gross, J., Schoffelen, J.M., Bowman, A. (2010). Spatiotemporal smoothing of single trial MEG data. Submitted to the *Journal of Neuroscience methods*.

Schnitzler, A., Gross, J. (2005). Normal and pathological oscillatory communication in the brain. *Nature reviews - Neuroscience* **6**, 285-296.

FIGURE 3. Results from Model 3 (varying location dipole model). On the left panel the map of the fitted values, with the white circle and white line indicating the location and orientation of the dipole. The central and right panels display a brain map of dipole orientations (black lines) for each trial (note that the estimated dipole location, here not marked, is placed in the middle of the black line). Central panel: trials with a right-left dipole. Right panel: trials with a back-front dipole.

A Bayesian adjustment of the modified profile likelihood

Ventura Laura[1], Racugno Walter[2]

[1] Department of Statistics, University of Padova, Italy, ventura@stat.unipd.it
[2] Department of Mathematics, University of Cagliari, Italy, racugno@unica.it

Abstract: We propose an adjustment of the modified profile likelihood based on a suitable matching prior on the parameter of interest only, i.e. a prior for which there is an agreement between frequentist and Bayesian inference. We show that the proposed modified profile likelihood has several desiderable inferential properties. Two examples are illustrated.

Keywords: Higher-order asymptotics; Matching prior; Nuisance parameter; Signed likelihood ratio statistic.

1 Introduction

Let us consider a model with likelihood function $L(\psi, \lambda) = L(\psi, \lambda; y)$, where ψ is a scalar parameter of interest, λ a d-dimensional nuisance parameter and $y = (y_1, \ldots, y_n)$ a random sample of size n. Standard first-order methods for inference about ψ are based on the profile likelihood $L_p(\psi) = L(\psi, \hat{\lambda}_\psi)$, with $\hat{\lambda}_\psi$ maximum likelihood estimator (MLE) of λ for fixed ψ, and can be seriously inaccurate, in particular when the dimension of λ is substantial relative to n. Starting from Barndorff-Nielsen (1983), various modifications of the form $L_{mp}(\psi) = L_p(\psi)M(\psi)$ have been proposed, for suitably defined correction terms $M(\psi)$; see Barndorff-Nielsen and Cox (1994, Chap. 8) and Severini (2000, Chap. 9) for detailed accounts. All the modifications are equivalent to second order and share the common feature of reducing the score bias to $O(n^{-1})$. However, the signed likelihood ratio statistic based on $L_{mp}(\psi)$ is standard normal only to first order, and can be inaccurate in models with many nuisance parameters (Sartori *et al.*, 1999).
In this paper we discuss a modification of $L_{mp}(\psi)$ from a new perspective based on recent advances on unified Bayesian and frequentist methods (see e.g. Ventura *et al.*, 2009, Ventura and Racugno, 2011). More precisely, as a convenient device to modify $L_{mp}(\psi)$ we use a suitable default prior on ψ only, which can be interpreted, from the frequentist point of view, as a non-negative weight function on ψ. The possibility of adjusting a likelihood function using priors, even if quite differently motivated, is suggested also in Efron (1993) and Liseo (1993).

Here, we focus on the class of strong matching priors for ψ derived from $L_{mp}(\psi)$ (Ventura *et al.*, 2009), i.e. priors for which there is an agreement between frequentist and Bayesian results and which validate the use of $L_{mp}(\psi)$ for Bayesian inference in the presence of nuisance parameters. We then investigate theoretically and numerically the modification of $L_{mp}(\psi)$ through the matching prior

$$\pi(\psi) \propto i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)^{1/2} \ , \tag{1}$$

where $i_{\psi\psi.\lambda}(\psi,\lambda) = i_{\psi\psi}(\psi,\lambda) - i_{\psi\lambda}(\psi,\lambda) i_{\lambda\lambda}(\psi,\lambda)^{-1} i_{\lambda\psi}(\psi,\lambda)$ is the partial information, with $i_{\psi\psi}(\cdot)$, $i_{\psi\lambda}(\cdot)$, $i_{\lambda\lambda}(\cdot)$, and $i_{\lambda\psi}(\cdot)$ blocks of the expected Fisher information $i(\psi,\lambda)$. The implied modified profile likelihood is thus defined as

$$L^*_{mp}(\psi) = L_{mp}(\psi)\ i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)^{1/2} \ . \tag{2}$$

In Section 3 we will show that $L^*_{mp}(\psi)$ has better inferential properties than $L_{mp}(\psi)$: the signed likelihood ratio statistic based on (2) is standard normal to second order, and the maximizer of (2) is a refinement of the MLE of ψ. Finally, two examples are illustrated in Section 4.

2 Background theory

Let us consider the modified profile likelihood of Barndorff-Nielsen (1983), given by $L_{mp}(\psi) = L_p(\psi) C(\psi)$, with $C(\psi) = (|j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)||j_{\lambda\lambda}(\hat{\psi},\hat{\lambda})|)^{1/2} / |\ell_{\lambda;\hat{\lambda}}(\psi,\hat{\lambda}_\psi)|$, $j_{\lambda\lambda}(\cdot)$ block $(\lambda\lambda)$ of the observed information $j(\psi,\lambda)$, $\ell_{\lambda;\hat{\lambda}}(\psi,\lambda) = \partial\ell(\psi,\lambda)/\partial\lambda\partial\hat{\lambda}^T$, $\ell(\psi,\lambda) = \log L(\psi,\lambda)$, and $(\hat{\psi},\hat{\lambda})$ MLE of (ψ,λ). Since $L_{mp}(\psi)$ depends only on y and ψ, it can be used also in the Bayesian framework as a genuine likelihood (Chang and Mukerjee 2006, Ventura *et al.*, 2009, Racugno *et al.*, 2010) to obtain the posterior distribution $\pi_{mp}(\psi|y) \propto i_{\psi\psi.\lambda}(\psi,\hat{\lambda}_\psi)^{1/2}\, L_{mp}(\psi)$.
Following standard Bayesian expansions, for $\pi_{mp}(\psi|y)$ a tail area approximation can be derived (Ventura and Racugno, 2011), of the form

$$\int_{-\infty}^{\psi_0} \pi_{mp}(\psi|y)\, d\psi \quad \doteq \quad \Phi(r^*_p) \ , \tag{3}$$

where $\Phi(\cdot)$ is the standard normal distribution and $r^*_p(\psi)$ is the modified signed likelihood ratio statistic $r^*_p(\psi) = r_p(\psi) + r_p(\psi)^{-1} \log(q(\psi)/r_p(\psi))$, with $r_p(\psi) = \text{sign}(\hat{\psi}-\psi)[2(\ell_p(\hat{\psi}) - \ell_p(\psi))]^{1/2}$, $\ell_p(\psi) = \log L_p(\psi)$,

$$q(\psi) = \frac{\ell'_p(\psi)}{|j_p(\hat{\psi})|^{1/2}} \frac{|i_{\psi\psi.\lambda}(\hat{\psi},\hat{\lambda})|^{1/2}}{|i_{\psi\psi.\lambda}(\psi,\hat{\lambda}_\psi)|^{1/2}} \frac{|\ell_{\lambda;\hat{\lambda}}(\psi,\hat{\lambda}_\psi)|}{|j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)|^{1/2} |j_{\lambda\lambda}(\hat{\psi},\hat{\lambda})|^{1/2}} \ ,$$

$j_p(\psi)$ profile observed information and $\ell'_p(\psi) = \partial\ell_p(\psi)/\partial\psi$. Since $r^*_p(\psi)$ corresponds to the expression derived in Barndorff-Nielsen and Chamberlin (1994), (1) is a strong matching prior (Fraser and Reid, 2002).

In view of (3), $H_\alpha = \{\psi : |r_p^*(\psi)| \leq z_{1-\alpha/2}\}$ is a high posterior density credible set for ψ with approximate frequentist validity $(1-\alpha)$, with z_α α-quantile of $\Phi(\cdot)$. Note that H_α is also an accurate likelihood-based confidence interval for ψ with approximate level $(1-\alpha)$ based on $r_p^*(\psi)$ (see, e.g., Severini, 2000, Chap. 7). Moreover, note that the posterior mode of $\pi_{mp}(\psi|y)$ can be computed as the solution in ψ of the estimating equation $r_p^*(\psi) = 0$, i.e. it coincides with the frequentist estimator defined as the zero-level confidence interval based on $r_p^*(\psi)$ (Giummolé and Ventura, 2002).

3 A new modified profile likelihood

The agreement in (3) suggests to modify $L_{mp}(\psi)$ as in (2) to define $L_{mp}^*(\psi)$. In this section we illustrate the properties of $L_{mp}^*(\psi)$. In particular, using results in Sartori *et al.* (1999), it can be shown that for $\ell_{mp}^*(\psi) = \log L_{mp}^*(\psi)$ we have

$$\ell_{mp}^*(\psi) = -\frac{1}{2}(r_p^*(\psi))^2 + O(n^{-1}) \ . \tag{4}$$

Indeed

$$\begin{aligned}
\ell_{mp}^*(\psi) &= \ell_{mp}(\psi) + \log \pi(\psi) \\
&= -\frac{1}{2}(r_{mp}(\psi))^2 + \log \pi(\psi) \\
&= -\frac{1}{2}(r_p(\psi))^2 - r_p(\psi)\left[\mathrm{NP} + \frac{1}{r_p(\psi)} \log \frac{1}{\pi(\psi)}\right] \\
&= -\frac{1}{2}(r_p(\psi) + \mathrm{NP} + \mathrm{INF})^2 + O(n^{-1}) \\
&= -\frac{1}{2}(r_p^*(\psi))^2 + O(n^{-1}) \ ,
\end{aligned}$$

where NP is the nuisance parameter adjustment NP$= -(1/r_p(\psi)) \log C(\psi)$ and INF is the information adjustment INF$= (1/r_p(\psi)) \log(q(\psi)C(\psi))/r_p(\psi)$ (Barndorff-Nielsen and Cox, 1994, Sect. 6.6). This shows that $\ell_{mp}^*(\psi)$ is equal, to second asymptotic order, to a r^*-type statistics, and the quantity $\pi(\psi) \propto i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)^{1/2}$ can thus be interpreted as a further adjustment to the profile likelihood.

In view of (4), for the proposed modified profile likelihood (2) we have that the associated signed likelihood ratio statistic $r_{mp}^*(\psi) = \mathrm{sgn}(\hat{\psi}_{mp}^* - \psi)[2(\ell_{mp}^*(\hat{\psi}_{mp}^*) - \ell_{mp}^*(\psi))]^{1/2}$, with $\hat{\psi}_{mp}^*$ maximizer of $\ell_{mp}^*(\psi)$, corresponds to $r_p^*(\psi)$ and thus is standard normal to second order. Moreover, $\hat{\psi}_{mp}^*$ can be computed as the solution of the estimating equation $r_p^*(\psi) = 0$, and thus is a refinement of the MLE, improving its small sample properties.

(n_x, n_y)		$\psi = 0.8$	$\psi = 0.9$	$\psi = 0.95$
(5,5)	$L^*_{mp}(\psi)$	0.952	0.949	0.949
	$L_{mp}(\psi)$	0.941	0.944	0.943
(10,10)	$L^*_{mp}(\psi)$	0.948	0.952	0.951
	$L_{mp}(\psi)$	0.944	0.946	0.947
(20,20)	$L^*_{mp}(\psi)$	0.949	0.949	0.950
	$L_{mp}(\psi)$	0.949	0.947	0.946
(30,30)	$L^*_{mp}(\psi)$	0.951	0.951	0.950
	$L_{mp}(\psi)$	0.948	0.949	0.949

TABLE 1. Coverage probabilities of 0.95% confidence intervals.

		$\psi = 0.8$		$\psi = 0.9$		$\psi = 0.95$	
(n_x, n_y)		bias	sd	bias	sd	bias	sd
(5,5)	$\hat{\psi}^*_{mp}$	0.012	(0.07)	0.010	(0.04)	0.006	(0.03)
	$\hat{\psi}_{mp}$	0.021	(0.07)	0.017	(0.04)	0.010	(0.04)
(10,10)	$\hat{\psi}^*_{mp}$	0.008	(0.07)	0.006	(0.02)	0.003	(0.02)
	$\hat{\psi}_{mp}$	0.010	(0.07)	0.008	(0.02)	0.005	(0.02)
(20,20)	$\hat{\psi}^*_{mp}$	0.004	(0.05)	0.003	(0.02)	0.001	(0.02)
	$\hat{\psi}_{mp}$	0.005	(0.05)	0.004	(0.02)	0.003	(0.02)
(30,30)	$\hat{\psi}^*_{mp}$	0.002	(0.04)	0.001	(0.02)	0.001	(0.01)
	$\hat{\psi}_{mp}$	0.003	(0.04)	0.002	(0.02)	0.001	(0.01)

TABLE 2. Bias (and standard deviations) of the MLEs of $L^*_{mp}(\psi)$ and of $L_{mp}(\psi)$.

4 Two examples

Example 1: We provide a simulation study of the proposed modified profile likelihood in the context of the exponential stress-strength model (Kotz *et al.*, 2003). In particular, we assume that X and Y are independent and exponentially distributed, with rates α and β, respectively. In this framework, the reliability parameter $\psi = P(X < Y)$ can be written as $\psi = \alpha/(\alpha + \beta)$. Let us consider the one-to-one transformation $\theta = (\psi, \lambda)$, with $\psi = \alpha/(\alpha + \beta)$ and $\lambda = \alpha + \beta$. The profile likelihood for ψ is $L_p(\psi) = \hat{\lambda}_\psi^{(n_x+n_y)} \psi^{n_x} (1 - \psi)^{n_y}$, with $\hat{\lambda}_\psi = (n_x + n_y)\hat{\lambda}\bar{x}/(n_y(\bar{x} + \bar{y})(1 - B\psi))$, $\hat{\psi} = \bar{y}/(\bar{x} + \bar{y})$, $\hat{\lambda} = (\bar{x}+\bar{y})/(\bar{x}\bar{y})$, $B = (n_y\bar{y} - n_x\bar{x})/(n_y\bar{y})$, and $\bar{x}$ and $\bar{y}$ sample means. Moreover, we have $j_{\lambda\lambda}(\psi, \lambda) = (n_x + n_y)/\lambda^2$. Simple calculations show that $L_{mp}(\psi) = L_p(\psi)\hat{\lambda}^2_\psi (n_x + n_y)^{-1/2}/\hat{\lambda}$ and that $i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)^{1/2} = 1/(\psi(1 - \psi))$. The proposed modified profile likelihood is thus

$$L^*_{mp}(\psi) = \psi^{n_x - 1}(1 - \psi)^{n_y - 1}(1 - B\psi)^{-(n_x+n_y)} \ .$$

The behaviour of $L^*_{mp}(\psi)$ is illustrated through a simulation study, based on 10000 Monte Carlo trials. Table 1 gives the empirical coverages for 95% confidence intervals from $L^*_{mp}(\psi)$ and from $L_{mp}(\psi)$. We observe that, even for small (n_x, n_y), $L^*_{mp}(\psi)$ has the correct frequentist coverages. Larger sample sizes $(n_x, n_y > 20)$ show, as one would expect, rather little differences between the results of the two procedures. Table 2 gives the bias and standard deviation of the MLEs of $L^*_{mp}(\psi)$ and $L_{mp}(\psi)$. It can be noted that $\hat{\psi}^*_{mp}$ exhibits a smaller bias than the maximum modified profile estimator. This result is due to the fact that maximizer of $L^*_{mp}(\psi)$ is a r^*_p-based estimator.

Example 2: Let us consider the scalar skew-normal model (Azzalini, 1985) with density function $p(y; \psi, \mu, \sigma) = (2/\sigma)\phi((y-\mu)/\sigma)\Phi(\psi(y-\mu)/\sigma)$, where $\phi(x)$ denote the $N(0, 1)$ density. Let ψ be the parameter of interest and

let $\lambda = (\mu, \sigma)$, with μ and σ unknown location and scale parameters, be the nuisance parameter. Estimation of the shape parameter ψ is a quite challenging problem since $L_p(\psi)$, as well as $L_{mp}(\psi)$, can be monotone increasing, giving an infinite MLE. Some recent solutions are Sartori (2006), Liseo and Loperfido (2006), and Cabras *et al.* (2010). In particluar, Cabras *et al.* (2010) give the expressions of the modified profile likelihood $L_{mp}(\psi)$ and of $\pi(\psi)$, which is shown to be proper and independent on λ.
We illustrate our proposal with a quite challenging data set for the estimation of ψ. In particular, consider the *Frontier* data set, available at the package `sn` of the `R` software, which is a random sample of size $n = 50$ from a skew-normal model, with $\mu = 0$, $\sigma = 1$ and $\psi = 5$. This dataset has some interest and has been analyzed in several papers since it leads to an infinite $\hat{\psi}$, with both $L_p(\psi)$ and $L_{mp}(\psi)$ monotone functions in ψ. Sartori (2006) obtains a modified maximum likelihood estimate equal to 6.24, while the maximum likelihood estimate from $L^*_{mp}(\psi)$ is $\hat{\psi}^* = 6.3$. Figure 1 shows the modified profile likelihoods $L_{mp}(\psi)$ for ψ, which is monotone, and $L^*_{mp}(\psi)$.

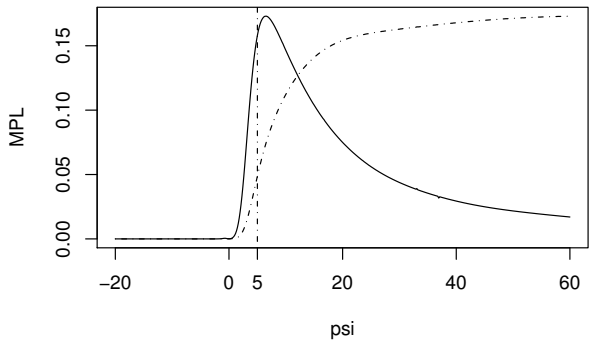


FIGURE 1. *Frontier* data: Plot of normalized $L_{mp}(\psi)$ (dashed) and $L^*_{mp}(\psi)$ (solid).

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.*, **12**, 171-178.

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Barndorff-Nielsen, O.E, Chamberlin, S.R. (1994). Stable and invariant adjusted directed likelihoods. *Biometrika*, **81**, 485–499.

Barndorff-Nielsen, O.E., Cox, D.R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.

Cabras, S., Castellanos, M.E., Racugno, W., Ventura, L. (2010). A matching prior for the shape parameter of the skew-normal distribution. Under revision.

Chang, H., Mukerjee, R. (2006). Probability matching property of adjusted likelihoods. *Statist. Probab. Lett.*, **76**, 838-842.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, **80**, 3-26.

Fraser, D.A.S., and Reid, N. (2002). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Plan. Inf.*, **103**, 263-285.

Giummolé, F., and Ventura, L. (2002). Practical point estimation from higher-order pivots. *J. Statist. Comput. Simul.*, **72**, 419-430.

Kotz, S., Lumelskii, Y., Pensky, M. (2003). *The Stress-Strength Model and Its Generalizations: Theory and Applications*. World Scientific, Singapore.

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika*, **80**, 295-304.

Liseo, B., Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution. *J. Statist. Plann. Inf.*, **136**, 373-389.

Racugno, W., Salvan, A., Ventura, L. (2010). Bayesian analysis in regression models using pseudo-likelihoods. *Comm. Stat. Th. Meth.*, **39**, 3444-3455.

Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *J. Statist. Plann. Inf.*, **136**, 4259-4275.

Sartori, N., Bellio, R., Salvan, A., and Pace, L. (1999). The directed modified profile likelihood in models with many nuisance parameters. *Biometrika*, **86**, 735-742.

Severini, T.A. (2000). *Likelihood Methods in Statistics*, Oxford: University Press.

Ventura, L., Cabras, S., and Racugno, W. (2009). Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *J. Amer. Stat. Assoc.*, **104**, 768–774.

Ventura, L., Racugno, W. (2011). Recent advances on Bayesian inference for $P(X < Y)$. *Bayesian Analysis*, to appear.

Bayesian Structured Additive Quantile Regression

Elisabeth Waldmann[1], Thomas Kneib[1]

[1] Institute of Mathematics, Carl von Ossietzky University Oldenburg, Carl-von-Ossietzky-Str. 9-11, D-26129 Oldenburg,
{elisabeth.waldmann, thomas.kneib}@uni-oldenburg.de

Abstract: Since Koenker first suggested using quantile regression in order to give a more detailled description of the conditional distribution in regression contexts (Koenker and Bassett (1978)), a lot of expansions to this concept have been made. By using the **A**symmetric **L**aplace **D**istribution (ALD) as an error distribution, quantile regression became accessible to Bayesian inference. A reformulation of the ALD using location-scale mixtures of normals transforms the problem into a Gaussian regression with offset (Yue and Rue (2011)). Based on these results, we want to explain the possibility to extend linear quantile regression by adding nonlinear and geoadditive effects to the predictor, the possibility to use the LASSO for shrinkage and selection and the inclusion of **D**irichlet **P**rocess **M**ixtures (DPM) for random effects and clustering aims. We will present the idea behind the theoretical calculations leading to the corresponding MCMC procedure and illustrate these in two different applications.

Keywords: asymmetric Laplace distribution; Bayesian quantile regression; MCMC; LASSO; Dirichlet process mixtures.

1 Quantile Regression

Quantile regression is a tool used to estimate the influence of a predictor on the quantiles of the conditional distribution of a dependent variable. One of the main advantages over mean regression is that this method permits to supply detailled information about the conditional distribution without specifying a parametric data distribution. The coefficients of the quantile regression can be estimated by minimizing sums of weighted absolute residuals, which is a completely nonparametric approach. The changes in characteristics compared to mean regression are the same as the changes from looking at the mean of a dataset to taking into account quantiles.
To make quantile regression feasible in the Bayesian context, the idea of working completely non-parametrically has to be quit. One possibility to do this is to use an auxiliary error distribution. The distribution we will use in this context is the asymmetric Laplace distribution (ALD). In order to make it suitable for MCMC issues we rewrite the ALD using a scale-location mixture. Datasets with very large covariate vectors are on hand

in many different fields. For this reason it is necessary to adapt selection and shrinkage tools to the idea of quantile regression. The most popular way to combine these both aspects of regularisation is the LASSO. Another type of data coming up in many cases are time series data or other data implying dependency. In such applications the idea of quantile regression has to be expanded by a term of random effects. DPMs are an idea to handle this kind of data which came up recently and implies the advantage of databased clustering. We want to show how Bayesian quantile regression can be made suitable for types of applications in which these influences are combined with other covariates like nonlinear and geoadditive effects.

2 Analyzed Data

2.1 LASSO in the Munich Rental Guide

The German tenancy law gives restrictions to the increase of rents and forces the landlords to keep the prize in a range that is common for flats which are comparable in size, location and quality. To make it easier for tenants and owners to assess if the rent is appropriate to the flat, Munich collects every year a big dataset of several flats and a list of characteristics as well as the price. From this dataset a regression model is generated which can be used by the inhabitants in order to check if the price of their flat is in a normal range. The collected data contains 241 covariates: size in square meters, year of construction, subquarter and a vector of categorical covariates, which consists of 238 characteristics of the flats, such as garden, type of kitchen or balcony. The predictor we used is:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}_{\text{cat}} + f_1(size) + f_2(year) + f_{\text{spat}}(region).$$

Size as well as year of construction are modeled nonlinear using cubic bayesian P-splines with random walk prior of second order, while subquarter is taken into the model as a spatial effect using a Markov random field prior. The parameters β_i of the influential variables are assigned a Bayesian LASSO prior.
An interesting question which arises is: will there be a lot of differences between the covariates which will be selected for the different quantiles?
In fact there are substantial differences between the results for selection as well as shrinkage behaviour in different quantiles. Furthermore nonlinear and spatial effects differ between the quantiles, too. As an example, see the nonlinear effect *size of the flat* in Figure 1. The picture shows the centralised curves for four different quantiles in comparison to the posterior interval of the median regression.

2.2 DPM in the LISA-Data

The following example deals with obesity among children measured in terms of the BMI. The LISA (Influences of **L**ife-style factors on the development

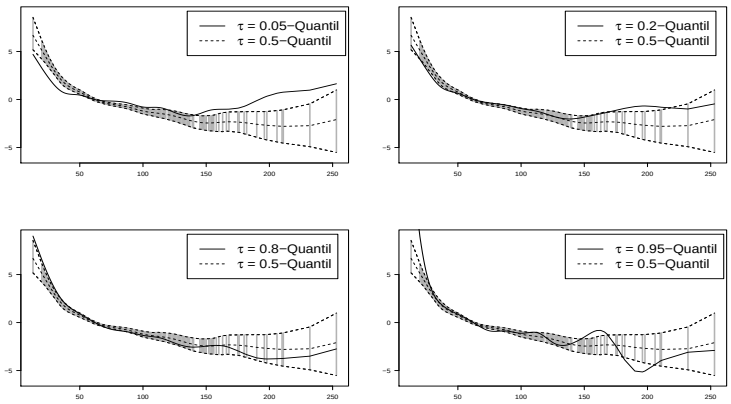


FIGURE 1. Solid line: effect of size of the flat on the non central quantiles, dashed lines: effect of size of the flat on the median and 95%-posterior interval, lightgrey lines in background: concentration of data

of the **I**mmune **S**ystem and **A**llergies in East and West Germany) study contains longitudinal data collected over 60 months at 9 points in time. Collected covariates are for example gender, nutrition until the age of 4 months (bottle or breastfed) and maternal smoking during pregnancy. An appropriate model for this data is:

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + f(t_{ij}) + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i + \varepsilon_{ij}$$

The first term consists of linear cause variables, while time is modeled nonlinear using Baysian P-splines again.
As the dataset consists of longitudinal data, a high dependency within data may be suspected. To allow for this fact we used random effects, denoted by $\boldsymbol{b}_i$. The above mentioned advantage of being able to consider clustering seems particularly useful as there might be different types of weight gaining children.
The results for the median regression are quite similar to these of the mean regression, with obvious differences in robustness. Figure 2 shows the BMI for four different individuals in median regression (on the left side) and 95%-quantile regression on the right. While the outliers for two children are ignored in the median regression they are obviously taken into account in the curve on the right side.
Comparison via the DIC showed, that the DPMModel performed better, than a model without the random effects as well as a model with normal Gaussian mixtures for the random effects.

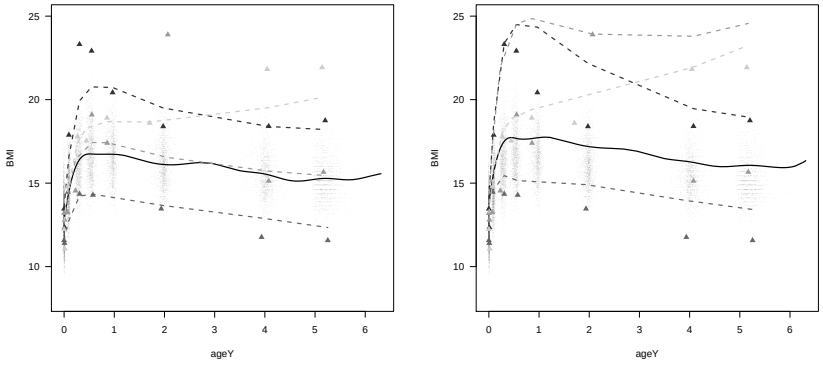


FIGURE 2. Effect of the age on the BMI in 50% – quantile (on the left) and 95% – quantile Regression (on the right), solid line: general effect of age, dashed lines: four different individual effects

3 Discussion

The gain of information by using quantile regression is obvious and there has been made much progress in amplifying the concept of Bayesian quantile regression by using the asymmetric Laplace distribution. Even complex statistical methods like the LASSO and the DPM can be added to the idea. Yet we have to state that using the ALD is a misspecification and hence we can only accept the results with reservation. Another task for the future is to find a way to avoid the high computation times of the MCMC techniques. One possibility might be the idea of the variational approximations.

Acknowledgments: Special thanks to Felix Heinzl for sharing his DPM-Code for mean regression and the German National Science Foundation (DFG) for financial support in the project *Structured additive quantile and expectile regression* (KN922/4-1).

References

Koenker, R.W. Bassett, J. (1978). Regression Quantiles. *Econometrica*, **46**, 33-50.

Yue, Y., Rue, H. (2011). Bayesian Inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis*, **55**, 84-96.

Groups within networks

Rober M West [1] [2], Paul G Dempster [2] , Justin Keen [2]

[1] Centre for Epidemiology and Biostatistics, University of Leeds, Worsley Building, Clarendon Way, Leeds LS2 9JT, UK
[2] Leeds Institute of Health Sciences, University of Leeds, Charles Thackrah Building, 101 Clarendon Road, Leeds LS2 9LJ, UK

Abstract: Structure is important for the mobilisation of knowledge within networks. The implementation of electronic health records provides a motivating example. One approach to assessing structure is through a position latent cluster model. This is adopted here and a cluster–classification algorithm is developed to provide a model at the cluster level rather than for relational ties.

Keywords: Social network analysis, position latent cluster models, knowledge mobilisation, electronic health records

1 Background

The mobilisation of knowledge through a large institution can have a great influence on efficiency in terms of getting things done. For example when establishing the implementation of electronic health records within a healthcare organistion, senior managers will initiate the mobilisation of knowledge which will progress through middle managers to those directly responsible for implementation. Formal management structures will influence mobilsation, Antonelli (1996), but there will also be further reasons for practical communication and development of knowledge. Informal networks, established practically, are of greater interest. Networks in healthcare are likely to be non-hierarchical because a diverse range of individuals with different perspectives and levels of responsibility are required to produce a workable solution to a particular problem.

Networks constrain or facilitate innovation and the spread of ideas by encouraging conformism to dominant perceptions of appropriate behaviour. Therefore likemindedness is a key to understanding the structure of the informal network, and might be captured by attributes of the network members (vertices). Considine *et al.* (2009) comment upon constraining innovation by encouraging conformism to dominant perceptions of appropriate behaviour.

Network data has particular challenges due to the dependence between ties. Exponential random graph models, Frank and Strauss (1986), are often employed following the result of Besag (1974). These are key steps in the development of modelling for networks.

Relevant to modelling of knowledge mobilisation, is a most interesting development provided by Hoff *et al.* (2002) with the concept of latent social space. The important concept is that participants are positioned such that proximity conveys an increased probability of a relational tie (link or edge) between participants. In practice there may be many potential explanatory variables for a tie to exist or not but the social space represents the first few (typically 2 or 3) principal components. Often explanatory variables are not collected, or an insufficient number are available so that the latent social space contributes an essential component of the analysis. It is possible though, and the focus here, that explanatory variables are available as attributes of the network participants or attributes of the ties.
Of further relevance is the development, recorded in Krivitsky and Handcock (2008), of clustering in latent social space. Here a mixture model approach is employed and participants are assigned probabilistically to clusters located at different centres within the social space. The existence of such clusters provide a plausible explanation for the effectiveness of mobilisation within some parts of an organisation as well as difficulties in mobilisation between those parts. Hence position latent cluster models are worthy of consideration, revealing potentially important structures within a network.
Explanation of structures is also desirable. These might be identified through either participant or tie attributes. Note however that the focus of exponential random graph models, of which position cluster models are a subset, is on modelling, thus explaining, the relational ties. The purpose of this work is to identify attributes of participants and ties that are most strongly associated with the clusters rather than the ties. Further, there is a method proposed to check if sufficient explanation has been achieved.

2 Method

The following cluster–classify algorithm for analysis is proposed for a network with vertex attributes:

1. Fit a position latent cluster model without specifying any attributes for vertices within the model. The number of clusters and the dimension of the social space are determined by the minimisation of the Bayesian Information Criterion (BIC).

2. From the best fit, determine a modal assignment of vertices to clusters.

3. Employ a classification tree to ascertain which attributes are associated with the defined clusters. In particular the first split within the tree might be taken.

4. Refit the position latent cluster model, using the vertex variables identified by the classification tree. Once more ascertain the optimum number of clusters by the minimisation of BIC.

5. Again use a classification tree and iterate the procedure until the optimum fit is with a single cluster.

The above algorithm can be modified to model attributes of ties rather than vertex attributes. In place of clusters of vertices, consider clusters of ties within clusters. Note that this analysis ignores the ties/edges between clusters.
It is envisaged that, although iteration is possible, only a few steps only would be undertaken, perhaps with only one classification. The classification step might be simplified by establishing just one attribute which maximises entropy between the clusters: thus being a single branching, or split, of the tree.

3 Example: monks

Hoff *et al.* (2002) and Kritisky and Handcock (2008) have investigated a subset of the well-known Monk data collected by Sampson (1968). Fitting a position latent cluster model with two-dimensional social space and without any attributes yields a three-cluster model which minimises the BIC. The cluster step is clear. The classification step is also extremely easy. The commonly available dataset has a single vertex attribute variable: namely the grouping determined by Sampson. This fits well in a classification tree. Returning to the position latent cluster model, fitting 2D social space and Sampson's groups yields a model for which the BIC is minimised by a single cluster: that is convergence is reached after a single classification step.
Note to reviewers: this example will almost certainly be replaced in the final version of this work.

4 Discussion

The modelling strategy expressed here focusses upon modelling at the cluster level rather than at the level of individual ties. In terms of likelihood, such an approach may not maximise the fit: overall the BIC found in later iterations may be greater than the BIC obtained at earlier steps. This is simply due to the model being specified in terms of relational ties rather than clusters. Nonetheless, network structure expressed here as clusters, has the greatest relevance to the interpretation of the influence of networks in knowledge mobilisation.

Acknowledgments: The work undertaken by the authors was supported by the National Institute for Health Research, grant number SDO EH239. This support was greatly appreciated.

References

Antonelli, C. (1996). Localised knowledge percolation processes and information networks. *Journal of Evolutionary Economics*,**6**, 281–295.

Besag J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society C*, **36**, 192–236.

Considine, M., Lewis, J.M., and Alexander, D. (2009). *Networks, innovation and public policy: Politicians, bureaucrats and the pathways to change inside government*. Houndsmills UK: Palgrave Macmillan.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, **81**, 832–842.

Hoff, P.D., Raftery, A.E., and Handcock, M.S. (2002) Latent space approaches to social network analysis. *Journal of the American Statistical Association*, **97**, 1090–1098.

Krivitsky, P.N. and Handcock, M.S. (2008). Fitting position latent cluster models for social networks with `latentnet`. *Journal of Statistical Software*, **24**, 1–23.

Sampson, S.F. (1968). *A novitiate in a period of change: an experimental and case study of relationships*. Ph.D. thesis Cornell University.

Robust mixture modelling of telemetry data in wildlife studies of home range

Bruce J. Worton[1], Chris R. Mclellan[1]

[1] School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK. email: Bruce.Worton@ed.ac.uk

Abstract: This paper considers the advantages of using mixtures of distributions, rather than the standard nonparametric approaches, for estimation in wildlife studies of home range. Mixtures are only used in a very limited way currently even though they were first proposed for such modelling in 1983. However, recent advances in the theory and computational techniques for mixtures of distributions mean that they may be used to analyse radio telemetry data and the models used to extract important ecological features of animal behaviour. We illustrate that robust mixture modelling is necessary as a common feature of telemetry data sets is that they contain outliers which lead to problems when determining appropriate mixture models. In particular, we investigate the use of mixtures of bivariate t distributions to take account of outliers, and study their properties. An application to brush rabbit telemetry data is used to show the advantages of using such heavy tailed mixtures.

Keywords: Finite mixture modelling; Robustness; Telemetry data.

1 Introduction

In wildlife studies of the home range behaviour of an animal, radio telemetry provides an effective and efficient sampling technique that can be used to collect large quantities of high quality data (White and Garrott 1990). This has brought new problems for data analysts involved in such studies: crude methods that were originally designed to analyze poor quality data produced by live-trapping often do not provide a sufficiently detailed summary of data to adequately answer behavioural and ecological questions of interest in a study. A common objective of home range studies is to describe an animal's use of space. It is convenient to consider an idealized probabilistic model of the way an animal uses its home range, and assume that the animal's (x, y)-position has a bivariate distribution over the plane for a specified time period; this distribution is known as the utilization distribution. Independent observations on the position of an animal, $\mathbf{x}_1 = (x_1, y_1)', \ldots, \mathbf{x}_n = (x_n, y_n)'$, which have been collected by radio-tracking can then be used to estimate the utilization density in various ways (White and Garrott 1990, Worton 1987),

2 Robust mixture modelling

Although Don and Rennolls (1983) proposed using finite mixtures of bivariate normal distributions to model locational home range data they have been little used. Standard methods for analysis of home range data are to estimate the utilization density by using a kernel-type estimator (Silverman 1986, Worton 1995a), or a convex-hull based estimator (Worton 1995b). However, mixture modelling has some very attractive properties that were discussed by Don and Rennolls in their paper. For example, Don and Rennolls did not estimate the means of the mixture components. Instead, the means were taken as *known* drey locations, but as the data sets were quite small they assumed circular normals which may be too restrictive in general.

In this paper we consider more flexible mixture modelling (McLachlan and Peel 2000), and show that it is fairly easy to fit such models using the R package `MCLUST` (Fraley and Raftery 2003). However, in applications to real locational telemetry data the use of t components in mixtures is advisable due to the presence of outliers, so here we propose use of the m-component model

$$f(\mathbf{x};\Psi) = \sum_{k=1}^{m} w_k f(\mathbf{x};\mu_k,\mathbf{\Sigma}_k,\nu_k),$$

where the kth component is bivariate t_{ν_k} $(\mu_k,\mathbf{\Sigma}_k)$ with degrees of freedom ν_k. This provides an attractive modelling approach as outliers are found to be a common feature of data resulting from animals occasionally exploring areas outside their usual home range to investigate them. This, however, leads to problems determining the number of components for fairly light-tailed mixture distributions such as the normal.

3 Simulation study

To investigate the properties of the robust t component mixtures procedures a simulation study was conducted. We present the results for two particular cases here, with sample size $n = 100$. In case I, the mixture density was taken as

$$\frac{1}{2}N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}5&-2\\-2&5\end{pmatrix}\right)+\frac{1}{2}N\left(\begin{pmatrix}5\\5\end{pmatrix},\begin{pmatrix}15&-1\\-1&2\end{pmatrix}\right),$$

with an outlier placed at $(12,-3)$. In case II, the mixture density was taken as

$$\frac{1}{4}t_4\left(\begin{pmatrix}5\\5\end{pmatrix},\mathbf{I}\right)+\frac{1}{2}t_4\left(\begin{pmatrix}0\\0\end{pmatrix},\mathbf{I}\right)+\frac{1}{4}t_4\left(\begin{pmatrix}0\\10\end{pmatrix},\mathbf{I}\right).$$

Mean BIC values produced are shown in Table 1. Mixtures of ts with $\nu = 2$ degrees of freedom were used, as a way of building in a robust approach to guard against the influence of outliers. However, we note that mixtures

TABLE 1. Mean BIC values for cases I and II.

No. comp. m	1	2	3	4	5	6
Case I	1131	1088	1100	1113	1124	1136
Case II	1202	1118	1080	1092	1100	1106

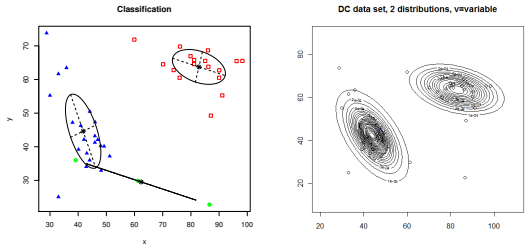


FIGURE 1. Best fitted mixture models for locational data on a female brush rabbit. Left panel: bivariate normal mixture (3 components); right panel: bivariate t mixture (2 components).

of ts with variable degrees of freedom produce improved fits, but lead to the same conclusions. For case I, the BIC identified the correct number of components using fixed degrees of freedom ($\nu = 2$) modelling in 470 of 500 simulated data sets; allowing for variable degrees of freedom gave 488 of 500.

4 Brush rabbit data analysis

Initially, finite bivariate normal mixture models were fitted to the locational data on a brush rabbit. However, the outliers in the data set lead to problems determining a satisfactory model. The left panel of Figure 1 shows the best fit as determined by BIC when using MCLUST. Outliers lead to problems with determining the number of mixture components in the normal case, but the right panel shows that a t mixture model has correctly identified the two clusters, without being unduly influenced by the outlying points. In each case these were the best models, based on BIC, over all possible models.

5 Conclusions

It is evident that finite mixture distributions provide very natural ways of modelling locational data on animals. They are particularly attractive as it is possible to build in *known* biological features of the home range, but

at the same time allow for features which are unknown at the start of the study. In contrast the standard approach of bivariate kernel smoothing is *only* able to highlight features, but does not attempt to model them. One area of interest would be building models that give descriptions of habitat and interaction with other animals. Therefore, we hope in the future that biologists will employ such mixture models more as they have advantages over nonparametric density estimation in investigating animal behaviour. With the large data sets it is now possible to collect, we have the opportunity to make important discoveries with regard to the ways animals use their environment. Future development of such models can incorporate environmental and habitat information explicitly into the mixture modelling.

Acknowledgments: CRM is supported by an EPSRC studentship. We thank Professor Keith Rennolls and Dr Bruce Don for extremely stimulating discussions on the concept of home range. Also, we would like to thank Professor K. Dixon for kindly providing access to the locational data set on brush rabbit.

References

Don, B.A.C., and Rennolls, K. (1983). A home range model incorporating biological attraction points. *Journal of Animal Ecology*, **52**, 69-81.

Fraley, C., and Raftery, A.E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *Journal of Classification*, **20**, 263-286.

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

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

White, G.C., and Garrott, R.A. (1990). *Analysis of Wildlife Radio-Tracking Data.* San Diego: Academic Press.

Worton, B.J. (1987). A review of models of home range for animal movement. *Ecological Modelling*, **38**, 277-298.

Worton, B.J. (1995a). Using Monte Carlo simulation to evaluate kernel-based home range estimators. *The Journal of Wildlife Management*, **59**, 794-800.

Worton, B.J. (1995b). A convex hull-based estimator of home-range size. *Biometrics*, **51**, 1206-1215.

Row-Column Association Models

Thomas W. Yee[1], Alfian F. Hadi[2]

[1] Department of Statistics, University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand,
[2] Statistics Department, Bogor Agricultural University, Bogor, 16680, Indonesia.

Abstract: This paper describes a statistical framework and software for fitting row-column association models (RCAMs) to two-way table responses. We consider some link function applied to the mean (say) of a cell equalling a row effect plus a column effect plus an interaction term. The interaction term is modelled as a reduced-rank regression (with complexities ranging from rank-1 and upwards), while the row and column (main) effects are handled using simple indicator variables. What sets apart this work from others is that our framework incorporates a very wide range of statistical models. For example, (i) log-link with Poisson counts is Goodman's RC model, (ii) zero-inflated Poisson distribution may be suitable with a two-way table with lots of zeros, (iii) identity-link with a double exponential (Laplace) distribution is akin to median polish, (iv) identity-link with normal errors is similar to two-way ANOVA with one observation per cell and allowing for modelling the interactions in a semi-complex manner, (v) log-link with negative binomial counts may help handle overdispersion relative to the Poisson model. New software within the first author's VGAM R package makes it very easy to fit a wide range of RCAMs to data. Altogether, the main result of this work is that RCAMs facilitates the analysis of two-way tables of many data types, therefore is potentially very useful in many areas of applied statistics.

Keywords: Main effects and interaction models; Reduced-rank regression; Two-way table; Vector generalized linear models; VGAM R package.

1 Introduction

Yee and Hastie (2003) introduced the class of reduced-rank vector generalized linear models (RR-VGLMs) which apply reduced-rank regression to the class of VGLMs. VGLMs cover a very wide range of statistical models, and its central algorithm involves iteratively reweighted least squares (IRLS) and Fisher scoring. It usually results in maximum likelihood estimation. A nontechnical introduction to VGLMs and RR-VGLMs is Yee (2010). In this paper we specialize the use of RR-VGLMs to two-way table responses. This may consist of continuous values, counts, proportions, or other data types. We wish to facilitate the modelling of main effects (row and column effects) plus possible interactions, while residing inside a statistical framework that can handle many data types of the responses. The

result makes it easy for the user as it provides a lot of flexibility relative to the slope of the learning curve.

2 RR-VGLM framework

Suppose our data comprises $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, for $i = 1, \ldots, n$, where $\boldsymbol{x}_i$ denotes the vector of explanatory variables for the ith observation and $\boldsymbol{y}_i$ is the response (possibly a vector). The first value of $\boldsymbol{x}_i$ is 1 for the intercept.
VGLMs are similar to ordinary GLMs but allow for multiple linear predictors. VGLMs handle M linear predictors (the dimension M depends on the model to be fitted) where the jth one is

$$\eta_j \;=\; \eta_j(\boldsymbol{x}) \;=\; \boldsymbol{\beta}_j^T \boldsymbol{x} \;=\; \sum_{k=1}^{p} \beta_{(j)k}\, x_k, \qquad j = 1, \ldots, M. \tag{1}$$

The η_j of VGLMs may be applied directly to parameters of a distribution, θ_j, rather than just to mean $\mu = E(Y)$ as for GLMs. In general,

$$\eta_j \;=\; g_j(\theta_j) \tag{2}$$

for some parameter link function g_j and parameter θ_j.
Bundling the linear predictors together gives

$$\boldsymbol{\eta} \;=\; \boldsymbol{\eta}(\boldsymbol{x}) \;=\; \begin{pmatrix} \eta_1(\boldsymbol{x}) \\ \vdots \\ \eta_M(\boldsymbol{x}) \end{pmatrix} \;=\; \mathbf{B}^T \boldsymbol{x} \;=\; \begin{pmatrix} \boldsymbol{\beta}_1^T \boldsymbol{x} \\ \vdots \\ \boldsymbol{\beta}_M^T \boldsymbol{x} \end{pmatrix} \tag{3}$$

where $\mathbf{B}$ is a $p \times M$ matrix of (sometimes too many) regression coefficients. In many situations the regression coefficients are related to each other. For example, some of the $\beta_{(j)k}$ may be equal, set to zero, or add up to a certain quantity. These situations may be dealt with by use of constraint matrices. VGLMs in general have

$$\eta_j(\boldsymbol{x}) \;=\; \sum_{k=1}^{p} \mathbf{H}_k\, \boldsymbol{\beta}_{(k)}^*\, x_k, \qquad j = 1, \ldots, M, \tag{4}$$

where $\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_p$ are known full-column rank constraint matrices, and $\boldsymbol{\beta}_{(k)}^*$ are vectors of unknown coefficients. With no constraints at all, $\mathbf{H}_1 = \mathbf{H}_2 = \cdots = \mathbf{H}_p = \mathbf{I}_M$. Then, for VGLMs,

$$\mathbf{B}^T \;=\; \begin{pmatrix} \mathbf{H}_1\boldsymbol{\beta}_{(1)}^* & \mathbf{H}_2\boldsymbol{\beta}_{(2)}^* & \cdots & \mathbf{H}_p\boldsymbol{\beta}_{(p)}^* \end{pmatrix}. \tag{5}$$

2.1 RR-VGLMs

Partition $\boldsymbol{x}$ into $(\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T$ (of dimension $p_1 + p_2 = p$) and $\mathbf{B} = (\mathbf{B}_1^T \; \mathbf{B}_2^T)^T$. If $\mathbf{B}_2$ has too many regression coefficients then we can reduce its number dramatically by a reduced-rank regression. RR-VGLMs then have

$$\boldsymbol{\eta} \;=\; \mathbf{B}_1^T \boldsymbol{x}_1 + \mathbf{B}_2^T \boldsymbol{x}_2 \tag{6}$$

where we approximate $\mathbf{B}_2$ by a reduced-rank regression

$$\mathbf{B}_2 \;=\; \mathbf{C}\,\mathbf{A}^T. \tag{7}$$

Here, $\mathbf{C}$ and $\mathbf{A}$ are $p_2 \times R$ and $M \times R$ respectively, and they are 'thin' because the rank R is low, e.g., $R = 1$ or 2. Thus

$$\boldsymbol{\eta} \;=\; \mathbf{B}_1^T \boldsymbol{x}_1 + \mathbf{A}\,\boldsymbol{\nu} \tag{8}$$

where $\boldsymbol{\nu} = \mathbf{C}^T \boldsymbol{x}_2$ is a vector of R latent variables.
To make the parameters unique, it is common to enforce corner constraints on $\mathbf{A}$. By default, the top $R \times R$ submatrix is fixed to be $\mathbf{I}_R$ and the remainder of $\mathbf{A}$ is estimated.

3 RCAMs

We initially use Goodman's RC(r) model to explain what a RCAM is. How does this model fit within the VGLM framework? Suppose $\mathbf{Y} = [(y_{ij})]$, a $n \times M$ matrix of counts. Goodman's model fits a reduced-rank type model to $\mathbf{Y}$ by firstly assuming $Y_{ij} \sim$ Poisson, and that

$$\log \mu_{ij} \;=\; \mu + \alpha_i + \gamma_j + \sum_{k=1}^{R} a_{ik}\, c_{jk}, \tag{9}$$

where $\mu_{ij} = E(Y_{ij})$ is the mean of the i-j cell. Identifiability constraints are needed in (9) for the row and column effects α_i and γ_j; we use corner constraints $\alpha_1 = \gamma_1 = 0$ here. The parameters a_{ik} and c_{jk} also need constraints, e.g., we use $a_{1k} = c_{1k} = 0$ for $k = 1, \ldots, R$. Then write (9) as

$$\log \mu_{ij} \;=\; \mu + \alpha_i + \gamma_j + \delta_{ij},$$

where the $n \times M$ matrix $\boldsymbol{\Delta} = [(\delta_{ij})]$ of interaction terms is approximated by the reduced rank quantity $\sum_{k=1}^{R} a_{ik} c_{jk}$.
Goodman's RC(R) fits within the VGLM framework by letting

$$\boldsymbol{\eta}_i \;=\; \log \boldsymbol{\mu}_i \tag{10}$$

where $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i)$ is the mean of the ith row of $\mathbf{Y}$. Then the matrix $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)^T$ fits into the RR-VGLM framework as follows. From Section 2, we obtain $\mathbf{B}_1^T \boldsymbol{x}_{1i} =$

$$\Big(\mu \mathbf{1}_M \;\; \alpha_2 \mathbf{1}_M \;\; \cdots \;\; \alpha_n \mathbf{1}_M \;\; \big(\mathrm{Diag}(\gamma_1, \ldots, \gamma_M)_{(-1)}\big)^T\Big) \begin{pmatrix} 1 \\ \boldsymbol{e}_{(-1)i} \\ \mathbf{1}_{M-1} \end{pmatrix} \tag{11}$$

TABLE 1. Some VGAM family functions useful in conjunction with `rcam()`. "GRC" stands for Goodman's RC model.

Family name	Comments
`alaplace2(0.5)`	Median polish when rank-0.
`normal1()`	Two-way ANOVA (one observation per cell).
`poissonff()`	GRC model.
`negbinomial()`	GRC with overdispersion wrt Poisson.
`zipoissonff()`	GRC with lots of 0's and/or structural 0's.

where a subscript "(-1)" means the first element or row is removed from the vector or matrix. This shows, for example, that the intercept and row score variables have $\mathbf{1}_M$ as their constraint matrices. Similarly, because $\mathbf{B}_2$ is approximated by $\mathbf{C}\mathbf{A}^T$, the ith row of $\boldsymbol{\Delta}$ will be approximated by $\boldsymbol{x}_{2i}^T\mathbf{C}\mathbf{A}^T$, or equivalently, $\boldsymbol{\Delta}$ is approximated by $\left(\boldsymbol{x}_{21}^T, \ldots, \boldsymbol{x}_{2n}^T\right)^T \mathbf{C}\mathbf{A}^T$. The desired reduced-rank approximation of $\boldsymbol{\Delta}$ can be obtained if $\boldsymbol{x}_{2i} = \boldsymbol{e}_i$ so that $\mathbf{I}_{p_2}\mathbf{C}\,\mathbf{A}^T = \mathbf{C}\,\mathbf{A}^T$. Note that

$$\boldsymbol{\Delta} \;=\; \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \widetilde{\boldsymbol{\Delta}} \end{pmatrix} \;\approx\; \mathbf{C}\,\mathbf{A}^T \;=\; \begin{pmatrix} \mathbf{0}^T \\ \mathbf{C}_{(-1)} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \left(\mathbf{A}_{(-1)}\right)^T \end{pmatrix},$$

that is, the first row of $\mathbf{A}$ consists of structural zeros which are 'omitted' from the reduced rank regression of $\boldsymbol{\Delta}$.

3.1 RCAMs

One could define RCAMs as a RR-VGLM with

$$\eta_{1ij} \;=\; \mu + \alpha_i + \gamma_j + \sum_{k=1}^{R} a_{ik}\, c_{jk}, \tag{12}$$

(cf. (9)). Note that (12) applies to the *first* linear/additive predictor; for models with $M > 1$ one can leave $\eta_2, \ldots, \eta_M$ unchanged. Of course, choosing η_1 for (12) is only for convenience. The software chooses $g_1^{-1}(\widehat{\eta}_1)$ as the fitted values of the model and these are returned by `fitted(rcamobject)` and the result should be the same dimension as the two-way table.
To summarize, RCAMs in general are RR-VGLMs where the first linear/additive predictor is modelled as the sum of a row effect, a col effect, and an interaction effect which is expressed as a reduced-rank regression. Table 1 summarizes a few possible RCAMs.

4 Data and software

The first author has written `rcam()` to fit RCAMs within his VGAM package. This function calls `vglm()` if the rank is zero, otherwise `rrvglm()`. In both

cases the dummy variables and constraint matrices are set up beforehand, corresponding to (11). The `family` argument of `rcam()` is passed into an argument of the same name in `vglm()`/`rrvglm()` to fit the desired model. Currently, it is important that the first linear/additive predictor η_1 corresponds to the mean or some parameter measuring central location. Consequently, `zipoissonff()` may be used rather than `zipoisson()` because the latter models the probability of a structural zero in η_1 whereas the former models the mean of the Poisson distribution. All other parameters are generally fitted with intercept-only, for example, the k parameter for the negative binomial NB(μ, k). And they may all be constrained to be equal over rows and columns of $\mathbf{Y}$.
In the future, ideally every possible VGAM `family` function will work with `rcam()`, however, whether the output makes sense or is sensible is another story.
The typical call for a median polish-type fit might be of the form

```
rcam0 <- rcam(auuc, alaplace2(tau = 0.5, intparloc = TRUE))
```

The software is currently undergoing development, therefore future changes to what is presented here are possible.

4.1 Crash data

The VGAM package has several two-way tables suitable for exploring RCAMs. These include `crashi`, `crashf`, `crashp`, and `alcoff`. In general these are a variety of reported crash data cross-classified by time (hour of the day) and day of the week, accumulated over 2009. Thus the data frames are 24×7 in dimension. The data include fatalities and injuries (by car), trucks, motor cycles, bicycles and pedestrians. There is some alcohol-related data too (Special thanks to Warwick Goold for help with the data.)
The matrix `alcoff` is the number of alcohol offenders caught from breath screening drivers, during the whole of 2009. Then

```
> rbind(head(alcoff, 2), tail(alcoff, 2))
   Monday Tuesday Wednesday Thursday Friday Saturday Sunday
0     121      98       165      324    827     1379   1332
1      97      92       157      278    619     1327   1356
22     90     143       345      765    976     1026    114
23    110     169       363      899   1265     1179    159
```

Here, the first row is from midnight to 1am, and the last row is for 11pm to midnight.

5 Example

We fit a rank-0 Goodman's RC model to `alcoff`. We preprocess the data by offsetting the data with respect to the hour. We say the *effective* day starts at 6am since partying at late night often spills over to the early

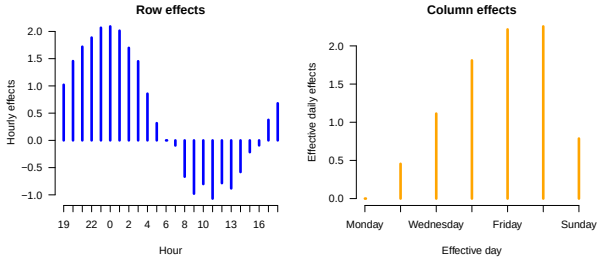


FIGURE 1. Hourly and effective daily effects of a Goodman's RC model fitted to `alcoff`. This is output from `plotrcam0()`.

morning. Hence effective `Monday` starts at 6am and finishes on Tuesday at the same time. The function `Rcam()` and/or `moffset()` enables us to create the effective day.
We fit the GRC model by

```
fit0 <- rcam(Rcam(moffset(alcoff, "6")), family = poissonff,
             rprefix = "Hours.24.", cprefix = "Days.")
```

Alternatively we could use `grc()`. Then applying `plotrcam()` gives Fig. 1 which plots the fitted main effects. The results agree with what is expected: the greatest number of alcohol-related offences occur on Friday and Saturday nights (and their following morning), and there is a gradual increase from Sunday/Monday to these peak days. Also, they are at their lowest in the late morning to lunchtime period.

References

Yee, T.W., and Hastie, T.J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, **3**, 15–41.

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

Use of Marginal Likelihoods in Statistical Inference

Kathryn Ziegler-Graham[1], Charles A. Rohde[2]

[1] St Olaf College, 1520 St Olaf Ave Northfield, MN USA (kziegler@stolaf.edu)
[2] Johns Hopkins University, 615 North Wolfe St Baltimore, MD USA (crohde@jhsph.edu)

Abstract: Marginal likelihoods are explored from the point of view of the evidential perspective. Specifically, we consider the case of a multidimensional likelihood where we are interested in obtaining evidence about only one parameter or a single function of parameters with the remaining parameters considered to be nuisance parameters. Several methods have been proposed to deal with the situation of nuisance parameters including orthogonal likelihoods, marginal likelihoods, conditional likelihoods, in addition to estimated and profile likelihoods. Using a marginal likelihood when available provides us with a solution to eliminating nuisance parameters. Although the marginal likelihood is not the "full" likelihood, it is a "true" likelihood since it is constructed from actual probability density or mass functions. Because the marginal likelihood is a true likelihood, as opposed to a profile or estimated likelihood, evidential properties automatically hold. In particular, the universal bound on misleading evidence holds.
The non-central t and F distributions are used to obtain marginal likelihoods for several important parameters including the variance, the effect size for two groups, the overlapping coefficient, the area under an ROC curve, and the shrinkage parameter in hierarchical models. These marginal likelihoods are true likelihoods. The probabilities of misleading and weak evidence can be obtained and the universal bound on the probability of observing misleading evidence applies. In addition use of reference priors allows for Bayesian analysis. The graphical display of parameter support and uncertainty provide clean alternatives to typically computationally intensive confidence interval calculations.

Keywords: Evidence, Marginal Likelihoods, Hierarchical Models

1 Marginal Likelihood for The Variance

One of the most important examples of a marginal likelihood is the likelihood for the variance. Specifically if $X_1, X_2, \ldots, X_n$ are iid each $\mathrm{N}(\mu, \sigma^2)$ then it is well known that

$$U = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \overset{d}{\sim} \text{Chi-square}(n-1)$$

The resulting marginal likelihood is

$$L_m(\sigma^2;t) = \left[\frac{\widehat{\sigma}_m^2}{\sigma^2}\exp\left\{1-\frac{\widehat{\sigma}_m^2}{\sigma^2}\right\}\right]^{\frac{n-1}{2}}$$

2 Effect Size for Two Groups

Suppose that we have two samples, one from a $\mathrm{N}(\mu_1,\sigma^2)$ and the other from a $\mathrm{N}(\mu_2,\sigma^2)$ with sample sizes n_1 and n_2 respectively. Define s^2, the pooled estimate of variance by

$$(n_1+n_2-2)s^2 = (n_1-1)s_1^2+(n_2-1)s_2^2$$

If we consider the distribution of

$$T = \frac{\overline{x}_2-\overline{x}_1}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

then T is non-central t with n_1+n_2-2 degrees of freedom and non-centrality parameter λ where

$$\lambda = \frac{\sqrt{n_1n_2}(\mu_2-\mu_1)}{\sigma\left(n_1+n_2\right)^{1/2}} = \sqrt{\frac{n_1n_2}{n_1+n_2}}\frac{\mu_2-\mu_1}{\sigma}$$

or

$$\lambda = \sqrt{\frac{n_1n_2}{n_1+n_2}}\delta$$

and δ is the effect size defined by

$$\delta = \frac{\mu_2-\mu_1}{\sigma}$$

It follows that we may obtain a likelihood for δ using the non-central t distribution.

3 Overlapping Coefficient

If f and g are two densities then the **overlapping coefficient** is defined as

$$\theta = \int_{-\infty}^{+\infty}\min[f(x),g(x)]dx$$

If f and g are each normal with common variance then

$$\theta = 2\Phi\left(\frac{\mu_1-\mu_2}{2\sigma}\right)$$

We can use the non-central t distribution to get a marginal likelihood for

$$\frac{\mu_1-\mu_2}{\sigma}$$

and hence a marginal likelihood for θ.

4 Area under the ROC curve

The area under the ROC curve (AUC ROC) is equivalent to

$$\text{AUC} = P(X < Y). \tag{1}$$

For the case where X and Y are independent and normally distributed (the binormal case) this area simplifies to

$$\text{AUC} = P(X < Y) = \Phi\left(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right) \tag{2}$$

Using the marginal likelihood for the two sample effect size and equation 2 we obtain a marginal likelihood for the area under the ROC curve.

5 "Evidence" in Hierarchical Models

Assume that we have p samples $\mathbf{y_i} = \mathbf{y_{i1}}, \mathbf{y_{i2}}, \ldots, \mathbf{y_{ir}}$ each of which is a realized value of $\mathbf{Y_i} = (\mathbf{Y_{i1}}, \mathbf{Y_{i2}}, \ldots, \mathbf{Y_{ir}})$ where $\mathbf{Y_i}$ is N $(\mathbf{1_r}\mu_\mathbf{i}, \mathbf{I_r}\sigma^\mathbf{2})$ and σ^2 is known. Assume that the $\mathbf{Y_i}$ are independent and also assume that the μ_i are independent each N(μ, σ_μ^2) where μ and σ_μ^2 are both known. This is the one-way random effects model with balanced data or equivalently a one level hierarchical model. The number of clusters is p and there are r observations per cluster.
Of interest is the evidence for $H_{i1} : \mu_i = \mu_{i1}$ vs $H_{2i} : \mu_i = \mu_{i2}$ and a likelihood for μ_i.
The evidence for μ_{i1} vs μ_{i2} is given by

$$\frac{\exp\left\{-\frac{\sum_{j=1}^{r}(y_{ij}-\mu_{i2})^2}{2\sigma^2} - \frac{(\mu_{i2}-\mu)^2}{2\sigma_\mu^2}\right\}}{\exp\left\{-\frac{\sum_{j=1}^{r}(y_{ij}-\mu_{i1})^2}{2\sigma^2} - \frac{(\mu_{i1}-\mu)^2}{2\sigma_\mu^2}\right\}}$$

which may be rewritten as

$$\exp\left\{\frac{r(\mu_{i2}-\mu_{i1})}{\sigma^2}\left[\overline{y}_{i+} - \frac{\mu_{i2}+\mu_{i1}}{2}\right] + \frac{(\mu_{i2}-\mu_{i1})}{\sigma_\mu^2}\left[\mu - \frac{\mu_{i2}+\mu_{i1}}{2}\right]\right\}$$

This expression is the product of two factors. The first,

$$\exp\left\{\frac{r(\mu_{i2}-\mu_{i1})}{\sigma^2}\left[\overline{y}_{i+} - \frac{\mu_{i2}+\mu_{i1}}{2}\right]\right\}$$

represents the evidence supplied by the data. It is exactly the evidence for μ_{i2} vs μ_{i1} supplied by a random sample of size r from a normal distribution with known variance.

The second,

$$\exp\left\{\frac{(\mu_{i2}-\mu_{i1})}{\sigma_\mu^2}\left[\mu-\frac{\mu_{i2}+\mu_{i1}}{2}\right]\right\}$$

represents what we might call 'model' or 'prior' evidence.
The 'likelihood' for μ_i which is proportional to

$$\exp\left\{\frac{-2r\overline{y}_{i+}\mu_i+r\mu_i^2}{2\sigma^2}-\frac{-2\mu\mu_i+\mu_i^2}{2\sigma_\mu^2}\right\}$$

or

$$\exp\left\{-\frac{\mu_i^2}{2}\left(\frac{r}{\sigma^2}+\frac{1}{\sigma_\mu^2}\right)-\frac{2\mu_i}{2}\left(\frac{r\overline{y}_{i+}}{\sigma^2}+\frac{\mu}{\sigma_\mu^2}\right)\right\}$$

If we define

$$\widetilde{\mu}_i=\frac{\frac{r\overline{y}_{i+}}{\sigma^2}+\frac{\mu}{\sigma_\mu^2}}{\frac{r}{\sigma^2}+\frac{1}{\sigma_\mu^2}}$$

Then the likelihood may be written as

$$\text{Lik}=\exp\left\{-\frac{(\mu_i-\widetilde{\mu}_i)^2}{2\widetilde{\sigma}^2}\right\}$$

where

$$\widetilde{\sigma}^2=\frac{1}{\frac{r}{\sigma^2}+\frac{1}{\sigma_\mu^2}}$$

Thus the 'likelihood' of μ_i is that of a normal distribution centered at $\widetilde{\mu}_i$ with scale parameter $\widetilde{\sigma}$. Note that

$$\widetilde{\mu}_i=\overline{y}_{i+}-\gamma(\overline{y}_{i+}-\mu)$$

where

$$\gamma=\frac{\frac{1}{\sigma_\mu^2}}{\frac{r}{\sigma^2}+\frac{1}{\sigma_\mu^2}}=\frac{1}{\frac{r\sigma_\mu^2}{\sigma^2}+1}=\frac{\sigma^2}{\sigma^2+r\sigma_\mu^2}$$

represents a "shrinkage factor".
The amount of shrinkage is determined by γ. Under the model assumed here it is known that

$$\frac{\text{MSB}/(\sigma^2+r\sigma_\mu^2)}{\text{MSE}/\sigma^2}=\frac{\text{MSB}}{\text{MSE}}\frac{\sigma^2}{\sigma^2+r\sigma_\mu^2}=\gamma\frac{\text{MSB}}{\text{MSE}}$$

has a central F distribution with $p-1$ and $p(r-1)$ degrees of freedom.
Building on the above we can obtain a likelihood for the shrinkage factor using the central F distribution. Since

$$Y=\gamma\,\frac{\text{MSB}}{\text{MSE}}$$

is F with $p-1$ and $p(r-1)$ degrees of freedom it follows that the density function of

$$X = \frac{\text{MSB}}{\text{MSE}} = Y/\gamma$$

is given by

$$f(x;\gamma) = \frac{m^{m/2}n^{n/2}}{B(m/2,n/2)} \frac{(\gamma x)^{\frac{m}{2}}}{(n+m\gamma x)^{(m+n)/2}}$$

The maximum occurs when

$$\widetilde{\gamma} = \frac{1}{x}.$$

The likelihood for γ can therefore be obtained using the F distribution. Note that a γ of 1 implies that there is complete shrinkage i.e. every sample mean is estimated by the population mean while small values of γ indicate lack of shrinkage.

References

Barndorff-Nielsen, O.E. and Cox, D.R. (1994) *Inference and Asymptotics* Chapman & Hall/CRC

Bernardo, J.M. and Juárez, M.A. (2003) Intrinsic Estimation in *Bayesian Statistics 7* Oxford: University Press

Cox, D.R. (1975) Partial Likelihood. Biometrika (62) 269-276

Edwards, A.W.F. (1972). *Likelihood* Cambridge University Press.

Freedman, D.P.,Pisani,R & Purves, R. (2007). *Statistics (4th ed.)* New York & London: WW Norton

Hacking, I. (1965). *Logic of Statistical Inference* Cambridge University Press

Royall, R.M. (1997). *The Likelihood Paradigm* Chapman & Hall/CRC

Index

IWSM 2011 Sponsors

We are very grateful to the following organisations for sponsoring IWSM 2011.,

- Conselleria d'Educació, Generalitat Valenciana
- GEeitEma Research group
- Universitat de València
- Servei d'Investigació, Universitat de València
- Facultat de Matemàtiques, Universitat de València
- Departament d'Estadística i Investigació Operativa, Universitat de València
- Biostatnet, Research network

IWSM 2011

FUNDACIÓ UNIVERSITAT EMPRESA
VNIVERSITAT DE VALÈNCIA

SMS
Statistical Modelling Society

VNIVERSITAT DE VALÈNCIA

GENERALITAT VALENCIANA

CONSELLERIA D'EDUCACIÓ

GEeitEma

VNIVERSITAT DE VALÈNCIA
Servei d'investigació

BIOstatnet

VNIVERSITAT DE VALÈNCIA
Facultat de Matemàtiques.

VNIVERSITAT DE VALÈNCIA
Departament d'Estadística i Investigació Operativa